Simulation Results: PhiMr Filter for Feature Selection

Contents

Prepare Working Environment 1
Continuous Features 1
Discrete Features with Weakly Correlated Signals 4
Discrete Features with Strongly Correlated Signals 6

We review results from our simulations for feature selection.

Prepare Working Environment

We begin by loading the relevant R packages:

```
# 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"))
```

All scenarios in this simulation setting share the following parameter values:

```
num_replicates  <- 1000
error_distribution <- 'normal'
signal_strength <- 1
values_for_signal_density <- c('sparse', 'dense')
values_for_sample_size <- c(50, 100, 150, 200)
values_for_error_corr_strength <- c(0, 0.5)
```

Continuous Features

This section includes scenarios where $\boldsymbol{X}$ is simulated as a continuous random vector.

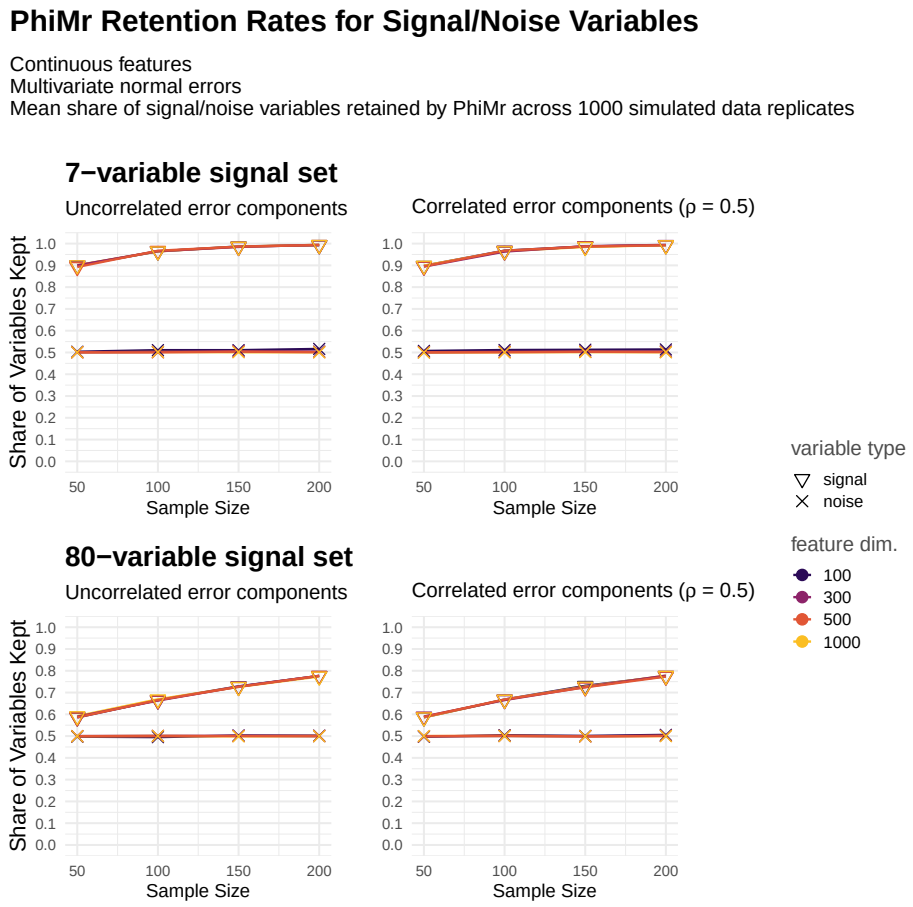
```
x_type <- 'cts'
source(file.path(dir_src, 'initialize_adaptive_across.R'))

# Create filenames and load plot data
files <- adaptiveAcrossPlotFiles()
load(files$plotdata)
```

Retention Rates for Signal/Noise Variables

We plot the mean value, across all data replicates, of the share of signal variables in the original feature set that appear in the testing subset selected by the PhiMr filter. On the same plot, we also include the mean share of noise variables selected by PhiMr for comparison.

```
makePhimrRetentionPlots(plotdata_phimr)
```



The mean share of noise variables retained was consistently 0.5 across all scenarios seen here.

Among the signal variables, the mean share of signal variables retained was greater than that of noise variables, with the magnitude of the difference increasing with sample size (horizontal axis) and with the number of signal variables (top vs. bottom panels).

Notably, the individual signal variables in the 28-variable signal set (top panels) have very different marginal signal strength from those in the 122-variable signal set (bottom panels); in our simulation design, a signal of similar strength was distributed over the signal variables in each signal set (specifically, for each response component j, the coefficient c_j for the effect function h_j was tuned such that over many observations of $\boldsymbol{X}$, the sample standard deviation of $h_j(\boldsymbol{X})$ was similar for both signal sets), resulting in stronger marginal signals for smaller signal sets.

From the bottom panels, we see that when the marginal signals were weak and sample size was small, PhiMr did not retain signal variables at a much higher rate than noise variables.

Feature dimension had no effect on the share of signal or noise variables kept in any of the scenarios for this setting.

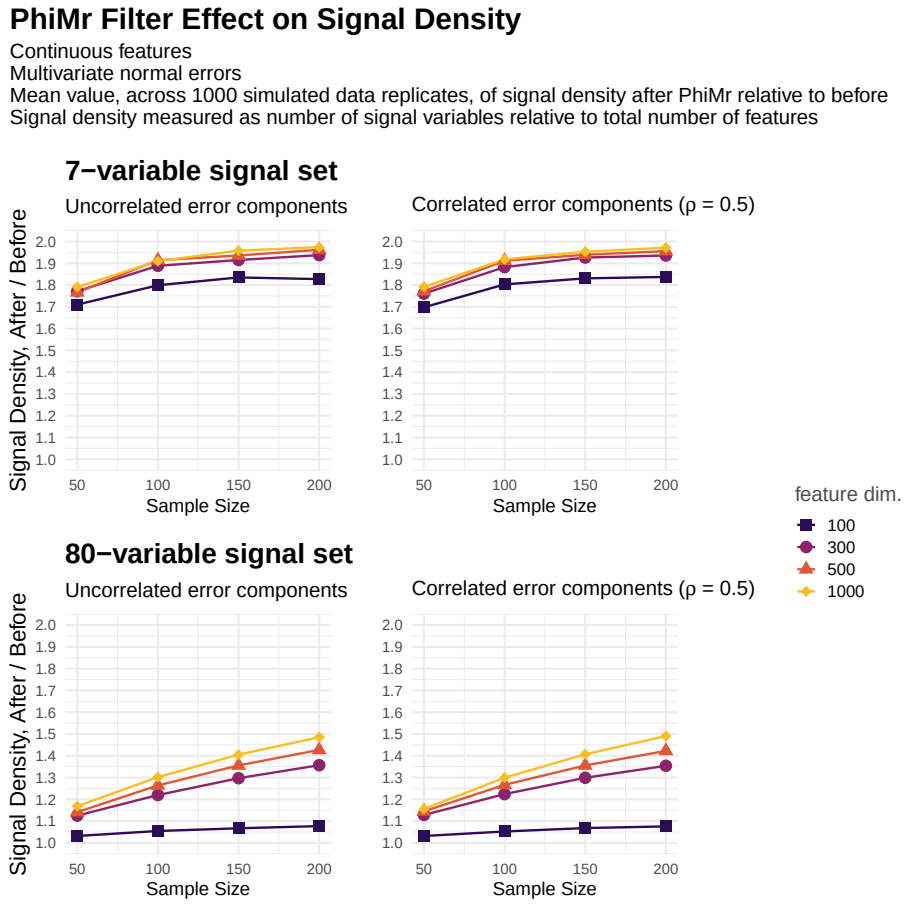
Correlation between components of the random error vector $\boldsymbol{\varepsilon}$ (left vs. right panels) did not appear to have

an effect on the share of noise or signal variables retained in this setting.

Effect of PhiMr on Signal Density

We plot the mean value, across all data replicates, of the signal density after applying PhiMr as a share of the signal density before applying PhiMr, where signal density is measured as the number of signal variables relative to the total feature dimension.

```
makePhimrDensityPlots(plotdata_phimr)
```



For reference, a value of 1 on the vertical axis corresponds to the testing subset selected by PhiMr having the same signal density as the full feature set; values greater than 1 correspond to a testing subset with greater signal density than the full feature set, while values less than 1 correspond to a testing subset with lower signal density than the full feature set.

Here we see that the mean value was greater than 1 in all scenarios considered here, indicating that PhiMr always yielded an improvement in signal density.

Sample size and marginal signal strength both influenced the increase in signal density in this setting, which is expected given their influence on the share of signals selected (and lack of influence on the share of noise selected) observed in the previous plot.

For a particular signal set and sample size, the increase in signal density grew with the feature dimension. This is expected given our other observations in this setting: if we denote the number of signals and total number of features selected by PhiMr as p^* and γ^*, respectively (and the initial number of signals by γ), the

signal density after applying PhiMr relative to before is

$$\frac{\gamma^*/p^*}{\gamma/p} = \frac{p\gamma^*}{p^*\gamma}.$$

We observed in the previous plot that for the current setting, when the number γ of signals is fixed, changing the feature dimension p did not affect the mean retention rates for signal or noise; thus, on average, the number γ^* of signals selected by PhiMr should also be unchanged. Furthermore, with γ fixed, increasing p solely increases the number of noise variables, and because these are not all selected by PhiMr, p^* should increase by less than p, leading to a greater value of the above ratio of signal density after PhiMr to before.

Looking at the scenarios with 80 signal variables (bottom panels), we notice that when $p = 100$, the signal density did not improve much after PhiMr; this is explained by the fact that there were only 20 noise variables to begin with, yielding an initial signal density already close to 1 and thus limiting the relative increase possible from PhiMr. In other scenarios, which had a more substantial share of noise variables and were thus more representative of the motivational setting for PhiMr, we saw PhiMr yield a substantive gain in signal density given a reasonable combination of sample size and marginal signal strength.

Discrete Features with Weakly Correlated Signals

This section includes scenarios where $\boldsymbol{X}$ is simulated as a discrete random vector representing additive-encoded SNP-set data. In this setting, the signal variables are comprised of SNPs among which pairwise correlations are typically mild.

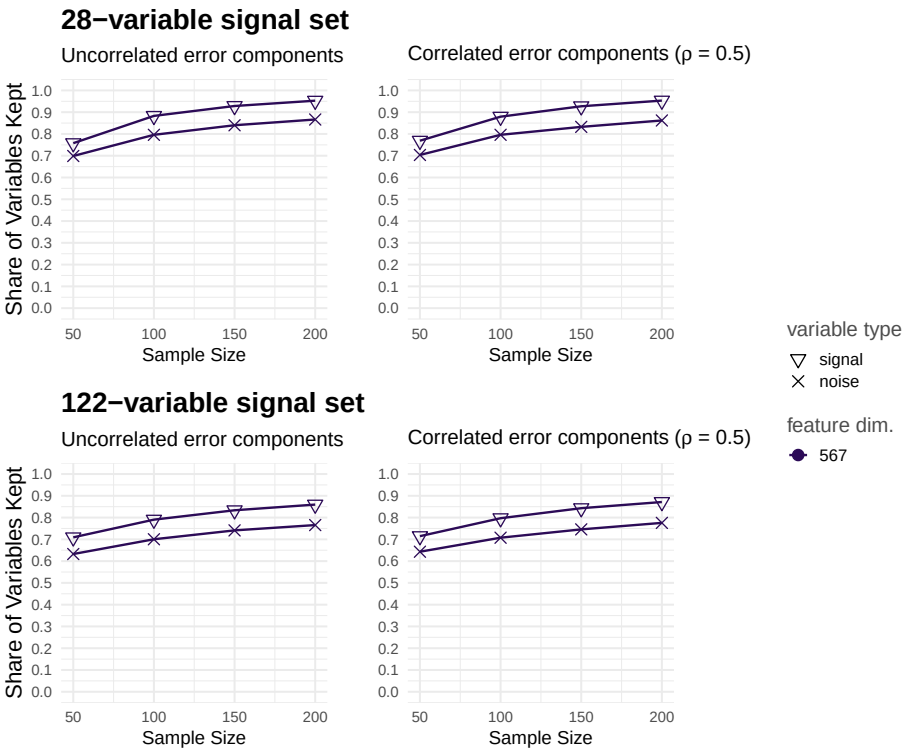
```
x_type <- 'snp'
signal_correlation <- "low"
source(file.path(dir_src, 'initialize_adaptive_across.R'))

# Create filenames and load plot data
files <- adaptiveAcrossPlotFiles()
load(files$plotdata)
```

Retention Rates for Signal/Noise Variables

We plot the mean value, across all data replicates, of the share of signal variables (respectively, noise variables) in the original feature set that appear in the testing subset selected by the PhiMr filter.

```
makePhimrRetentionPlots(plotdata_phimr)
```



Comparing this setting to the previous setting with continuous features, we see that on average, a higher share of noise variables were retained by PhiMr: in the continuous setting, the average share of noise retained was consistently 0.5 in all scenarios; here, the average share of noise kept ranges from 0.65 to 0.85. Whereas this value was constant across scenarios in the continuous setting, here it increased with both the sample size as well as the marginal signal strength.

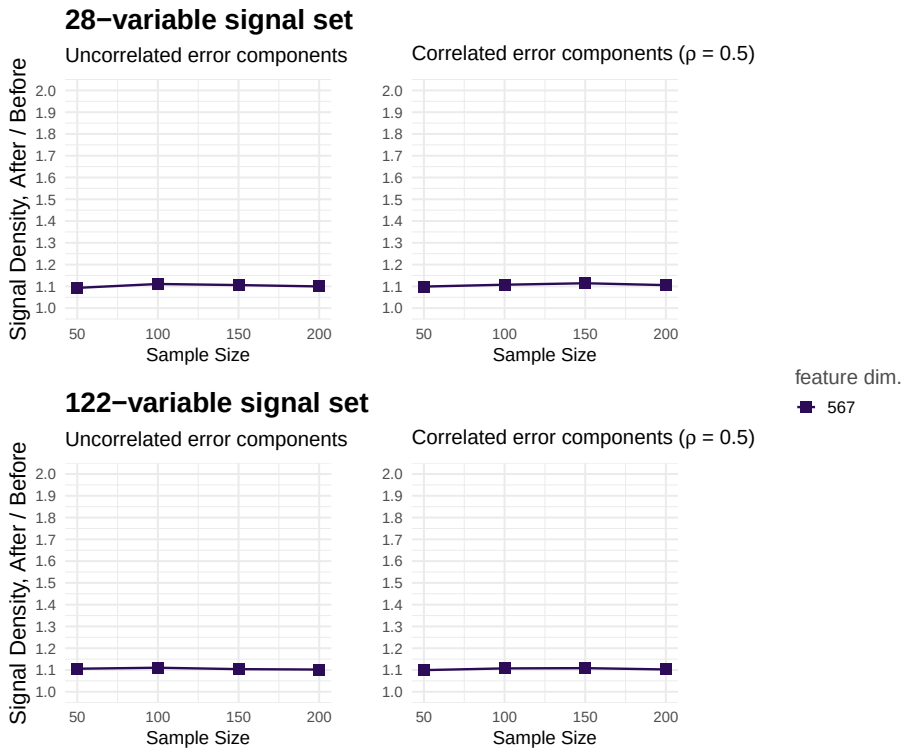
While a greater share of signal variables were retained on average as compared to noise variables, the relative and absolute differences here were substantially less than those observed in the continuous setting. The mean share of signals retained was affected by sample size and marginal signal strength similarly as for noise variables, with the result being that across signal sets and sample sizes, the average difference between the share of signals kept and the share of noise kept (the vertical distance between the two lines) was roughly constant.

As in the continuous setting, the correlation among random error components did not show an effect on PhiMr's behavior.

Effect of PhiMr on Signal Density

We plot the mean value, across all data replicates, of the signal density after applying PhiMr as a share of the signal density before applying PhiMr, where signal density is measured as the number of signal variables relative to the total number of feature variables.

```
makePhimrDensityPlots(plotdata_phimr)
```



Here we see that PhiMr always yielded an improvement in signal density, but this improvement was consistently modest, with the signal density increasing by roughly one tenth of its original value, on average.

This average relative increase was constant across sample sizes and signal sets, similar to the average difference between the share of signals kept and the share of noise kept (as observed in the previous plot).

Discrete Features with Strongly Correlated Signals

This section includes scenarios where $\boldsymbol{X}$ is simulated as a discrete random vector representing additive-encoded SNP-set data. In this setting, the signal variables are comprised of SNPs among which pairwise correlations are typically very strong.

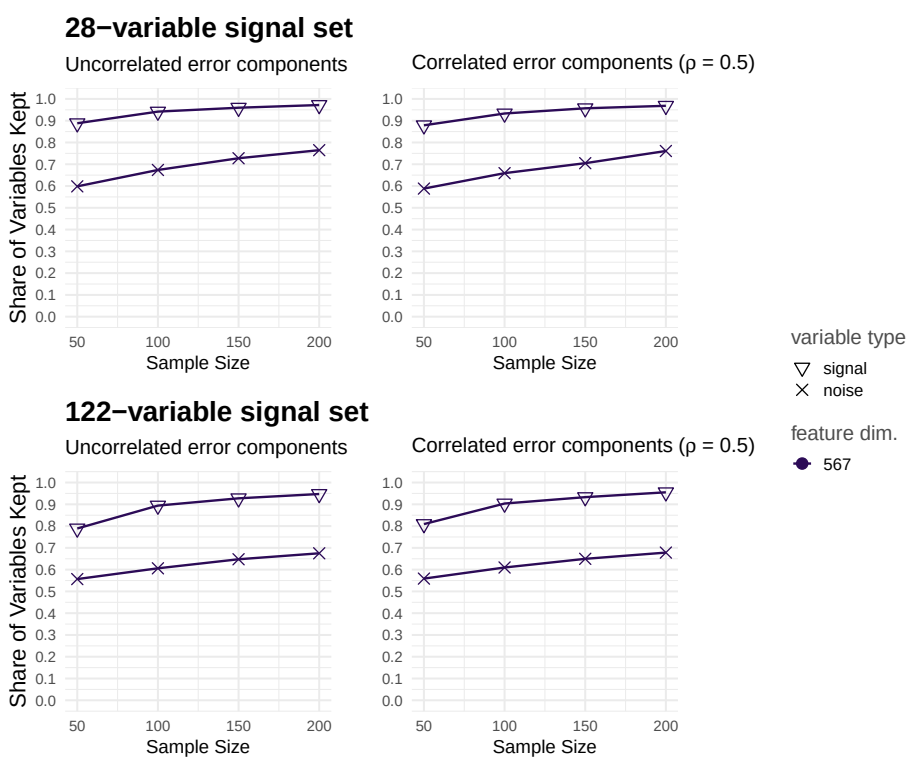
```
x_type <- 'snp'
signal_correlation <- "high"
source(file.path(dir_src, 'initialize_adaptive_across.R'))

# Create filenames and load plot data
files <- adaptiveAcrossPlotFiles()
load(files$plotdata)
```

Retention Rates for Signal/Noise Variables

We plot the mean value, across all data replicates, of the share of signal variables (respectively, noise variables) in the original feature set that appear in the testing subset selected by the PhiMr filter.

```
makePhimrRetentionPlots(plotdata_phimr)
```



Interestingly, when we contrast these results with those for the previous setting where the signal variables were weakly correlated, we notice that here, where the signal variables were strongly correlated, the average share of noise variables retained by PhiMr was lower, while the average share of signal variables retained was higher, resulting in a greater average difference between the two values.

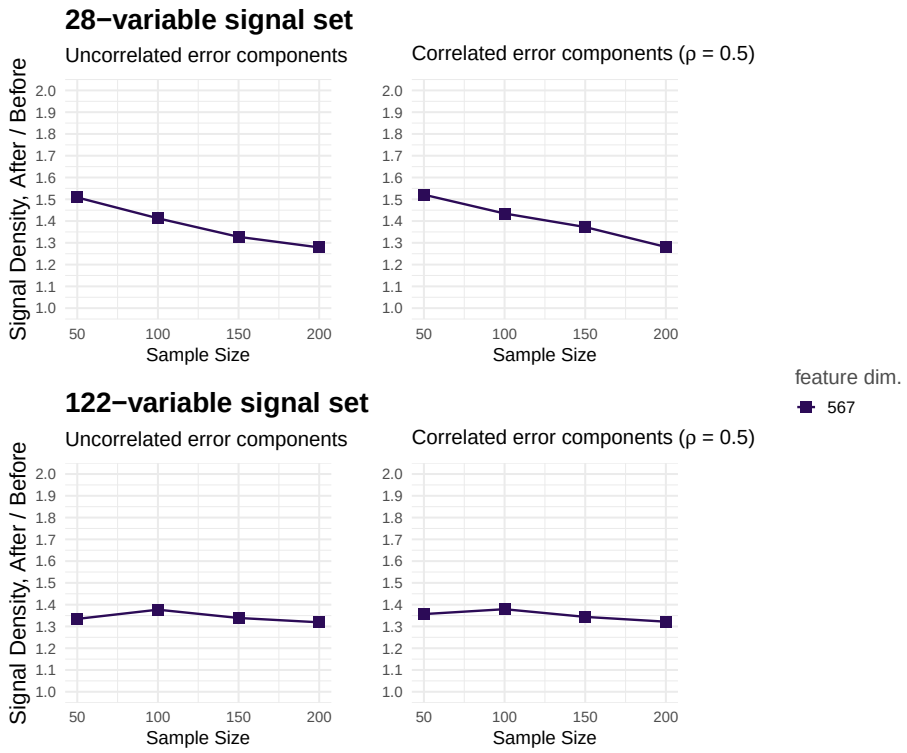
As in the other discrete setting, the shares of signal and noise kept were both similarly affected by sample size and marginal signal strength, though here the share of signals kept is already high enough to where we see diminishing marginal returns from the increases in sample size at the upper end of the range considered here.

As in other settings, we see no effect of the correlation strength among random error components on PhiMr's behavior.

Effect of PhiMr on Signal Density

We plot the mean value, across all data replicates, of the signal density after applying PhiMr as a share of the signal density before applying PhiMr, where signal density is measured as the number of signal variables relative to the total number of feature variables.

```
makePhimrDensityPlots(plotdata_phimr)
```



Here the average improvement in signal density is considerably greater than we observed in the discrete setting with weakly-correlated signal variables, consistent with the greater average differences between the share of signals kept and the share of noise kept.

Interestingly, for the 28-variable signal, the increase in signal density became more modest as sample size increased; this appears likely attributable to faster rate at which the share of noise variables selected increased with sample size relative to the same rate for signals (as seen in the top panels of the previous plot).

The 122-variable signal set (bottom panels) shows a more modest version of this trend, with the gain in signal density diminishing at a much slower rate, and with the difference in retention rates between signal and noise less pronounced (bottom panels of previous plot).