Simulation Results: AMKAT P-value Estimator For PhiMr

Contents

Introduction

These simulations were designed to explore the distribution of our proposed P-value estimator for AMKAT for testing with the PhiMr filter. The estimator is defined as

$$\tilde{P}_{T_{\mathcal{S}},Q} = \frac{\left(\sum_{b=1}^{B} I(\overline{T}_{\mathcal{S},Q} \leq \ddot{T}_{\mathcal{S},b})\right) + 1}{B},$$

where $\overline{T}_{\mathcal{S},Q}$ is the sample mean of Q AMKAT test statistics computed with PhiMr, $\ddot{T}_{\mathcal{S},b}$ are AMKAT permutation statistics computed with PhiMr, B is the number of permutations and I is the indicator function.

We are particularly interested in the effect of the number Q of test statistics $T_{\mathcal{S}}$ and the number B of permutation statistics $\ddot{T}_{\mathcal{S}}$ used when estimating the P-value under the PhiMr filter.

In each scenario for this setting, a single set of data was generated; using this fixed data, we computed 256,000 values of the test statistic $T_{\mathcal{S}}$ and 2,500,000 permutation statistics $\ddot{T}_{\mathcal{S}}$ using the PhiMr filter; we also computed 2,500,000 permutation statistics $\ddot{T}$ without the PhiMr filter, as well as the value of the associated test statistic T, which is nonrandom on a fixed data set.

All statistics other than T were equally partitioned into 500 batches, each containing 512 values of $T_{\mathcal{S}}$, 5000 values of $\ddot{T}_{\mathcal{S}}$ and 5000 values of $\ddot{T}$. For each batch, different variations of AMKAT were performed using different subsets of the statistics: first, the common value of T and the first B values of $\ddot{T}$ were used to compute a P-value estimate $\hat{P}_T$ for AMKAT without the PhiMr filter, where B ranged from 100 to 5000; next, the first Q values of $T_{\mathcal{S}}$ and the first B values of $\ddot{T}_{\mathcal{S}}$ were used to compute a P-value $\hat{P}_{T_{\mathcal{S}},Q}$ under PhiMr, where Q ranged from 1 to 512 and B ranged from 100 to 5000. To assess the variability of the PhiMr filter itself (rather than the associated P-value) on the fixed data set, for each component of $\boldsymbol{X}$ we recorded the proportion of the 256,000 PhiMr applications for which the component was included in the selected subset.

For all scenarios in this setting, we simulated $\boldsymbol{\epsilon}$ as multivariate normal. We considered the same signal sets and effect functions as those used in our power simulations.

Our presentation of results for the first data set contains additional remarks and commentary regarding our proposed P-value estimator, its motivations, and the motivational context for the exploration conducted in this simulation setting. Presentation of results for subsequent data sets will assume knowledge of this context.

Prepare Working Environment

```
# Load relevant packages
pkgs_to_load <- c('dplyr', 'ggplot2', 'tidyr', 'viridis', 'cowplot')
lapply(X = pkgs_to_load, FUN = library, character.only = TRUE)

# Define directories for scripts and data
dir_main <- dirname(dirname(rstudioapi::getActiveDocumentContext()$path))
dir_src <- file.path(dir_main, 'source_scripts')
source(file.path(dir_src, 'define_directories.R'))
```

```
# Define custom functions used for plotting
source(file.path(dir_src, "define_plot_functions.R"))
source(file.path(dir_src, "define_plot_settings.R"))
```

Continuous Features

We initialize the simulation setting for the case where the components of $\boldsymbol{X}$ are continuous:

```
x_type <- 'cts'
source(file.path(dir_src, "initialize_adaptive_within.R"))
```

Data Set #1

Data Set #1 was generated using the following scenario parameters:

```
n <- 30 # sample size
p <- 100 # feature dimension
signal_density <- 'sparse' # 7 signal variables
error_correlation_strength <- 0.5 # mixed directions
```

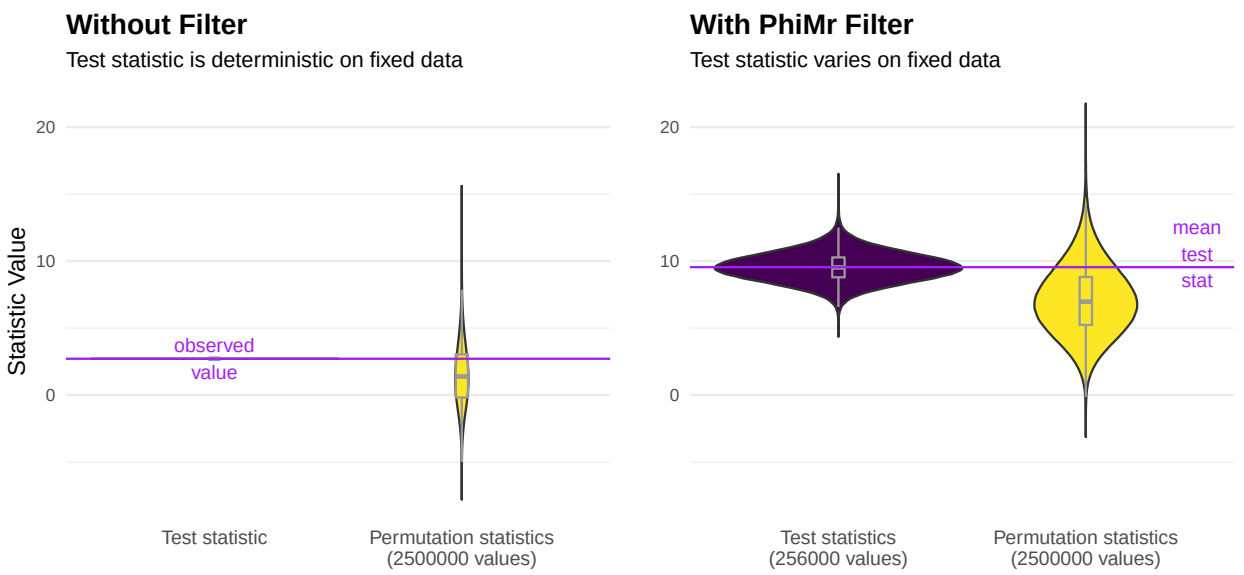
We load the test and permutation statistics generated on this data set:

```
pv_files <- pvaluePlotFiles()
load(pv_files$stats)
load(pv_files$pvalues)
```

Distribution of Test and Permutation Statistics

We first plot the distributions of the test statistics and permutation statistics with and without the PhiMr filter:

```
makeStatDistrPlots(
  test_stats, perm_stats, perm_stats_no_filter, test_stat_no_filter)
```



The histogram for each sample of statistics has been smoothed using kernel density estimation and mirrored to produce a violin plot. It should be noted that for a fixed data set, the unknown underlying distribution is actually a discrete distribution. The choice to display smoothed versions of the empirical densities was made for visual practicality, as the plot is primarily intended to serve as a conceptual aid to help motivate and illustrate our proposed method of P-value estimation with the PhiMr filter.

The left-hand plot, which displays the distribution of AMKAT permutation statistics generated without the PhiMr filter, corresponds to a traditional permutation test, with a test statistic that is deterministic on our fixed data set. Visually, the P-value estimate can be conceptualized as the upper-tail area of the permutation statistic distribution that is cut off by the observed test statistic value (the the true and estimated P-values are actually based on discrete distributions, but the discrepancy is not crucial to the idea we wish to illustrate).

The right-hand plot displays, for our data set, the distribution of permutation statistics generated with the PhiMr filter, as well as the distribution of test statistics generated with PhiMr. Because of the randomness introduced by the permutation-based methodology of the PhiMr filter, the test statistic is no longer fully determined by the data, but instead exhibits random variation as seen in the plot. This provides an illustrative depiction of the source of additional variation in the P-value estimate when using the PhiMr filter. For this data set, the distribution of the test statistic is unimodal and symmetric, with the probability mass spread across a considerable range of the permutation statistic distribution; when drawing a single test statistic to use as a tail-area cutoff value for the permutation statistic distribution, this cutoff value, and thus the resulting P-value estimate, could vary considerably irrespective of the number of permutation statistics drawn, which is quite a different situation from a traditional permutation test.

Looking at the plot for this particular data set, we can observe that the test statistic distribution on the right-hand plot appears centered at a cutoff value located farther out in the tail of the respective permutation distribution than is the case for the left-hand plot, implying that on average, AMKAT will yield a lower P-value estimate with the PhiMr filter than without it. However, the spread of the test statistic distribution in the right-hand plot suggests that this outcome may be highly inconsistent across repeated applications of AMKAT on the same data set. Even if AMKAT with PhiMr is more capable, on average, of discriminating between the null and alternative, this may be of little use if it cannot do so with reasonable consistency. For certain data and a particular significance level, we could find ourselves in a situation where performing

AMKAT with PhiMr could be highly inconsistent in terms of concluding whether or not to reject H_0, which is a particularly troubling possibility.

The plot suggests that in order to address this separate source of variation in the P-value estimate introduced by the PhiMr filter, we should seek to reduce the variation in the test statistic value under PhiMr. One way to do this would be to increase the number of permutations used by PhiMr. Currently, PhiMr makes its determinations using only a single permuted copy of the data; by considering more permuted copies and including a way to incorporate the information from the multiple copies into a single decision (by majority rule, etc.), its random variation could be reduced. However, this would scale up the computational complexity of our test by passing the increased cost of PhiMr on to each permutation statistic generated; depending on how many additional permutations are needed by PhiMr to adequately reduce test statistic variation, this could greatly compromise the practical usability of AMKAT with PhiMr.

Another possible strategy, and the one we have ultimately considered (formally defined at the beginning of this document), is depicted in the plot; by generating multiple test statistic values and using a measure of their center as the cutoff value for the permutation statistic distribution, we can expect more consistent P-value estimates compared to using only a single value. This has the added benefit of scaling independently from the number of permutation statistics used, potentially offering much greater flexibility in trading increases in complexity for reductions in the variation of the P-value estimator.

A concern with such a strategy is that, in some sense, it moves even further away from a traditional permutation test, in that the permutation statistics are no longer being generated according to the exact same procedure as the test statistic value that is being compared to their empirical distribution; this is deliberate, as doing so would require each permutation statistic value being the average of multiple statistics generated on the same set of permuted data, which would be similar in terms of effect and computational cost to increasing the number of permutations used by the PhiMr filter.

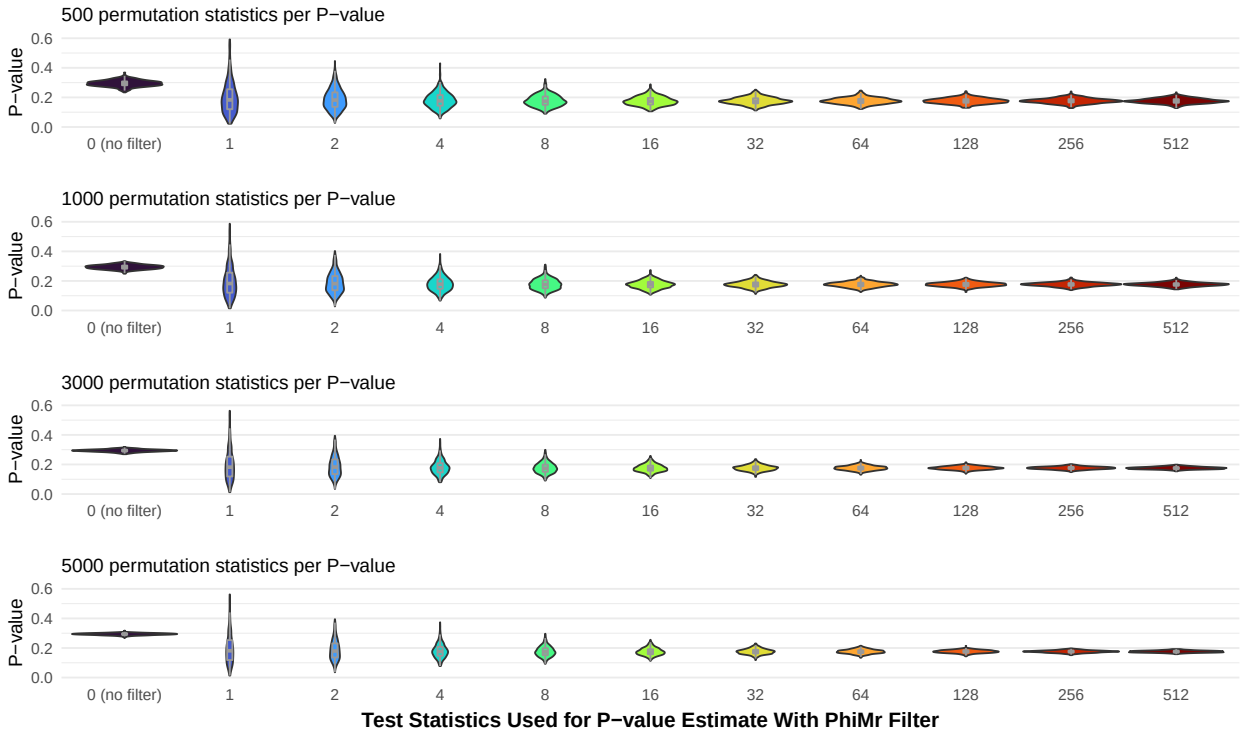
Given that we are choosing not to modify our method for computing permutation statistics in the same manner as that for computing our test statistic, it is crucial to investigate whether this proposed method of P-value estimation under PhiMr is able to adequately control the probability of a type I error. This is addressed in our simulations for size.

In the remaining plots for this setting, we continue to investigate how the variation in our proposed P-value estimator for AMKAT with the PhiMr filter behaves compared to the traditional estimator used by AMKAT without the PhiMr filter, and how this behavior depends on the number of test statistics and number of permutation statistics used. One of our objectives across the scenarios in this simulation setting is to form general recommendations for the number of test statistics and permutation statistics to use, which requires a better understanding of the tradeoffs involved in increasing one versus the other from different starting combinations of values.

Distribution of P-values

After partitioning the test and permutation statistics into batches and using them to estimate P-values, we plot the distributions of the resulting P-values.

```
makeViolinQPlots(pvalues, title_settings = title_violinQ(num_replicates))
```



Each row in the plot corresponds to a particular number B of permutation statistics used to estimate a P-value; each column corresponds to a particular number Q of test statistics used to estimate the P-value under PhiMr (where 0 corresponds to AMKAT without the PhiMr filter).

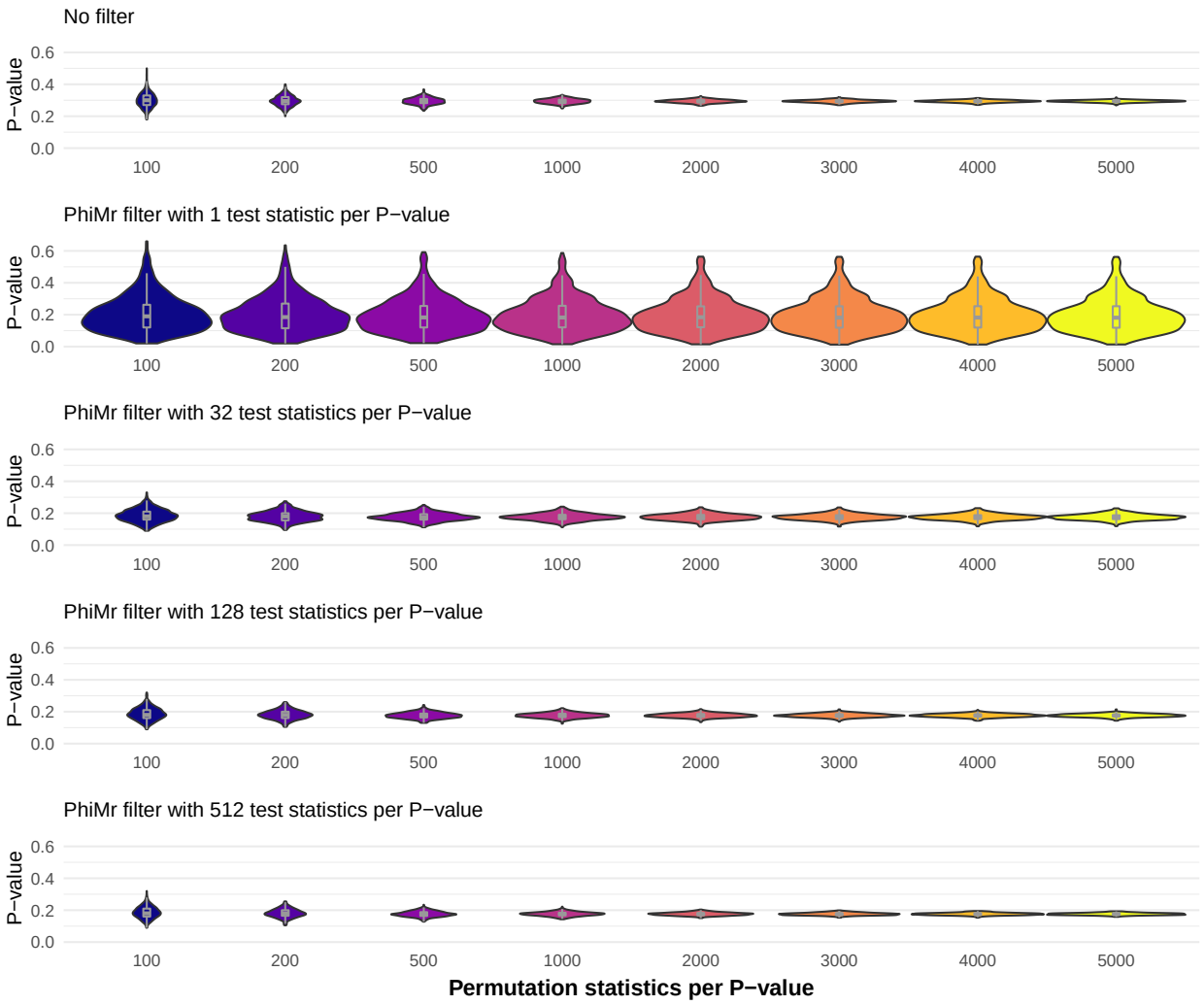
From each row of the plot, it is visibly clear for this data set that when using PhiMr for testing subset selection along with our proposed P-value estimator $\tilde{P}_{T_S,Q}$, the number Q of test statistics has a huge effect on the spread of the estimator's distribution regardless of the number B of permutation statistics used. When compared to AMKAT without PhiMr, the distribution's variance appears much greater at low values of Q but comparable at higher values. This effect of Q on the variance of $\tilde{P}_{T_S,Q}$ appears more pronounced at greater numbers B of permutation statistics.

One notable observation is the fact that the distribution of $\tilde{P}_{T_S,Q}$ appears to converge as $Q \to \infty$ and $B \to \infty$, with a stable center and decreasing variance; furthermore, for sufficient Q (e.g., $Q = 512$) the rate of convergence with respect to B appears similar to that for AMKAT without the PhiMr filter.

For this data, the P-value estimator distribution appears centered at a lower P-value with the PhiMr filter compared to without.

We now consider the same plot, but this time with values of B arranged horizontally and those of Q arranged vertically (also considering slightly different sets of values for each):

```
makeViolinBPlots(pvalues, title_settings = title_violinB(num_replicates))
```



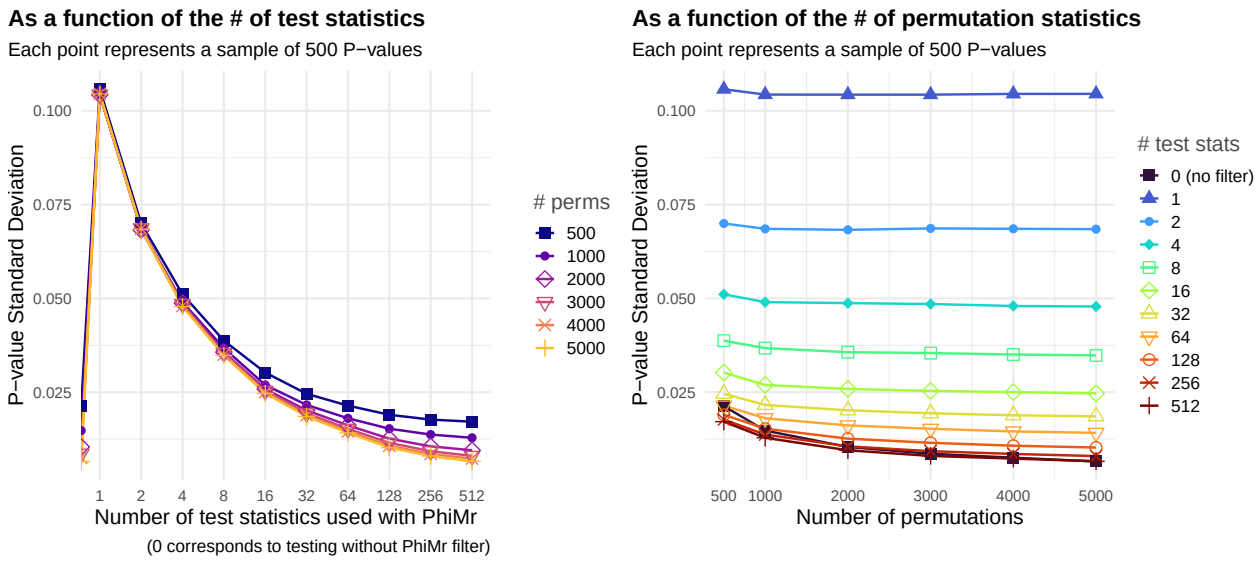
From this plot, we notice that at low values of Q (e.g., second row), the variance of $\tilde{P}_{T_S,Q}$ does not noticeably decrease as we increase the number B of permutation statistics, while at greater values of Q (e.g., bottom row), the effect of B resembles that for the case without the PhiMr filter (top row).

Similarly, we notice that the effect of Q on the variance of $\tilde{P}_{T_S,Q}$ becomes more pronounced at higher values of B (as clearly seen when comparing the leftmost column of the plot, where $B = 100$, to the rightmost column where $B = 5000$).

P-value Mean and Standard Deviation

To further examine the effect of Q and B on the distribution for the P-value estimator $\tilde{P}_{T_S,Q}$, we plot the sample standard deviation of the P-values for each variation of AMKAT below:

```
makePVLinePlots(pvalues, num_replicates)
```



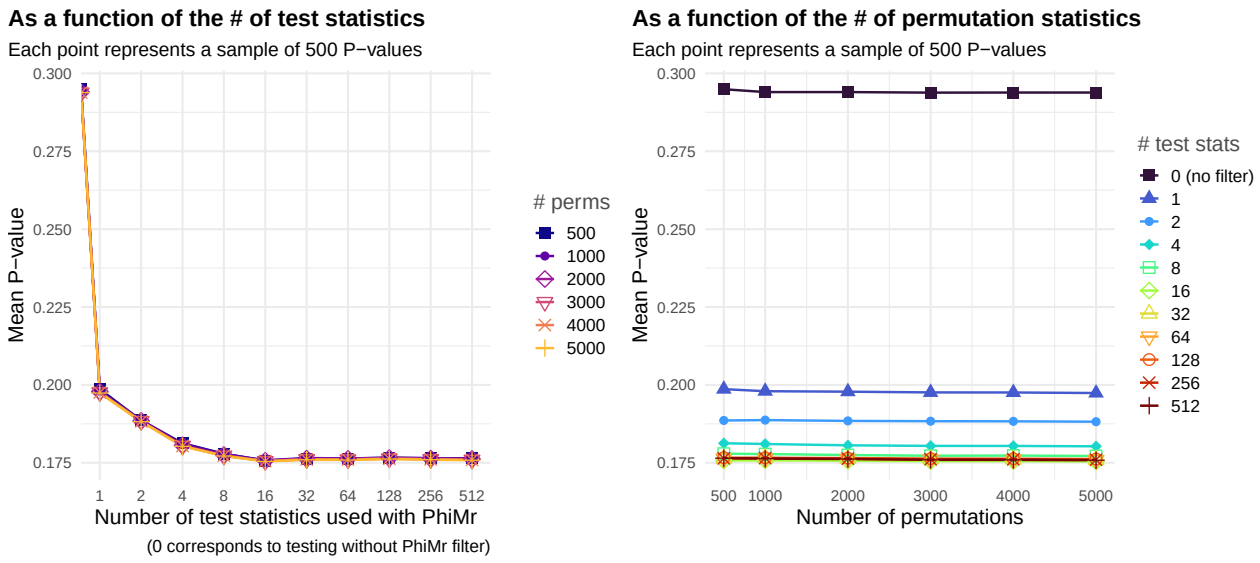
The left-hand plot shows the sample standard deviation across different values of Q (for each value of B), while the right-hand plot shows the sample standard deviation as a function of B (for each value of Q). The influence of Q relative to that of B is quite substantial here; notably, the standard deviation converges quite rapidly with respect to Q, suggesting that an excessive number of test statistics is not needed to achieve an acceptable degree of variation in P-value.

For this data set, $\tilde{P}_{T_S,Q}$ at $Q \geq 256$ had comparable standard deviation to the P-value estimator for AMKAT without the PhiMr filter.

We also see other trends observed earlier: the standard deviation decreases steadily as both Q and B increase; also, at $Q = 1$, the standard deviation is near-constant with respect to B, while at high values of Q the relationship more closely matches that for AMKAT without the PhiMr filter.

Next, we consider the same plot, but this time for the sample mean of each P-value distribution:

```
makePVLinePlots(pvalues, num_replicates, summary_stat = mean,
                title_settings =
                  title_pvline(
                    joint_title = "AMKAT Mean P-value on a Fixed Data Set"))
```



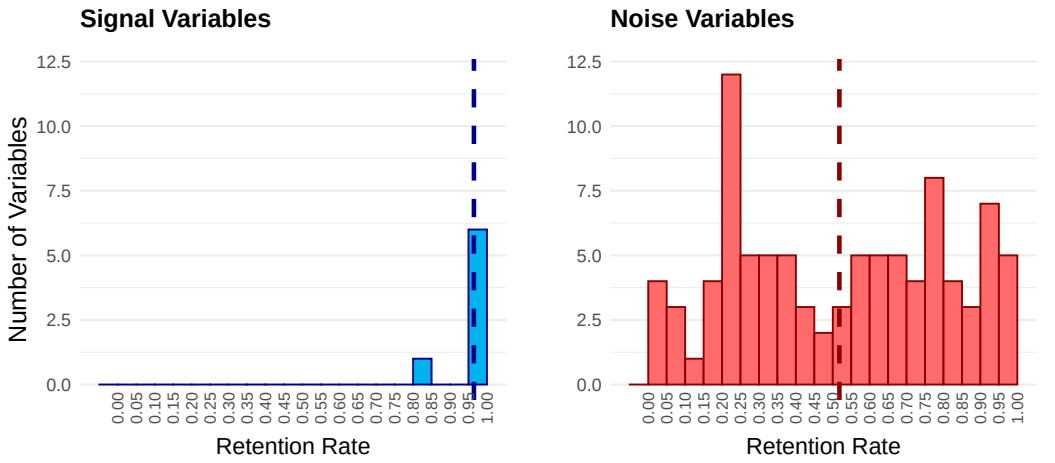
From the right-hand plot we see that the mean appears stable across values of B, both with PhiMr at each value of Q as well as without PhiMr.

From the left-hand plot, we see that the mean P-value is considerably lower for AMKAT with PhiMr across all values of Q when compared to AMKAT without PhiMr ($Q = 0$). The mean P-value steadily decreases until around $Q = 16$, beyond which it remains stable.

Distribution of Variable Retention Rates by PhiMr

To help inform our understanding of the observed differences in behavior between AMKAT with and without the PhiMr filter for this particular data set, we plot a histogram (for all signal variables and all noise variables, respectively) of the variable retention rate by PhiMr across the 256,000 values of T_S (each of which involves a separate application of PhiMr to this set of data).

```
makePhimrHistograms(feature_select_rates,
                    getSignalIndices(x_type, signal_density))
```



For this data set, which included 7 signal variables and 93 noise variables, most signal variables were retained at a very high rate across the 256,000 PhiMr applications, with a mean rate above 0.95 across all 7 signals, which is while the retention rate for noise variables appears distributed fairly uniformly over rates from 0 to 1, with a mean retention rate close to 0.5.

The high average retention rate for the signal variables relative to the noise variables for this data set suggests that the PhiMr filter is able to consistently improve the ratio of signal to noise for this data set; this is consistent with the observation for this data set that PhiMr yielded smaller P-values on average than AMKAT without PhiMr.

Data Set #2

This data set was generated using the following scenario parameters:

```
n <- 50 # sample size
p <- 150 # feature dimension
signal_density <- 'sparse' # 7 signal variables
error_correlation_strength <- 0.5 # mixed directions
```

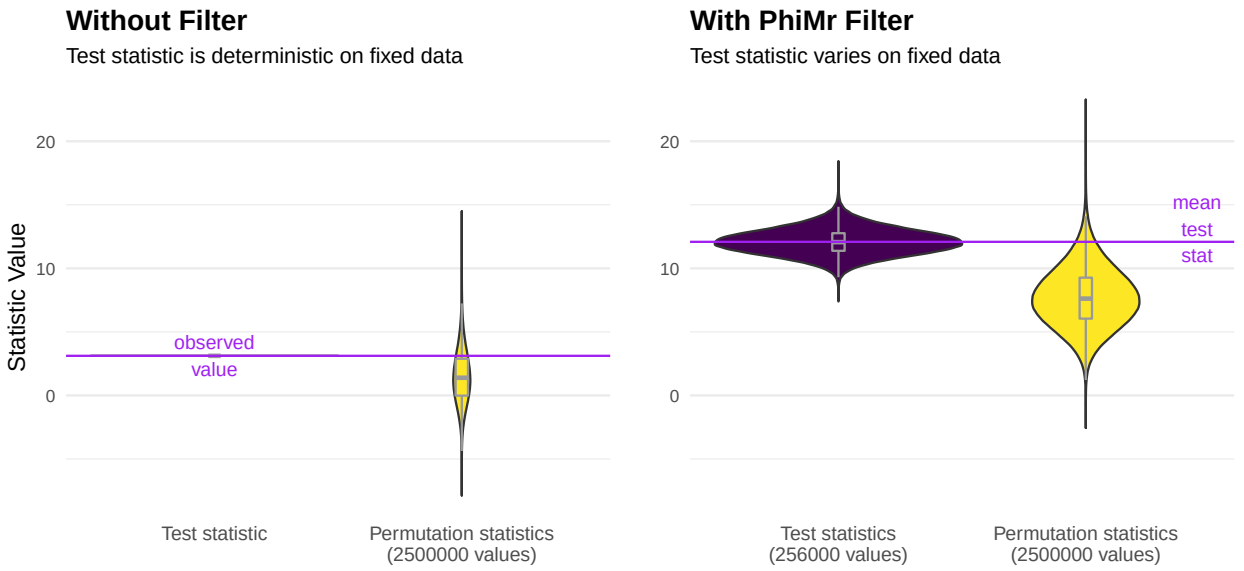
We load the test and permutation statistics generated on this data set:

```
pv_files <- pvaluePlotFiles()
load(pv_files$stats)
load(pv_files$pvalues)
```

Distribution of Test and Permutation Statistics

We first plot the distributions of the test statistics and permutation statistics with and without the PhiMr filter:

```
makeStatDistrPlots(
  test_stats, perm_stats, perm_stats_no_filter, test_stat_no_filter)
```



Here the distribution of the test statistic with PhiMr (right panel) is unimodal and symmetric as seen in the previous data set.

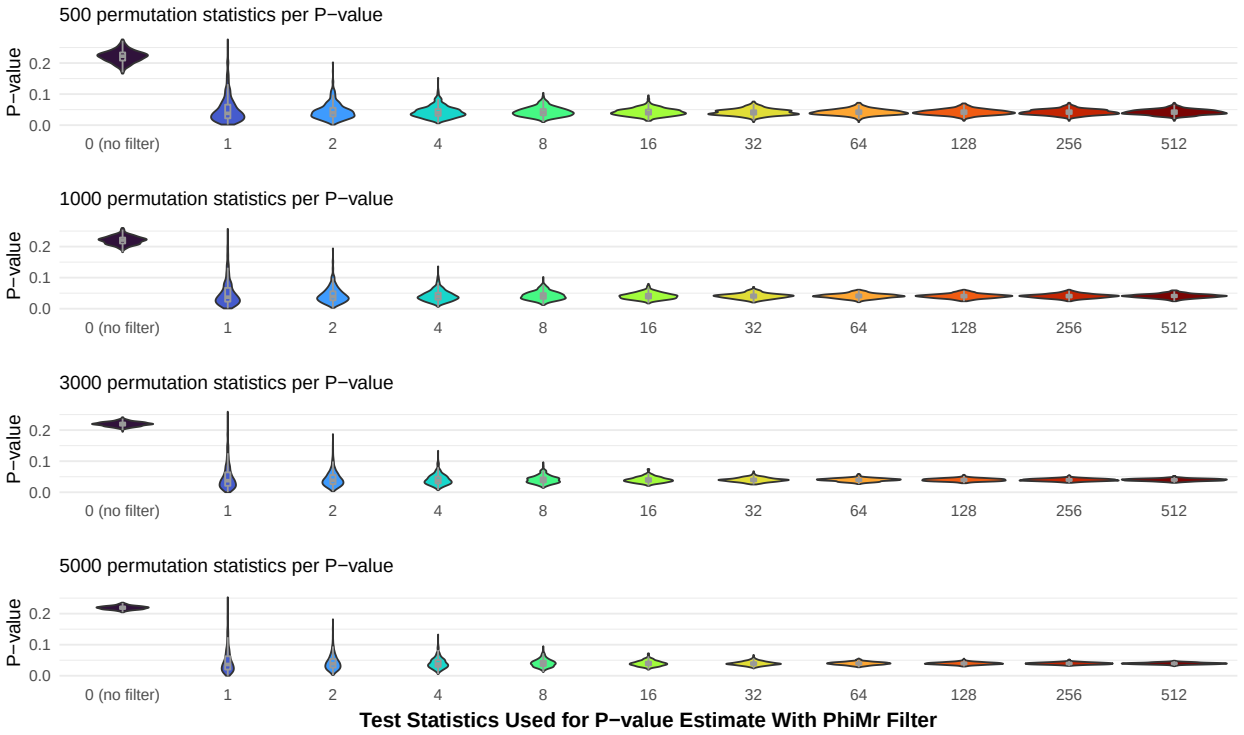
From the boxplots overlayed atop the violin plots, we can see that without the PhiMr filter (left panel), the observed test statistic T is roughly at the third quartile of the sample distribution for the permutation statistics $\ddot{T}$, implying a P-value estimate of roughly 0.25.

For the case with PhiMr (right panel), the sample mean of the 256,000 values of the test statistic $T_{\mathcal{S}}$ lies significantly further into the upper tail of the respective permutation statistic distribution, suggesting a substantially smaller P-value estimate using $\tilde{P}_{T_{\mathcal{S}},Q}$ (i.e., with the PhiMr filter), on average; additionally, for this data set, most of the values of $T_{\mathcal{S}}$ appear to lie beyond the third quartile of the distribution for $\ddot{T}_{\mathcal{S}}$, implying that using PhiMr consistently results in a lower P-value for this data set, even at a small number Q of test statistics.

Distribution of P-values

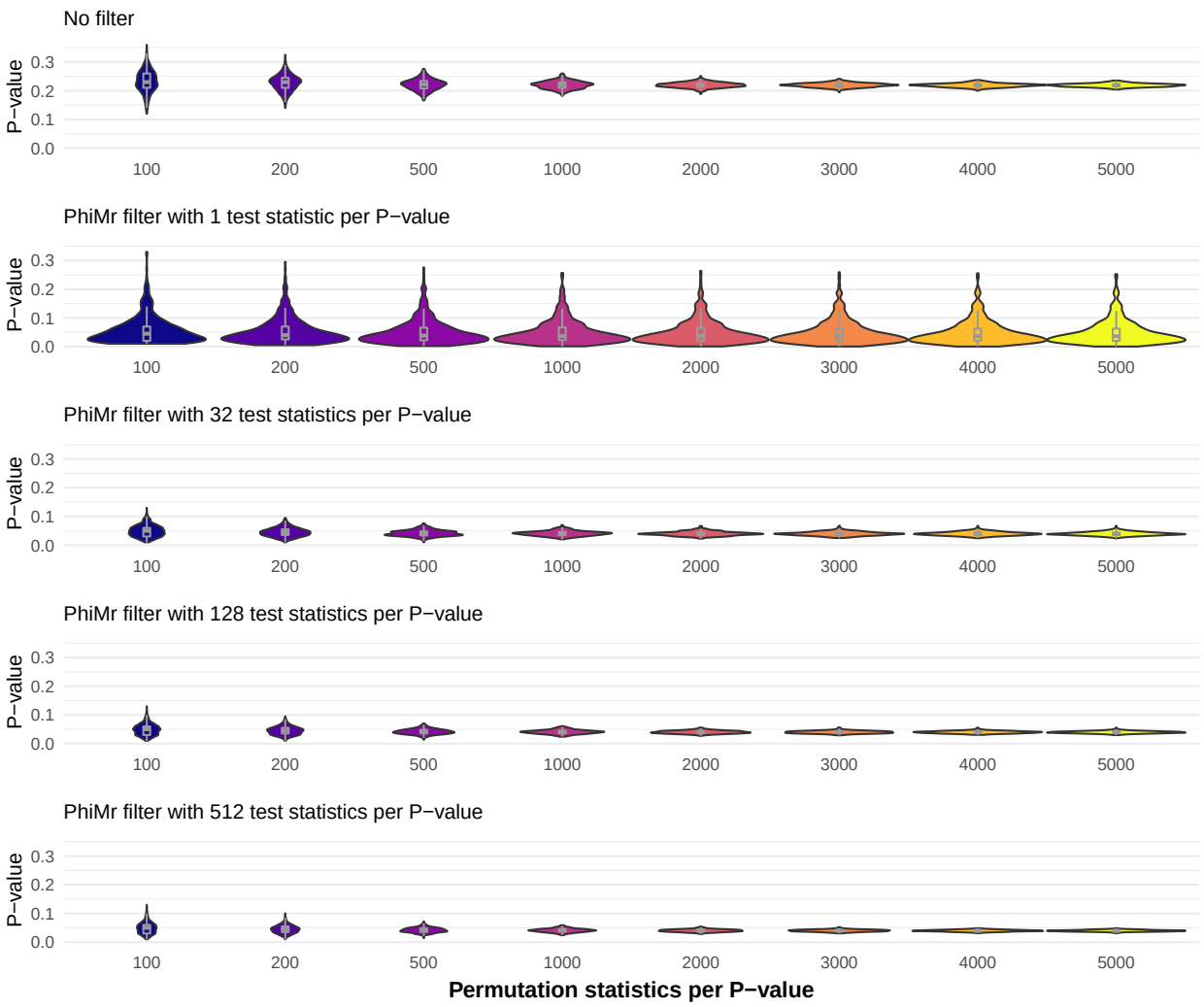
After partitioning the test and permutation statistics into batches and using them to estimate P-values, we plot the distributions of the resulting P-values.

```
makeViolinQPlots(pvalues, title_settings = title_violinQ(num_replicates))
```



Here we a situation not starkly different from that seen for the previous data set. P-values are smaller with PhiMr than without; the variance with PhiMr is greater than without PhiMr for small values of Q, but as Q increases variance with PhiMr steadily and substantially decreases; and the distribution of the P-values with PhiMr appears to converge as B and Q tend to infinity.

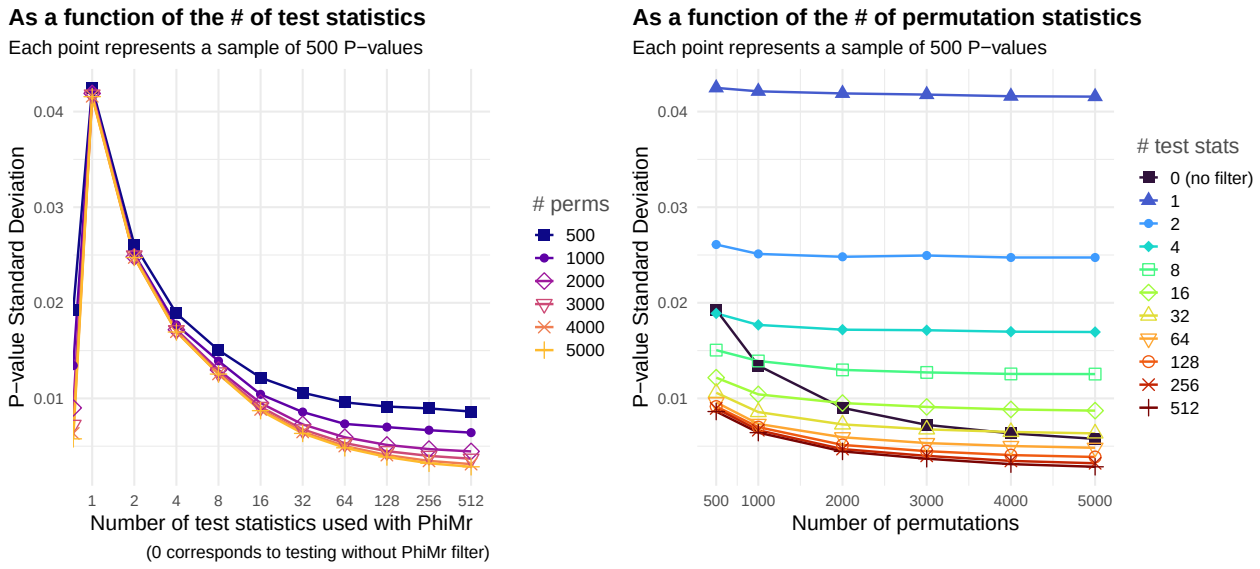
```
makeViolinBPlots(pvalues, title_settings = title_violinB(num_replicates))
```



For this plot, which more clearly highlights the effect of B, we again notice, similar to the previous data set, that the number B of permutation statistics has little effect on the variance of the P-value when $Q = 1$ (second row), while at greater values of Q (subsequent rows) we notice a clear reduction in variance as B increases.

P-value Mean and Standard Deviation

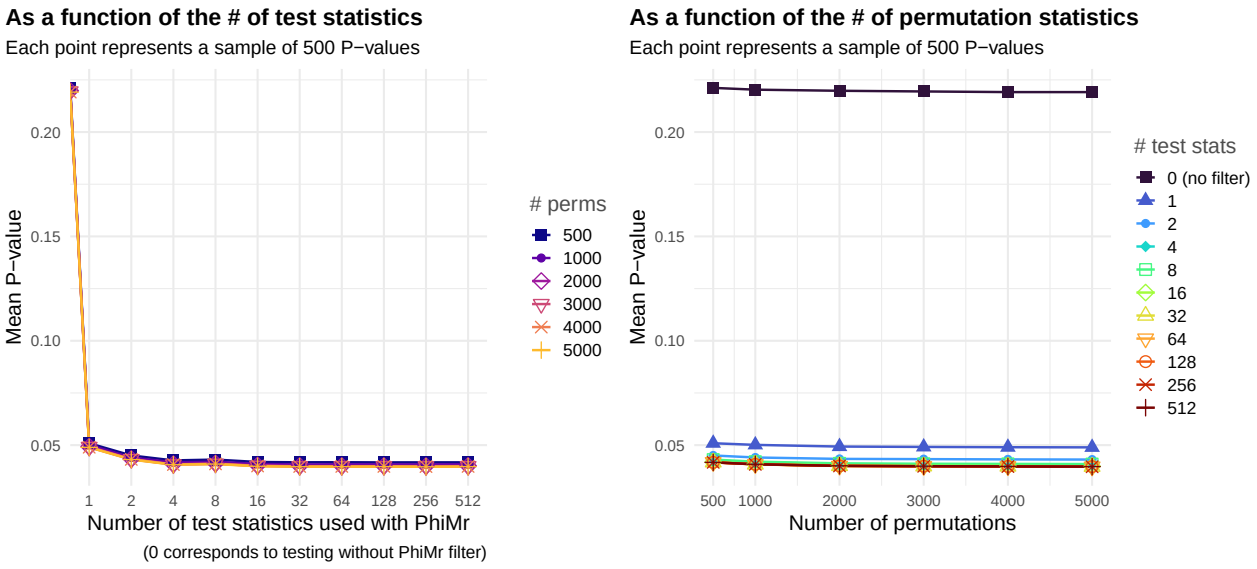
```
makePVLinePlots(pvalues, num_replicates)
```



From the right panel we notice that for $Q \geq 64$, the P-value standard deviation with PhiMr is less than that without PhiMr at all numbers B of permutation statistics considered.

At very low values of Q, the standard deviation of the P-value with PhiMr is substantially greater than that without PhiMr, similar to the behavior seen in the previous data set. This reinforces the case to recommend using the estimator $\tilde{P}_{T_S,Q}$ and averaging a sufficient number Q of test statistics when testing with the PhiMr filter.

```
makePVLinePlots(pvalues, num_replicates, summary_stat = mean,
                title_settings =
                  title_pvline(
                    joint_title = "AMKAT Mean P-value on a Fixed Data Set"))
```



Consistent with our observations from the first plot depicting the test statistic and permutation statistic distributions, the mean P-value without the PhiMr filter is close to 0.25. Meanwhile, the mean P-value with PhiMr was 0.05 at $Q = 1$ and decreased to around 0.04 at $Q = 4$, remaining stable as Q (and B) increased further:

```
# Mean P-value with PhiMr at Q=4 and B=1000
mean(pvalues["1000", "PF-4", ])
```

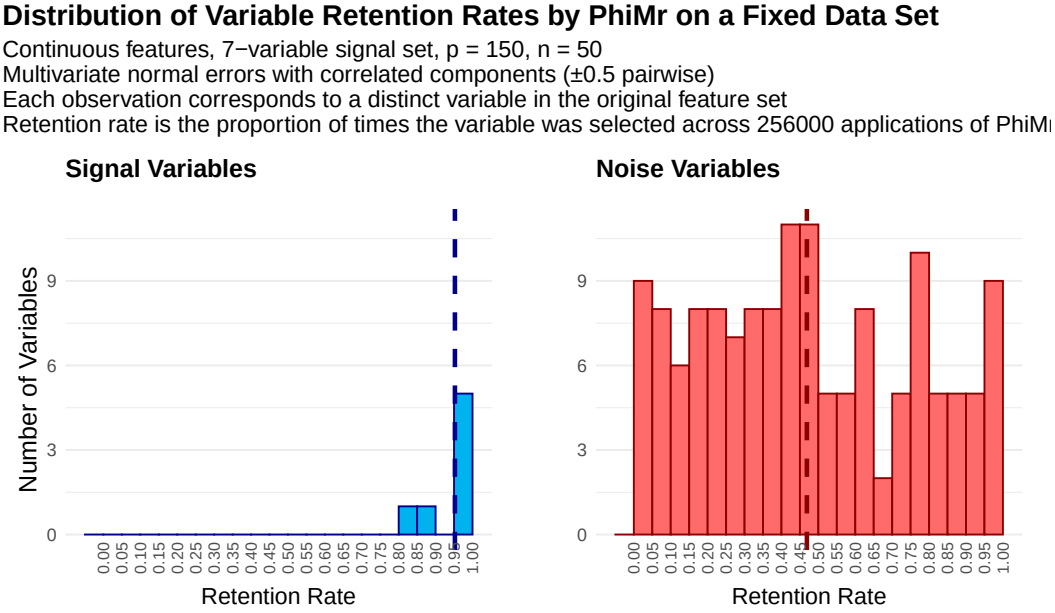
```
## [1] 0.04179
```

```
# Mean P-value with PhiMr at Q=512 and B=5000
mean(pvalues["5000", "PF-512", ])
```

```
## [1] 0.0397012
```

Distribution of Variable Retention Rates by PhiMr

```
makePhimrHistograms(feature_select_rates,
                    getSignalIndices(x_type, signal_density))
```



This data set contains 7 signal variables and 143 noise variables. The situation here very closely resembles that seen for the previous data set; the mean retention rate for the signals is close to 1, while the mean retention rate for the noise variables is around 0.5, with a uniform distribution of rates from 0 to 1. We again see this difference in mean retention rate between signal and noise being consistent with the difference in mean P-value observed between AMKAT with and without the PhiMr filter.

Data Set #3

This data set was generated using the following scenario parameters:

```
n <- 50 # sample size
p <- 200 # feature dimension
signal_density <- 'sparse' # 7 signal variables
error_correlation_strength <- 0.5 # mixed directions
```

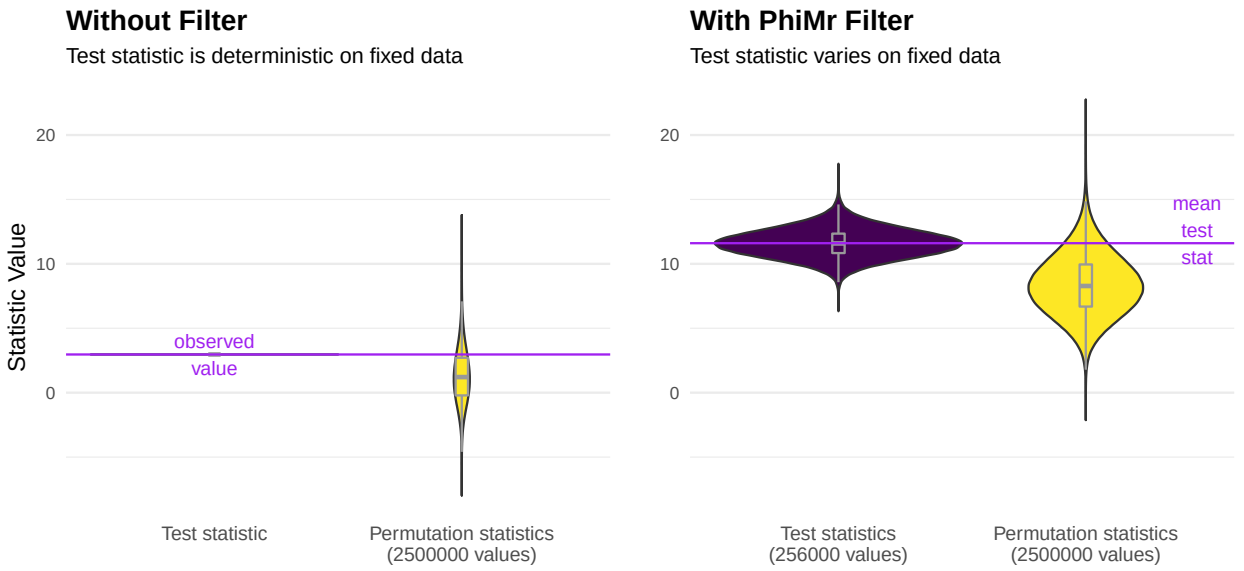
We load the test and permutation statistics generated on this data set:

```
pv_files <- pvaluePlotFiles()
load(pv_files$stats)
load(pv_files$pvalues)
```

Distribution of Test and Permutation Statistics

We first plot the distributions of the test statistics and permutation statistics with and without the PhiMr filter:

```
makeStatDistrPlots(
  test_stats, perm_stats, perm_stats_no_filter, test_stat_no_filter)
```

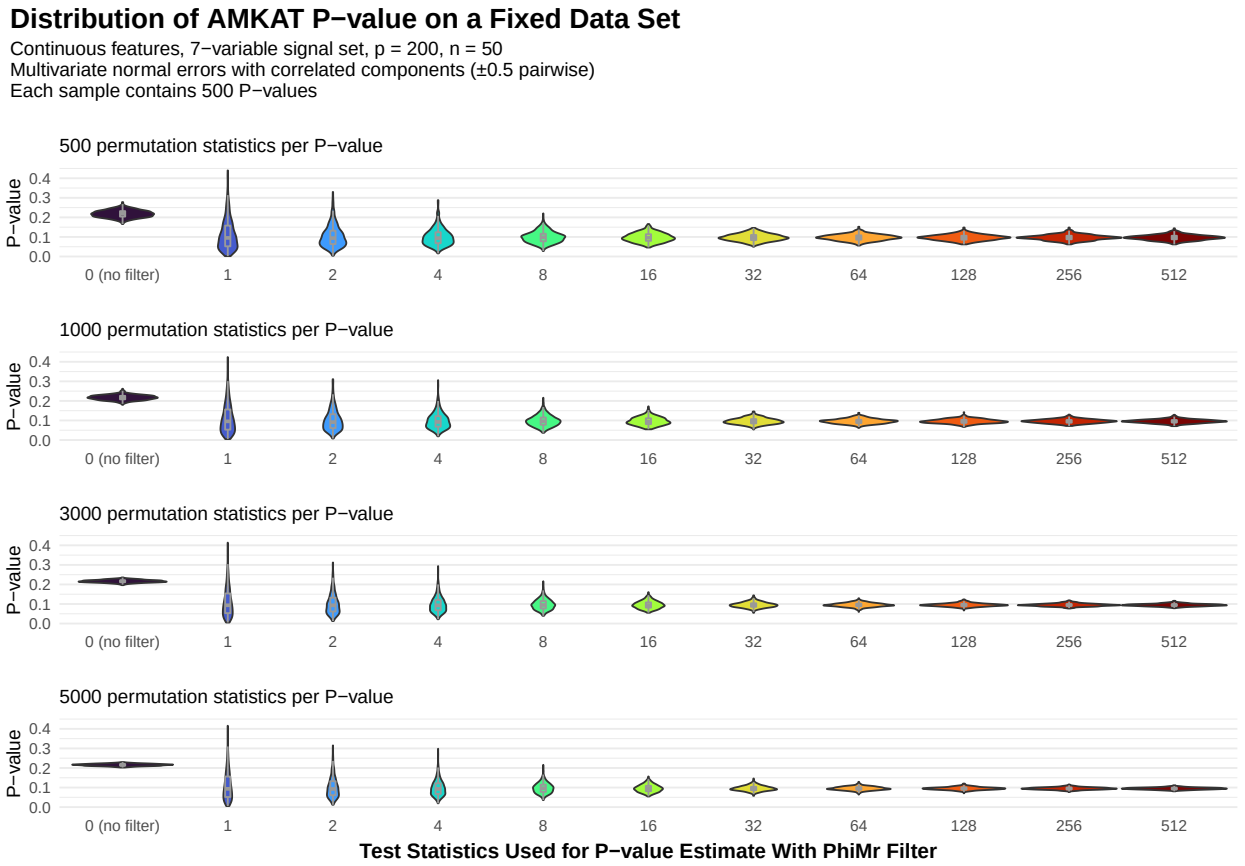


The distributions in both plots are similar to their counterparts seen in the previous data set, though here in the right panel, the distribution of test statistics is not centered quite as far into the tail of the permutation statistic distribution, and not as many of the values appear to fall beyond the third quartile.

Distribution of P-values

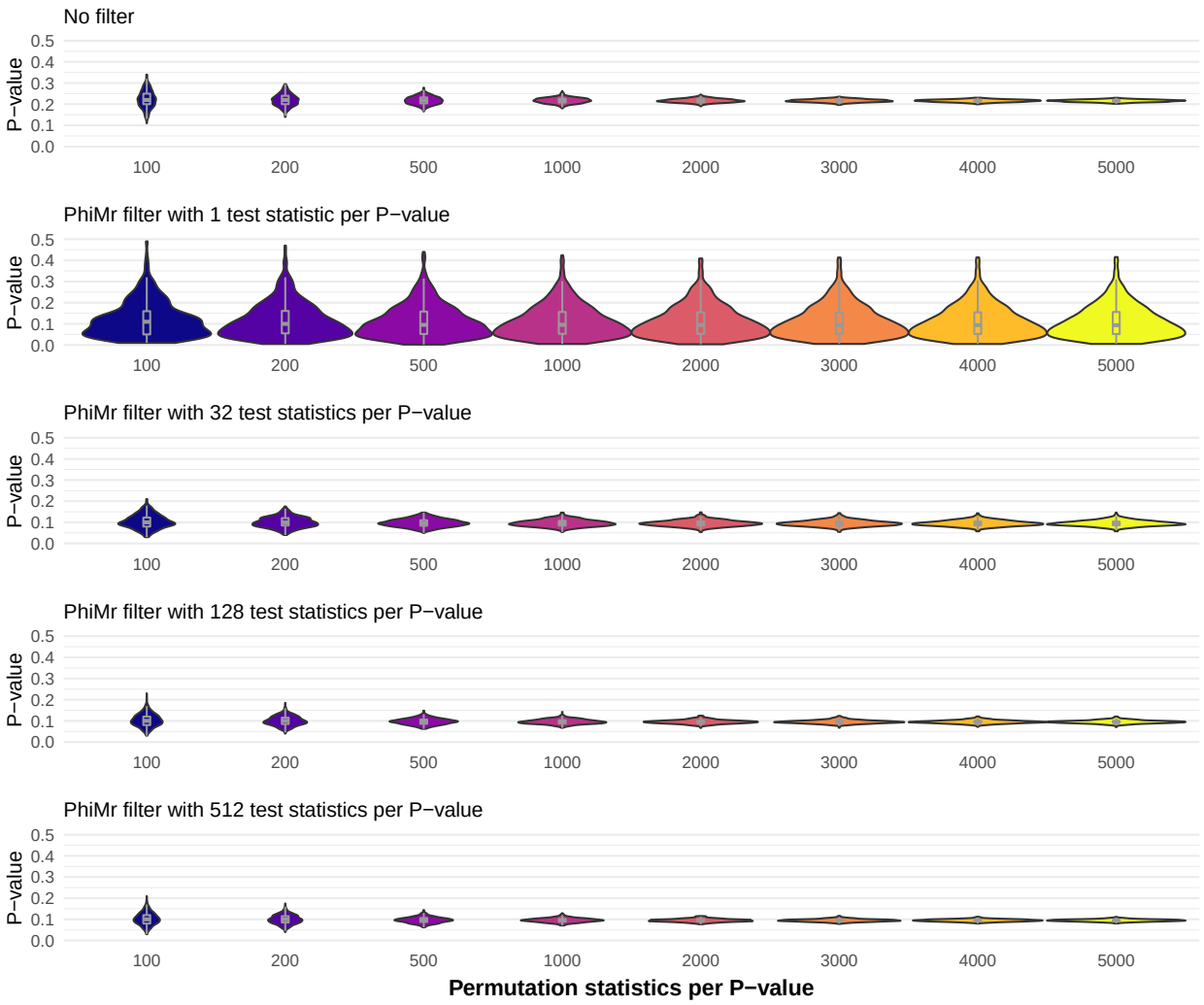
After partitioning the test and permutation statistics into batches and using them to estimate P-values, we plot the distributions of the resulting P-values.

```
makeViolinQPlots(pvalues, title_settings = title_violinQ(num_replicates))
```



We see trends similar to those observed in the previous data set: P-values are smaller with PhiMr than without; the variance with PhiMr is greater than without PhiMr for small values of Q, but as Q increases variance with PhiMr steadily and substantially decreases; and the distribution of the P-values with PhiMr appears to converge as B and Q tend to infinity.

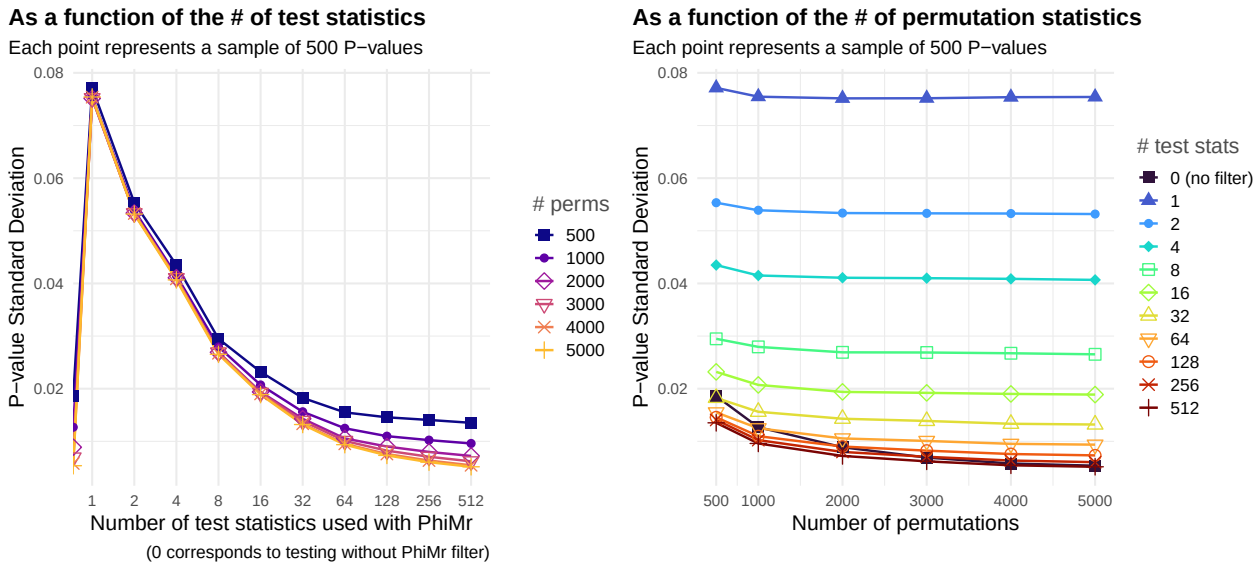
```
makeViolinBPlots(pvalues, title_settings = title_violinB(num_replicates))
```



For this plot, which more clearly highlights the effect of B, we again notice, similar to the previous data set, that the number B of permutation statistics has little effect on the variance of the P-value when $Q = 1$ (second row), while at greater values of Q (subsequent rows) we notice a clear reduction in variance as B increases.

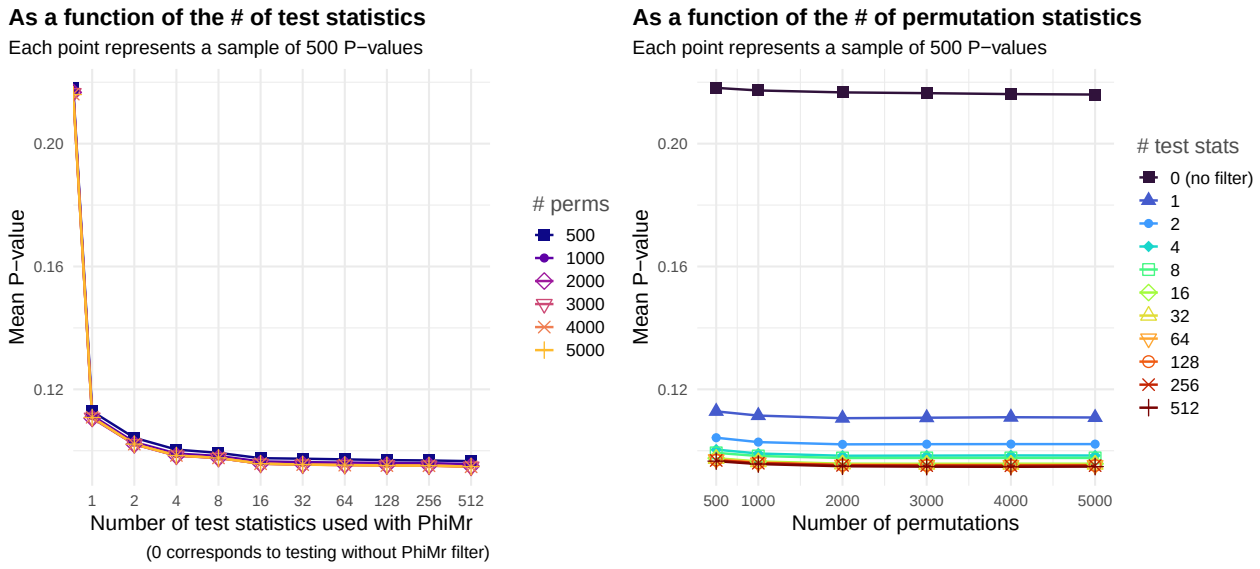
P-value Mean and Standard Deviation

```
makePVLinePlots(pvalues, num_replicates)
```



Here the difference between AMKAT with and without PhiMr at high values of Q is minimal, resembling the case for data set #1 more than that for the previous data set (#2), where the P-value with PhiMr had lower standard deviation for sufficiently large Q. At $B \leq 3000$, the standard deviation with PhiMr is lower given sufficient Q, but the difference is relatively modest. For small Q, the standard deviation is substantially greater with PhiMr than without.

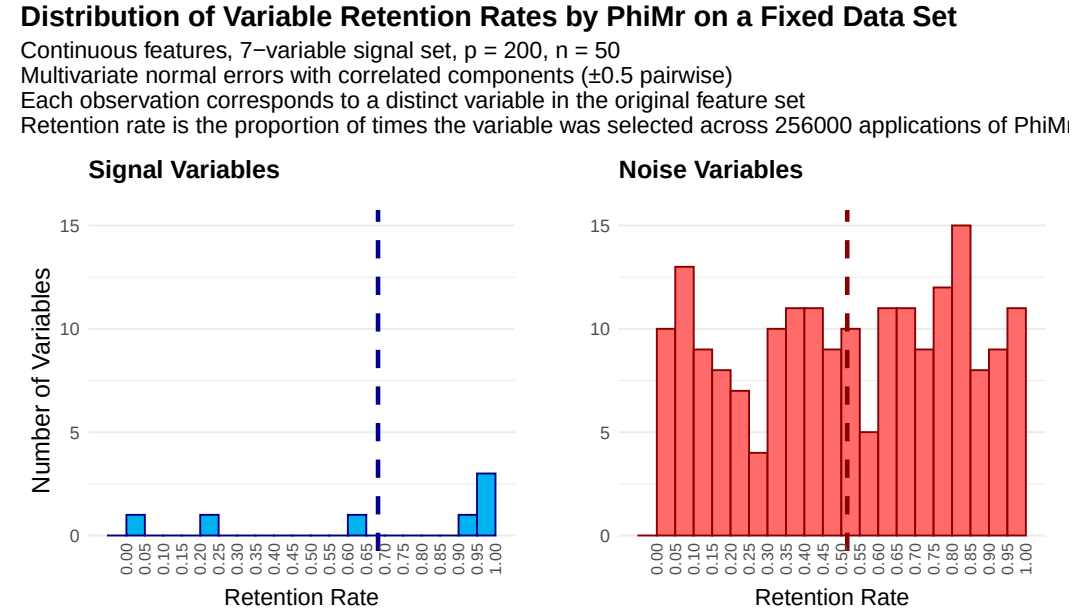
```
makePVLinePlots(pvalues, num_replicates, summary_stat = mean,
                title_settings =
                  title_pvline(
                    joint_title = "AMKAT Mean P-value on a Fixed Data Set"))
```



Here the situation is similar to that seen in the previous data set, with the mean P-value much lower with PhiMr than without, and converging rapidly with respect to Q.

Distribution of Variable Retention Rates by PhiMr

```
makePhimrHistograms(feature_select_rates,
                    getSignalIndices(x_type, signal_density))
```



This data set contains 7 signal variables and 193 noise variables. Interestingly, as compared to the first two

data sets, the mean retention rate for the signals was significantly lower, with two of the signal variables being selected at very low rates:

```
# Mean retention rate across all signals
mean(feature_select_rates[getSignalIndices(x_type, signal_density)])
```

```
## [1] 0.6817985
```

Nonetheless, the mean retention rate for the signals, at around 0.68, was still greater than that for the noise, for which the distribution of retention rates was consistent with the previous data sets.

Data Set #4

This data set was generated using the following scenario parameters:

```
n <- 100 # sample size
p <- 500 # feature dimension
signal_density <- 'dense' # 80 signal variables
error_correlation_strength <- 0.5 # mixed directions
```

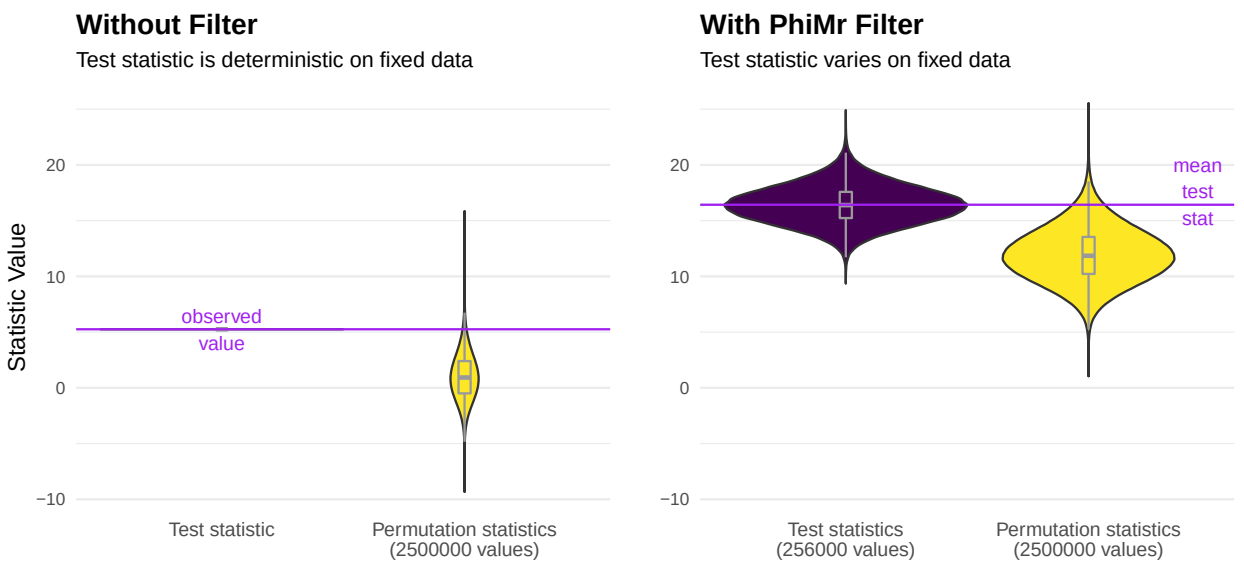
We load the test and permutation statistics generated on this data set:

```
pv_files <- pvaluePlotFiles()
load(pv_files$stats)
load(pv_files$pvalues)
```

Distribution of Test and Permutation Statistics

We first plot the distributions of the test statistics and permutation statistics with and without the PhiMr filter:

```
makeStatDistrPlots(
  test_stats, perm_stats, perm_stats_no_filter, test_stat_no_filter)
```

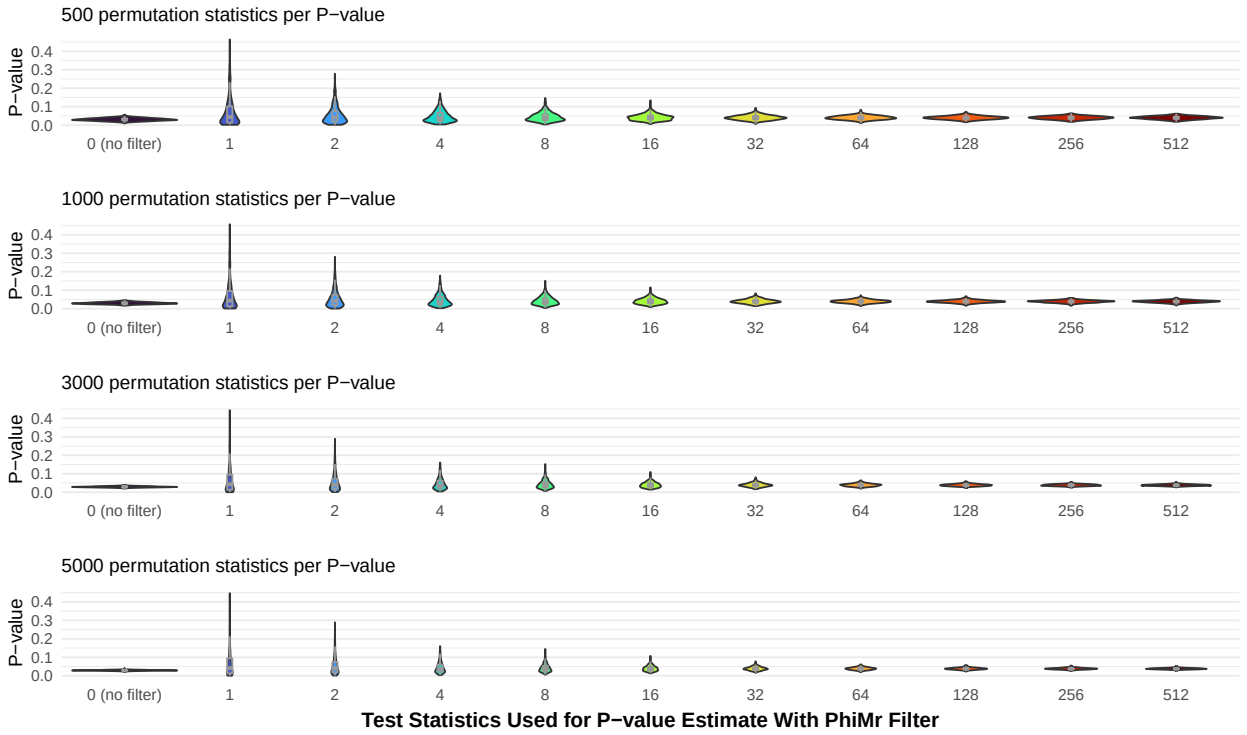


Here the test statistic T without the PhiMr filter (left panel) lies relatively far into the tail of the corresponding permutation statistic distribution, while most of the test statistics $T_{\mathcal{S}}$ generated with PhiMr (right panel) appear to lie relatively closer to the center of their corresponding permutation statistic distribution, suggesting that AMKAT will yield lower P-values without PhiMr than with PhiMr for this data set.

Distribution of P-values

After partitioning the test and permutation statistics into batches and using them to estimate P-values, we plot the distributions of the resulting P-values.

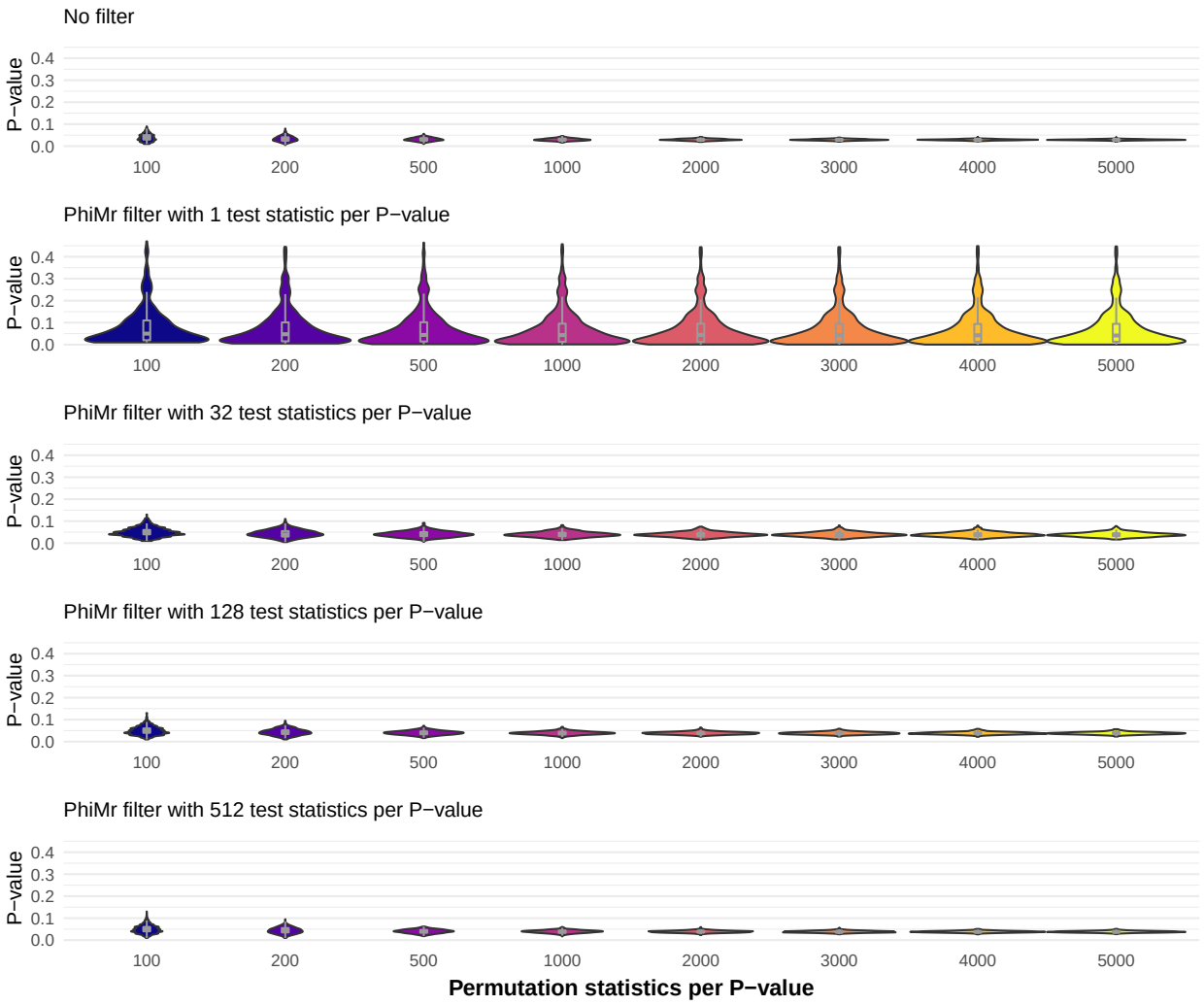
```
makeViolinQPlots(pvalues, title_settings = title_violinQ(num_replicates))
```



We see trends similar to those observed in previous data sets: the variance with PhiMr is greater than without PhiMr for small values of Q, but as Q increases variance with PhiMr steadily and substantially decreases, and the distribution of the P-values with PhiMr appears to converge as B and Q tend to infinity.

Different from previous data sets, the mean P-value for AMKAT without the filter appears potentially lower than that for AMKAT with PhiMr.

```
makeViolinBPlots(pvalues, title_settings = title_violinB(num_replicates))
```



For this plot, which more clearly highlights the effect of B, we again notice, similar to previous data sets, that the number B of permutation statistics has little effect on the variance of the P-value when $Q = 1$ (second row), while at greater values of Q (subsequent rows) we notice a clear reduction in variance as B increases.

P-value Mean and Standard Deviation

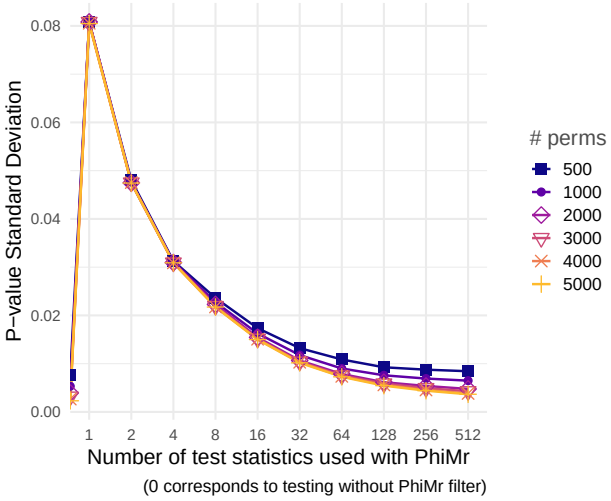
```
makePVLinePlots(pvalues, num_replicates)
```

Standard Deviation of AMKAT P−value on a Fixed Data Set

Continuous features, 80−variable signal set, p = 500, n = 100
Multivariate normal errors with correlated components (±0.5 pairwise)

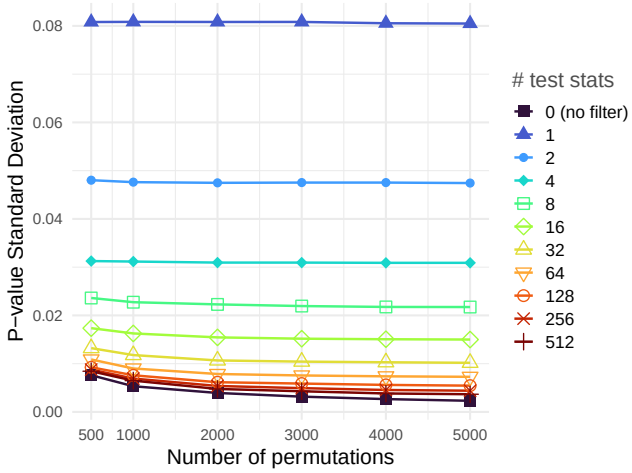
As a function of the # of test statistics

Each point represents a sample of 500 P−values



As a function of the # of permutation statistics

Each point represents a sample of 500 P−values



For this data set, the P-value standard deviation with PhiMr was much greater than that without PhiMr when the number Q of test statistics was small, but was comparable when Q was large. The number B of permutation statistics had relatively little impact on the P-value standard deviation, especially with PhiMr at low values of Q.

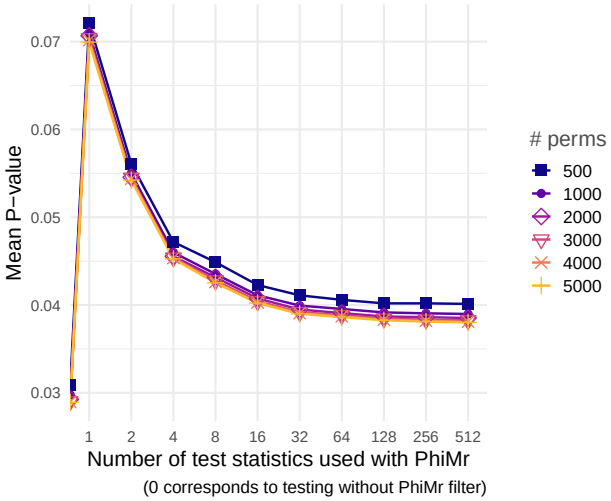
```
makePVLinePlots(pvalues, num_replicates, summary_stat = mean,
                title_settings =
                  title_pvline(
                    joint_title = "AMKAT Mean P-value on a Fixed Data Set"))
```

AMKAT Mean P−value on a Fixed Data Set

Continuous features, 80−variable signal set, p = 500, n = 100
Multivariate normal errors with correlated components (±0.5 pairwise)

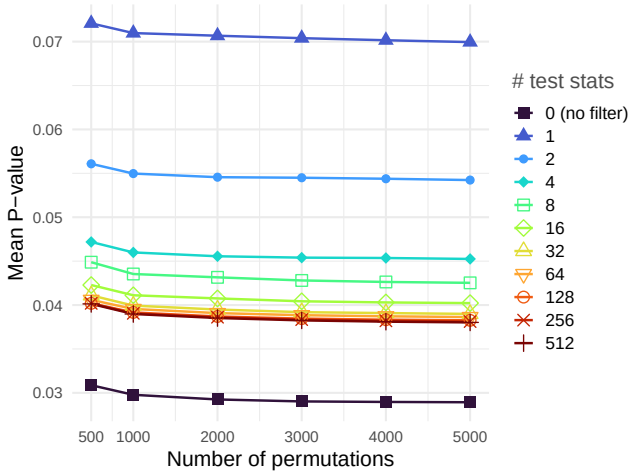
As a function of the # of test statistics

Each point represents a sample of 500 P−values



As a function of the # of permutation statistics

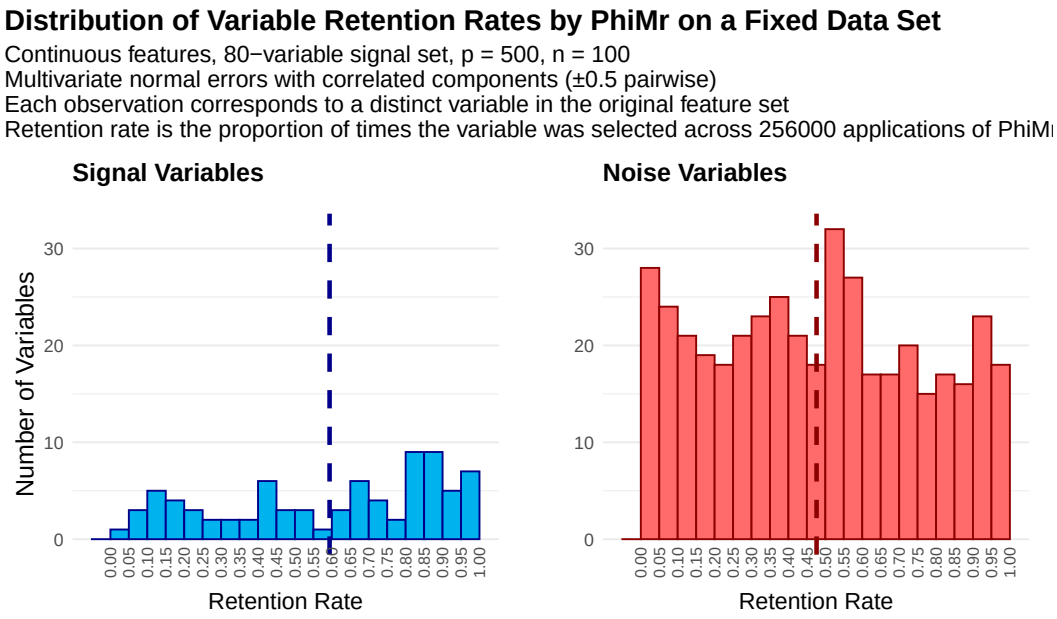
Each point represents a sample of 500 P−values



The mean P-value without PhiMr was lower (around 0.03) than that with PhiMr (around 0.04 for $Q \geq 16$ and $B \geq 1000$); here the mean P-value with PhiMr converged more slowly with respect to Q, and so the value of Q had a more substantial impact on the mean P-value over the values considered in the simulation. The number B of permutation statistics had very little impact on mean P-value, with the effect similar across variations of AMKAT.

Distribution of Variable Retention Rates by PhiMr

```
makePhimrHistograms(feature_select_rates,
                    getSignalIndices(x_type, signal_density))
```



This data set differs from the previous three in the number of signal variables; here there are 80 signal variables and 420 noise variables. The strength of the overall signal here is similar to that for the 7-variable signal; the similar signal strength distributed across a far greater number of variables results in extremely weak marginal signals and far greater difficulty in discriminating signal from noise.

This is reflected in the signal retention rate distribution for the signal variables (left panel), for which the mean retention rate was around 0.60 and a substantial share of signals were retained at very low rates.

The distribution for noise is similar to that seen in previous data sets, being fairly uniform with a mean retention rate near 0.5. The difference in the mean retention rate between the two groups is substantially lower here than was typically observed when the marginal signals were relatively strong, and a substantial share of signals are often lost after applying PhiMr; this is consistent with the observations for this data set of the PhiMr filter leading to higher P-values, on average, compared to AMKAT without PhiMr.

Discrete Features

We initialize the simulation setting for the case where the components of $\boldsymbol{X}$ represent discrete additive-encoded (0-1-2 quantitative minor allele count) single-nucleotide polymorphism (SNP) data:

```
x_type <- 'snp'
p <- 567
source(file.path(dir_src, "initialize_adaptive_within.R"))
```

Data Set #5

This data set was generated using the following scenario parameters:

```
n <- 45 # sample size
signal_density <- 'sparse' # 28 signal variables
signal_correlation <- 'high' # correlations between signal variables
error_correlation_strength <- 0.5 # mixed directions
```

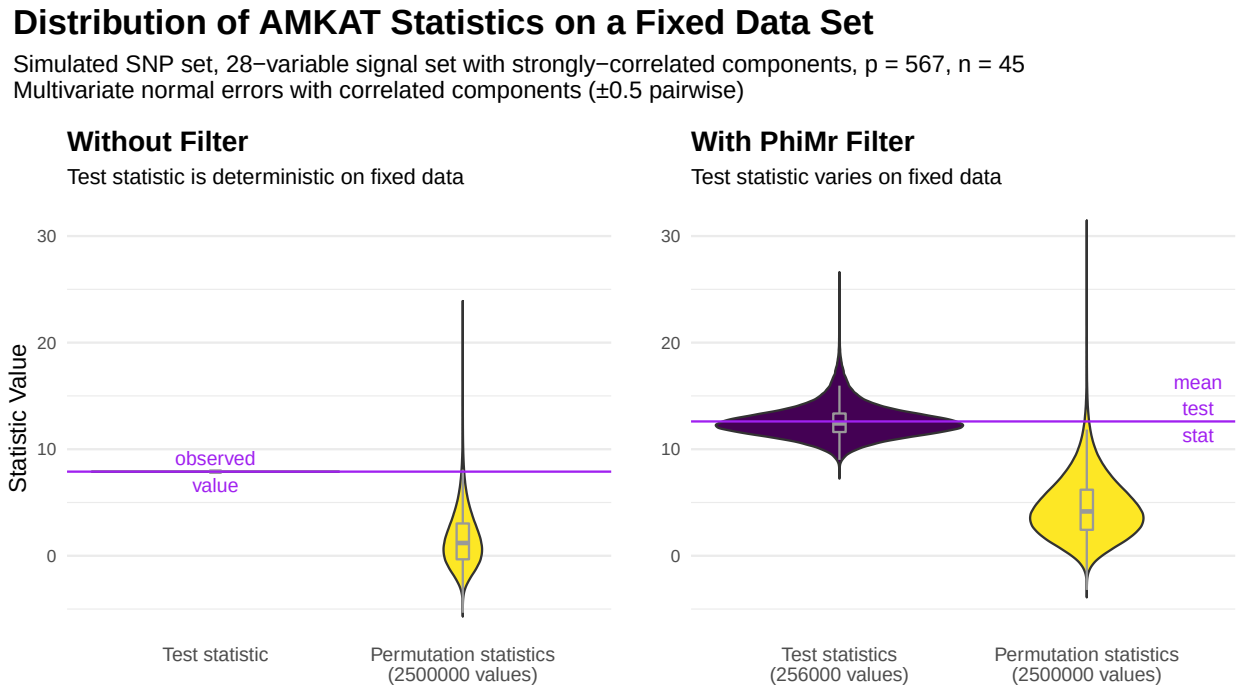
We load the test and permutation statistics generated on this data set:

```
pv_files <- pvaluePlotFiles()
load(pv_files$stats)
load(pv_files$pvalues)
```

Distribution of Test and Permutation Statistics

We first plot the distributions of the test statistics and permutation statistics with and without the PhiMr filter:

```
makeStatDistrPlots(
  test_stats, perm_stats, perm_stats_no_filter, test_stat_no_filter)
```

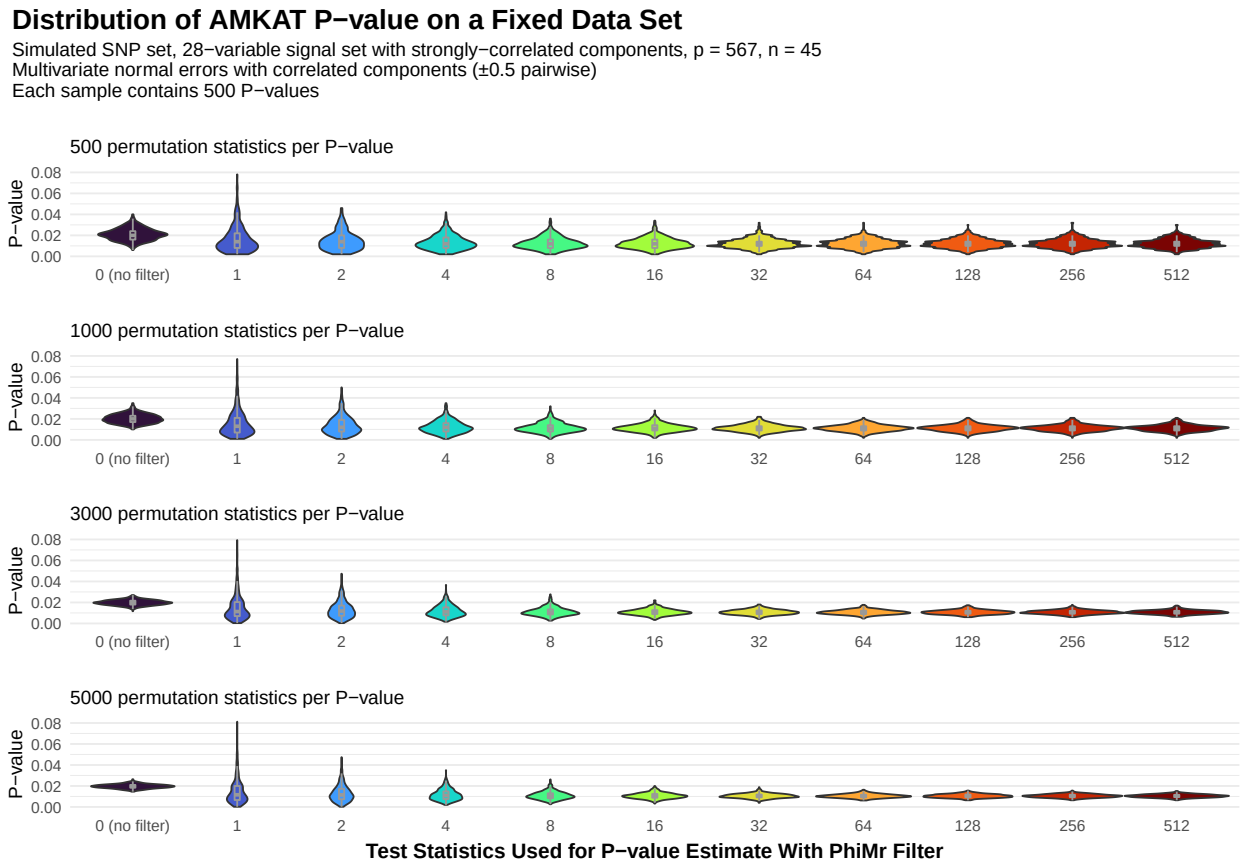


Here the test statistic with PhiMr has a unimodal distribution as was observed for the data sets where $\boldsymbol{X}$ was continuous, although here the distribution of the test statistic exhibits a slight right-skew; the permutation statistic distributions also appear more heavily skewed than we observed for continuous features. Both variations of AMKAT appear to result in similar P-values for this data set based on the plot.

Distribution of P-values

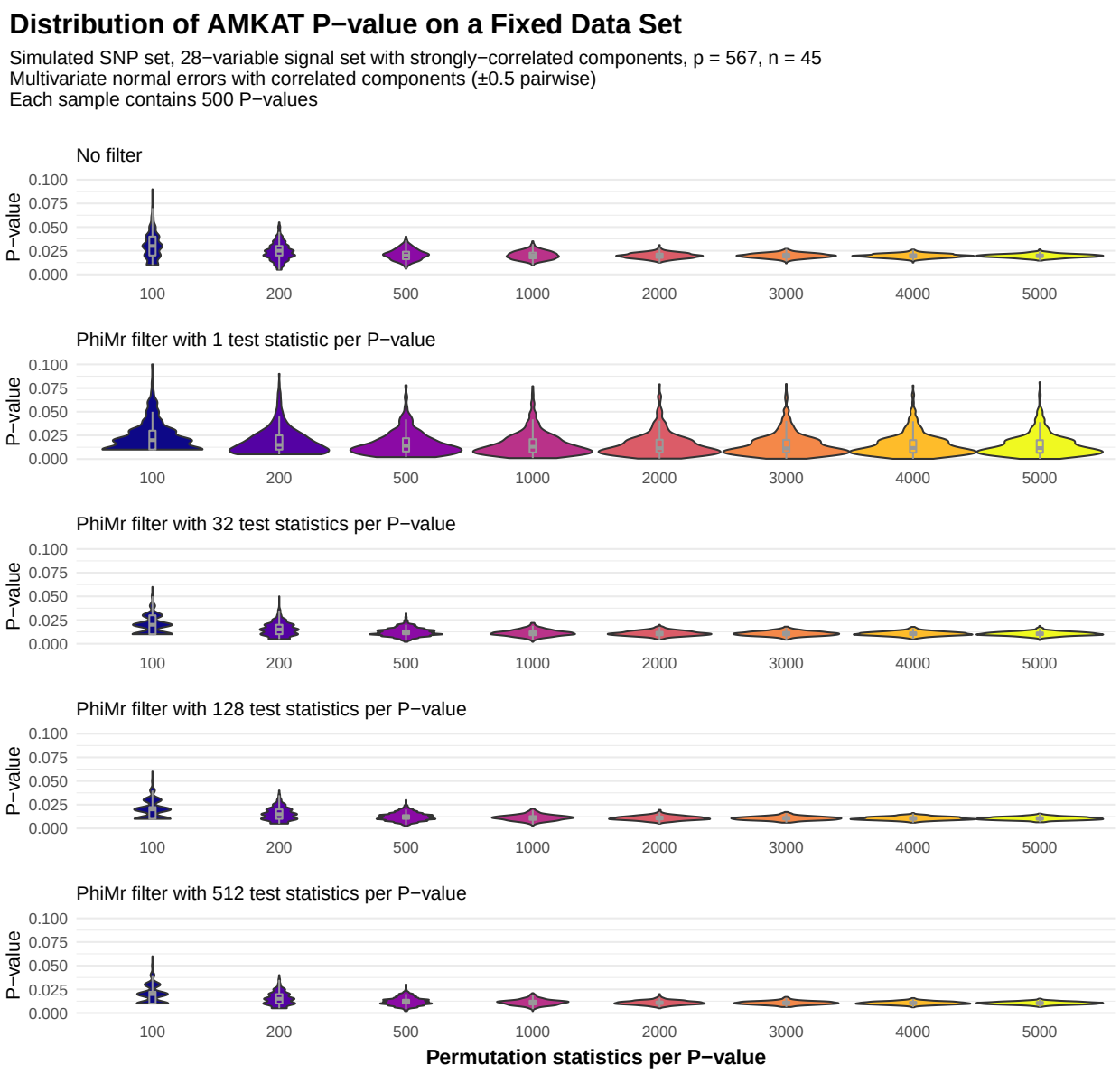
After partitioning the test and permutation statistics into batches and using them to estimate P-values, we plot the distributions of the resulting P-values.

```
makeViolinQPlots(pvalues, title_settings = title_violinQ(num_replicates))
```



The effect of Q and B on the P-value distribution with PhiMr is similar here was was observed for the data sets in the case with continuous features. For this data, the mean P-value does appear to be slightly lower with PhiMr than without.

```
makeViolinBPlots(pvalues, title_settings = title_violinB(num_replicates))
```



We again see patterns consistent with those observed in previous data sets, with B having little effect on the variance of the P-value with PhiMr at very low values of Q.

P-value Mean and Standard Deviation

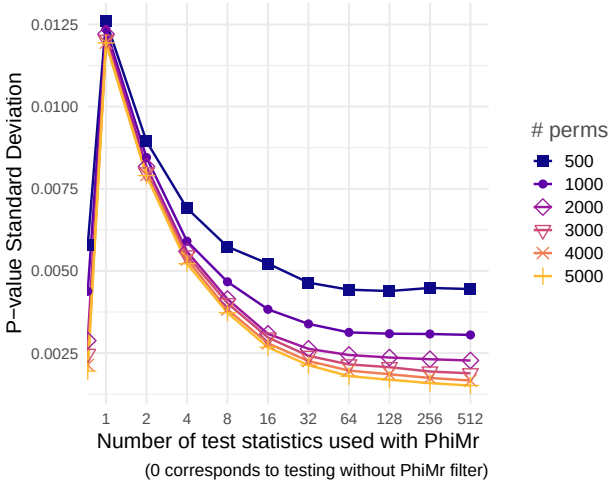
```
makePVLinePlots(pvalues, num_replicates)
```

Standard Deviation of AMKAT P−value on a Fixed Data Set

Simulated SNP set, 28−variable signal set with strongly−correlated components, p = 567, n = 45
Multivariate normal errors with correlated components (±0.5 pairwise)

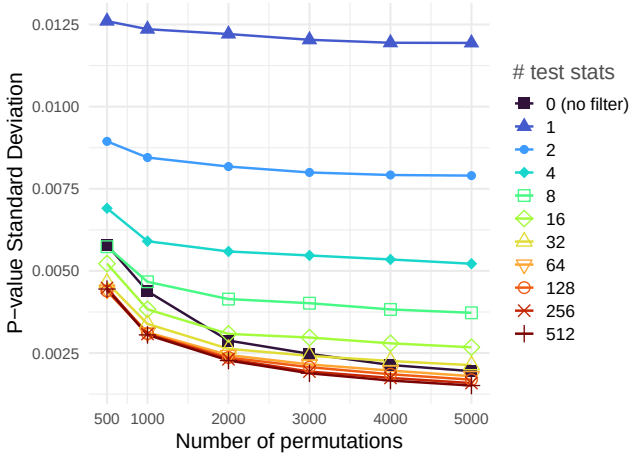
As a function of the # of test statistics

Each point represents a sample of 500 P−values



As a function of the # of permutation statistics

Each point represents a sample of 500 P−values



As in the case for continuous features, Q has a relatively greater impact on P-value standard deviation with PhiMr as compared to B, with most of this impact occurring at relatively modest values of Q.

Similar to data set #2, the P-value standard deviation with PhiMr is lower than that without PhiMr for sufficiently large Q.

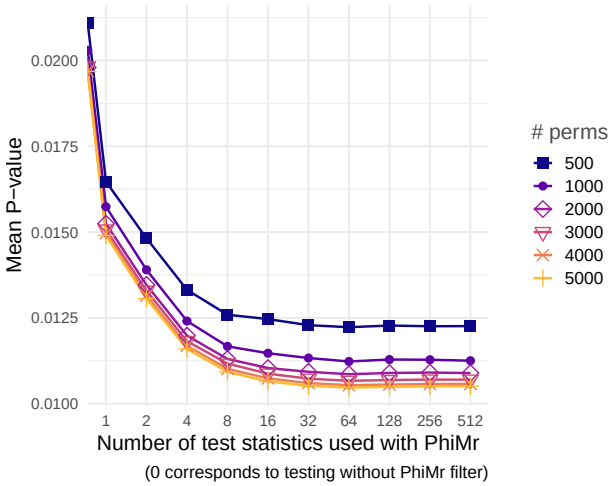
```
makePVLinePlots(pvalues, num_replicates, summary_stat = mean,
                title_settings =
                  title_pvline(
                    joint_title = "AMKAT Mean P-value on a Fixed Data Set"))
```

AMKAT Mean P−value on a Fixed Data Set

Simulated SNP set, 28−variable signal set with strongly−correlated components, p = 567, n = 45
Multivariate normal errors with correlated components (±0.5 pairwise)

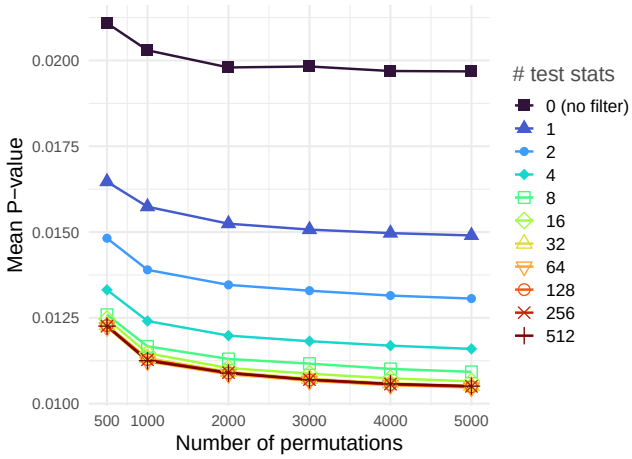
As a function of the # of test statistics

Each point represents a sample of 500 P−values



As a function of the # of permutation statistics

Each point represents a sample of 500 P−values



The mean P-value with PhiMr is substantially below that without PhiMr at all values of Q, and is around half of that without PhiMr for $Q \geq 16$:

```
# Mean P-value without PhiMr
mean(pvalues["2000", "NF", ])
```

```
## [1] 0.019796
```

```
# Mean P-value with PhiMr
mean(pvalues["2000", "PF-16", ])
```

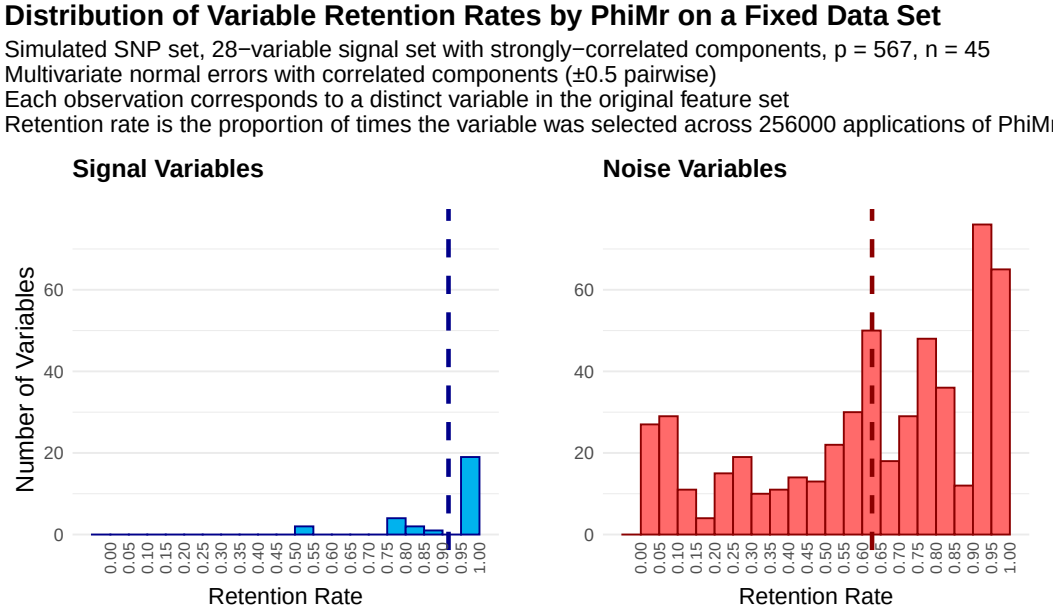
```
## [1] 0.011038
```

```
mean(pvalues["2000", "PF-512", ])
```

```
## [1] 0.010892
```

Distribution of Variable Retention Rates by PhiMr

```
makePhimrHistograms(feature_select_rates,
                    getSignalIndices(x_type, signal_density))
```



This data set contains 28 signal variables and 539 noise variables. The mean retention rate was quite high among the signal variables, at around 0.9. Among the noise variables, the mean retention rate was higher than for data sets with continuous features, at around 0.6, but PhiMr nonetheless appeared to consistently remove a significant share of noise while retaining the vast majority of signals, which is consistent with the smaller P-values observed on average with PhiMr compared to without.

Data Set #6

This data set was generated using the following scenario parameters:

```
n <- 50 # sample size
signal_density <- 'sparse' # 28 signal variables
signal_correlation <- 'high' # correlations between signal variables
error_correlation_strength <- 0.5 # mixed directions
```

We load the test and permutation statistics generated on this data set:

```
pv_files <- pvaluePlotFiles()
load(pv_files$stats)
load(pv_files$pvalues)
```

Distribution of Test and Permutation Statistics

We first plot the distributions of the test statistics and permutation statistics with and without the PhiMr filter:

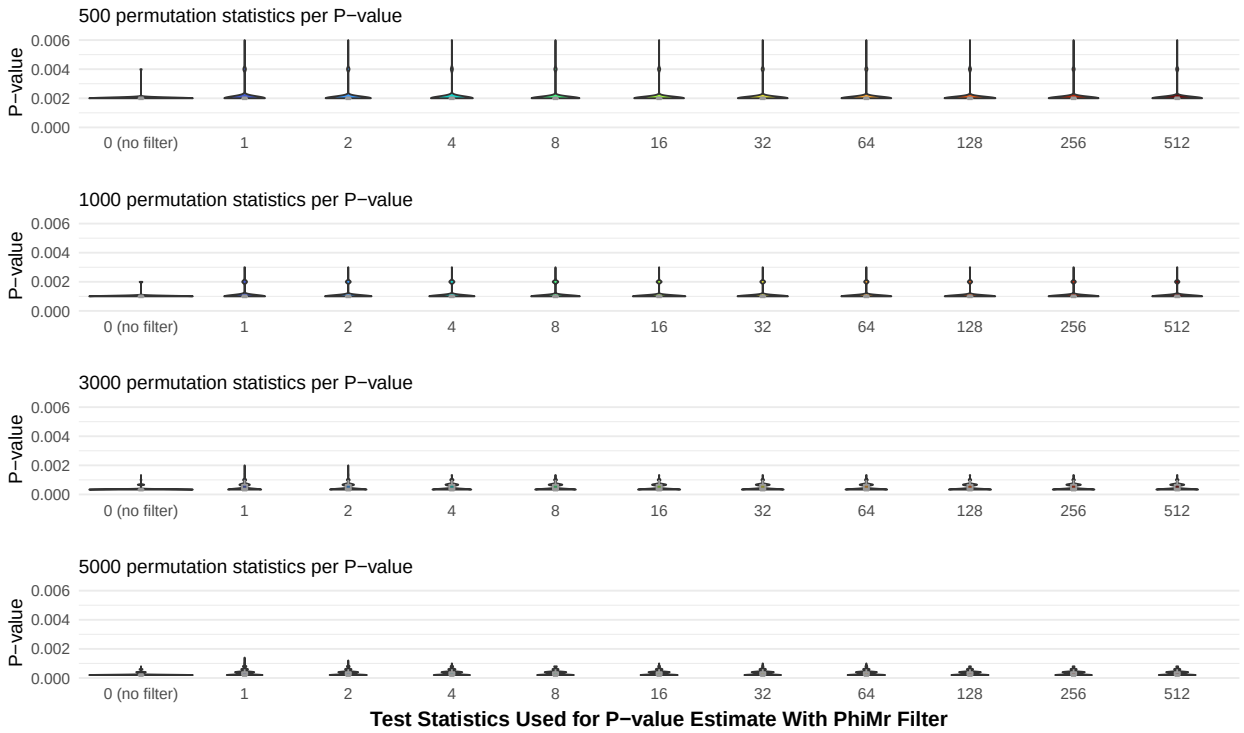
```
makeStatDistrPlots(
  test_stats, perm_stats, perm_stats_no_filter, test_stat_no_filter)
```

The test statistic distribution with PhiMr (right panel) appears more symmetric than in the previous data set, while the permutation statistic distributions are still prominently skewed. Both AMKAT variations appear to result in a very small P-value for this data set.

Distribution of P-values

After partitioning the test and permutation statistics into batches and using them to estimate P-values, we plot the distributions of the resulting P-values.

```
makeViolinQPlots(pvalues, title_settings = title_violinQ(num_replicates))
```

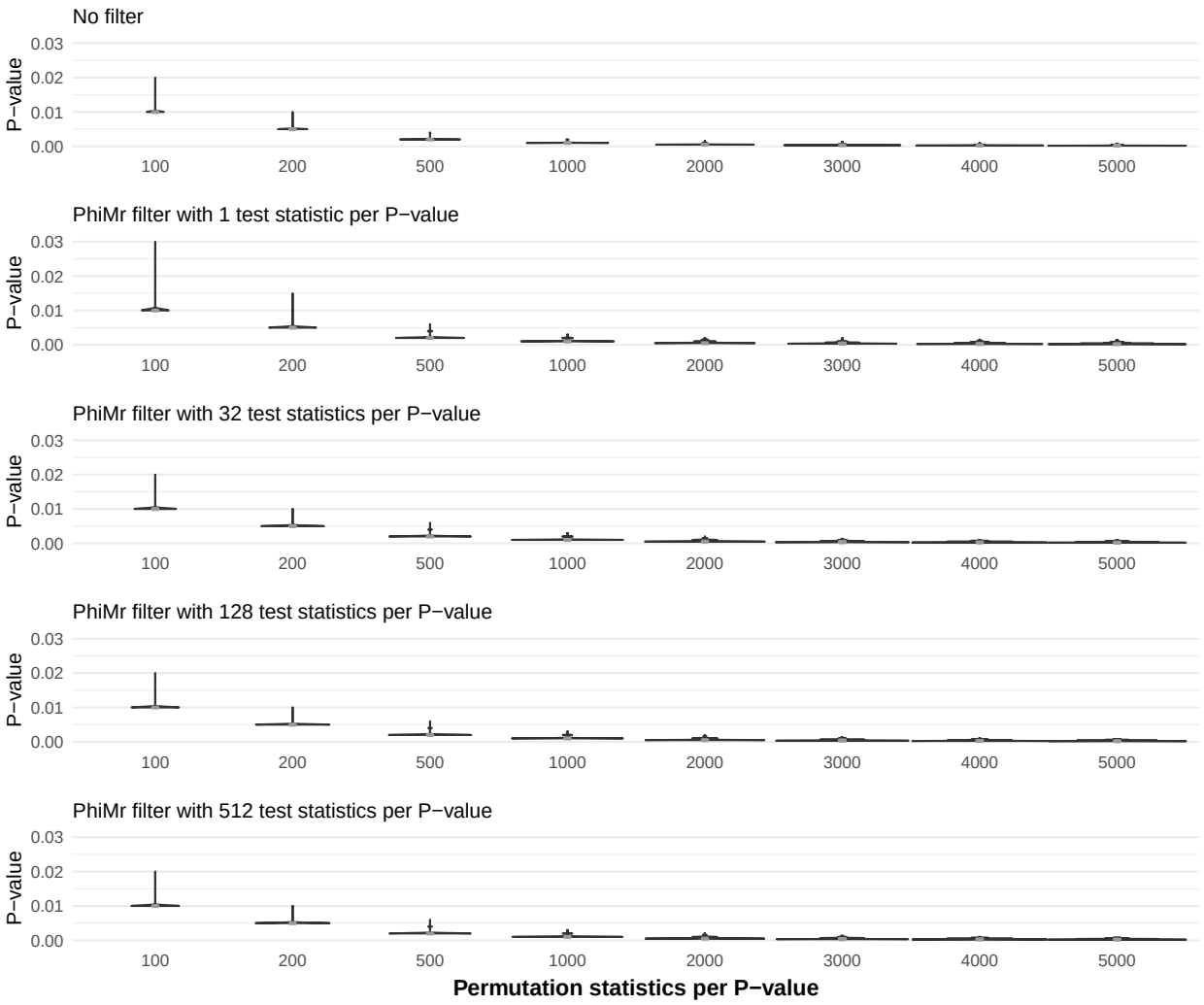


Here the P-value distributions are close to being degenerate, with the vast majority of P-values equal to $1/B$ (the smallest possible P-value estimate after the pseudo-count adjustment) and each distribution containing only a few distinct P-values (the actual distribution is discrete with values spaced in increments of $1/B$).

The P-value distributions with PhiMr are all quite similar across values of Q. Here B has the greater impact due to its effect on the resolution of the estimator's discrete distribution, allowing for a smaller minimum P-value after the pseudo-count adjustment and smaller increments between values.

From previous data sets, we have seen that the effect of B on the distribution's center is typically modest relative to the effect of Q; the importance of B here is due to the very small P-values for this data set, which causes the pseudo-count value and the granularity of the estimate to become highly-influential determinants of the distribution.

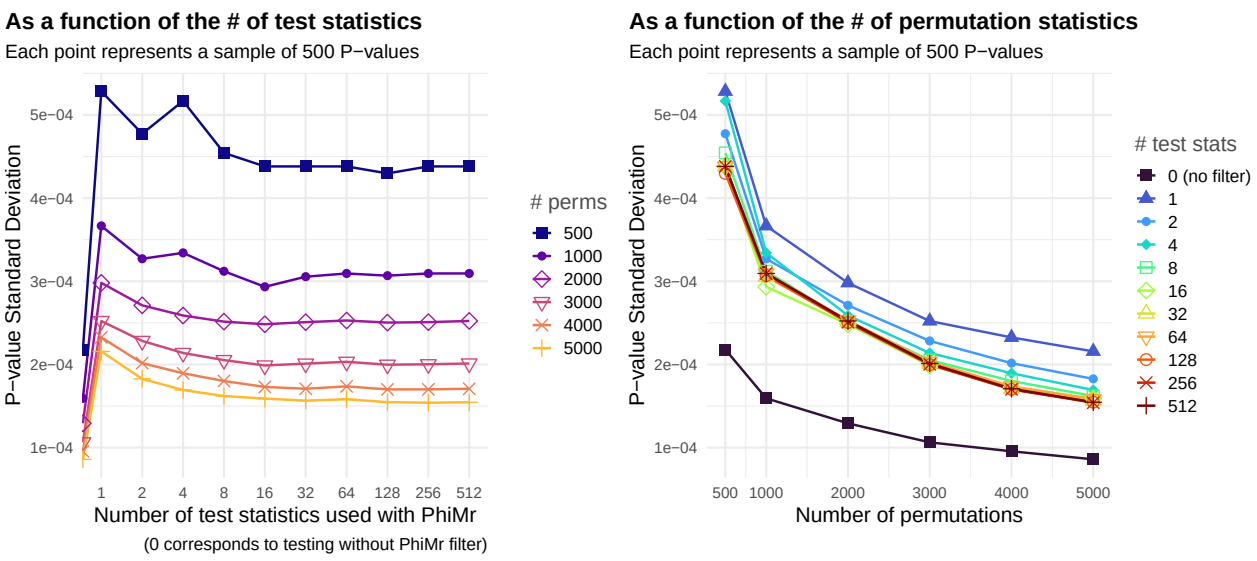
```
makeViolinBPlots(pvalues, title_settings = title_violinB(num_replicates))
```



Here we can more clearly see the effect of B on the distribution's center via the minimum imposed by the pseudo-count adjustment of $1/B$.

P-value Mean and Standard Deviation

```
makePVLinePlots(pvalues, num_replicates)
```

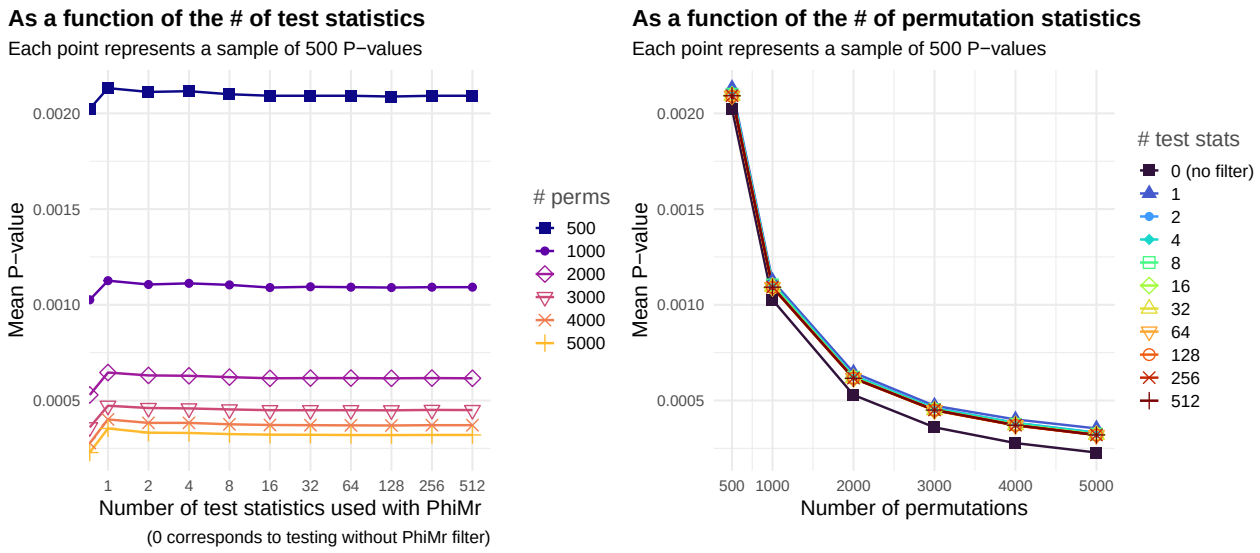


Looking at the range of values displayed on the vertical axis, the P-value standard deviation across all AMKAT variations was very small, as seen in the two previous plots.

Comparing the left and right panels, we see again that the influence of B was greater than that of Q when using PhiMr.

From the right panel, it is clear that the P-value standard deviation is lower without PhiMr than with PhiMr for this data set.

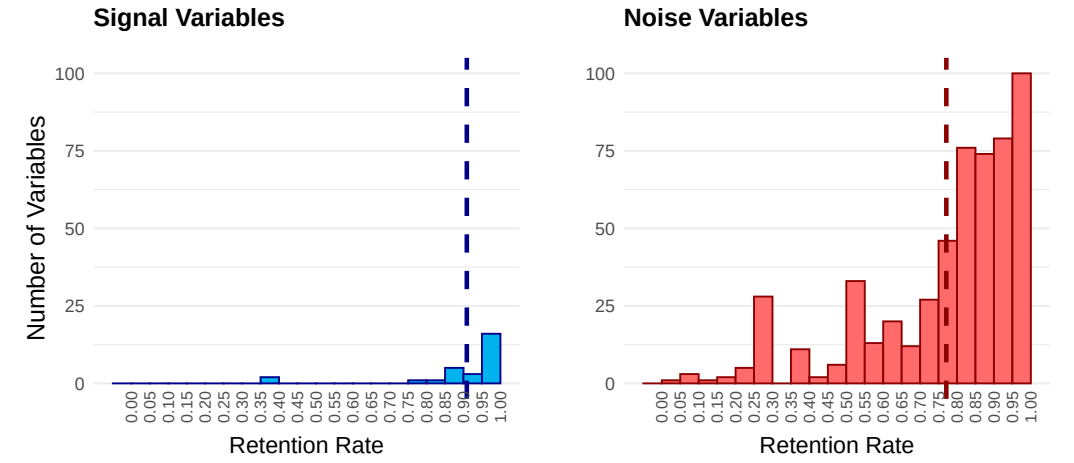
```
makePVLinePlots(pvalues, num_replicates, summary_stat = mean,
                title_settings =
                  title_pvline(
                    joint_title = "AMKAT Mean P-value on a Fixed Data Set"))
```



Comparing the two panels shows that Q has little to no effect on the mean P-value, while B has a substantial effect; again, this appears to owe largely to the effect of B on the minimum P-value via the pseudo-count adjustment of $1/B$ (as well as, to a lesser extent, the effect of B on the granularity of the P-value estimate).

Distribution of Variable Retention Rates by PhiMr

```
makePhimrHistograms(feature_select_rates,
                    getSignalIndices(x_type, signal_density))
```



This data set contains 28 signal variables and 539 noise variables.

While the mean retention rate among the signal variables was high and similar to that seen in the previous data set, the mean retention rate among the noise variables was considerably greater for this data set than for the previous, with the majority of noise variables being retained at high rates. There is still a substantial difference in the mean retention rate between the two groups, but as seen in the previous plots, the P-value estimate with PhiMr had a higher mean and standard deviation than that without PhiMr. It can be noted that PhiMr still yielded P-value distributions with very small mean and standard deviation for this data set.

Data Set #7

This data set was generated using the following scenario parameters:

```
n <- 50 # sample size
signal_density <- 'sparse' # 28 signal variables
signal_correlation <- 'low' # correlations between signal variables
error_correlation_strength <- 0.5 # mixed directions
```

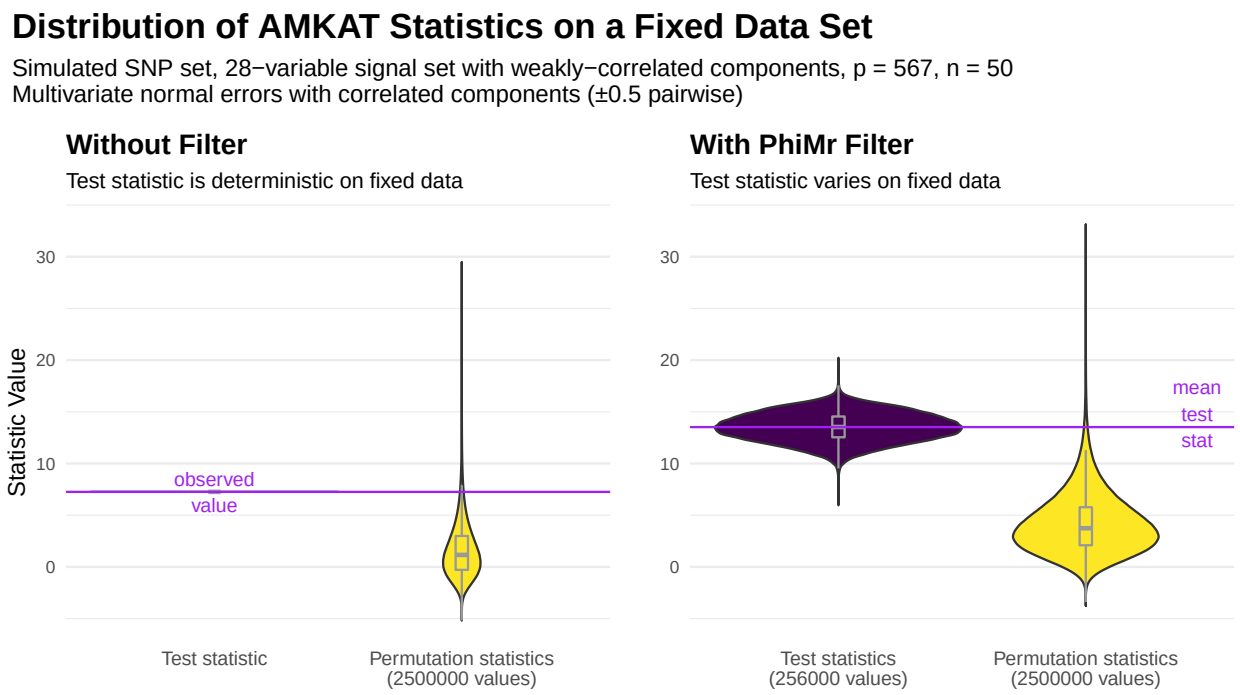
We load the test and permutation statistics generated on this data set:

```
pv_files <- pvaluePlotFiles()
load(pv_files$stats)
load(pv_files$pvalues)
```

Distribution of Test and Permutation Statistics

We first plot the distributions of the test statistics and permutation statistics with and without the PhiMr filter:

```
makeStatDistrPlots(
  test_stats, perm_stats, perm_stats_no_filter, test_stat_no_filter)
```

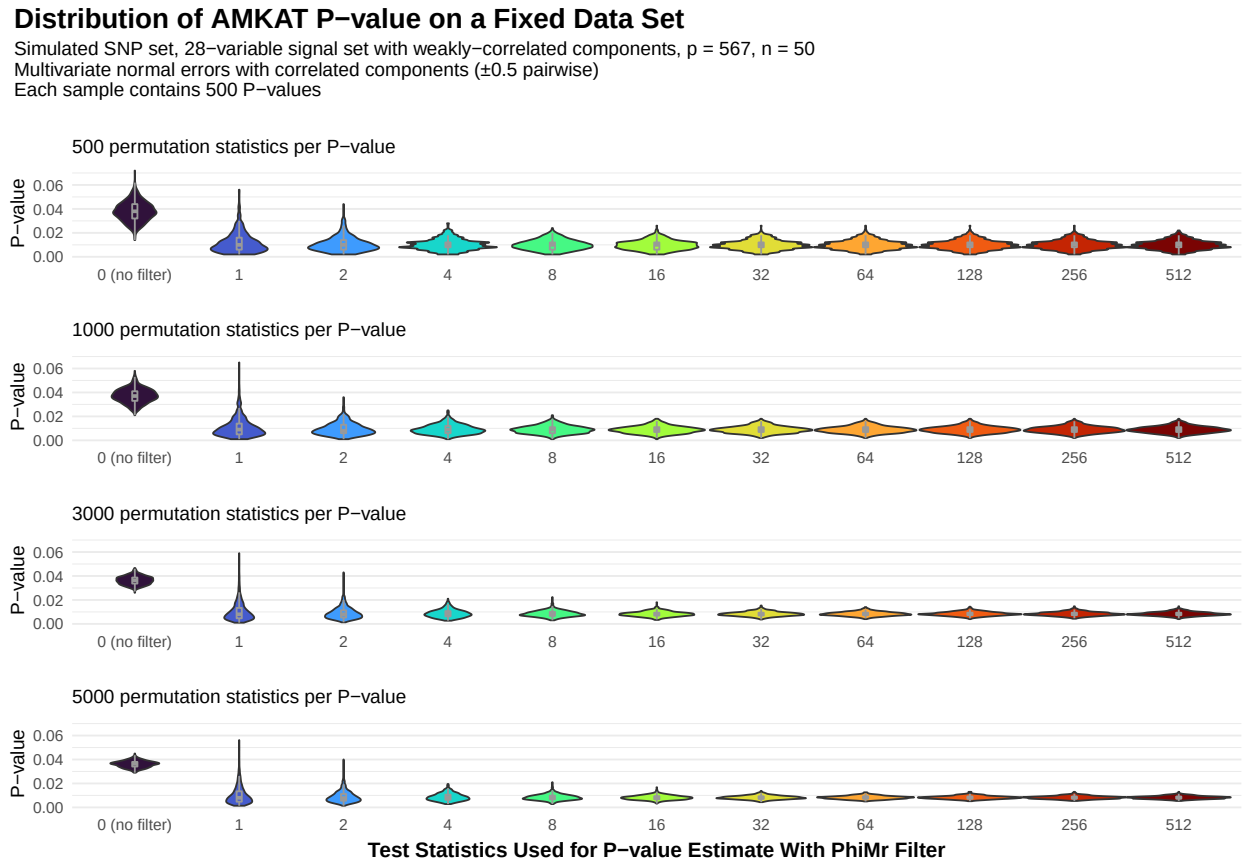


The distributions of test and permutation statistics seen here are similar to those seen in previous data sets. Here the PhiMr filter appears to result in a smaller P-value on average (its mean lies farther into the tail of its respective permutation statistic distribution as compared to the case without PhiMr shown in the left panel).

Distribution of P-values

After partitioning the test and permutation statistics into batches and using them to estimate P-values, we plot the distributions of the resulting P-values.

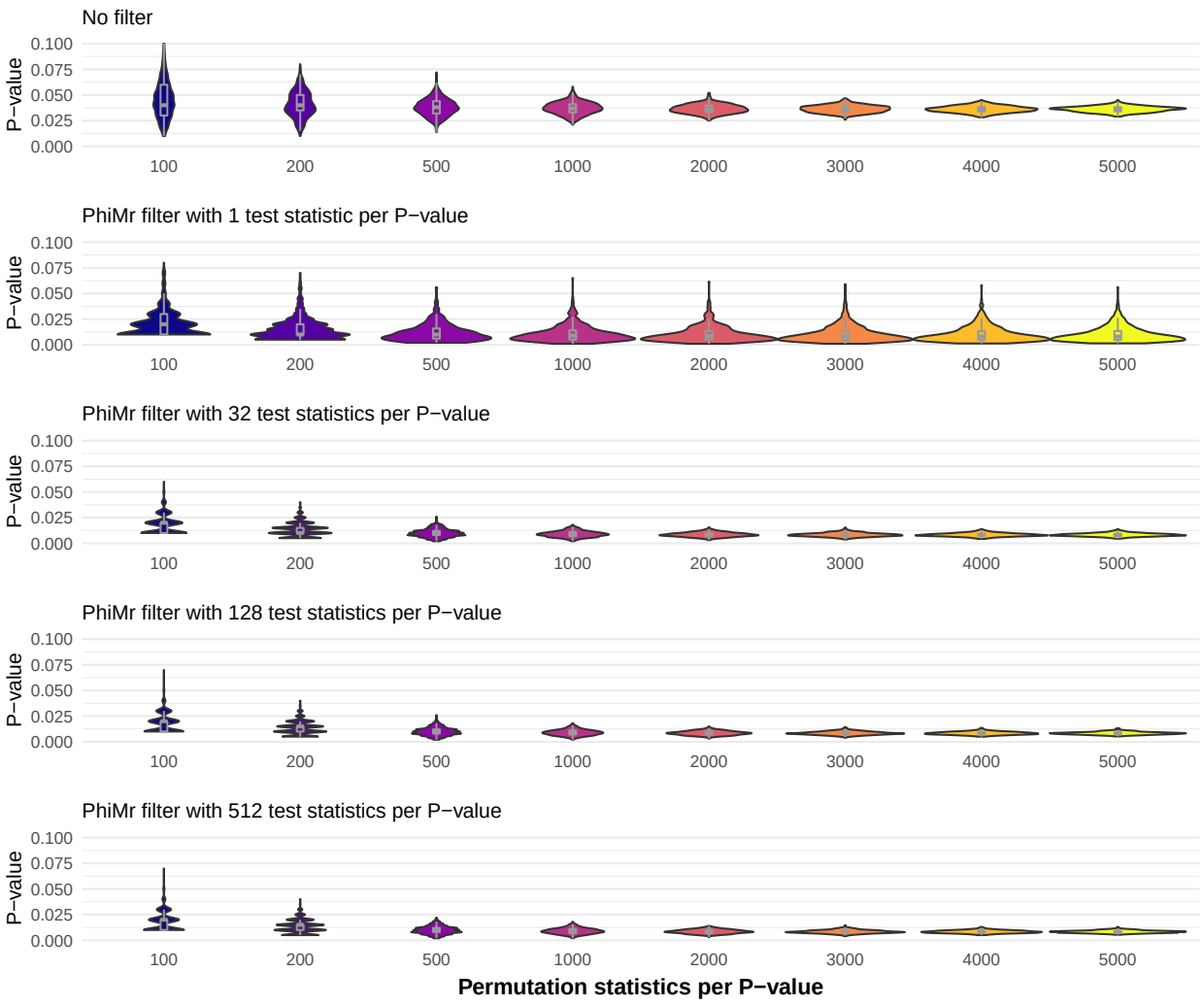
```
makeViolinQPlots(pvalues, title_settings = title_violinQ(num_replicates))
```



Here we see patterns which have been common to most previous data sets: as Q increases variance with PhiMr steadily and substantially decreases, and the distribution of the P-values with PhiMr appears to converge as B and Q tend to infinity.

Interestingly, in each row, the variance of the distribution without PhiMr does not necessarily appear to be less than the variance of the distribution with PhiMr when using only $Q = 1$ test statistics; typically the distribution with PhiMr requires greater values of Q to achieve comparable variance to the case without PhiMr. Here, even at modest values of Q, PhiMr appears to achieve P-value distributions with considerably less variation compared to testing without PhiMr.

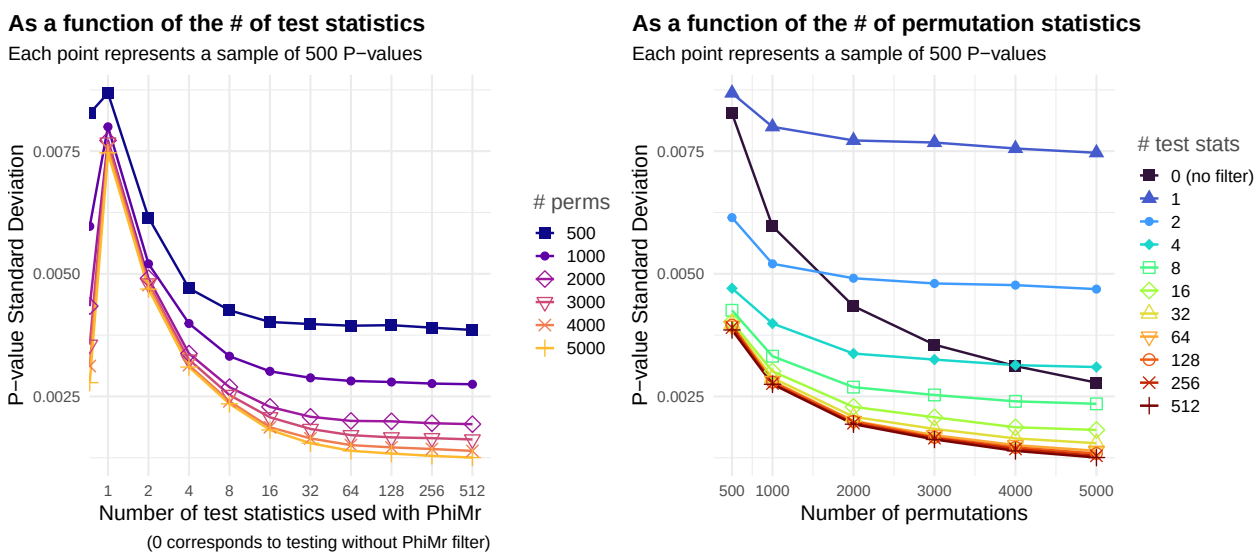
```
makeViolinBPlots(pvalues, title_settings = title_violinB(num_replicates))
```



Similar to the typical case for previous data sets, B has little apparent effect on P-value variance with PhiMr at $Q = 1$, but shows a noticeable effect once Q grows beyond a moderate value.

P-value Mean and Standard Deviation

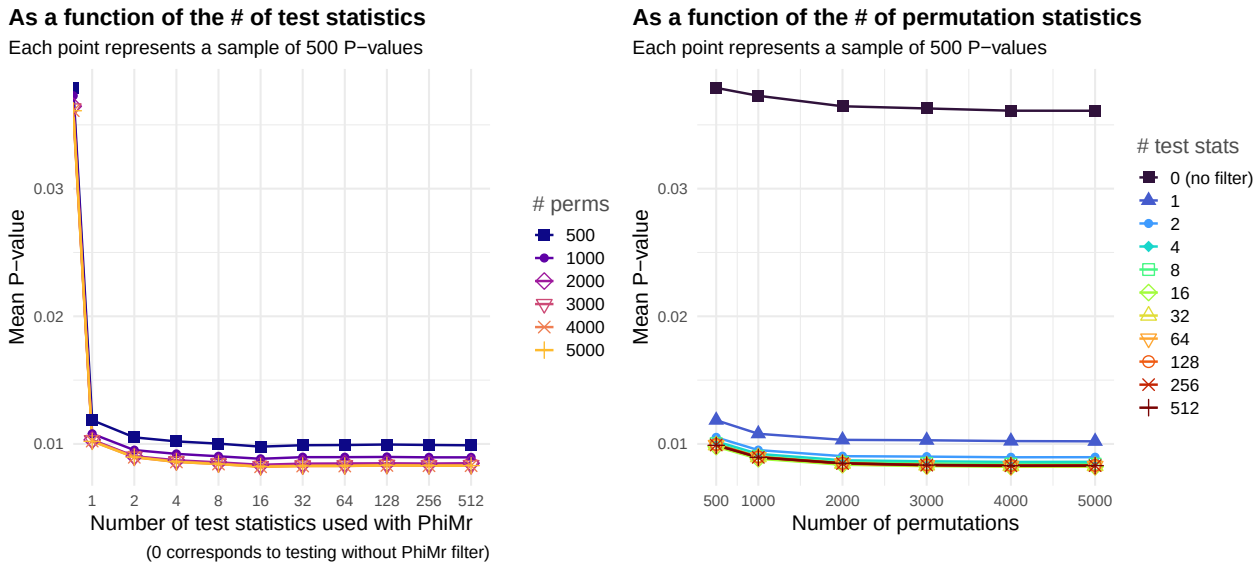
```
makePVLinePlots(pvalues, num_replicates)
```



From this graph, we can verify that at $Q = 1$, the P-value standard deviation with PhiMr exceeds that without PhiMr, which has typically been the case in previous data sets.

With as few as $Q = 8$ test statistics, however, PhiMr attains lower P-value variation compared to a filter-free AMKAT at all values of B seen in the plot.

```
makePVLinePlots(pvalues, num_replicates, summary_stat = mean,
                title_settings =
                  title_pvline(
                    joint_title = "AMKAT Mean P-value on a Fixed Data Set"))
```

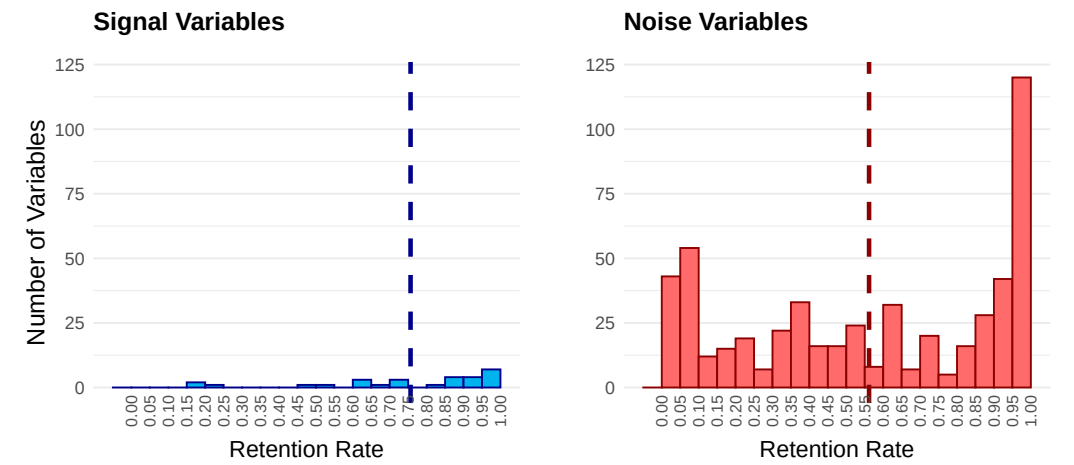


Here we see that the mean P-value for this data set is not substantially influenced by either of Q or B.

Here the mean P-value is far lower for all variations of AMKAT with PhiMr as compared to AMKAT without PhiMr.

Distribution of Variable Retention Rates by PhiMr

```
makePhimrHistograms(feature_select_rates,
                    getSignalIndices(x_type, signal_density))
```



This data set contains 28 signal variables and 539 noise variables.

For this data set, we see a lower mean retention rate across the signal variables as compared to data sets 5 and 6. We also see a lower mean retention rate across the noise variables.

One difference in design between the signal variables in this data set compared to those in data sets 5 and 6 is the correlation scheme among the signals: in data set 5 and 6, the signals were selected from among components of the SNP set that typically exhibited very strong pairwise correlation; in this data set, the signals corresponded to SNP-set components with milder correlations. This could potentially explain some of the difference in retention rates for the signal variables.

Because signals come from different SNP-set locations in this data set compared to the previous two, their correlations with the noise variables may also differ. This could explain why many more noise variables were retained more frequently for the data sets with stronger correlation among the signal variables, assuming that the signal variables in those data sets were also more strongly correlated with the noise variables.

Because signals come from different SNP-set locations in this data set compared to the previous two, the minor allele frequency (MAF) of each signal variable may also be different. The MAF corresponds to the proportion of minor alleles present in the population or reference data for a particular SNP, and thus affects the probability of different genotype values (0, 1 or 2) being observed in the simulated data at a particular index of $\boldsymbol{X}$. Rarity of minor alleles in the data for a particular signal SNP can cause difficulty in detecting its association with the response, due to low variation in its genotype data. This could potentially further explain the difference in retention rates for the signal variables if the signal set with more strongly-correlated components also had higher minor allele frequencies.

Data Set #8

This data set was generated using the following scenario parameters:

```
n <- 50 # sample size
signal_density <- 'dense' # 122 signal variables
signal_correlation <- 'high' # correlations between signal variables
error_correlation_strength <- 0 # independent errors
```

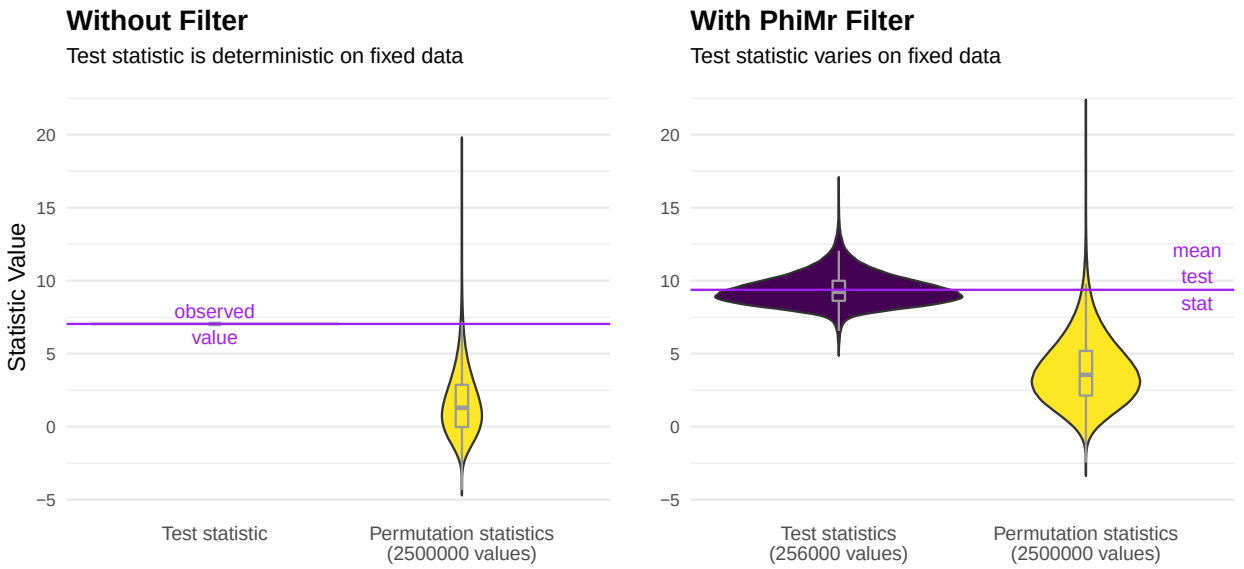
We load the test and permutation statistics generated on this data set:

```
pv_files <- pvaluePlotFiles()
load(pv_files$stats)
load(pv_files$pvalues)
```

Distribution of Test and Permutation Statistics

We first plot the distributions of the test statistics and permutation statistics with and without the PhiMr filter:

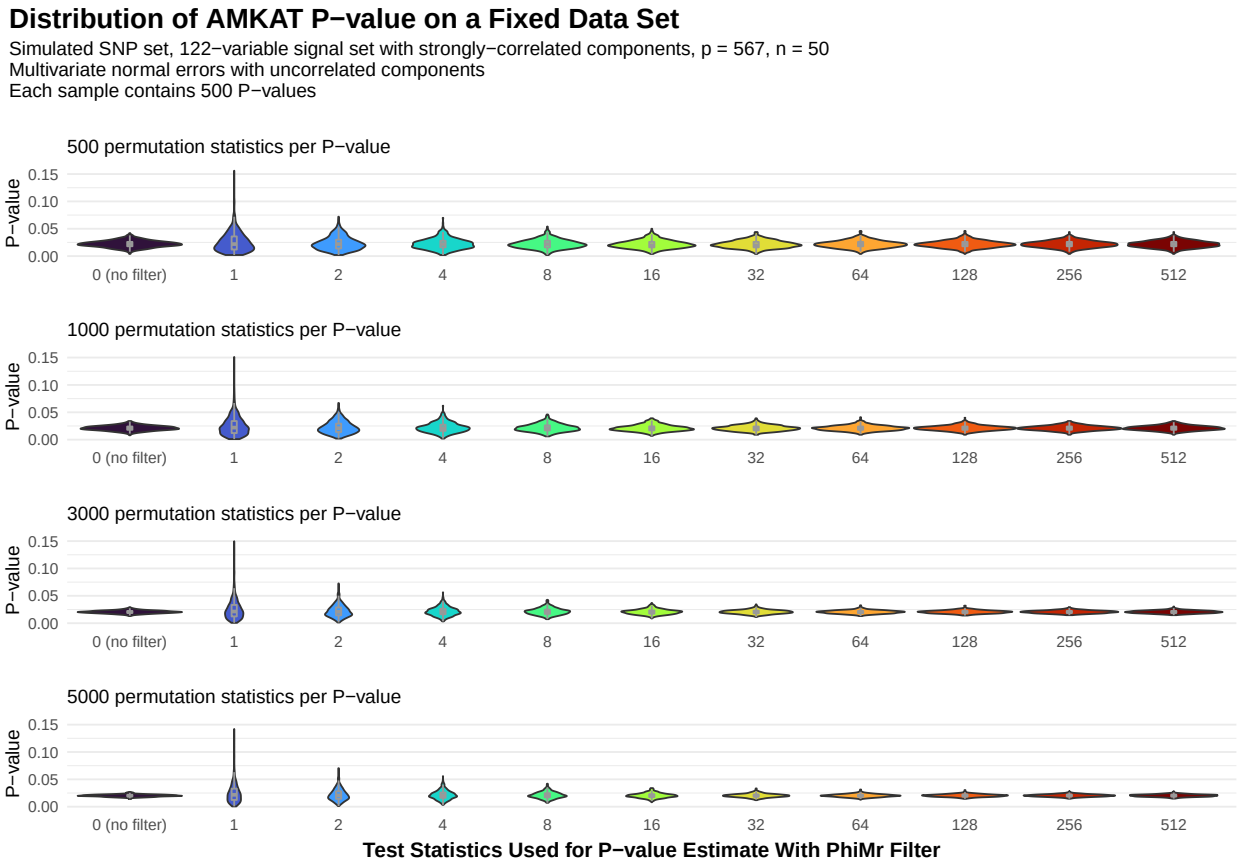
```
makeStatDistrPlots(
  test_stats, perm_stats, perm_stats_no_filter, test_stat_no_filter)
```



Distribution of P-values

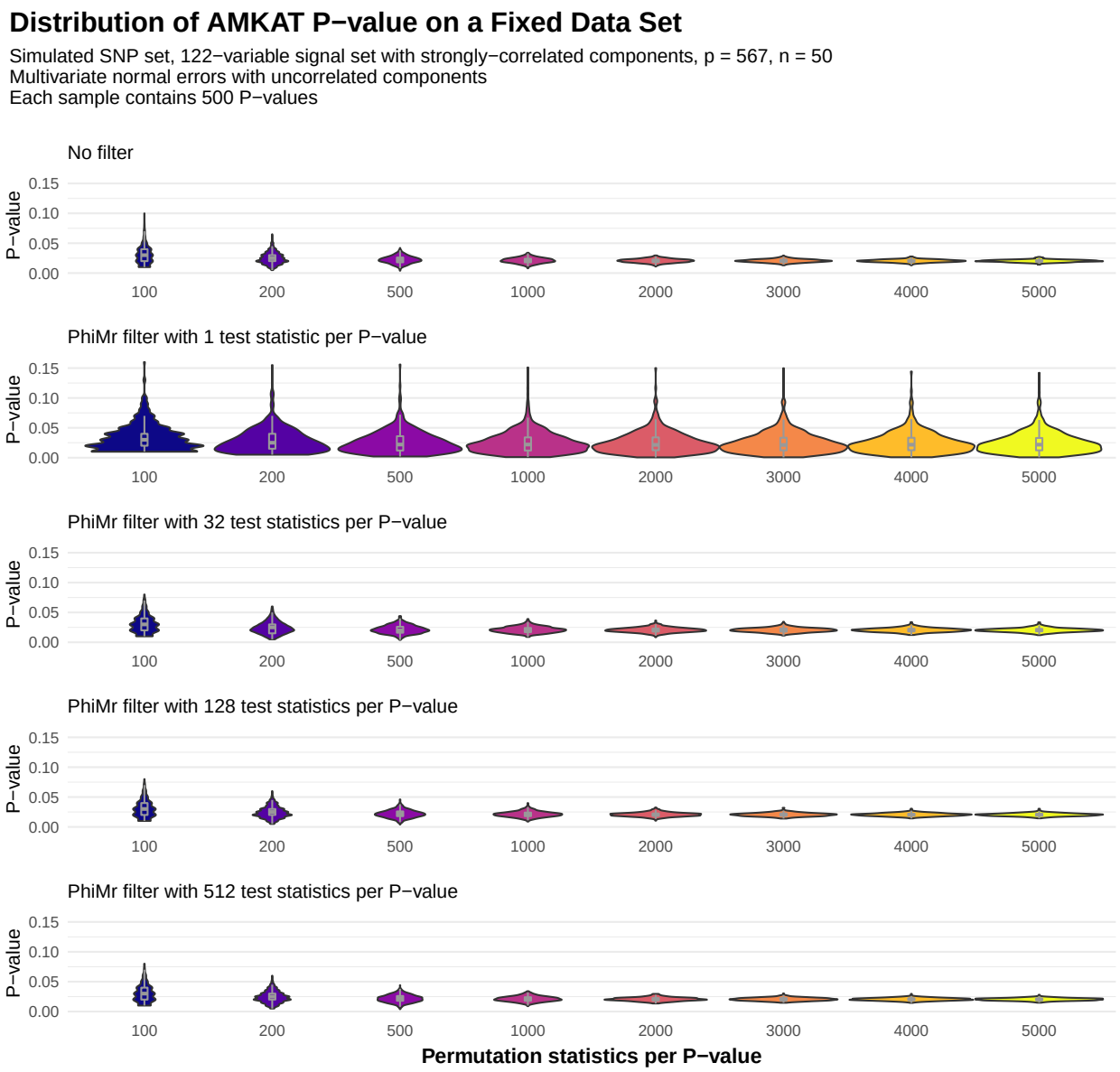
After partitioning the test and permutation statistics into batches and using them to estimate P-values, we plot the distributions of the resulting P-values.

```
makeViolinQPlots(pvalues, title_settings = title_violinQ(num_replicates))
```



The patterns here are similar to those observed in previous data sets.

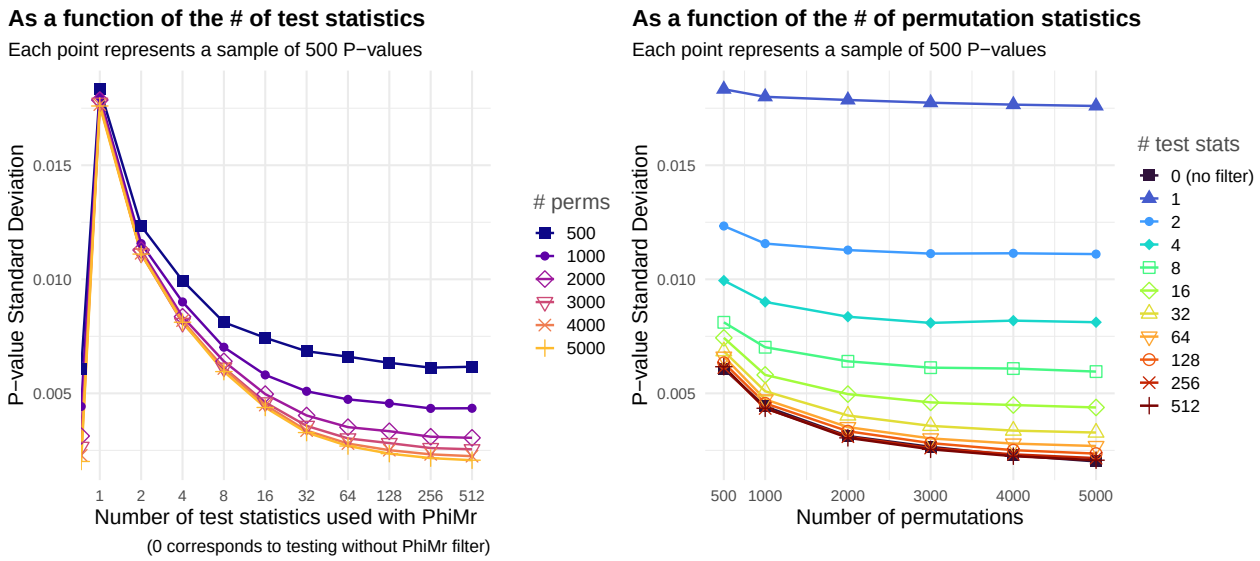
```
makeViolinBPlots(pvalues, title_settings = title_violinB(num_replicates))
```



We again observe similar patterns as for previous data sets.

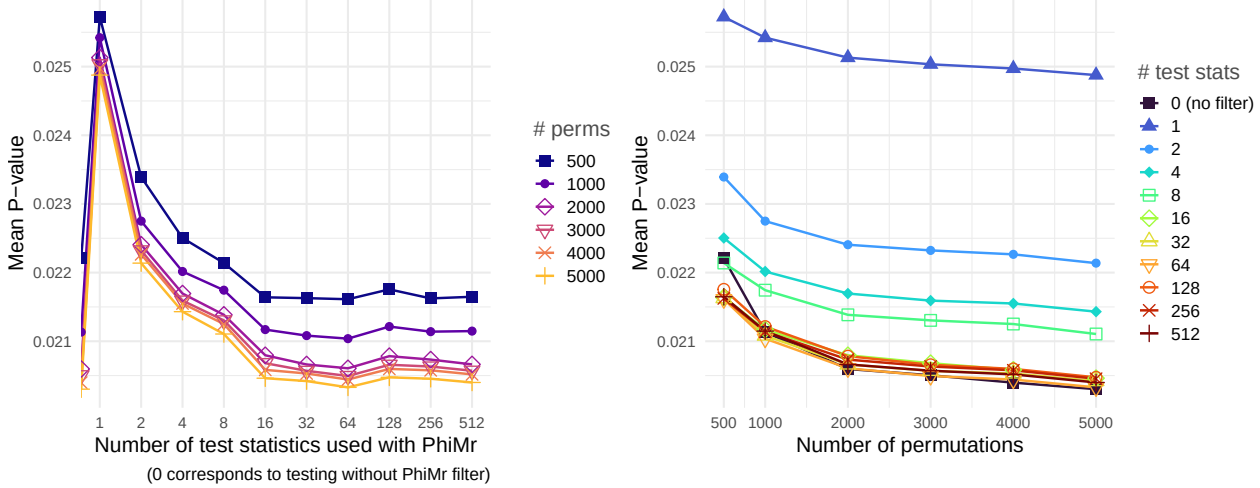
P-value Mean and Standard Deviation

```
makePVLinePlots(pvalues, num_replicates)
```



The P-value standard deviation of AMKAT with PhiMr when $Q = 512$ is approximately equal to that of AMKAT without PhiMr, with this pattern holding across values of B.

```
makePVLinePlots(pvalues, num_replicates, summary_stat = mean,
                title_settings =
                  title_pvline(
                    joint_title = "AMKAT Mean P-value on a Fixed Data Set"))
```

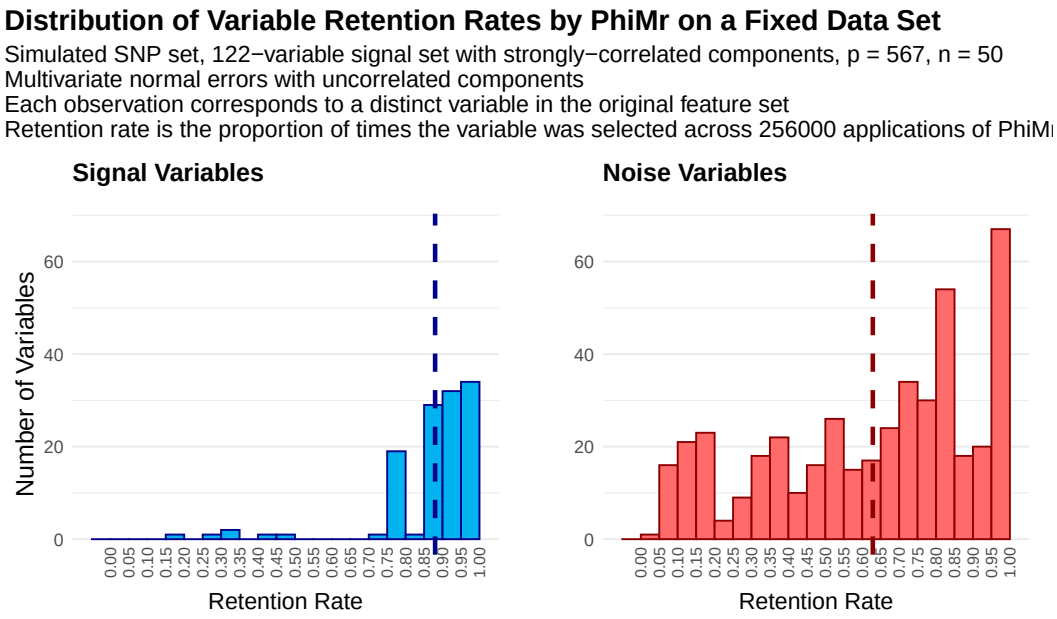


At greater values of Q, the mean P-value with PhiMr is very close to that without PhiMr, as seen in the right-hand panel.

When using PhiMr, the mean P-value appears to converge more slowly with respect to Q here than we observed for previous data sets (left panel). Based on this, it may be prudent to use greater values of Q (e.g., 500 or 1000) in favor of the more modest values that were sufficient in most other data sets to yield P-value distributions with highly-consistent centers.

Distribution of Variable Retention Rates by PhiMr

```
makePhimrHistograms(feature_select_rates,
                    getSignalIndices(x_type, signal_density))
```



This data set contains 122 signal variables and 445 noise variables. Like data sets 5 and 6, the signal variables consist of strongly-correlated SNPs. Similar to those data sets, the mean retention rate among the signals was high, with the retention rate among noise variables appearing to have a similar distribution to that seen in data set 5.

Data Set #9

This data set was generated using the following scenario parameters:

```
n <- 50 # sample size
signal_density <- 'dense' # 122 signal variables
signal_correlation <- 'low' # correlations between signal variables
error_correlation_strength <- 0 # independent errors
```

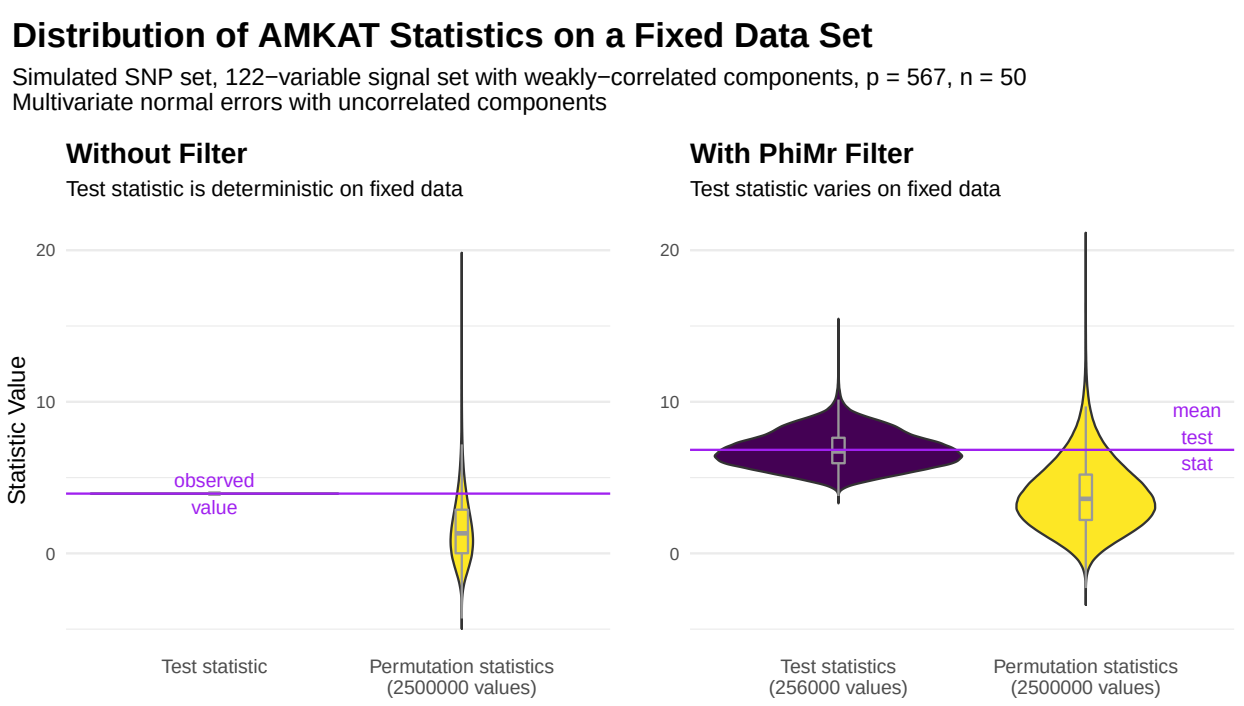
We load the test and permutation statistics generated on this data set:

```
pv_files <- pvaluePlotFiles()
load(pv_files$stats)
load(pv_files$pvalues)
```

Distribution of Test and Permutation Statistics

We first plot the distributions of the test statistics and permutation statistics with and without the PhiMr filter:

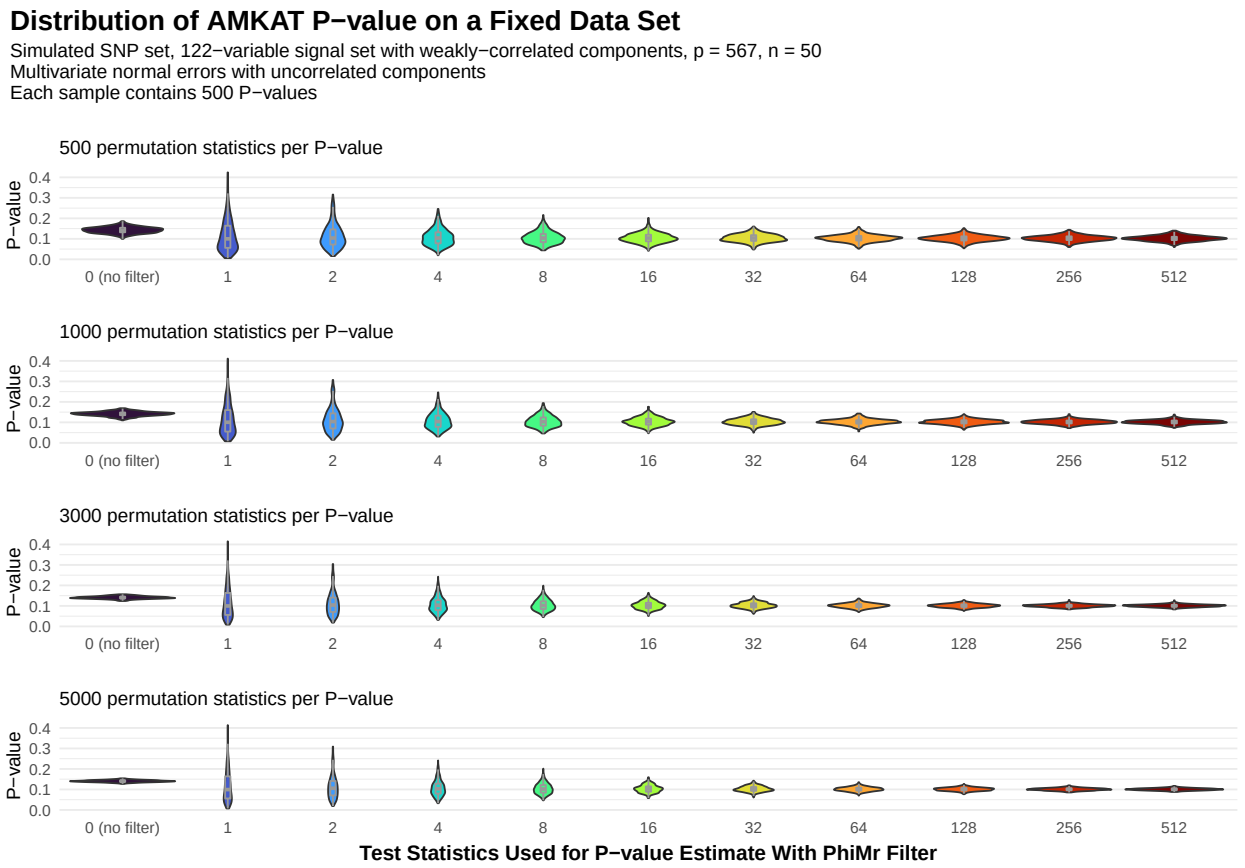
```
makeStatDistrPlots(
  test_stats, perm_stats, perm_stats_no_filter, test_stat_no_filter)
```



Distribution of P-values

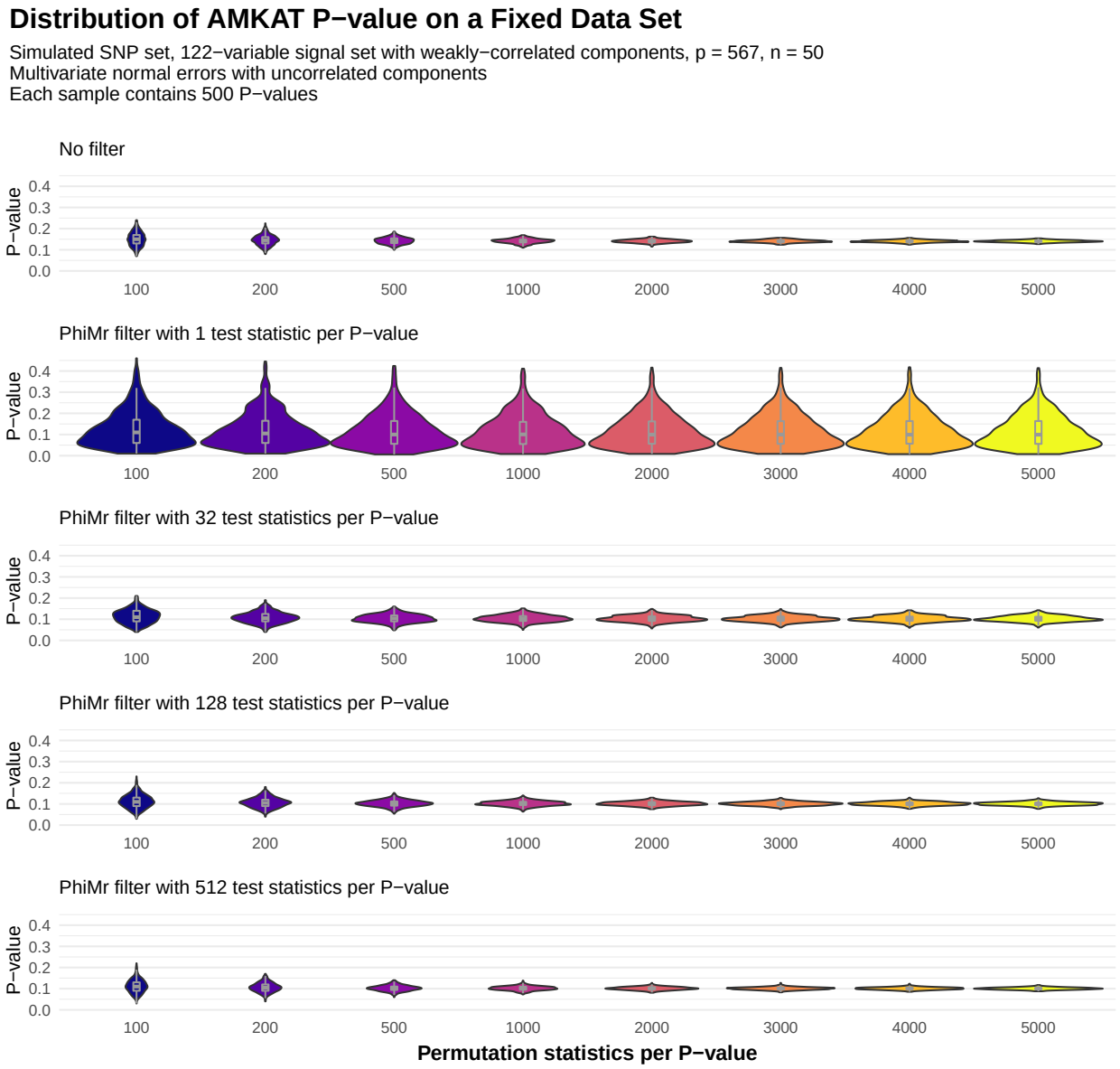
After partitioning the test and permutation statistics into batches and using them to estimate P-values, we plot the distributions of the resulting P-values.

```
makeViolinQPlots(pvalues, title_settings = title_violinQ(num_replicates))
```



We observe patterns similar to those for other data sets. PhiMr appears to yield a modestly lower mean P-value than AMKAT without PhiMr for this data.

```
makeViolinBPlots(pvalues, title_settings = title_violinB(num_replicates))
```



As typically observed for other data sets, B has little effect on P-value variance for $Q = 1$, but shows an impact once Q exceeds a modest value.

P-value Mean and Standard Deviation

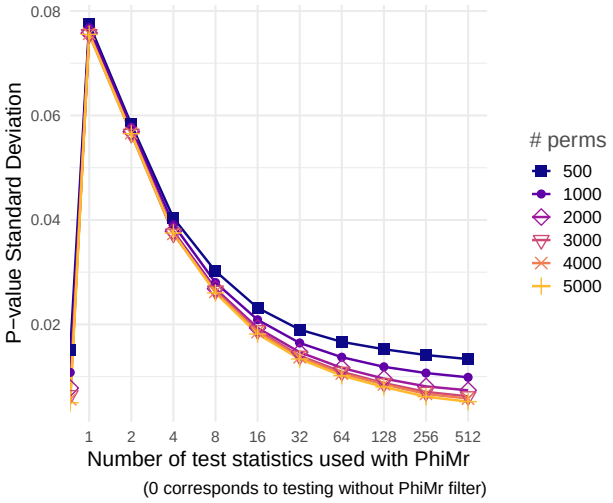
```
makePVLinePlots(pvalues, num_replicates)
```

Standard Deviation of AMKAT P−value on a Fixed Data Set

Simulated SNP set, 122−variable signal set with weakly−correlated components, p = 567, n = 50
Multivariate normal errors with uncorrelated components

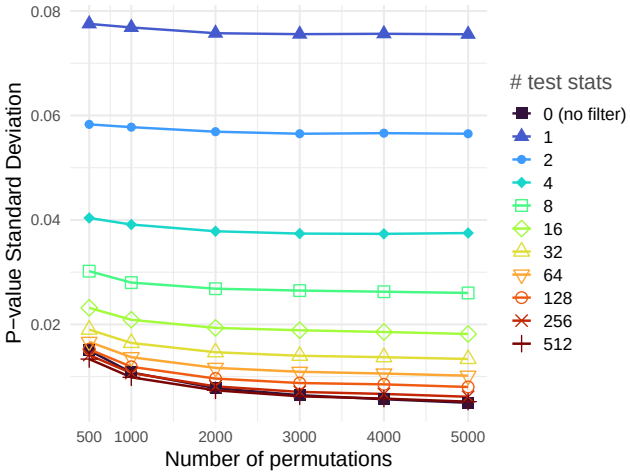
As a function of the # of test statistics

Each point represents a sample of 500 P−values



As a function of the # of permutation statistics

Each point represents a sample of 500 P−values



P-value standard deviation is much higher with PhiMr than without PhiMr for small Q, while for large Q it is approximately equal.

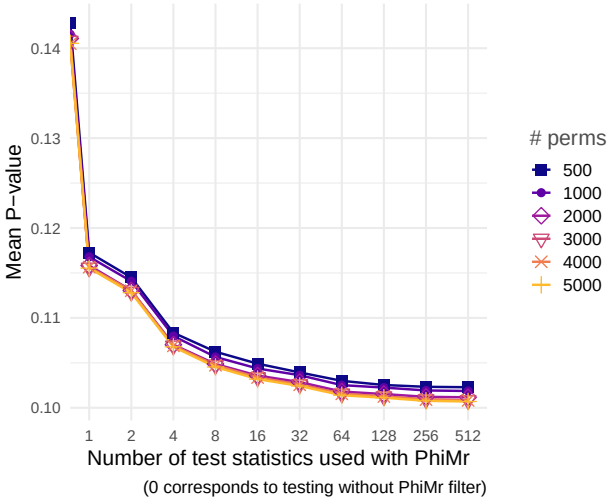
```
makePVLinePlots(pvalues, num_replicates, summary_stat = mean,
                title_settings =
                  title_pvline(
                    joint_title = "AMKAT Mean P-value on a Fixed Data Set"))
```

AMKAT Mean P−value on a Fixed Data Set

Simulated SNP set, 122−variable signal set with weakly−correlated components, p = 567, n = 50
Multivariate normal errors with uncorrelated components

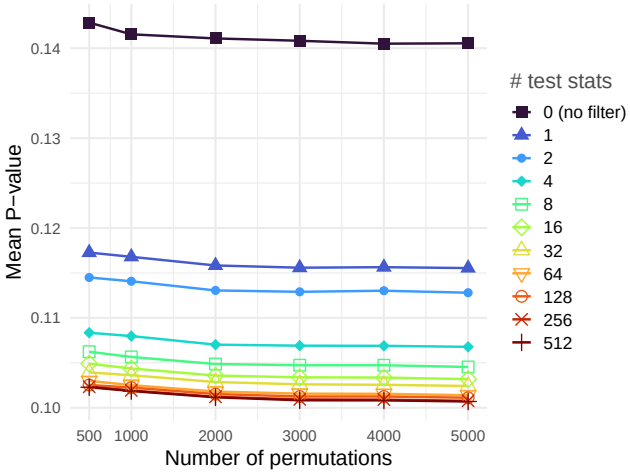
As a function of the # of test statistics

Each point represents a sample of 500 P−values



As a function of the # of permutation statistics

Each point represents a sample of 500 P−values

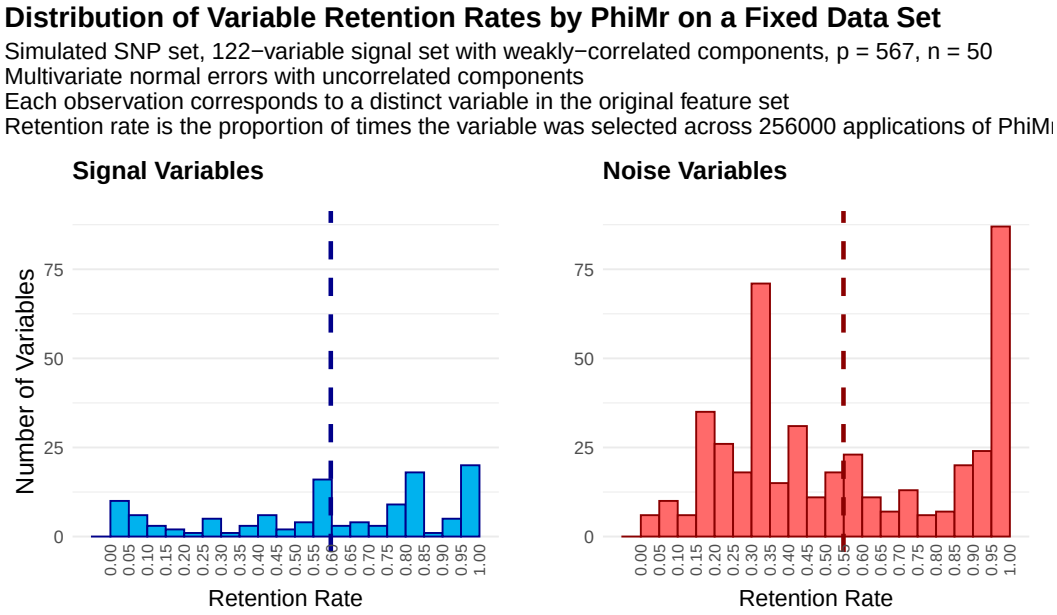


The mean P-value with PhiMr is moderately lower than without PhiMr.

The mean P-value decreases noticeably as Q increases starting from low values, becoming near-constant for $Q \geq 64$. B shows little influence on the mean P-value even for large Q and for AMKAT without PhiMr.

Distribution of Variable Retention Rates by PhiMr

```
makePhimrHistograms(feature_select_rates,
                    getSignalIndices(x_type, signal_density))
```



This data set contains 122 signal variables and 445 noise variables. Like data set 7, the signal variables are comprised of mildly-correlated SNPs. Similar to our observations for that data set, the mean retention rate here is much lower than it was when the signal set was of equal size but comprised of strongly-correlated SNPs (data set 8); another similar comparison is made for the noise variables, which show a lower mean rate of retention than we observed in data set 8.