TASK

Compare performance of several well-established model selection methods. The methods are lasso, forward / backward stepwise selection based on p-values, forward / backward stepwise selection based on Akaike information criterion, forward / backward stepwise selection based on Bayesian information criterion and forward / backward stepwise selection based on cross-validation. Use several performance metrics including root-mean-square error and percentage of correctly identified true predictors.

SOLUTION

VARIABLE SELECTION IN SIMULATED SETTING

The purpose of this study is to compare several estimation methods for linear models. As performance characteristics we choose 4 measures:

1) TP: percentage of variables selected out of the universe of true predictors,
2) FP: percentage of variables selected out of the universe of false (noise) predictors,
3) CE: root-mean-square error of an estimate of a regression coefficient,
4) VA: percentage of the variation in the true signal captured by the estimated model.

The estimation methods that we try are

1) forward stepwise regression based on p-values and increases in R^2,
2) backward stepwise regression based on p-values,
3) forward stepwise regression based on Akaike Information Criterion (AIC),
4) backward stepwise regression based on Akaike Information Criterion ,
5) forward stepwise regression based on Bayesian Information Criterion (BIC),

6) backward stepwise regression based on Bayesian Information Criterion,

7) forward stepwise regression based on cross-validation (CV),

8) backward stepwise regression based on cross-validation,

9) Lasso, where the shrinkage parameter is chosen based on the Cp statistic (an equivalent of AIC).

To compare the 9 methods we do the following. We define the specification of an experiment as

- the number of "true" predictors in the experiment (variables that influence the dependent variable and are considered as predictors in a regression),
- the correlation between each two true predictors,
- the number of "false" predictors (irrelevant variables which we try in a regression),
- the correlation between each two false predictors,
- the signal-to-noise ratio,
- the degrees of freedom of the t-distribution of the residual, where $+\infty$means that the residual is Normal.

For each specification we simulate many independent sets of regression coefficients. Each set corresponds to a separate linear model. We run each estimation method on each model, compute the coefficients estimates and determine the 4 performance scores (TP, FP, CE and VA). Then, for each experiment specification, we average the performance scores over the models to get an estimate of the performance measure (separately for TP, FP, CE and VA). We calculate the standard error of the estimate as the standard deviation of the performance scores divided by the square root of the number of models.

Next we examine each performance measure separately. For each experiment specification we calculate the 95% confidence intervals corresponding to the 9 estimation methods and see if they overlap.

The simulation results are presented later in this document. In the next section we describe the whole range of simulated models mathematically.

MODEL

Each simulated model has the following form:

$$Y = Beta\_0 + Beta\_1 * X\_1 + ... + Beta\_p * X\_p + E, \quad (1)$$

where X_1, ..., X_p are standard normal variables and each two of them have correlation Rho_X. Residual E has either t-distribution with V degrees of freedom or standard normal distribution (which is equivalent to the t-distribution having $+\infty$ degrees of freedom). The signal is defined as the predictable part of equation (1), namely (Beta_0 + Beta_1 * X_1 + ... + Beta_p * X_p). The signal-to-noise ratio is defined as

$$S/N = Var[Beta\_0 + Beta\_1 * X\_1 + ... + Beta\_p * X\_p] / Var[E].$$

We call variables X_1, ..., X_p the "true" predictors. Unfortunately, when we build a linear model for Y, we do not know the exact list of factors that influence Y. Therefore variables X_1, ..., X_p are only part of the universe of "candidate" predictors. In addition to X's, the universe has variables Z_1, ..., Z_q, which are completely uncorrelated with Y. They represent nuisance, giving way to potential errors in model selection. We call variables Z_1, ..., Z_q the "false" predictors. We assume that each of them has standard normal distribution and each two have correlation Rho_Z.

Together values of p, Rho_X, q, Rho_Z, S/N and V constitute the specification of a random experiment. In each independent realization of such experiment we simulate true regression coefficients (Beta_1, ..., Beta_q) as independent standard normal variables. The estimation methods are run on centered versions of Y, X_1, ..., X_p, Z_1, ..., Z_q. Therefore, the value of Beta_0 does not matter and is set to 0.

In each experiment we simulate independent realizations of (Y, X_1, ..., X_p, Z_1, ..., Z_q) to create a sample of size N. Suppose that a given estimation method produces estimates (Beta_1_Hat, ..., Beta_p_Hat, Gamma_1_Hat, ..., Gamma_q_Hat) of the coefficients of (X_1, ..., X_p, Z_1, ..., Z_q). We know that the true coefficients are (Beta_1, ..., Beta_p, 0, ..., 0). Therefore we can define the performance characteristics in the following way:

TP = percentage of variables selected out of the universe of true predictors = (number of non-zero Beta_i_Hat) / p * 100,

FP = percentage of variables selected out of the universe of false (noise) predictors = (number of non-zero Gamma_i_Hat) / q * 100,

CE = root-mean-square error of an estimate of a regression coefficient =

$$= sqrt([Sum\_\{i=1,..,p\} (Beta\_i\_Hat - Beta\_i)^2 + Sum\_\{i=1,..,q\} (Gamma\_i\_Hat - 0)^2] / [p+q]),$$

VA = percentage of the variation in the true signal captured by the estimated model =

= R^2 of regressing the true signal (Beta_1 * X_1 + ... + Beta_p * X_p) on the estimated signal (Beta_1_Hat * X_1 + ... + Beta_p_Hat * X_p + Gamma_1_Hat * Z_1 + ... + Gamma_q_Hat * Z_q) =

= [(Beta_1_Hat, ..., Beta_p_Hat, Gamma_1_Hat, ..., Gamma_q_Hat)' * [joint covariance of X and Z] * (Beta_1, ..., Beta_p, 0, ..., 0)]^2 /

/ [(Beta_1_Hat, ..., Beta_p_Hat, Gamma_1_Hat, ..., Gamma_q_Hat)' * [joint covariance of X and Z] * (Beta_1_Hat, ..., Beta_p_Hat, Gamma_1_Hat, ..., Gamma_q_Hat) * (Beta_1, ..., Beta_p)' * [joint covariance of X] * (Beta_1, ..., Beta_p)].

RESULTS

It's not a big deal for any estimation method to deliver accurate results on a large sample, where there are at least 20 observations per candidate predictor. Therefore, we choose to work with the sample size 100. It is easier to contrast the accuracy of various estimation methods on a relatively small sample. On the other hand, we simulate 200 models for each experiment specification to achieve tight confidence intervals for each performance characteristic. Note that our code allows for generating several independent samples per each random linear model but we chose to generate just one, saving the computational resource for generating many different models.

For Forward / Backward Cross-Validation, we choose 5 as the number of folds. This choice allows for a decent number of observations to be used at each training round (4/5 * 100 = 80)

Any tested experiment specification is a modification of the following base specification:

p = 5,
Rho_X = 0.5,
q = 7,
Rho_Z = 0.5,
S/N = 1,
V = +∞.

At each stage of the analysis, we substitute 1 or 2 parameters with the whole array of possible values and study the variability of the performance characteristics with respect to these 1 or 2 parameters. <*The complete set of results has been provided to the client as R files, 2D plots, 3D plots and R infrastructure capable of testing various specifications. Here we display only the 2D plots, which capture the gist of our findings*.>

The 2D plots visualizing the results are placed into the appendix. The lines correspond to the performance characteristics of different estimation methods. The dots visualize the 95% confidence intervals of the performance characteristics. By looking at the confidence intervals we can see if one method is significantly better than another.

Based on the simulation results we make the following conclusions:

- In terms of TP, CE and VA, uniformly over most parameter specifications, Lasso performs better than Forward / Backward Cross-Validation, and Forward / Backward Cross-Validation performs substantially better than the other methods, based on p-values, AIC and BIC.
- In terms of FP, uniformly over most parameter specifications, Lasso performs substantially worse than Forward / Backward Cross-Validation and Forward / Backward Cross-Validation performs substantially worse than the other methods.
- The results for Forward / Backward P-values, Forward / Backward AIC and Forward / Backward BIC are comparable. The relative differences between the performance characteristic are not big. Nonetheless, sometimes the differences are statistically significant. We will describe some of these cases in detail below.
- In terms of TP, Forward / Backward BIC systematically underperforms Forward / Backward AIC, choosing too simplistic models on a small data set of 100 observations.
- In terms of FP, Forward / Backward AIC systematically underperforms Forward / Backward BIC, choosing too liberal models, containing noisy, useless predictors.
- In terms of CE and VA, for most specifications, there is no statistically significant difference between Forward / Backward P-values, Forward / Backward AIC and Forward / Backward BIC.
- The performance is best when there is only one true predictor to select. The performance deteriorates rapidly with the number of true predictors increasing from 1 to 10. Lasso and Forward / Backward Cross-Validation are more robust to this increase than the other methods.
- The performance is insensitive to the correlation between true predictors unless this correlation is really high, e.g. 0.7 - 0.9.
- The 9 estimation methods have approximately the same sensitivity to increases in the signal-to-noise ratio, except for increases at the lower end of the signal-to-noise, where Forward / Backward Cross-Validation and Lasso benefit more.

- The lower the residual degrees of freedom are, the fatter are the tails. This has detrimental effect on the performance of all 9 estimation methods, as extreme random values are incorrectly classified as significant effects. Forward / Backward Cross-Validation and Lasso are especially badly hit when the degrees of freedom go down from 3 to 1.

- In most cases, there is no statistically significant difference in the performance of the forward and backward versions of the same method.

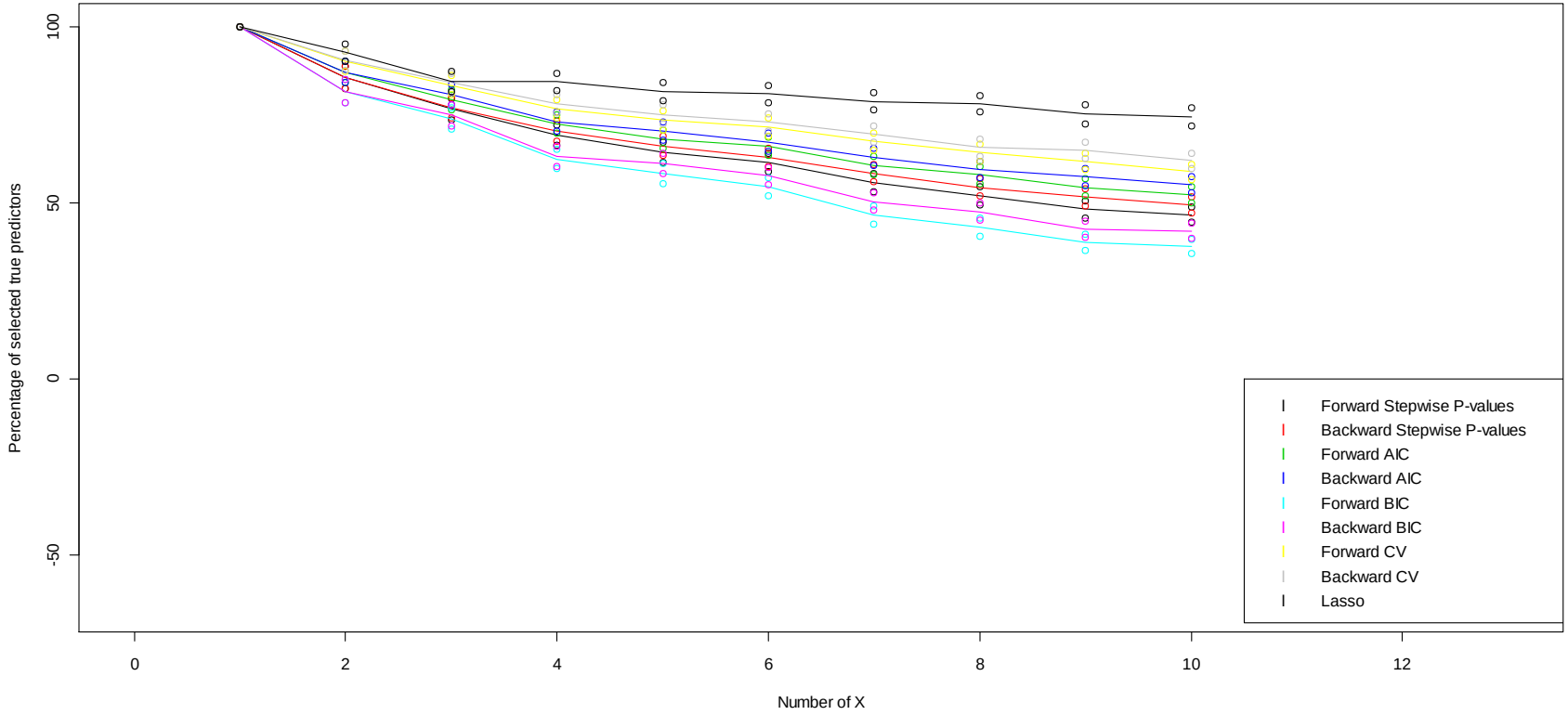
Overall, Lasso is the top performer. The key rationale here is "soft shrinkage" that Lasso implements. The remaining methods implement "hard elimination". When there are many useful predictors they may come out as only marginally significant on a relatively small sample. 8 methods out 9 would try to combine the information from these marginal predictors in a sub-optimal fashion, allowing only one path through them and discarding some useful information on the way. Lasso combines the marginal predictors in an optimal way according to a specific metric. The metric, perfect or not, ensures that the information from no single predictor is ignored completely.

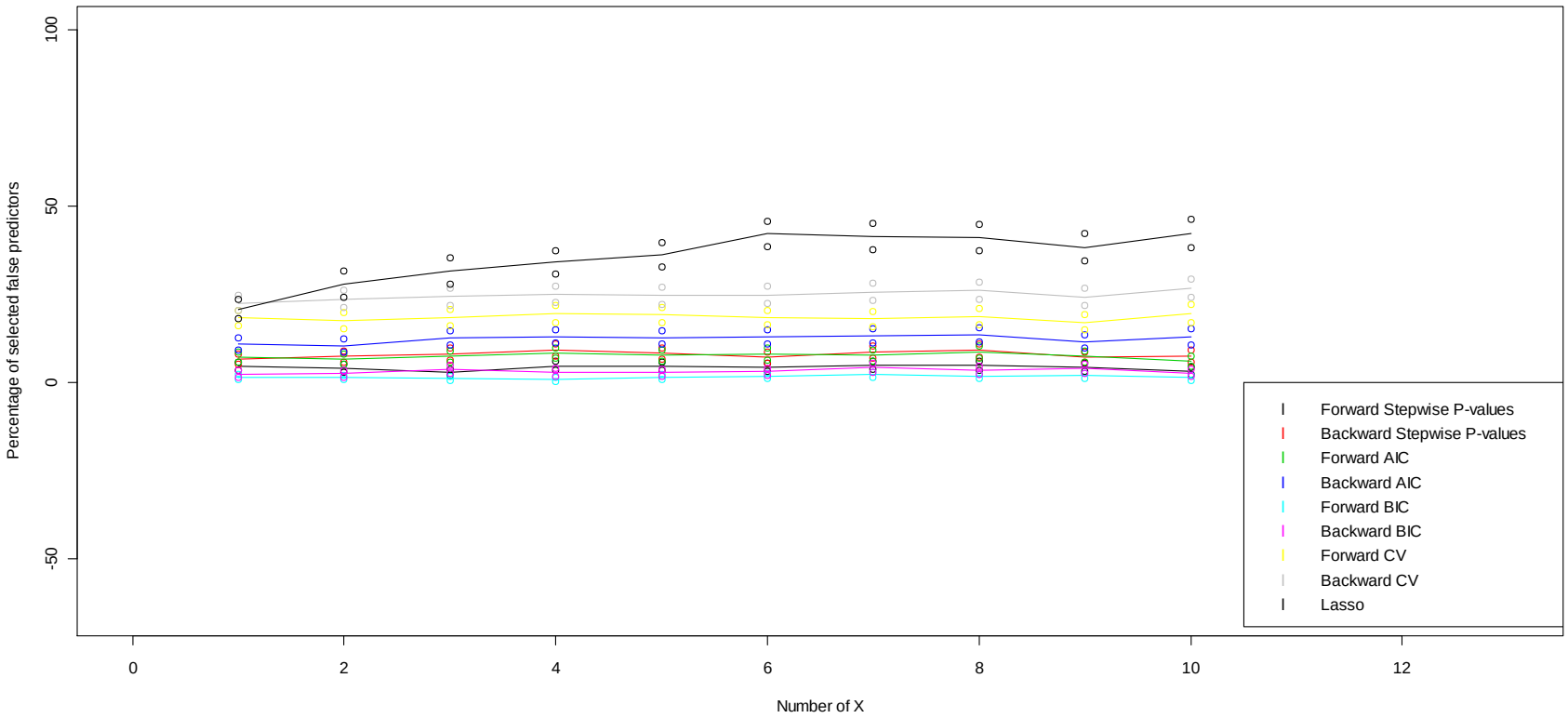
The Forward and Backward Cross-Validation share the 2nd and 3rd spots and are substantially better than the remaining methods. One possible explanation is that the significance of a fit improvement is tested out-of-sample in cross-validation. It is never formally tested when AIC or BIC is applied. The reasons for superiority of cross-validation over Forward / Backward P-values are still to be studied. One possible lead is that, at the testing stage, cross-validation is capable of detecting and downplaying extreme values that showed up in the training sample out of pure randomness. Forward / Backward P-values has no way of downplaying such values and takes every observation equally seriously.

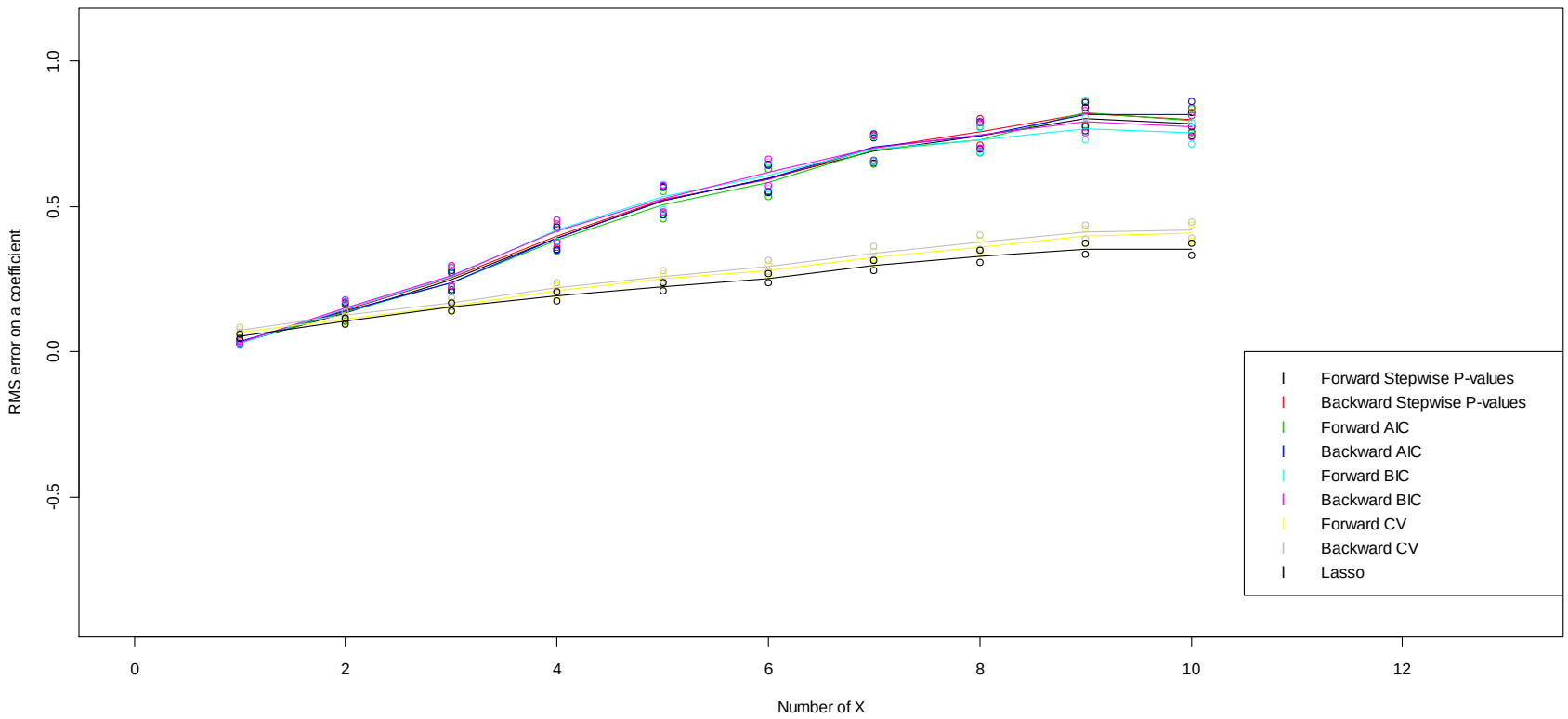
It has been proven that, on large samples, BIC tends to choose models simpler than the truth and, on small samples, AIC may choose models more complicated than the truth. Hence BIC underperforming AIC in TP and AIC underperforming BIC in FP has been expected.

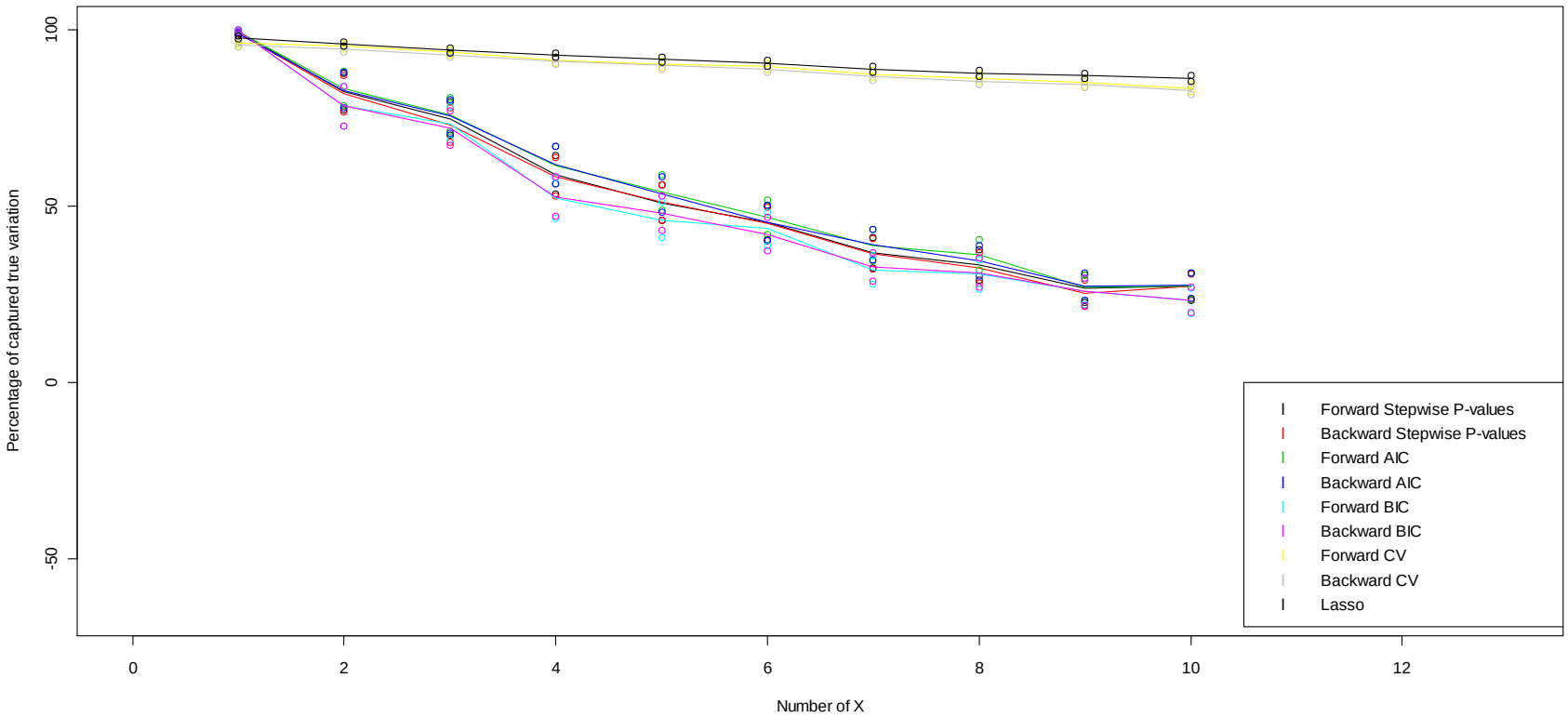
APPENDIX: 2D PLOTS

Variation with respect to p:

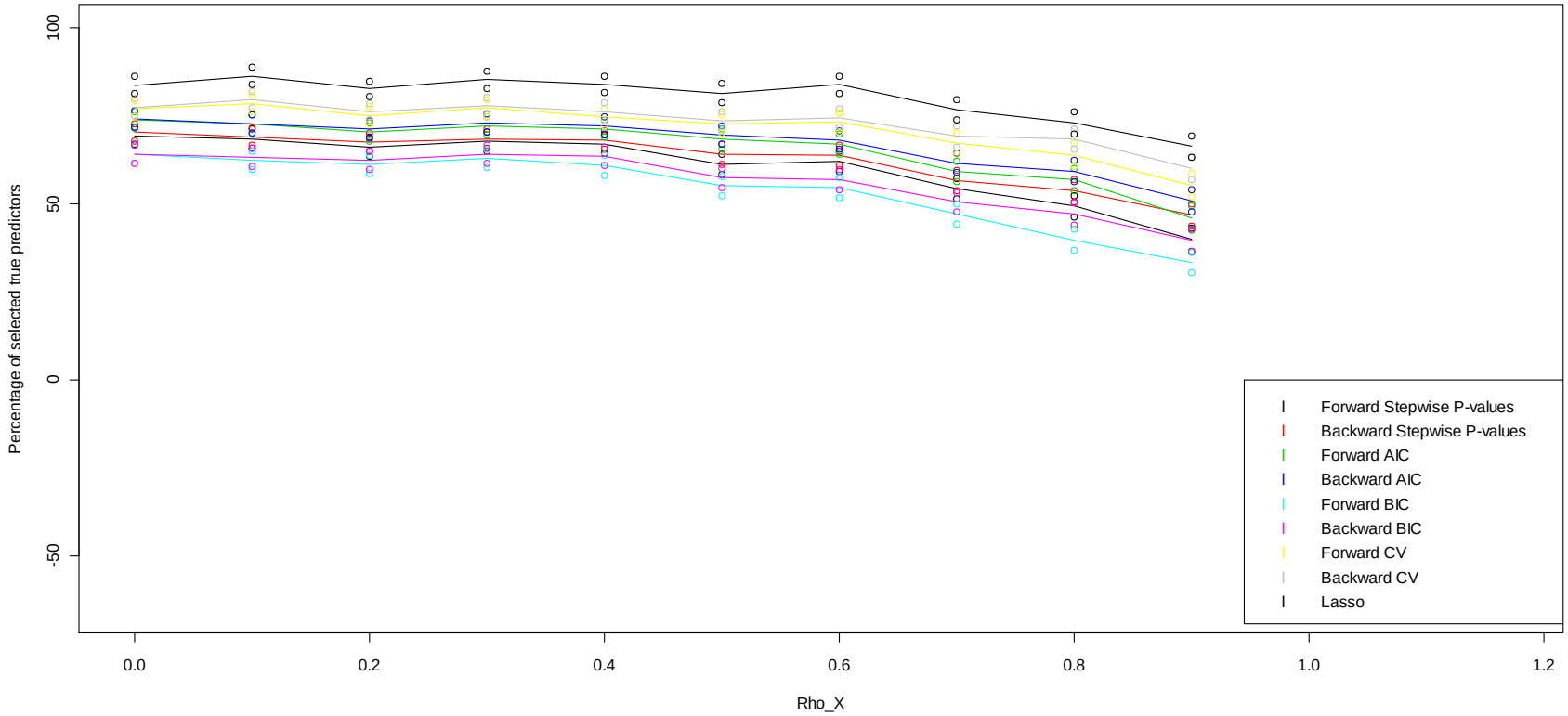


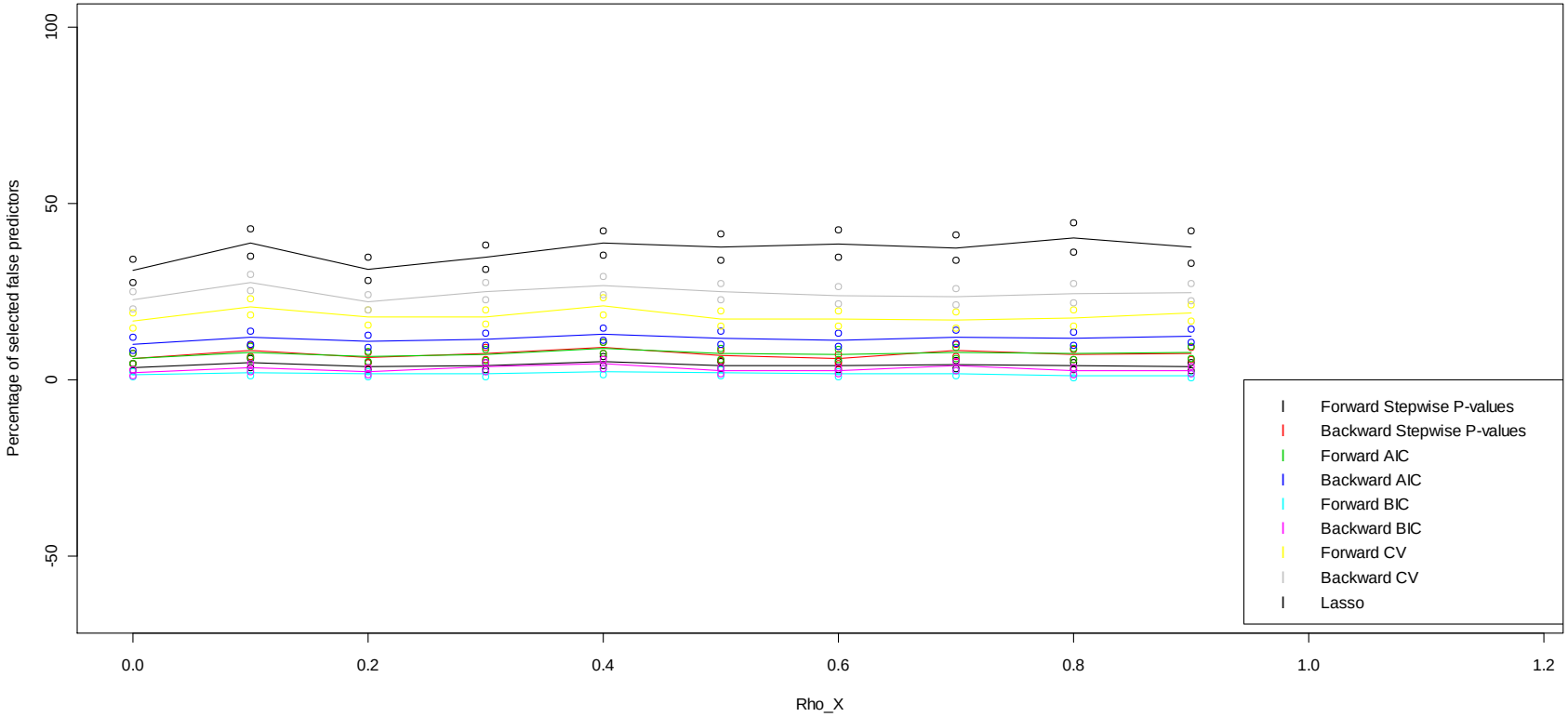

Percentage of selected false predictors
100
50
0
-50
0
2
4
6
8
10
12
Number of X
Forward Stepwise P-values
Backward Stepwise P-values
Forward AIC
Backward AIC
Forward BIC
Backward BIC
Forward CV
Backward CV
Lasso

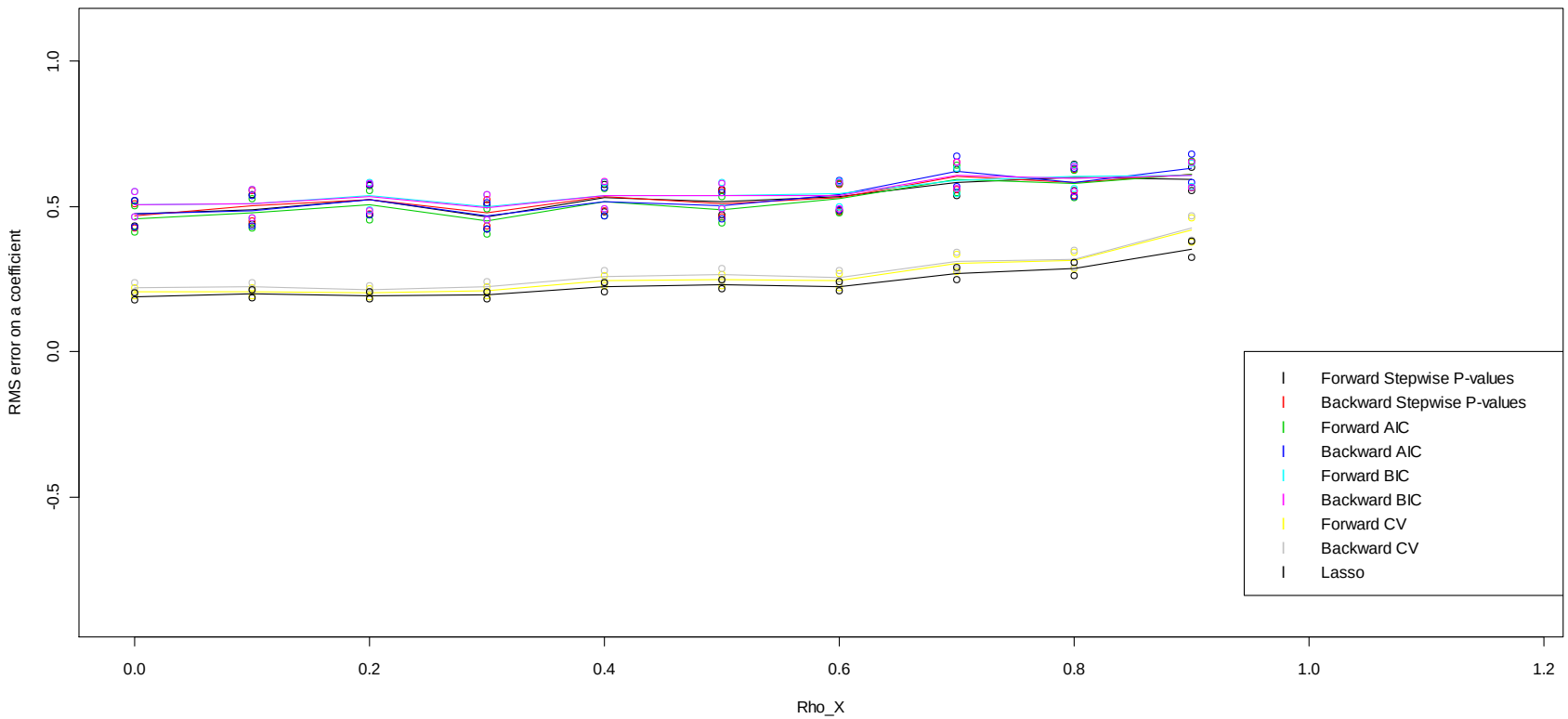

RMS error on a coefficient
1.0
0.5
0.0
-0.5
0
2
4
6
8
10
12
Number of X
Forward Stepwise P-values
Backward Stepwise P-values
Forward AIC
Backward AIC
Forward BIC
Backward BIC
Forward CV
Backward CV
Lasso

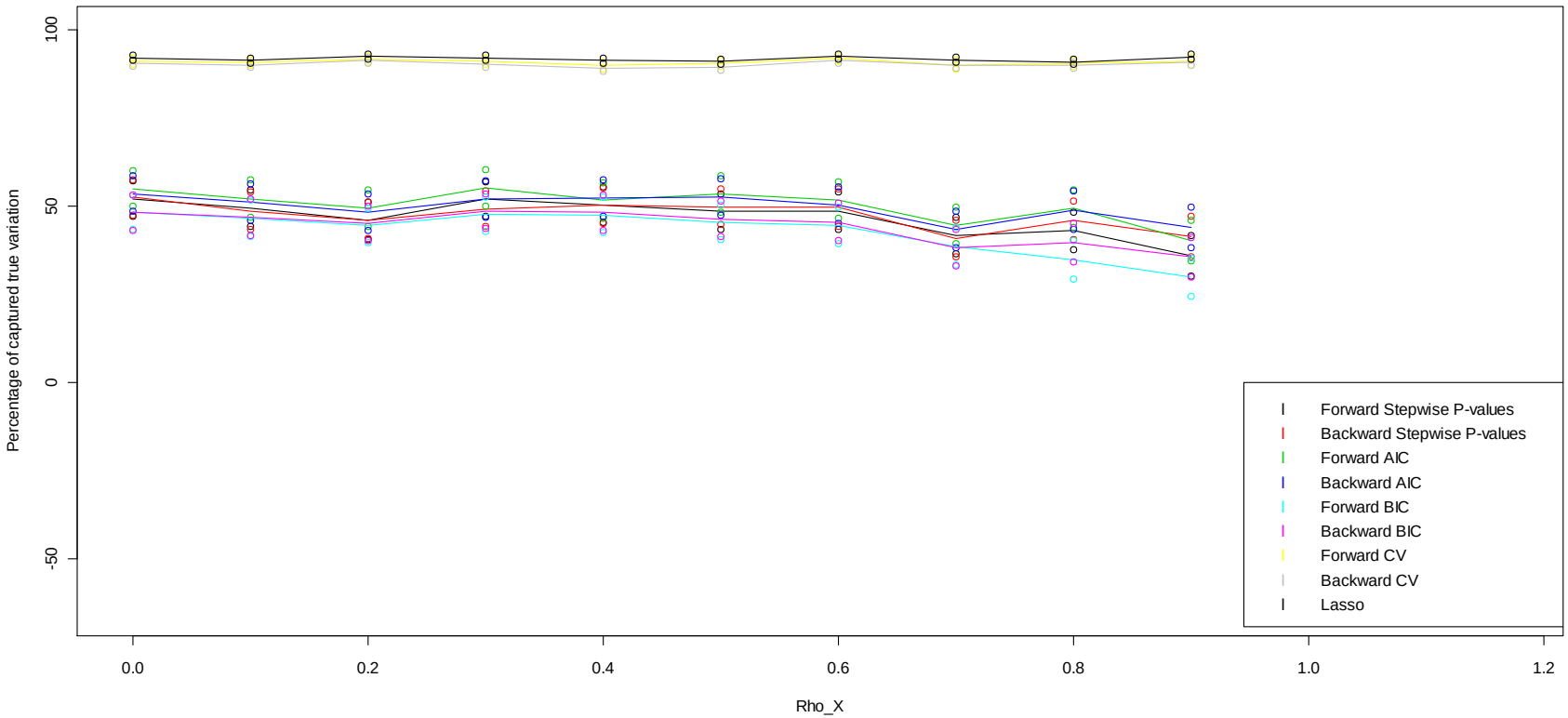

Percentage of captured true variation
100
50
0
-50
0
2
4
6
8
10
12
Number of X
Forward Stepwise P-values
Backward Stepwise P-values
Forward AIC
Backward AIC
Forward BIC
Backward BIC
Forward CV
Backward CV
Lasso

Variation with respect to Rho_X:

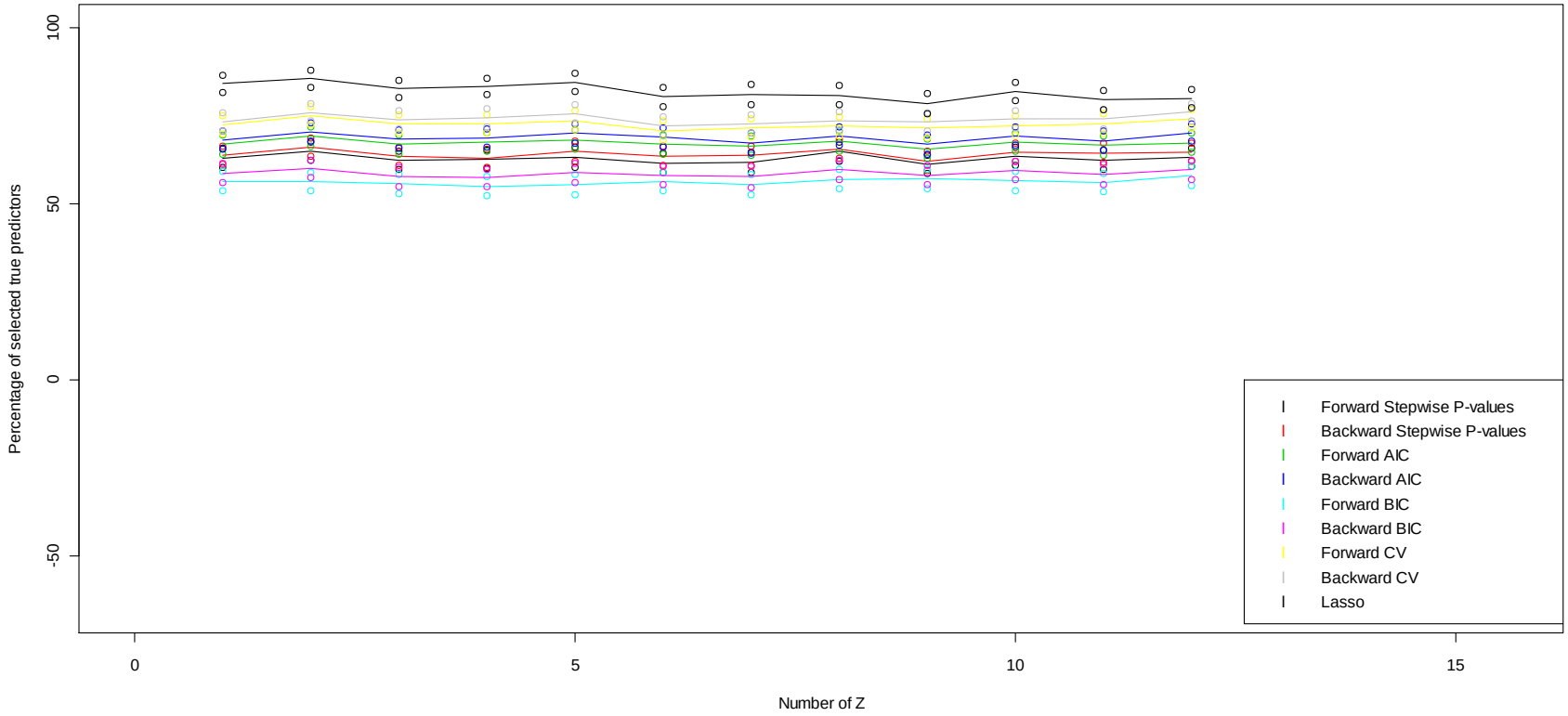


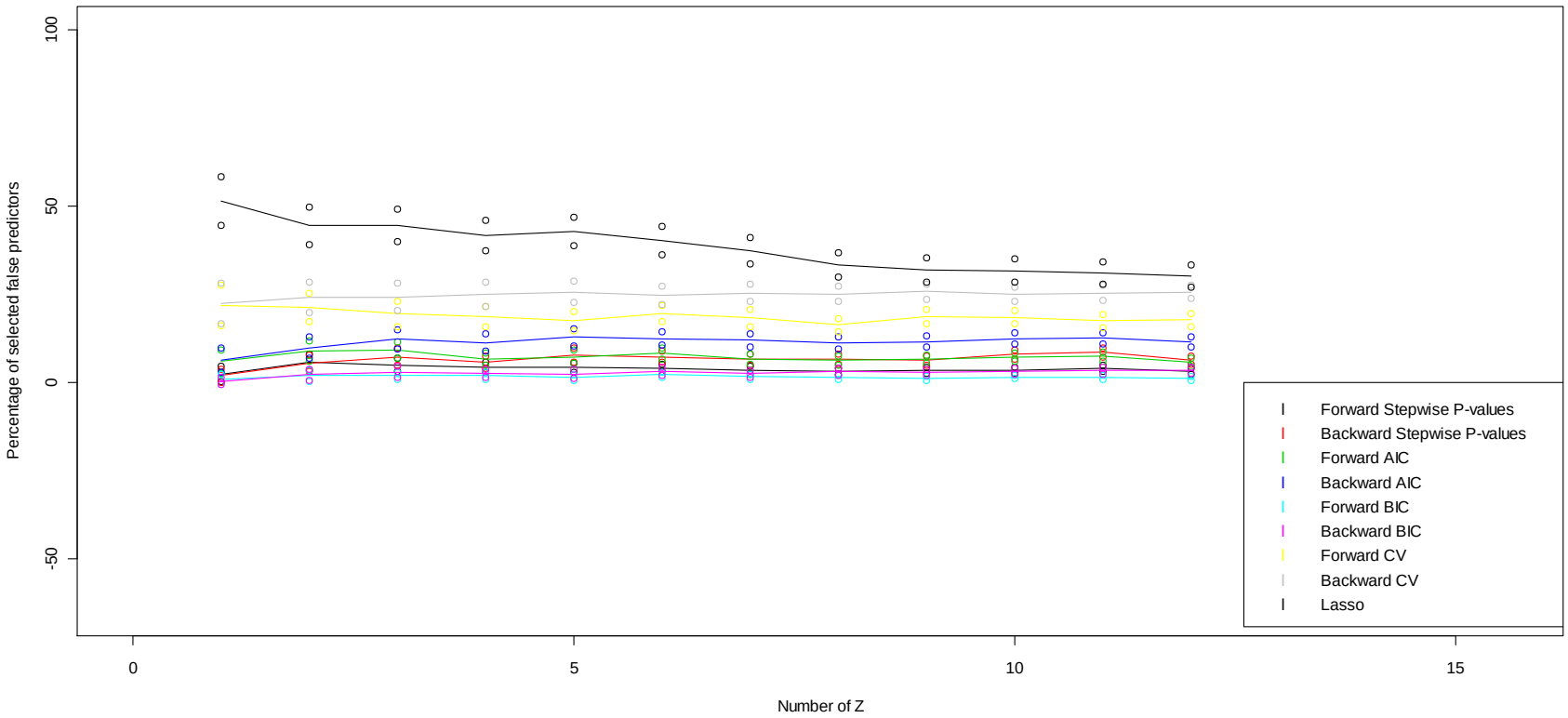

Percentage of selected false predictors
100
50
0
-50
0.0
0.2
0.4
0.6
0.8
1.0
1.2
Rho_X
Forward Stepwise P-values
Backward Stepwise P-values
Forward AIC
Backward AIC
Forward BIC
Backward BIC
Forward CV
Backward CV
Lasso

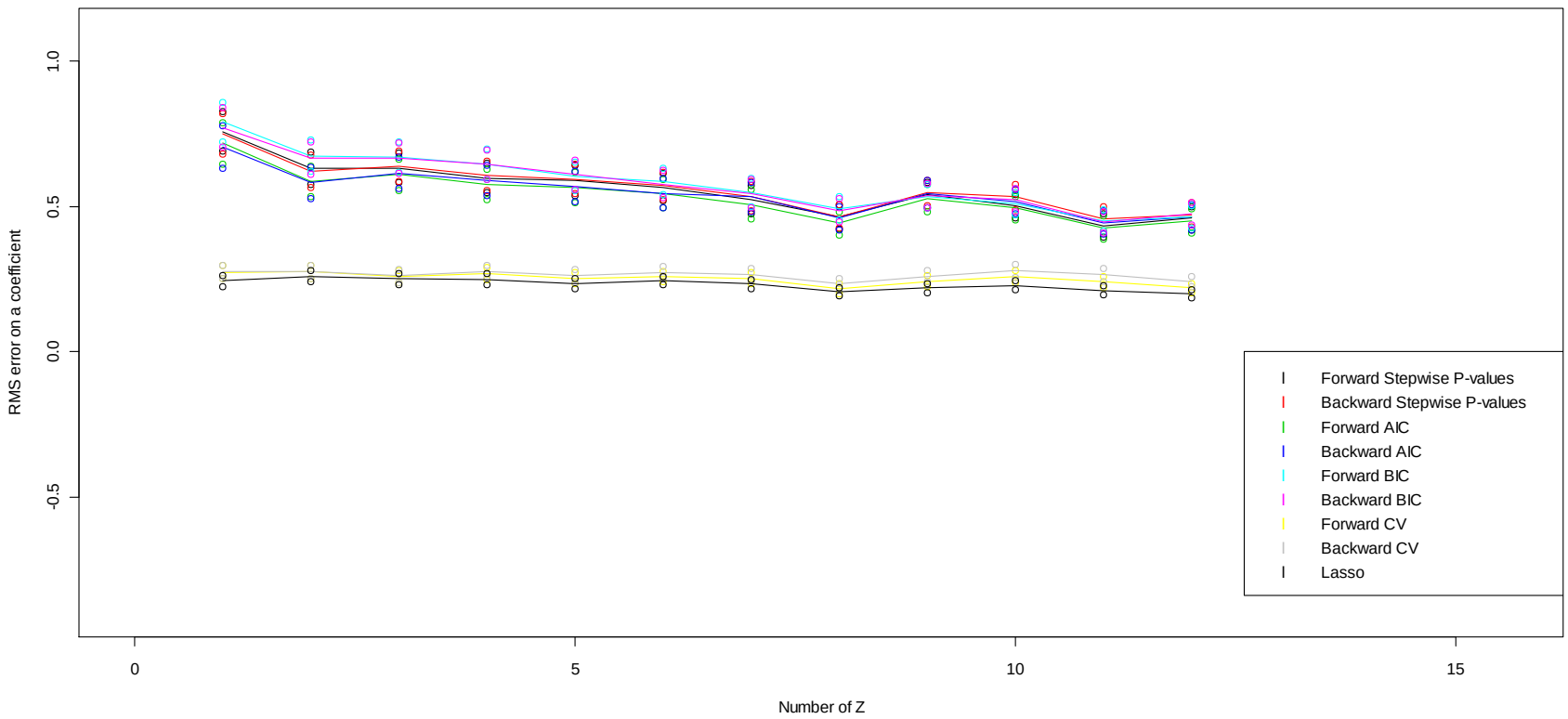

RMS error on a coefficient
1.0
0.5
0.0
-0.5
0.0
0.2
0.4
0.6
0.8
1.0
1.2
Rho_X
Forward Stepwise P-values
Backward Stepwise P-values
Forward AIC
Backward AIC
Forward BIC
Backward BIC
Forward CV
Backward CV
Lasso

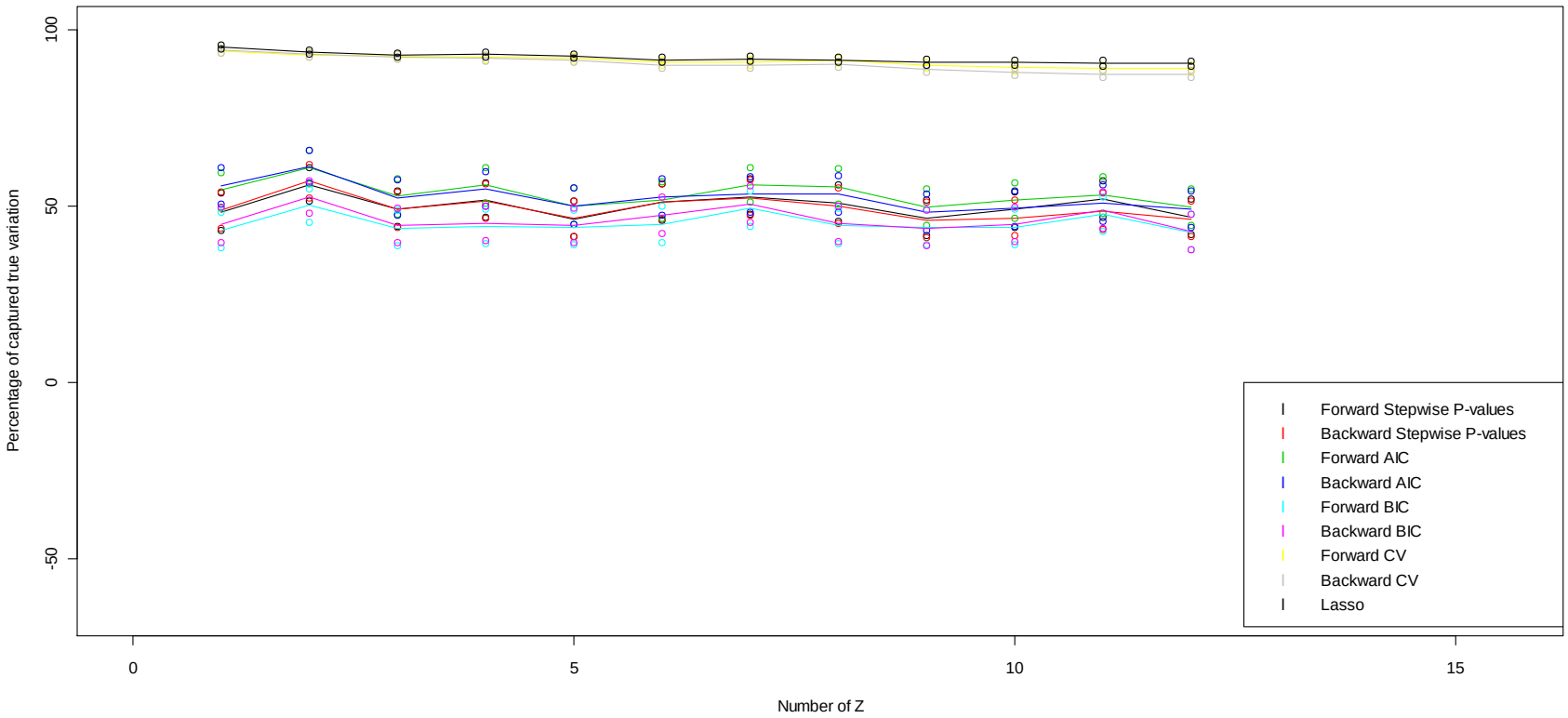

Percentage of captured true variation
100
50
0
-50
0.0
0.2
0.4
0.6
0.8
1.0
1.2
Rho_X
Forward Stepwise P-values
Backward Stepwise P-values
Forward AIC
Backward AIC
Forward BIC
Backward BIC
Forward CV
Backward CV
Lasso

Variation with respect to q:

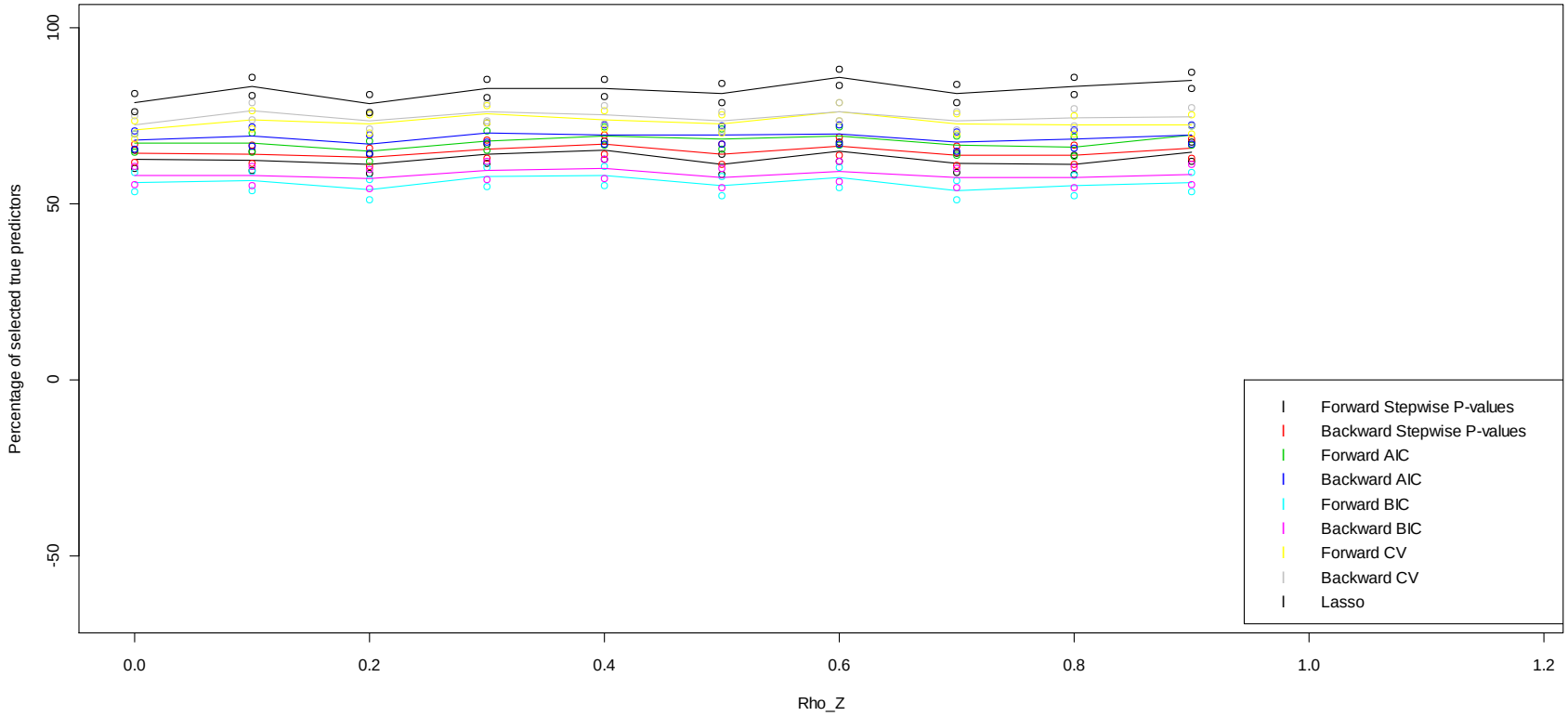


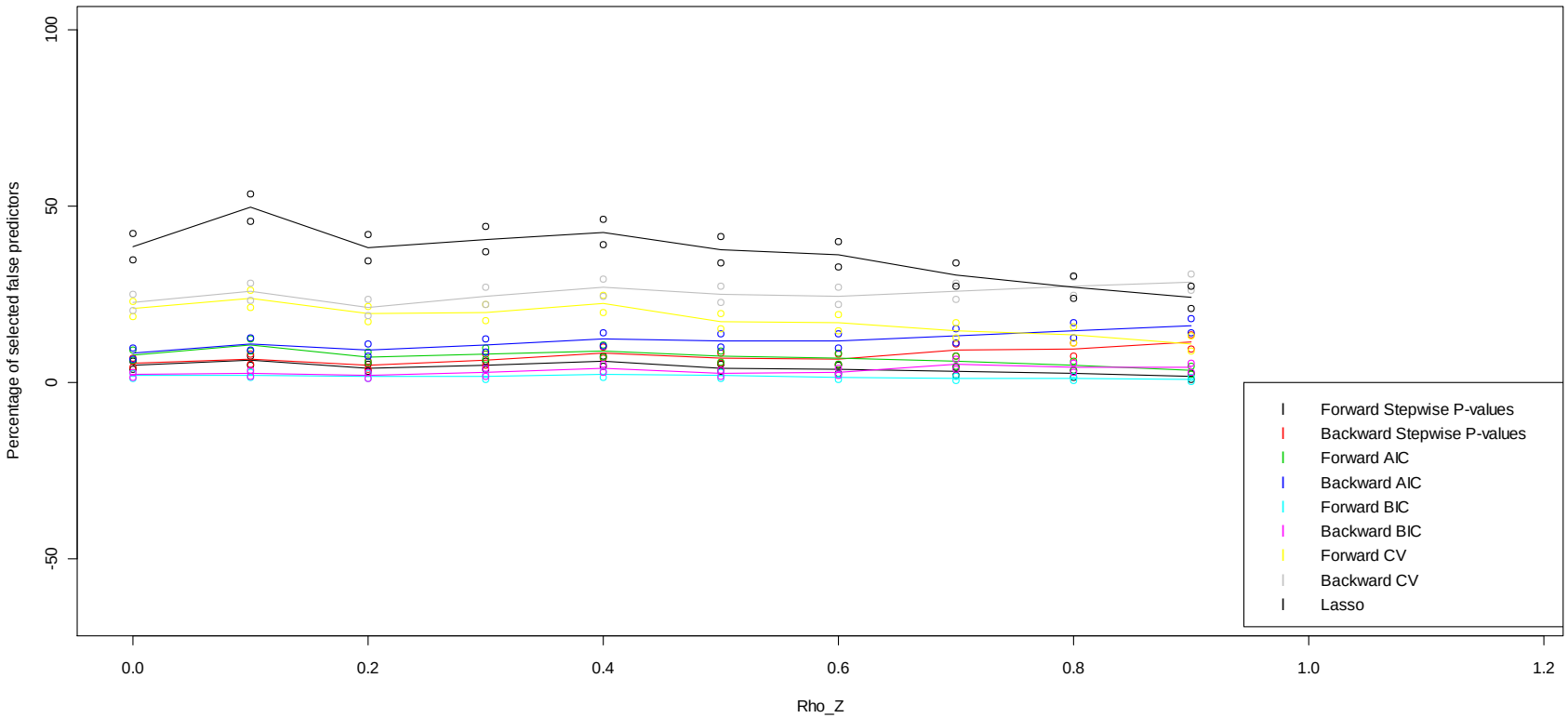

Percentage of selected false predictors
100
50
0
-50
0
5
10
15
Number of Z
Forward Stepwise P-values
Backward Stepwise P-values
Forward AIC
Backward AIC
Forward BIC
Backward BIC
Forward CV
Backward CV
Lasso

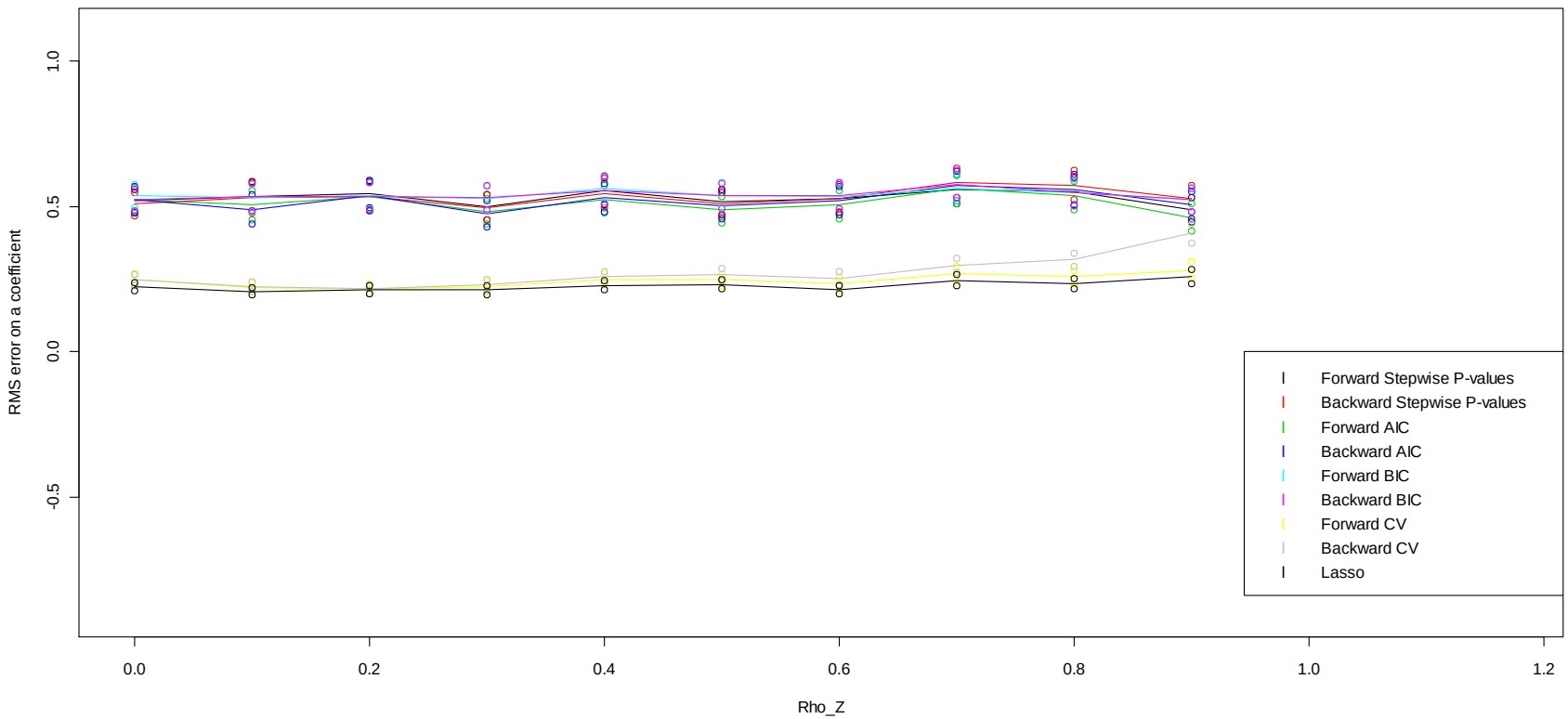

RMS error on a coefficient
1.0
0.5
0.0
-0.5
0
5
10
15
Number of Z
Forward Stepwise P-values
Backward Stepwise P-values
Forward AIC
Backward AIC
Forward BIC
Backward BIC
Forward CV
Backward CV
Lasso

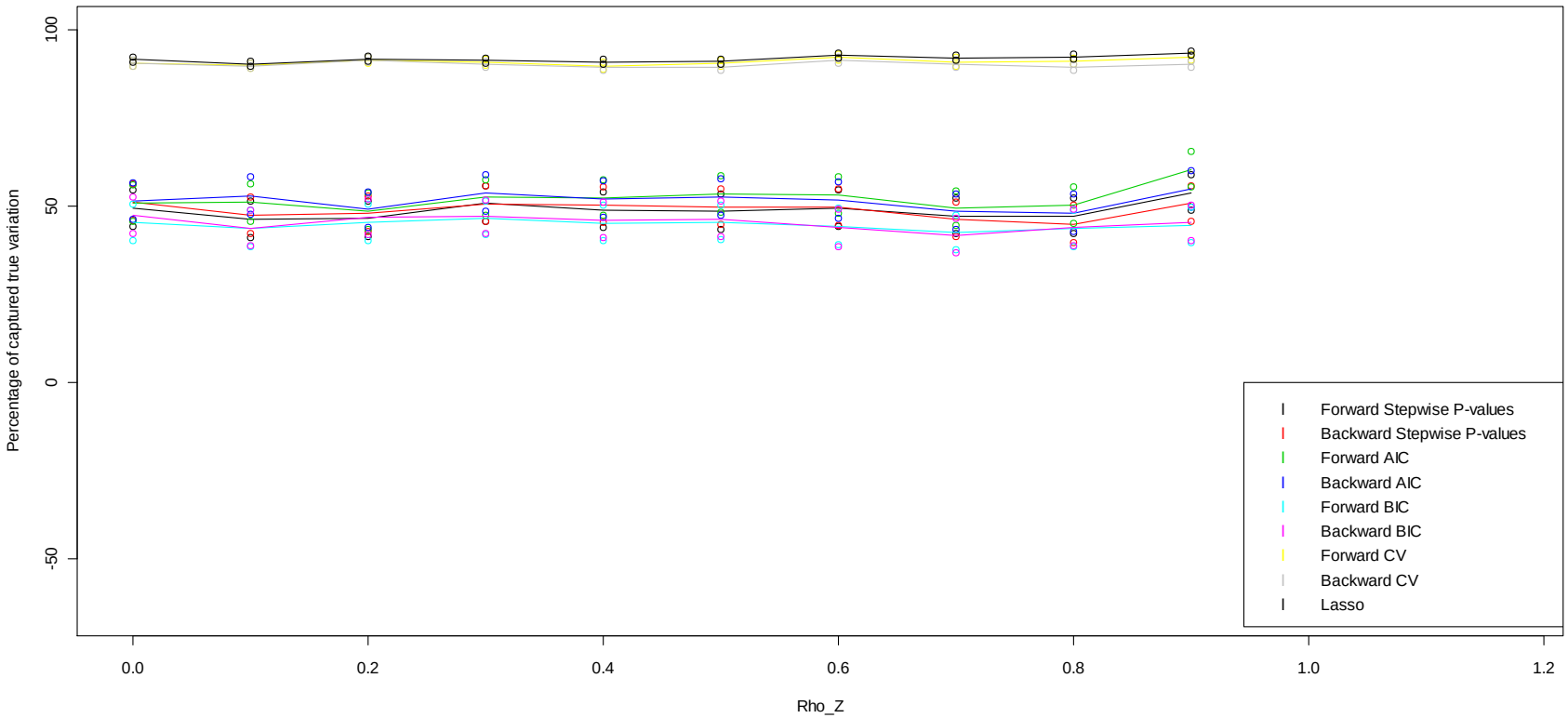

Percentage of captured true variation
100
50
0
-50
0
5
10
15
Number of Z
Forward Stepwise P-values
Backward Stepwise P-values
Forward AIC
Backward AIC
Forward BIC
Backward BIC
Forward CV
Backward CV
Lasso

Variation with respect to Rho_Z:

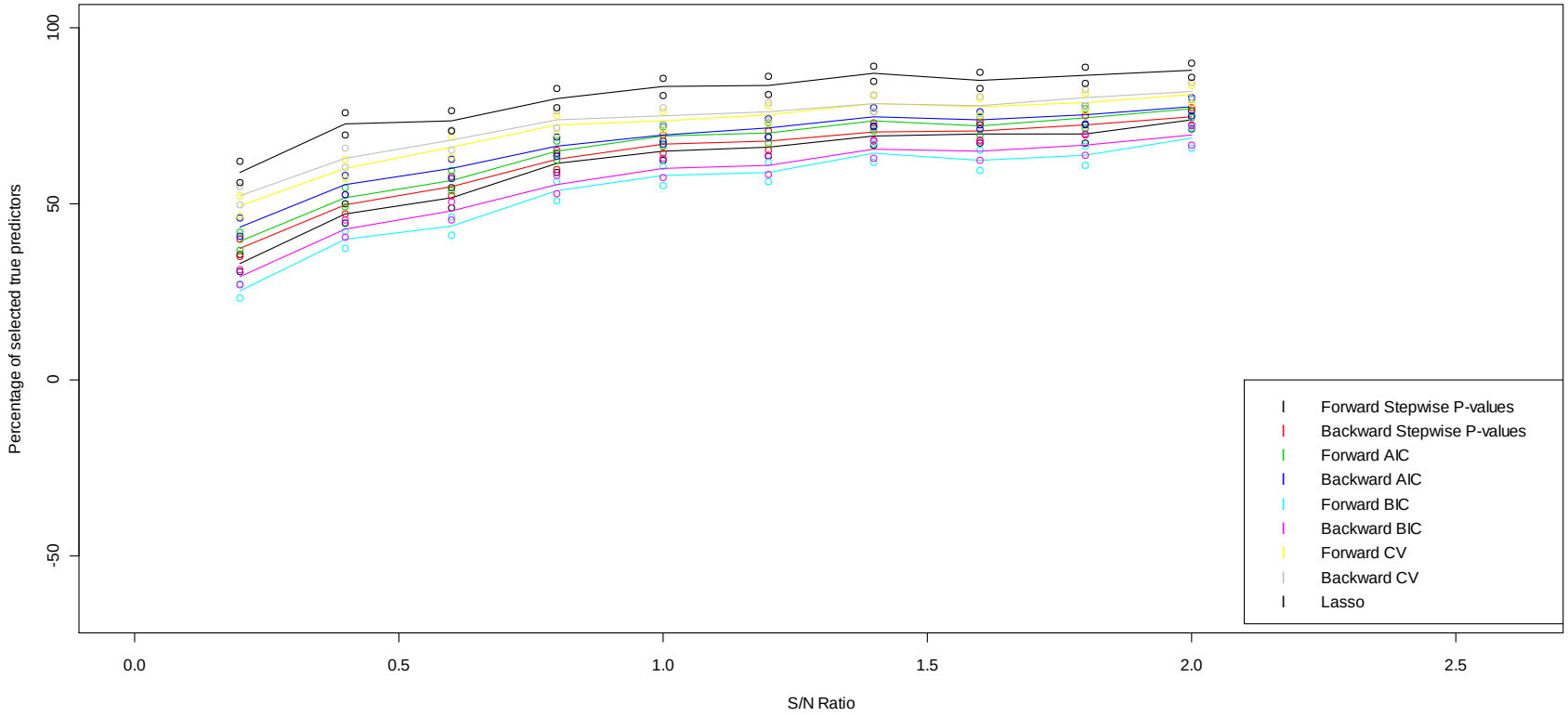


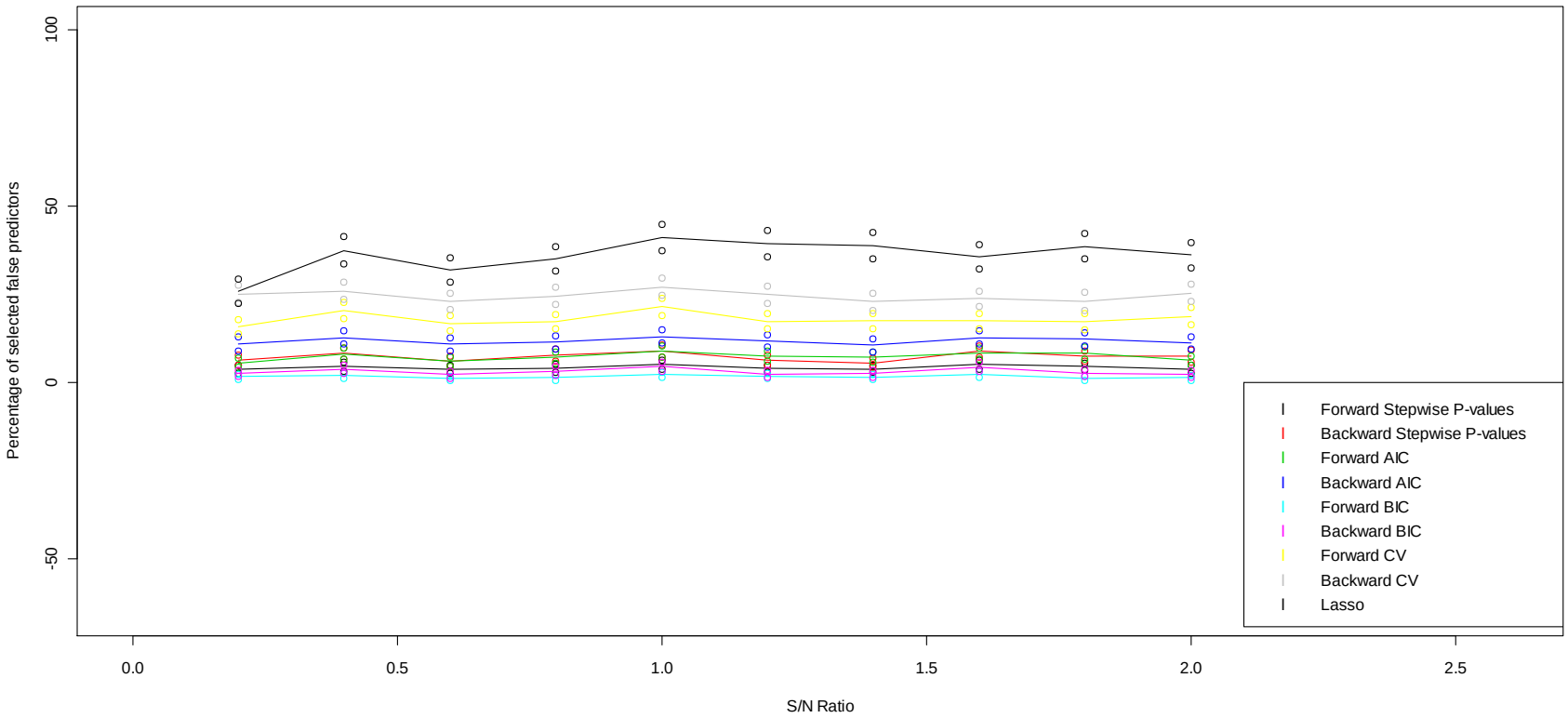

Percentage of selected false predictors
100
50
0
-50
0.0
0.2
0.4
0.6
0.8
1.0
1.2
Rho_Z
Forward Stepwise P-values
Backward Stepwise P-values
Forward AIC
Backward AIC
Forward BIC
Backward BIC
Forward CV
Backward CV
Lasso

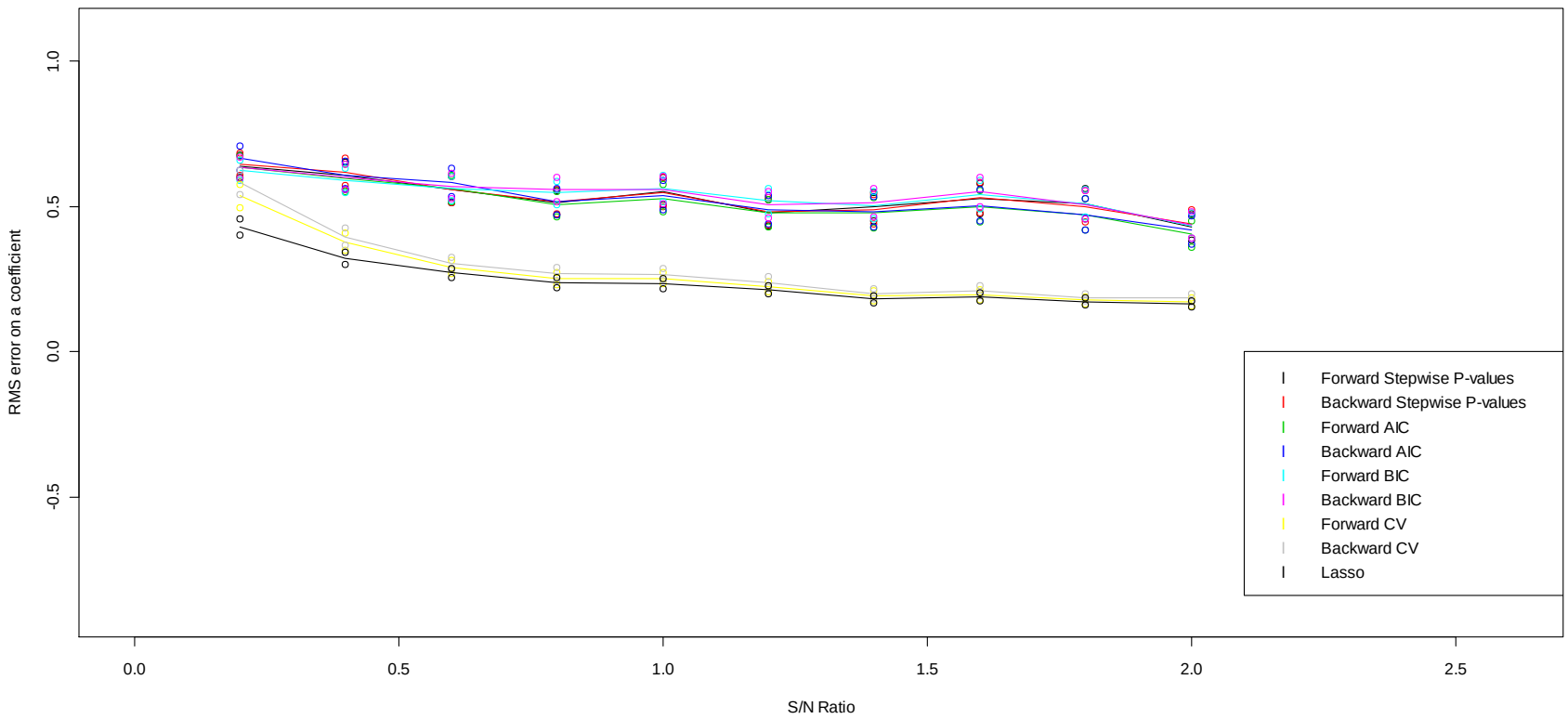

RMS error on a coefficient
1.0
0.5
0.0
-0.5
0.0
0.2
0.4
0.6
0.8
1.0
1.2
Rho_Z
Forward Stepwise P-values
Backward Stepwise P-values
Forward AIC
Backward AIC
Forward BIC
Backward BIC
Forward CV
Backward CV
Lasso

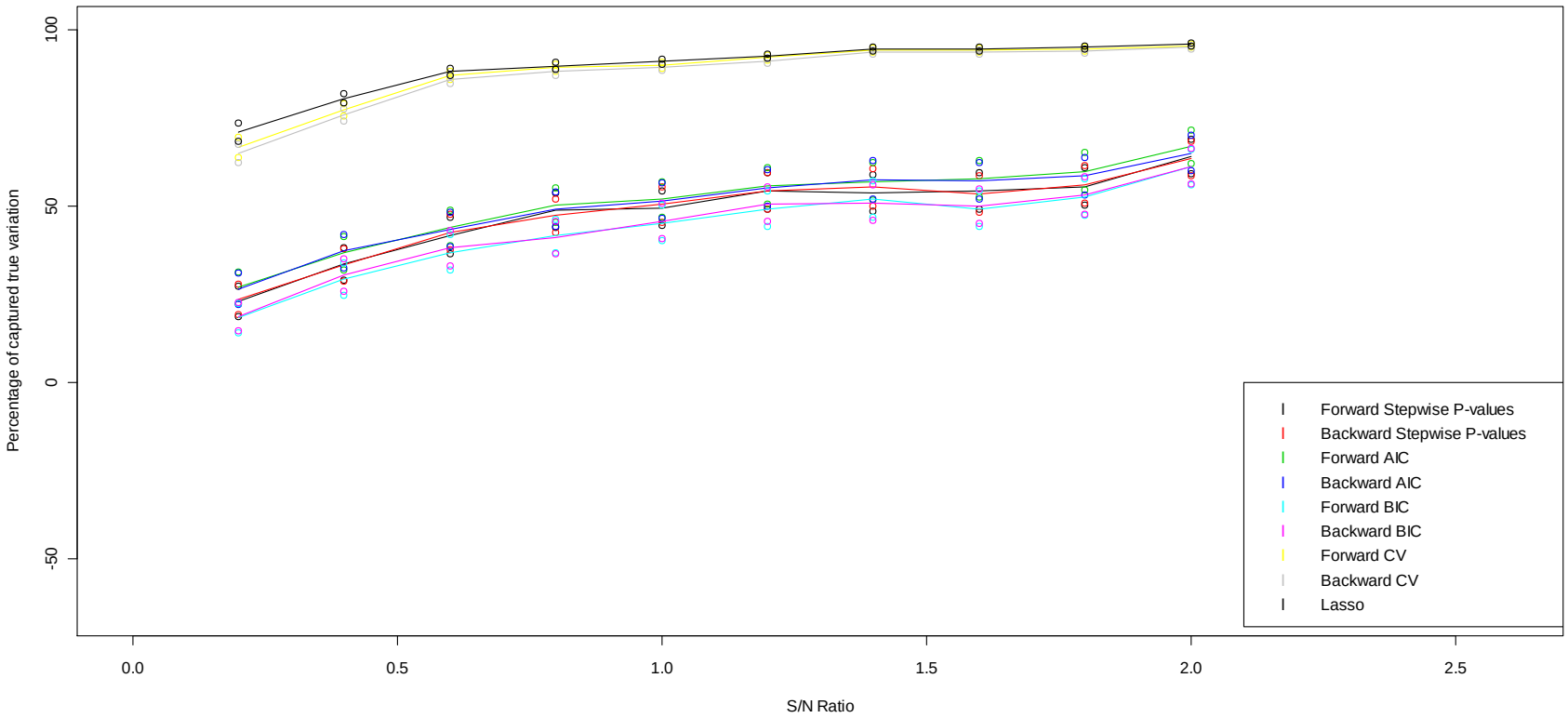

Percentage of captured true variation
100
50
0
-50
0.0
0.2
0.4
0.6
0.8
1.0
1.2
Rho_Z
Forward Stepwise P-values
Backward Stepwise P-values
Forward AIC
Backward AIC
Forward BIC
Backward BIC
Forward CV
Backward CV
Lasso

Variation with respect to S/N:

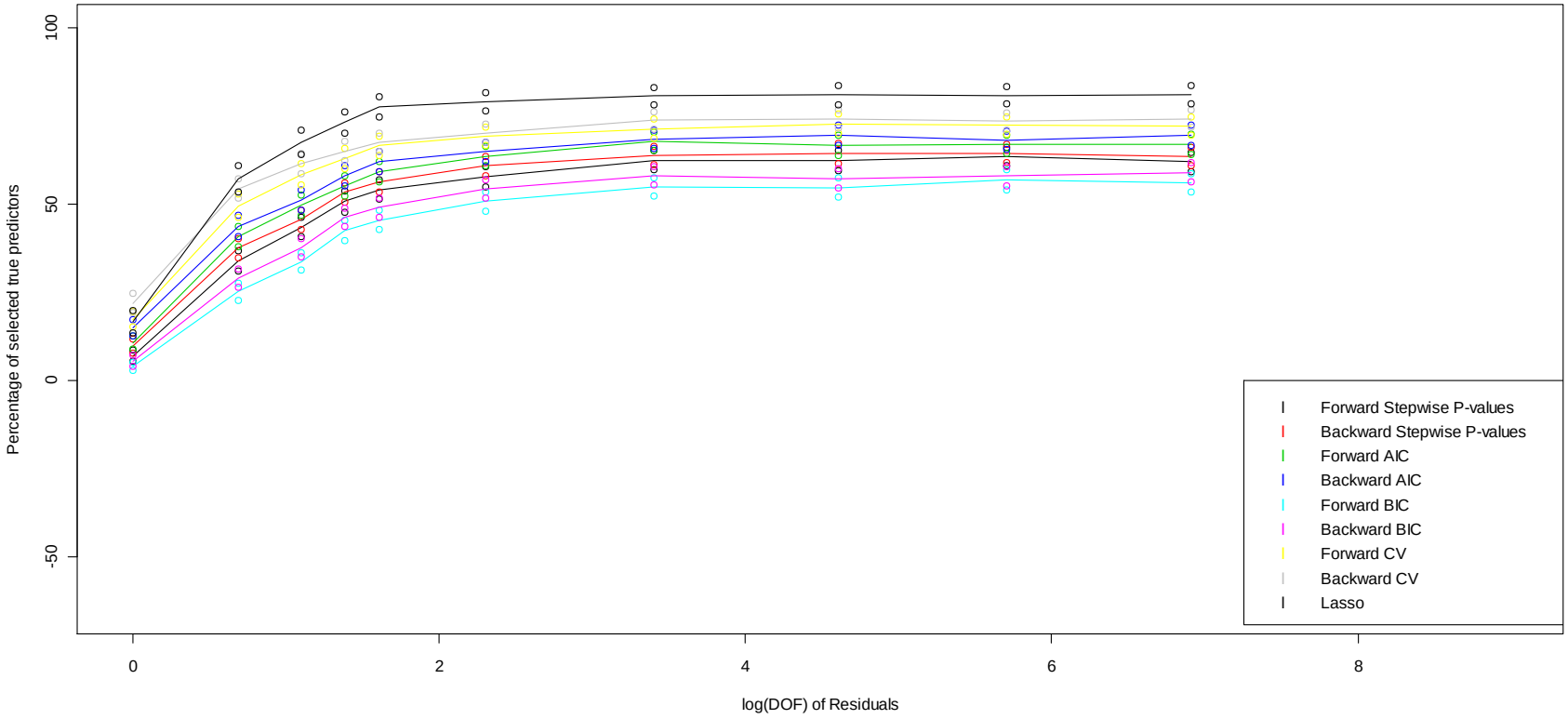


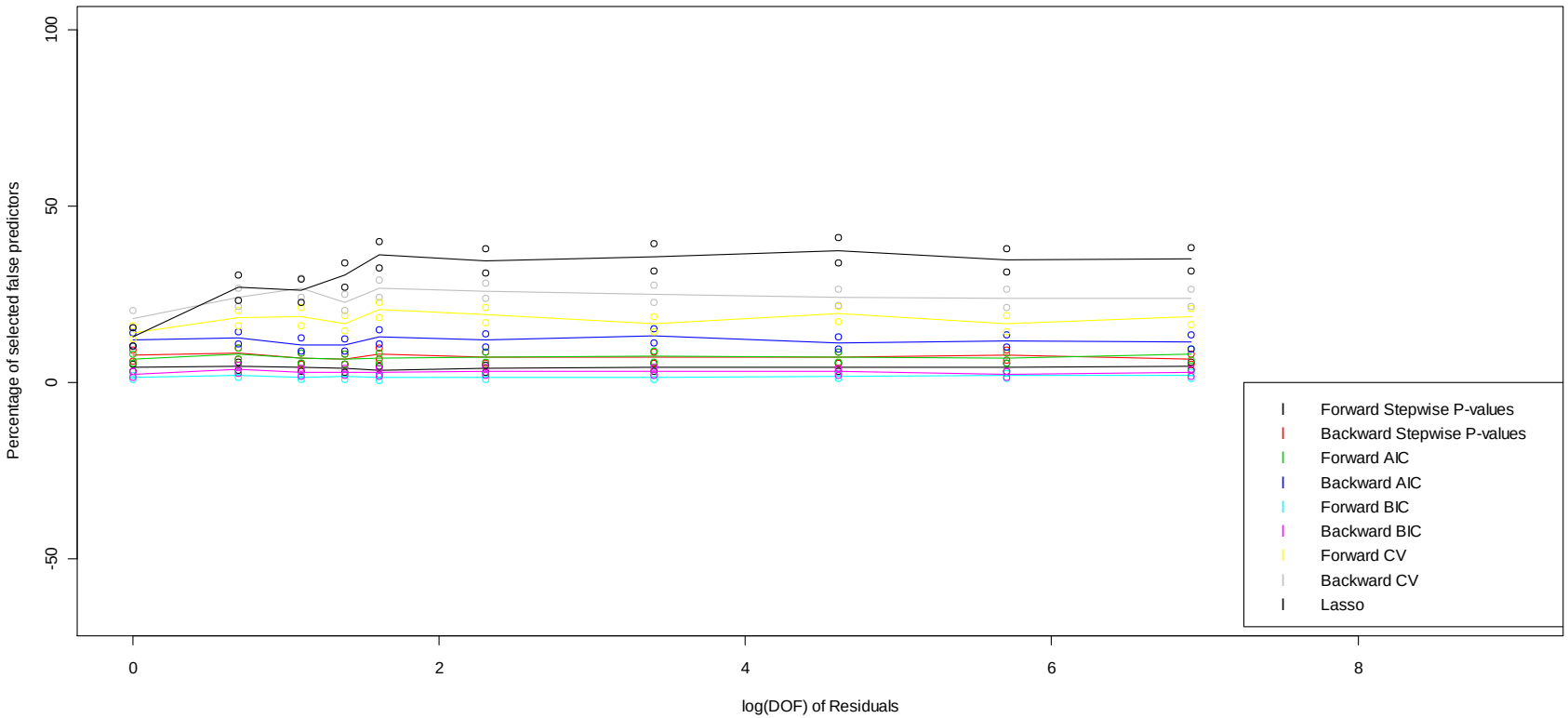

Percentage of selected false predictors
100
50
0
-50
0.0
0.5
1.0
1.5
2.0
2.5
S/N Ratio
Forward Stepwise P-values
Backward Stepwise P-values
Forward AIC
Backward AIC
Forward BIC
Backward BIC
Forward CV
Backward CV
Lasso

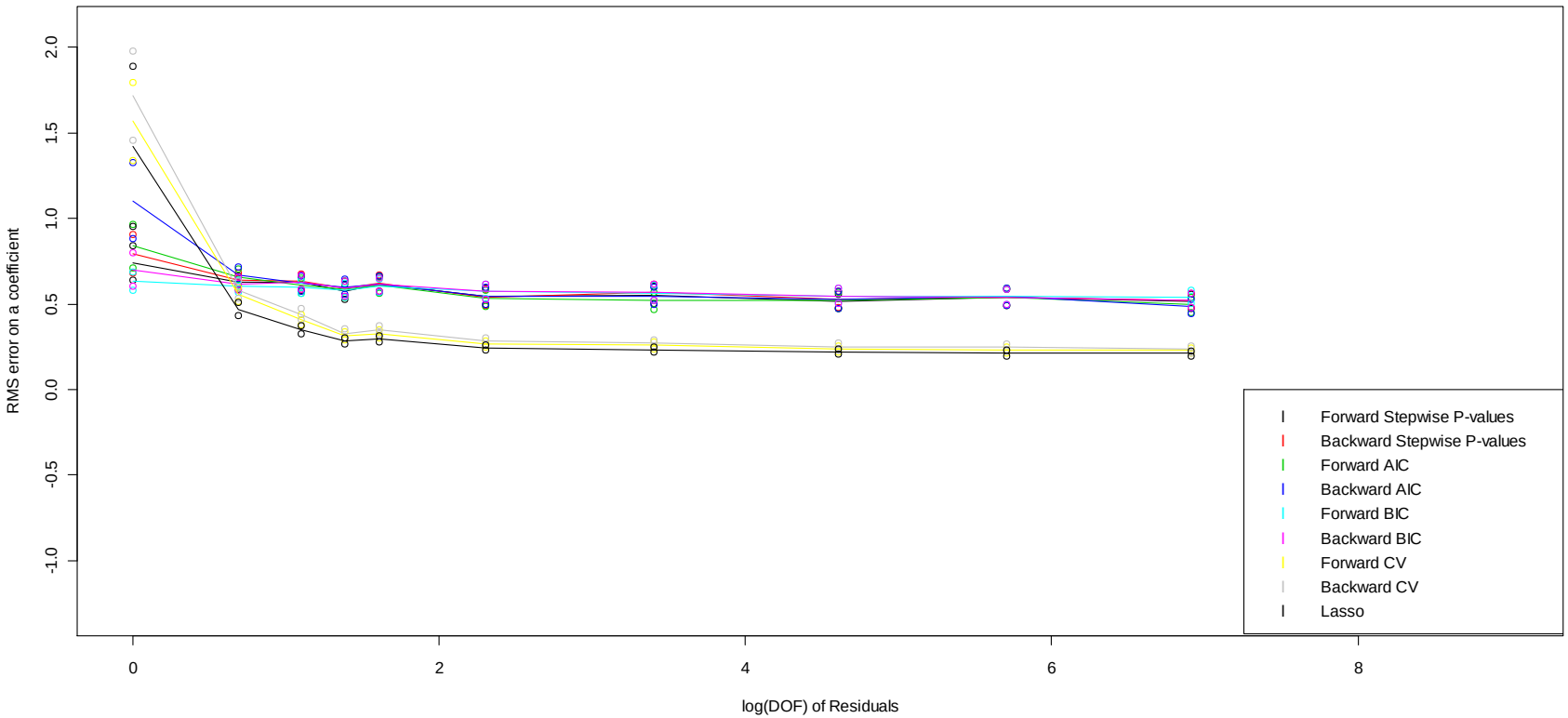

RMS error on a coefficient
1.0
0.5
0.0
-0.5
0.0
0.5
1.0
1.5
2.0
2.5
S/N Ratio
Forward Stepwise P-values
Backward Stepwise P-values
Forward AIC
Backward AIC
Forward BIC
Backward BIC
Forward CV
Backward CV
Lasso


Percentage of captured true variation
100
50
0
-50
0.0
0.5
1.0
1.5
2.0
2.5
S/N Ratio
Forward Stepwise P-values
Backward Stepwise P-values
Forward AIC
Backward AIC
Forward BIC
Backward BIC
Forward CV
Backward CV
Lasso

Variation with respect to V:




Percentage of selected false predictors
100
50
0
-50
0
2
4
6
8
log(DOF) of Residuals
Forward Stepwise P-values
Backward Stepwise P-values
Forward AIC
Backward AIC
Forward BIC
Backward BIC
Forward CV
Backward CV
Lasso


RMS error on a coefficient
2.0
1.5
1.0
0.5
0.0
-0.5
-1.0
0
2
4
6
8
log(DOF) of Residuals
Forward Stepwise P-values
Backward Stepwise P-values
Forward AIC
Backward AIC
Forward BIC
Backward BIC
Forward CV
Backward CV
Lasso