Feature Selection and Predictive Performance in Incomplete, Higher-Dimensional Datasets

Scenario 38: Summary Results

Introduction

The overall goal of this project is to evaluate the performance of various methods in terms of feature selection and predictive accuracy when applied to incomplete, smaller-sample, higher-dimensional datasets with the possibility of high pairwise correlation among the features. The supervisor is either continuous or binary and the features are multivariate normal. This markdown file presents the results of scenario 38 (of 40 total scenarios). Full results can be found at the associated RShiny page or in the associated paper.

Four methods are considered, and their respective details are incorporated into the associated markdown files. As a brief overview, the base case method is Multiple Imputation Random LASSO ("MIRL", Liu, et. al., 2016) which utilizes the Multivariate Imputation by Chained Equations ("MICE") algorithm for the imputation of missing data and an enhanced Random LASSO process for feature selection and coefficient estimation. Alternative method 1 ("MFRL") instead utilizes Missing Forest for the imputation portion of the algorithm (MissForest, Stekhoven and Buhlmann, 2012). MissForest was found to outperform KNN and MICE using various medical datasets, but with smaller amounts of missing data than used in this study. Alternative methods 2 and 3 originate from the observation of similarities between MIRL and Random Forest. Alternative method 2 ("RF") replaces the Random LASSO portion of the base method with Random Forest and utilizes the na.roughfix imputation option instead of MICE. Alternative method 3 is similar to 2 but instead utilizes the imputed data from the MissForest algorithm ("MFRF").

Scenario Parameters

100 training and testing datasets consisting of 200 observations each were simulated using the characteristics highlighted below. The supervisor is a function of the first 10 features and a normal error term and is either continuous or binary. The features are multivariate normal random variables with the selected pairwise correlation pattern. The number of noise features is equal to the total number of features identified in the scenario parameters minus 10.

Table 1: Selected scenario parameters

Supervisor	No. Features	Pct. Complete Cases	Correlation
Continuous	35	50%	Low (.2)
Binary	**60**	**25%**	Mid (.6)
	110		**Mixed-High (.3, .75-.85)**
	210		
	260		

Missing Data

The simulated datasets are designed to incorporate a significant amount of missing data with approximately 50% or 75% of observations containing at least one missing feature. The black lines in the chart below indicate the values that are missing from the first of the 100 simulated datasets. Note that a good portion of the missing data lies in features 5 and 10 which were used to generate the supervisor. The “1-SimulateDatasets.RMD” file contains further details and the programming for the simulated data.

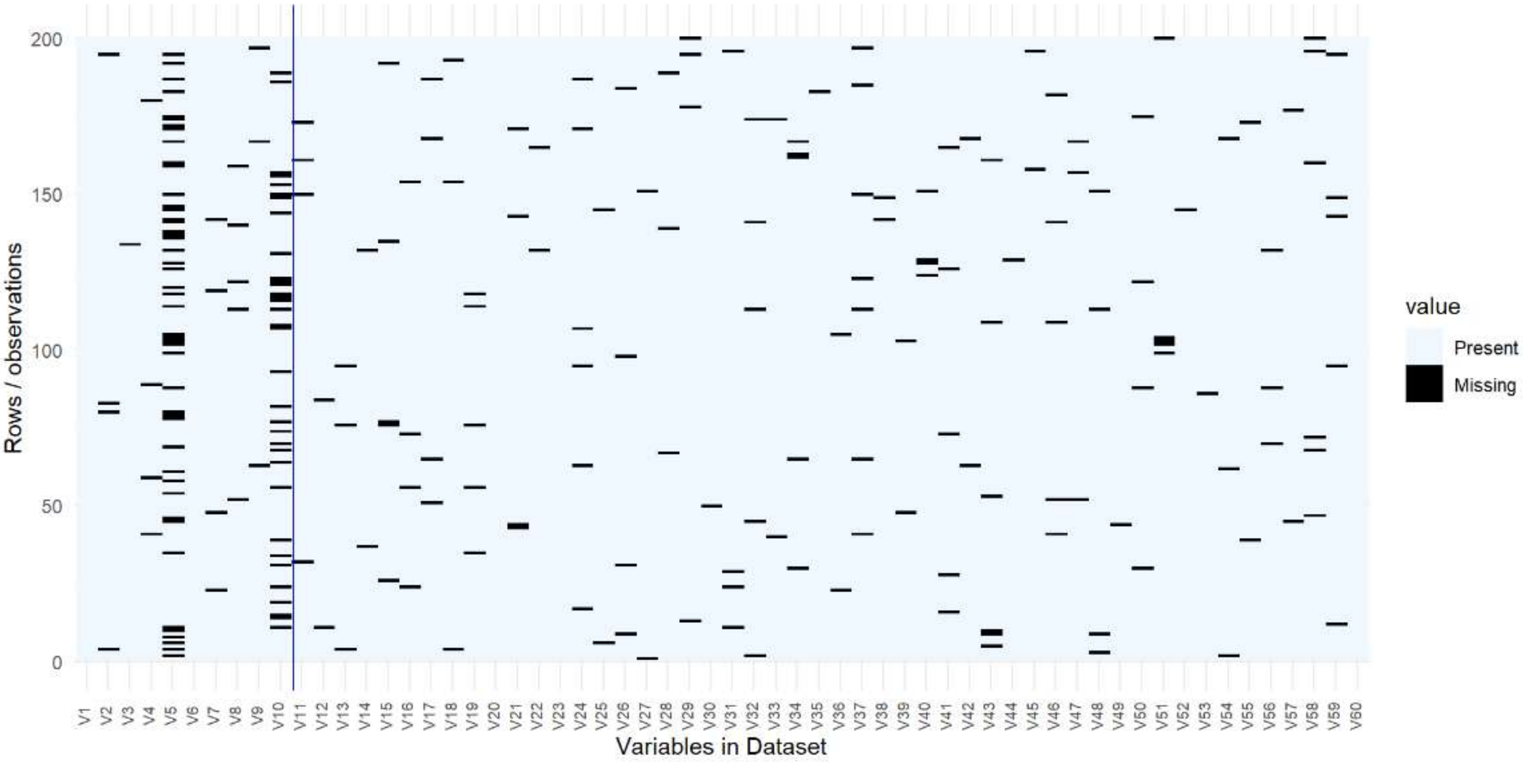

Fig. 1: Graphical depiction of present/missing features by observation

Feature Correlation

An example pairwise correlation structure for the features is shown below. The size and color of the dots represent the correlation between two simulated features, with the bigger circle and darker blue color indicating stronger positive correlation. Note that in this scenario, several features have very high levels of correlation (.75 to .85), of which, features 3-6 were included in the generation of the supervisor.

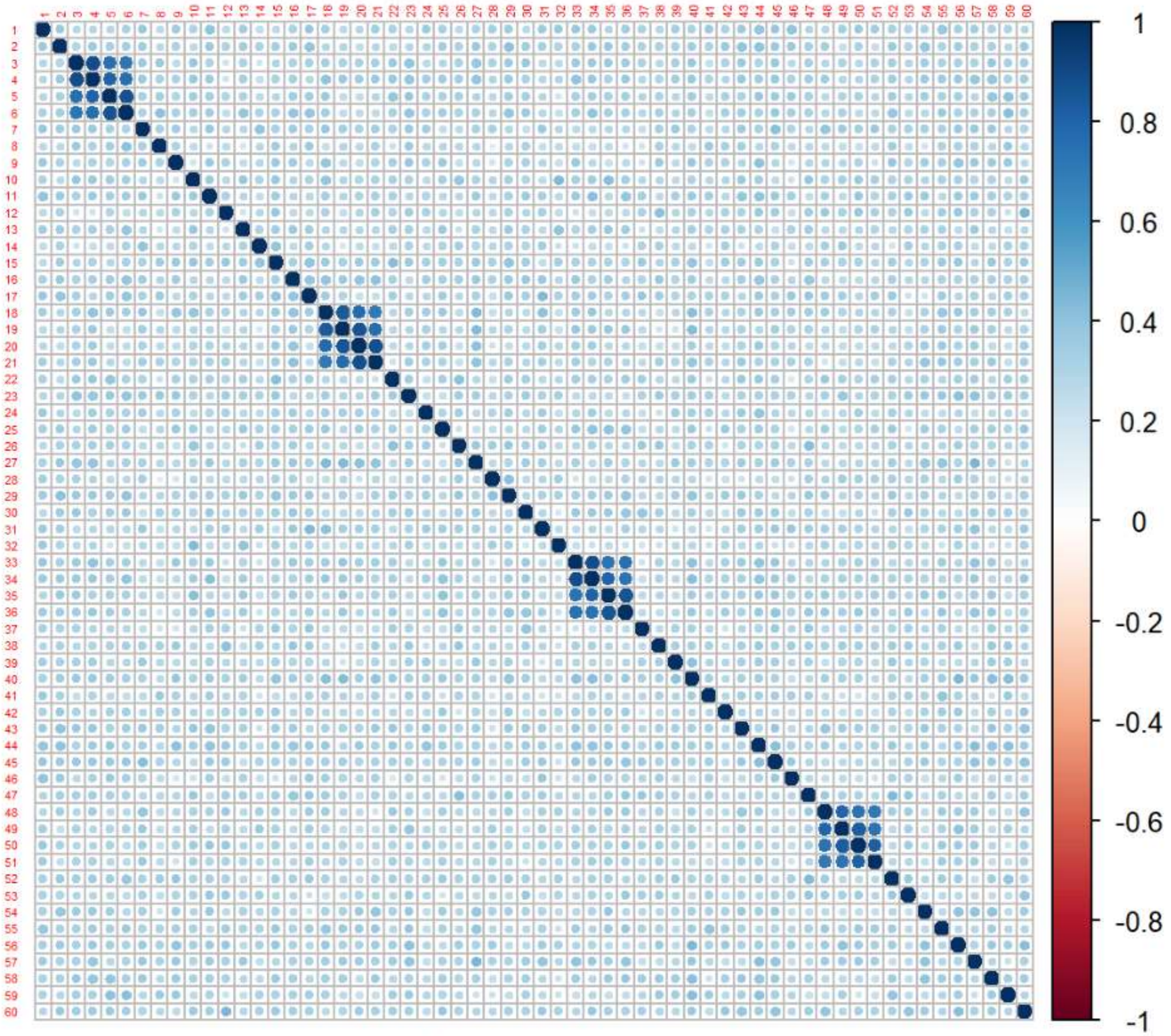

Fig. 2: Example pairwise feature correlation

Results Comparison

Feature Selection

Five feature selection performance metrics are used. The first is the number of times a truly associated feature is selected from the 100 separate datasets and is shown in Table 2, below. This table also lists each of the 10 features used to generate the supervisor along with its coefficient and association.

Overall, in this scenario, all 4 methods do very well in picking up the true features with the highest proportion of missing data and highest coefficients (#5 and 10). More generally, the Random LASSO methods ("RL") outperform the random forest methods ("RF"). MIRL slightly outperforms MFRL in identifying the features with the highest coefficients (#3-5,8-10), but MFRL does a better job at picking up the features with the smaller coefficients (#1,2,7). Neither do a good job at picking up the negatively associated feature #6, which does not contain any missing data, but is highly correlated with features 3 through 5. Interestingly, both RF based methods do a very good job in selecting feature #6, indicating that a combination of a RL and RF approach could be useful for feature selection.

Table 2: Feature Selection Performance Metrics

	Association	MIRL	MFRL	RF	MFRF
Coefficient: 0.5					
5	Positive	98	94	96	99
10	Negative	100	99	94	98
Coefficient: 0.4					
4	Positive	99	99	100	100
9	Negative	100	99	91	88
Coefficient: 0.3					
3	Positive	94	89	100	100
8	Negative	90	89	73	70
Coefficient: 0.2					
2	Positive	72	75	24	24
7	Negative	68	70	45	46
Coefficient: 0.1					
1	Positive	40	46	18	17
6	Negative	18	18	92	90

Single Performance Metrics

Matthew's Correlation Coefficient

Following Liu, et. al. (2016), the second feature selection performance metric is Matthew's Correlation Coefficient ("MCC", Matthews, 1975). The formula is included in "2-MIRL.RMD". MCC is a single, consolidated measure of the confusion matrix. Its value ranges from -1 (incorrectly identified all features) to +1 (correctly identified all 10 associated features and the noise features) with a higher positive value indicating better feature selection performance.

Note: following Liu, et. al. (2016), two versions of MCC are shown: *Top-10* and *CV Threshold*. In Top-10 the methods assume there are 10 true features and thus 10 features are selected (Figure 3). In CV Threshold, the number of true features is unknown and the methods determine the number of features to select (Figure 4). Thus, the CV Threshold method better reflects real world implementations. Both RL methods outperform the RF methods, with MIRL slightly outperforming MFRL but still within the one-standard error band.

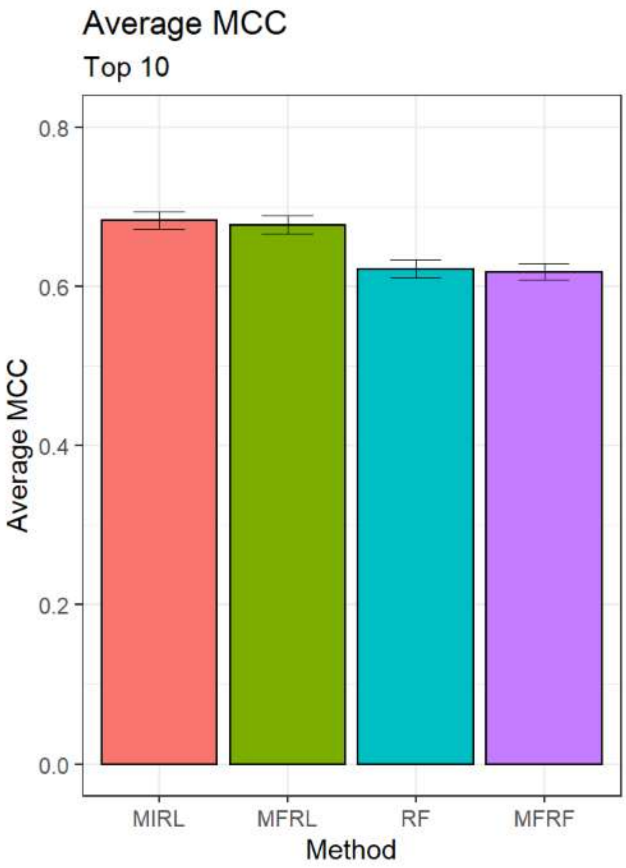

Fig. 3

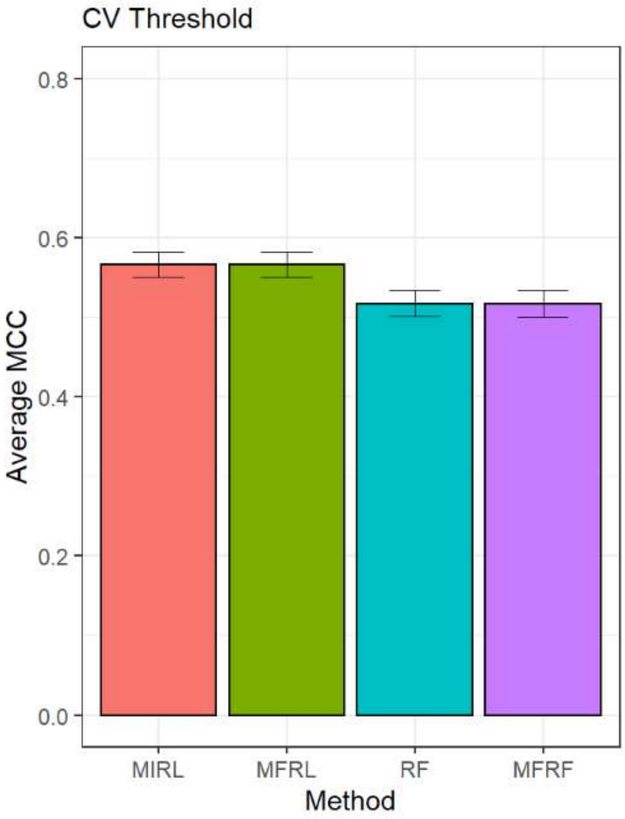

Fig. 4

True Positives

The average number of true positives, over 100 simulated datasets, measures how many features each method correctly selects as being truly associated with the supervisor. Again, the RL methods outperform the MF methods, while the performance of MIRL and MFRL are similar. The results for the CV Threshold method are slightly better than in Top 10 as the Threshold method usually results in the selection of more than 10 features, increasing the chance that the truly associated features are chosen.

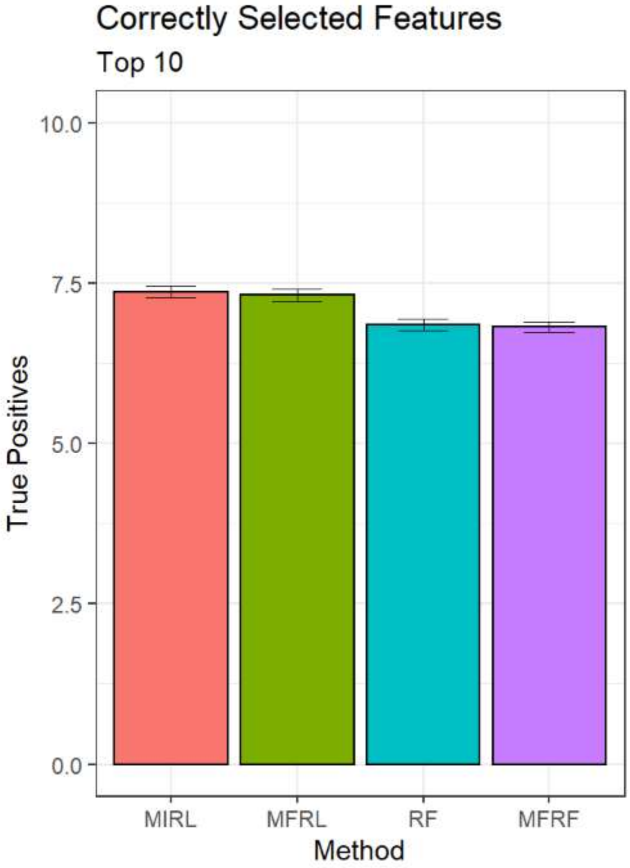

Fig. 5

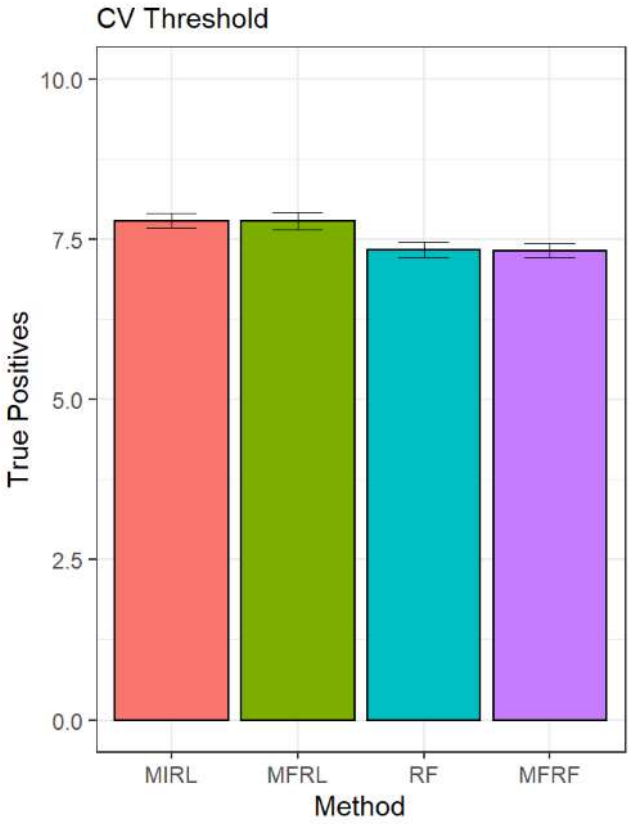

Fig. 6

False Positives

The average number of false positives measures how many noise features are incorrectly selected. The average number of false positives between the Top 10 and CV Threshold scenarios is quite different as in many instances, the CV Threshold scenario selects more than 10 features. Again, RL performance exceeds RF but in the CV Threshold scenario, performance is within the one standard error band.

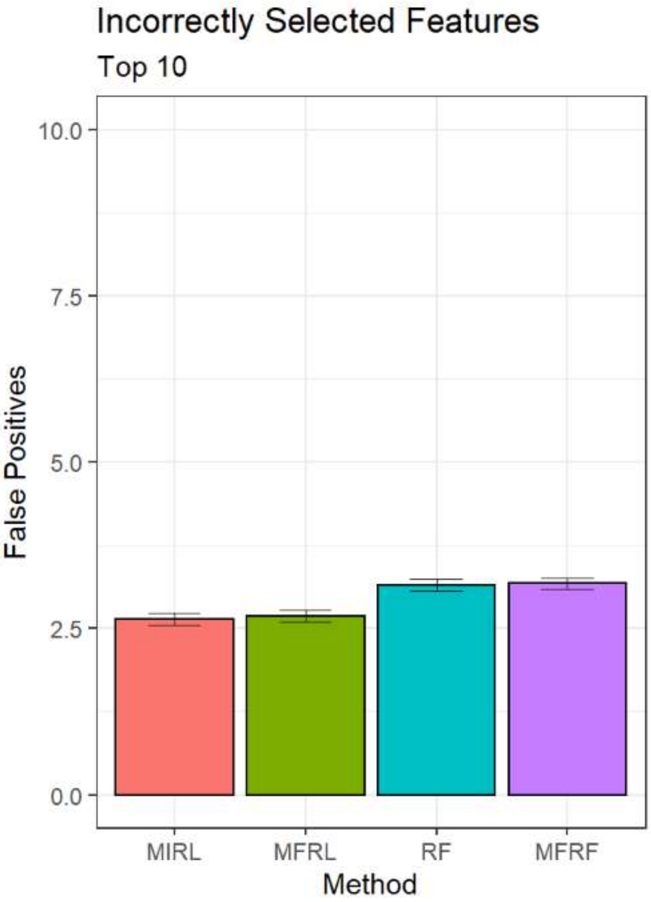

Fig. 7

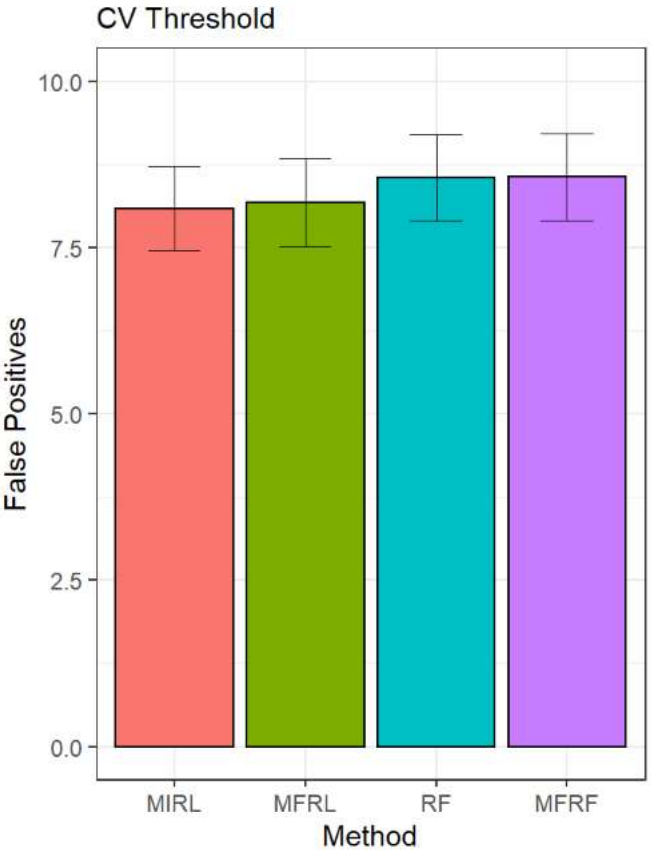

Fig. 8

False Negatives

The average number of false negatives measures how many features that are truly associated with the supervisor are missed by each method. Both MIRL and MFRL perform similarly.

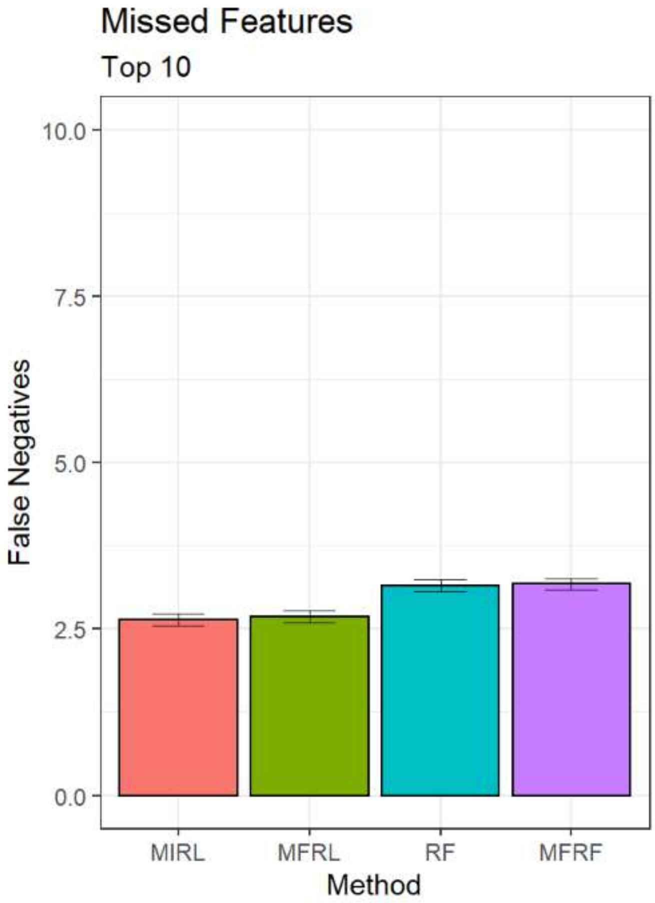

Fig. 9

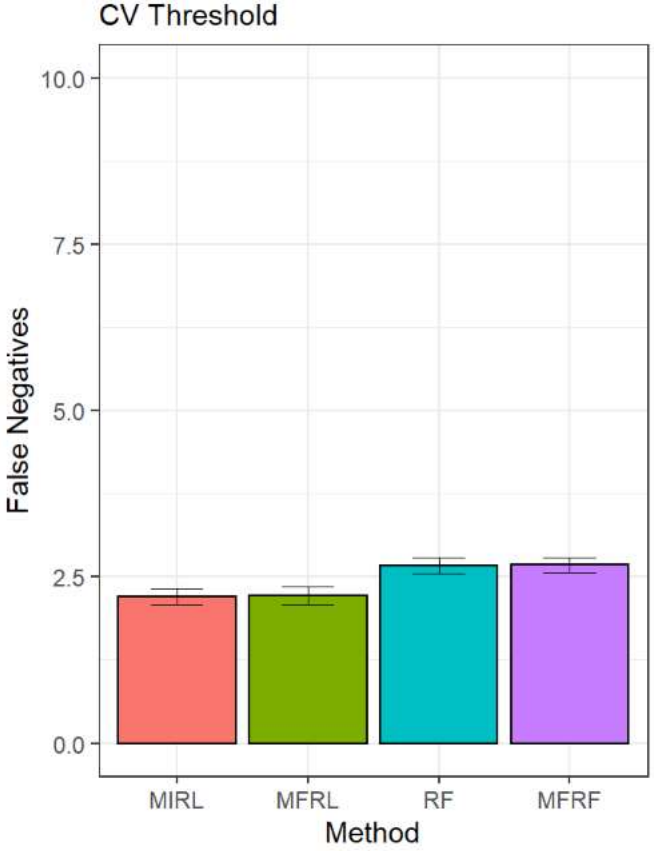

Fig. 10

Predictive Performance

$RMSE$

Predictive performance, for scenarios with a continuous supervisor, is compared using root mean squared error (RMSE). Lower values indicate better predictive performance. The RL based methods perform significantly better than the RF based methods, as found in other scenarios. MIRL slightly outperforms MFRL in this scenario. The predictive performance was similar between the Top 10 and CV Threshold scenarios.

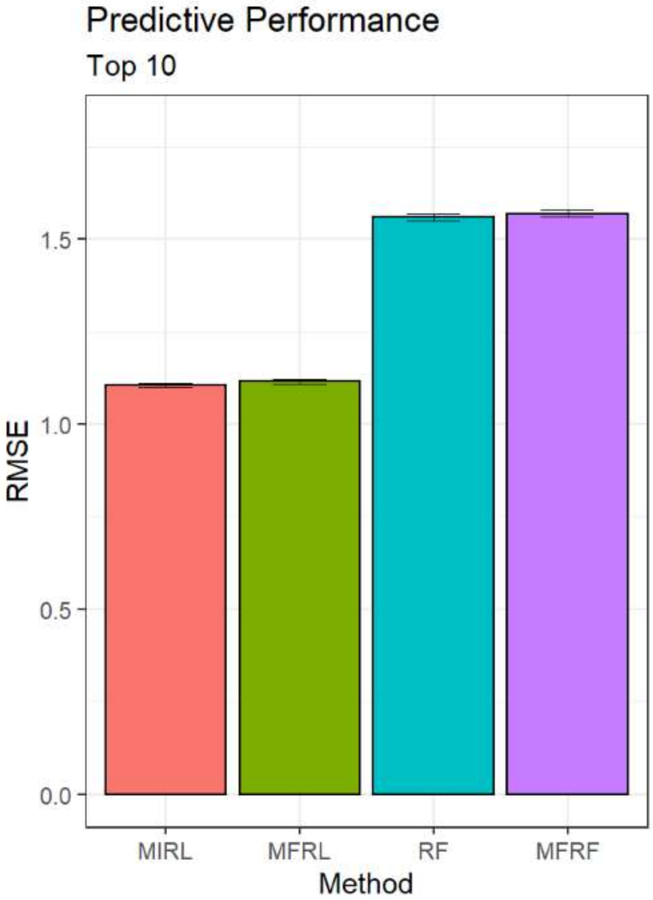

Fig. 11

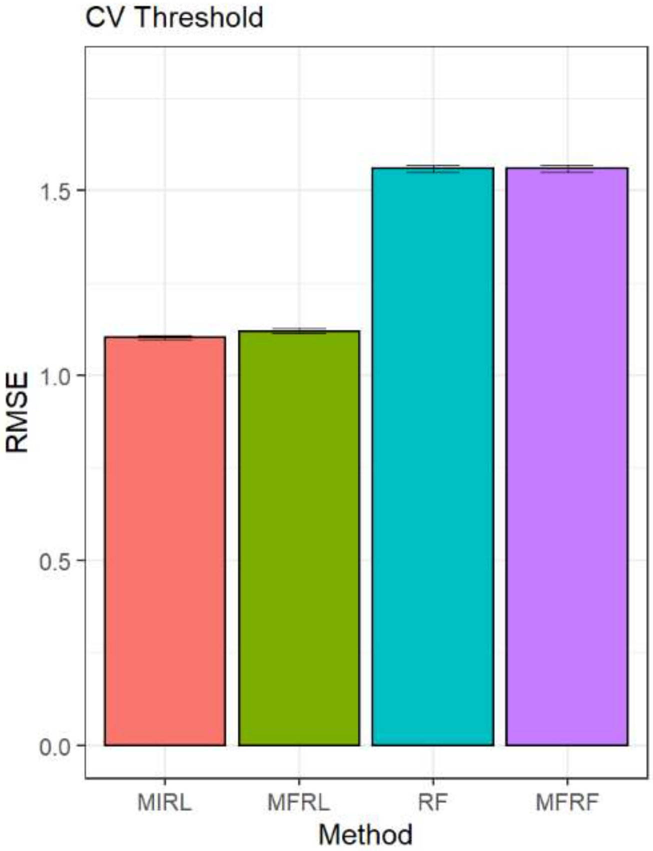

Fig. 12

A jigger plot of all 20,000 test supervisor values (from all 100 simulated test datasets) overlayed with a boxplot is shown below (Figure 13) to provide a basis for evaluating the magnitude of the RMSE. 50% of the test supervisor values lie in the (-1,1) range and a little over 99% lie in the (-4,4) range.

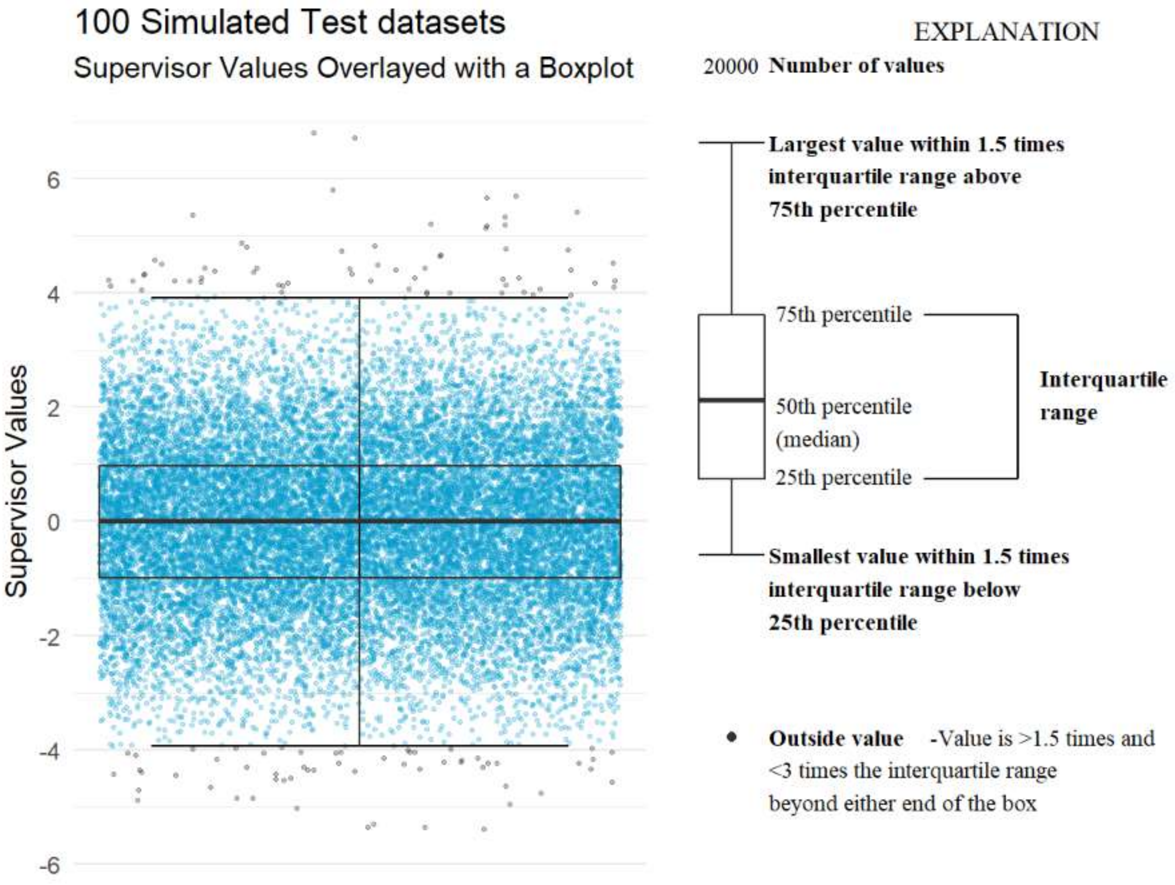

Fig. 13

Conclusion

In this scenario, MIRL slightly outperformed MFRL (and the other alternative RF based methods) in most of the performance metrics, but not by much. MFRL performed better in identifying the features with smaller coefficients (#1,2,7) and the RF based methods were best at picking up the negatively associated, and highly correlated feature #6, indicating that a combination of RL and RF methods for feature selection may be the best approach for datasets matching these scenario parameters. High proportions of missing data in features 5 and 10 didn't prohibit any of the methods from correctly selecting those features.

Full analysis and results, along with references, are contained in the associated paper and a summary of all scenario results are contained in the RShiny page.