Memorial Sloan-Kettering Cancer Center

Memorial Sloan-Kettering Cancer Center, Dept. of Epidemiology & Biostatistics Working Paper Series

Year 2017 *Paper* 34

Optimized Variable Selection Via Repeated Data Splitting

Marinela Capanu[*] Colin B. Begg[†]

Mithat Gonen[‡]

[*]Memorial Sloan-Kettering Cancer Center, capanum@mskcc.org

[†]Memorial Sloan-Kettering Cancer Center, beggc@mskcc.org

[‡]Memorial Sloan-Kettering Cancer Center, gonenm@mskcc.org

Optimized Variable Selection Via Repeated Data Splitting

Marinela Capanu, Colin B. Begg, and Mithat Gonen

Abstract

We introduce a new variable selection procedure that repeatedly splits the data into two sets, one for estimation and one for validation, to obtain an empirically optimized threshold which is then used to screen for variables to include in the final model. Simulation results show that the proposed variable selection technique enjoys superior performance compared to candidate methods, being amongst those with the lowest inclusion of noisy predictors while having the highest power to detect the correct model and being unaffected by correlations among the predictors. We illustrate the methods by applying them to a cohort of patients undergoing hepatectomy at our institution.

Optimized variable selection
via repeated data splitting

Marinela Capanu Colin B. Begg Mithat Gönen
Memorial Sloan Kettering Cancer Center,
New York, USA
email: capanum@mskcc.org

Abstract

We introduce a new variable selection procedure that repeatedly splits the data into two sets, one for estimation and one for validation, to obtain an empirically optimized threshold which is then used to screen for variables to include in the final model. Simulation results show that the proposed variable selection technique enjoys superior performance compared to candidate methods, being amongst those with the lowest inclusion of noisy predictors while having the highest power to detect the correct model and being unaffected by correlations among the predictors. We illustrate the methods by applying them to a cohort of patients undergoing hepatectomy at our institution.

KEY WORDS: variable selection; regression; data splitting; empirical threshold; screening

1 Introduction

Regression models have become the primary engine for many data analyses, attempting to explain the variability in the dependent variable by identifying independent predictors. Due to their popularity, simultaneous model selection and estimation continues to be one of the most important problems in applied statistics. It is frequently the case that there are many predictors to choose from and the literature does not provide a uniform message about which technique to use to select the variables to include in a model. Sample sizes are also limiting, often preventing the inclusion of all the candidate predictors in the final model. Therefore it is not surprising that many variable selection approaches have been proposed over the years including significance filtering, stepwise methods and penalized regression (see Harrell 2001; Miller 2002; Hastie et al. 2009, among others). None of these methods have become the standard, go-to method, suggesting that there is still substantial room for improvement.

An ideal variable selection method would include all “true” predictors in the model while excluding the “noisy” predictors, i.e. those with little or no independent association with the dependent variable. However, these two objectives clash with each other as they correspond to the type II and type I errors in hypothesis testing. As such, there is a tradeoff between maximizing selection of true predictors (full model) and minimizing selection of noisy predictors (intercept only model) and a good variable selection procedure has to strike a balance between these two extremes with the goal of obtaining an accurate yet parsimonious model.

Another fundamental problem encountered in variable selection is the tension between variable selection and estimation due to the bias-variance tradeoff that comes with model complexity: more complex models will have higher variance and lower bias, while the opposite happens as the model complexity is decreased: simpler models will have higher bias and lower variance (Hastie et al. 2009). An optimal model will have both low variance and low bias. Since the expected prediction error on new data can be decomposed into a bias component, a variance component, and noise (which is beyond our control), by choosing the model complexity to trade bias off with variance we, in effect, minimize the test error. This is a desirable goal for a model selection procedure since we want the model to predict future observations as well as it predicts the observed data. A mistake to be avoided when estimating the prediction error is to calculate it on the same data used to estimate the model parameters as this can result in overfitting, making the predictions look unrealistically good and degrading the ability of the model to generalize to a new dataset.

In this article we propose a new technique designed to optimize variable selection. The method initially involves repeated splitting of the data into a training and validation dataset, then for each split fitting a sequence of nested models on the training data (from the simplest model containing the intercept to the full model), validating the corresponding fitted models in the validation dataset, and finally choosing the empirical variable selection threshold α^* by minimizing the validation prediction errors averaged over the different splits. The final model is selected to include all predictors significant at the α^*-level. By choosing α^* to minimize the validation sum of squares (as opposed to using an arbitrary p-value) averaged over repeated validation datasets splits to avoid overfitting, we expect our procedure to find models that are both accurate and reliable.

In Section 2 we describe some commonly used model selection techniques and introduce the proposed methods. In Section 3 we report an extensive simulation study to evaluate the methods. In Section 4 we apply the methods to a real dataset and conclude with some final remarks in Section 5.

2 Methods

Consider the following linear regression model, where the i^{th} observation response variable, Y_i, is regressed on the p-dimensional covariate factor $\mathbf{X}_i = (1, X_{i,1}, \cdots, X_{i,p})$, for each $i = 1, \cdots, n$

$$E[Y_i|\mathbf{X}_i] = \mathbf{X}_i\boldsymbol{\beta} = \beta_0 + \beta_1 X_{i,1} + \cdots + \beta_p X_{i,p}, \tag{1}$$

where $\boldsymbol{\beta}^T = (\beta_0, \beta_1, \ldots, \beta_p)$ is the parameter vector. We assume the n observations are mutually independent and the responses are Gaussian, $Y_i|\mathbf{X}_i \sim N(\mathbf{X}_i\boldsymbol{\beta}, \sigma^2)$.

2.1 Some current variable selection procedures

2.1.1 Screening by p-values

A common model selection approach that researchers prefer due to its simplicity is to perform a two-stage approach in which, in the first stage, variables are screened based on their p-values (either from the individual univariate analyses or based on their marginal p-value from a model containing all predictors under investigation) followed by a second stage in which those predictors meeting a pre-specified threshold of significance (typically 0.05) in the first stage are included in the final model. One issue with this approach is that the threshold used for screening is arbitrary and could lead to an arbitrary model and to potential artifacts as pointed out by Freedman (1983) and Freedman and Pee (1989) who showed that depending on what screening level is used spuriously large F statistics and inflated Type I error can be obtained. Another shortcoming of this screening technique is that it attempts to find the best model based solely on the basis of significance testing without defining a measure of what represents a good model followed by finding the best model by optimizing this measure (Dziak et al. 2005).

2.1.2 BERDS: Backward elimination via repeated data splitting

Some other conventional variable selection procedures involve stepwise testing such as backward elimination (BE), forward selection (FS) and stepwise selection (SS). These depend on one or two predetermined thresholds, α_{stay} and α_{entry} that determine which variables are removed or added to the model. For example, the BE procedure starts with the full model containing all predictors and deletes variables from the model one by one (starting with the one that has the highest p-value greater than α_{stay}) until all the predictors remaining in the model yield F statistics significant at the prespecified α_{stay}. The FS procedure reverses the backward algorithm by starting with no variables in the model and adding variables one by one based on their contribution to the model, adding first the variable that has the largest F statistic that is significant at the α_{entry} level, and repeating the process by calculating the F statistics for the variables that have not been included in the model and checking against the α_{entry} cutoff, until no new predictors can be added to the model. Once a variable has been added to the model it remains in the model throughout the selection process. The stepwise selection combines the BE and FS but allows variables to be added or removed at each stage, so variables that are added early in the process do not necessarily stay in the final model. The SS proceeds just like the forward selection by adding predictors in the model one at a time, requiring the F statistic of the variable to be included to be significant at the α_{entry} cutoff. However after a variable is added the stepwise selection evaluates the F statistics for the currently selected variables and removes any that do not meet the criterion of having an F statistic significant at the α_{stay} threshold.

The issues with stepwise variable selection techniques have been discussed in numerous articles and a nice summary is provided by Harrell (2001). Some of these problems involve the bias of the regression coefficients and their standard errors, sensitivity of the chosen model to collinearity among the predictors, and the number of candidate predictors affecting the number of noisy variables being selected in the final model.

To resolve some of the issues encountered by stepwise approaches, Thall et al. (1992) introduced the backward elimination via cross-validation (BECV) technique. This method employs K-fold cross-validation to first find an empirical threshold α^* by minimizing the backward elimination cross-validation sum of squares, then performing backward elimination on the entire dataset using α^* as the α_{stay} threshold. BECV was shown to be effective at screening out noisy predictors at the expense of failing to include true predictors in many settings due to the fact that the cross-validation sum of squares is very sensitive to the particular K partitions the data is split into. Moreover, it was noticed that the cross-validation sum of squares displays a high degree of local variation and thus may have multiple local minima. To improve upon these shortcomings, Thall et al. (1997) introduced the backward elimination via repeated data splitting method (BERDS) that modifies BECV by performing repeated data splitting in place of K-fold cross-validation. Specifically, first the data are partitioned in two sets, half for estimation (E) and half for validation (V) purposes and an exhaustive backward elimination is performed on the estimation data. For each p-value threshold used in BE, the corresponding model from E is fitted in V and the validation sum of squares is computed. This process is repeated m times and the α threshold minimizing a (trimmed) average of the m sums of squares is chosen as the empirical cutoff to be used as α_{stay} when applying BE on the full dataset to obtain the final model. BERDS also modified the BECV by using a trimmed mean of the m objective functions as well as truncating the domain of the thresholds α to reduce the variability in the objective function described earlier. The authors showed that BERDS was superior to BECV with smaller model error, higher probability of excluding noisy variables and of selecting each of several independent true predictors.

2.1.3 Shrinkage methods: LASSO and elastic net

As an alternative to handle variable selection and dimensionality reduction in the case of large p small n datasets, regularization and shrinkage techniques have been developed. Some of the most commonly used shrinkage techniques that also perform variable selection are the LASSO and the elastic net. The elastic net solves the following regularization:

$$\min_{\beta_0,\beta}\left(\frac{1}{2n}\sum_{i=1}^{n}(Y_i-\beta_0-X_i\boldsymbol{\beta})^2+\lambda_1 P_{\lambda_2}(\boldsymbol{\beta})\right), \qquad (2)$$

where

$$P_{\lambda_2}(\boldsymbol{\beta}) = \frac{(1-\lambda_2)}{2}\|\boldsymbol{\beta}\|_2^2 + \lambda_2\|\boldsymbol{\beta}\|_1 = \sum_{j=1}^{p}\left(\frac{(1-\lambda_2)}{2}\beta_j^2 + \lambda_2|\beta_j|\right), \qquad (3)$$

with λ_1 nonnegative and λ_2 taking values between 0 and 1. The elastic net penalty is controlled by λ_2 and bridges the gap between Lasso ($\lambda_2 = 1$) and ridge regression ($\lambda_2 = 0$), with values in between 0 and 1 interpolating between the L^1 norm of $\boldsymbol{\beta}$ and the squared L^2 norm of $\boldsymbol{\beta}$. The regularization parameter λ_1 controls the amount of shrinkage (0 is no penalty and ∞ is complete penalty) and thus as λ_1 increases Lasso sets more coefficients to zero resulting in a more parsimonious model. Elastic net performs better than Lasso in the presence of correlated predictors for which Lasso tends to pick one and discard the others (Zou and Hastie 2005). LASSO and the elastic net were designed to deal with large p small n settings, and even though in this paper we focus attention on small to moderate p settings, we included these shrinkage techniques in our simulation study since they are widely used.

2.2 Selection Threshold OPtimized Empirically via Splitting (STOPES)

We combined the simplicity of the screening by p-values approach described in Section 2.1.1 with the advantages of deriving an empirical threshold as in BERDS and propose a new variable selection procedure that uses attributes of the two methods by first deriving an empirically optimized cutoff and then using this cutoff as a screening threshold to determine which variables to include in the final model. Specifically, we first randomly split the data into halves, one half for estimation and one half for validation. We then obtain the p-values for each simple regression containing one of the p predictors, fitted on the estimation data, and sort them in increasing order. Starting with the model containing only an intercept, we then fit the model containing the most significant predictor, followed by the model containing the top two most significant predictors, and so on until we fit the full model containing all predictors. For each of these models we obtain the predicted values and compute the corresponding prediction error sum of squares based on the validation dataset (called validation sum of squares for simplicity). This process is repeated m times and the optimized cutoff α^* is found to be the threshold that minimizes the objective function defined as the average of the validation sums of squares over the m repeated splits.

The challenge in minimizing this objective function lies in the fact that it displays a high degree of local variability. To smooth the function and alleviate the impact of such local fluctuations we truncated the domain of α to reduce local variation near 0 and above 0.5 and used a trimmed averaged for the sums of squares (as in Thall et al. (1997)). Moreover, the shape of the objective function is of such nature that typically after several steep drops, the function gradually decreases with smaller drops or rapidly rises back up again. Consequently, rather than choosing the global minimum as the cutoff, the desired threshold was chosen as the first local minimum following the last steep drop in $SSV(\alpha)$ to reduce the effect of such local variation. Thall et al. (1992) observed the same behavior and proposed as a compromise to minimize such an irregularly shaped function, the ".25-s rule" in which α^* is defined to be the smallest value of α such that $SSV(\alpha^*) \leq SSV(\alpha^0) + .25s$ where α^0 is the true minimum of the objective function $SSV(\alpha)$ and s^2 is an empirical estimate of the variance of $SSV(\alpha^0)$.

Our preliminary simulations in which different multipliers other than .25 were used, indicated that the ".25-s" rule functioned satisfactorily. We adopted it in our algorithm for determining the optimized cutoff α^* and denoted this method as STOPES-min.

While the ".25-s" rule was observed to perform well (see Section 3), its arbitrariness led us to investigate alternatives to identify the optimal cutpoint α^* in the presence of irregular local fluctuations and proposed for this purpose employing the Pruned Exact Linear Time (PELT) which is an exact search algorithm introduced by Killick et al. (2012) to search for changepoints. Note that we also investigated other changepoint algorithms such as binary segmentation (Scott and Knott 1974) and segment neighborhoods algorithms (Auger and Lawrence 1989) but PELT was selected for its superior performance. Our proposed variable selection technique employs PELT for the minimization step and is denoted STOPES (Selection Threshold OPtimized Empirically via Splitting).

The proposed algorithm is described step by step below:

1. Randomly split the data into halves: one half for estimation (E) and one half for validation (V).

2. For $j = 1, \cdots, p$, using E, fit each univariable regression model $E(Y|X_j) =$

$\beta_{0j} + \beta_j X_j$. Let α_j be the p-value for testing $H_{0j} : \beta_j = 0$. Order the p-values in increasing order, $\alpha_{(1)}, \cdots, \alpha_{(p)}$ and let $\alpha_{(p+1)} = 1$.

   (a) For each $k = 1, \cdots, (p+1)$, use $\alpha_{(k)}$ as the threshold that screens which variables are included in the model and fit the corresponding nested models on E:

$$E[Y_i | \mathbf{X}_{i,\mathbf{S_k}}] = \mathbf{X}_{i,\mathbf{S_k}} \boldsymbol{\beta}_{\boldsymbol{S_k}}, \qquad i \in E \tag{4}$$

   where $\mathbf{S_k} = \{j : \alpha_j < \alpha_{(k)}\}$. In other words, $\mathbf{S_1}$ would result in a model containing only the intercept, $\mathbf{S_2}$ would add to the intercept the most significant predictor, $\mathbf{S_k}$ would result in a model containing all predictors for which $\alpha_j < \alpha_{(k)}$, and $\mathbf{S_{p+1}}$ would correspond to the full model. Let $\hat{\boldsymbol{\beta}}_{\boldsymbol{S_k,E}}$ be the corresponding fitted coefficient vector from the regression in (4).

   (b) For each $k = 1, \cdots, (p+1)$, obtain the validation sum of squares corresponding to the particular (E, V) split and to the threshold $\alpha_{(k)}$:

$$SS_{E,V}(\alpha_{(k)}) = \sum_{i \in V} [Y_i - \mathbf{X}_{i,\mathbf{S_k}} \hat{\boldsymbol{\beta}}_{\boldsymbol{S_k,E}}]^2, \tag{5}$$

   where $\mathbf{S_k}$ and $\hat{\boldsymbol{\beta}}_{\boldsymbol{S_k,E}}$ are defined as in the step above. Note that the validation sum of squares is a step function with jumps at each p-value $\alpha_{(1)}, \cdots, \alpha_{(p)}$.

3. Repeat steps (1) and (2) m times.

   (a) At each of $r = 1, \cdots, m$ record the observed p-values and the corresponding validation sums of squares: $\alpha_{1,r}, \cdots, \alpha_{p,r}$ and $SS_{E_r,V_r}(\alpha_{(k),r})$, for $k = 1, \cdots, (p+1)$. Sort in increasing order the $p * m$ vector α containing the observed p-values at each of the m splits and use the step function definition (5) of the validation sums of squares to interpolate the sums of squares at each of these $p * m$ observed thresholds: for each $r = 1, \cdots, m$ let this $p * m$ vector of validation sums of squares be denoted by $SS_r(\alpha)$.

   (b) At each split, denote $\alpha_{min,r} = \min\{\alpha_1, \cdots, \alpha_p\}$ and $\alpha_{max,r} = \max\{\alpha_1, \cdots, \alpha_p\}$ for $r = 1, \cdots, m$. Let α_{min} be the 90^{th} quantile of $\{\alpha_{min,1}, \cdots, \alpha_{min,m}\}$ and α_{max} be the 10^{th} quantile of $\{\alpha_{max,1}, \cdots, \alpha_{max,m}\}$. Truncate the range of α between $\alpha_{min}, \alpha_{max}$ and take a 20% trimmed mean of $\{SS_1(\alpha), \cdots, SS_m(\alpha)\}$ over this truncated range of α. Denote this truncated and trimmed mean as $SSV(\alpha)$ which is the objective function to be minimized.

4. (a) STOPES-min: Employ the ".25-s" rule to minimize $SSV(\alpha)$: α^* is defined to be the smallest value of α such that $SSV(\alpha^*) \leq SSV(\hat{\alpha}^0) + .25s$ where $\hat{\alpha}^0$ is the observed global minimum of the objective function $SSV(\alpha)$ and s^2 is defined as the variance of $\{SS_{E_1,V_1}(\hat{\alpha}_0), \cdots, SS_{E_m,V_m}(\hat{\alpha}_0)\}$.
   (b) STOPES: Employ the PELT algorithm to identify the first change-point in $SSV(\alpha)$ and denote that as α^*.

5. Include in the final model all predictors significant at the α^*-level and fit the model on the full dataset.

3 Simulation Study

3.1 Simulation Design

We conducted extensive simulations to investigate the properties of the proposed methods. We generated the covariates $\mathbf{X}_1, \cdots, \mathbf{X}_p$ and the residuals, $\epsilon = \mathbf{Y} - E(\mathbf{Y}|\mathbf{X})$ as iid N(0, 1). The different configurations studied are chosen to mimic typical datasets encountered in practice for moderate sample sizes and numbers of predictors (see Table 1). Specifically, we varied the sample size, n, to be 100, 200, or 300 and assumed 10, 20, or 30 noisy predictors in the model. Models $M1$ through $M4$ allowed scenarios in which either a single true predictor, or 2 true predictors, or 3 true predictors were included in the model, while models $M5$-$M10$ included 3 true predictors. Model 1 assumed no correlation among the predictors, while models $M2$, $M3$ and $M4$ assumed the noisy predictors correlated with each other with correlation $\rho_n = 0.3, 0.6, 0.8$ respectively (and uncorrelated with the true predictors). Models $M5$, $M6$, and $M7$ allowed 2 of the 3 true predictors to be correlated with each other with correlation $\rho_t = 0.3, 0.6, 0.8$ respectively, while models $M8$, $M9$, and $M10$ further assumed that all noisy predictors are correlated with each other with correlation $\rho_n = 0.6$. The values of the β parameters (displayed in Table 2) were chosen based on empirical calibration studies such that for each particular configuration a 0.05 level t-test rejects the null hypothesis that $\beta = 0$ with 0.99 probability (see Thall et al. 1992).

We have also studied the null hypothesis under which none of the p predictors were related to the response Y (i.e. $\beta = 0$, model $M0$ in Table 1) and have reported the proportion of predictors that were significant at the 0.05 level in the final regression. This proportion should be close to 0.05 if the significance tests of regression coefficients of the variables included in the final model preserved a true Type I error rate of 5%. Note that this quantity is not available for LASSO and the elastic net as there is no statistical significance testing produced for these methods.

We compared the performance of the different methods investigated in terms of: (1) average number of noisy predictors included in the final model; (2) probability of including all true predictors in the selected model; (3) proportion of times the selected model was the correct model; (4) bias of the coefficients β; and (5) the model error $\frac{1}{n}\sum_{i=1}^{n}[\hat{Y}_i - E(Y_i|\mathbf{X}_i)]^2$, where $\hat{Y}_i$ is the i^{th} fitted value (estimated response). For the model error, we report the average ratios between the model error (as defined above) of the selected model and the model error of the true model (i.e. from fitting a model that includes only the true predictors). Operating characteristics were also evaluated by increasing the sample size while keeping the true model fixed to assess the asymptotic behaviour of the methods.

We conducted 1000 simulations for each scenario and averaged results across simulations to compare the operating characteristics of the methods investigated. Computations were performed in R using **Ubuntu system**. We used the glmnet R package to fit Lasso with cross-validation for the selection of λ while we assumed $\alpha = 0.5$ when fitting the elastic net. For the implementation of the proposed method STOPES, we have used the R package, changepoint with the cpt.meanvar function. BERDS was implemented by calling in R the C functions provided at `https://biostatistics.mdanderson.org/softwaredownload`.

3.2 Simulation Results

As seen in Table 3, for all methods the significance tests of regression coefficients of the final model preserve the Type I error at 5%, with BERDS being very slightly anti-conservative, while the other methods exhibiting a somewhat conservative behaviour for scenarios in which the ratio of sample size relative to the number of predictors (n/p) is the smallest ($m = 30$).

Simulation results are graphically summarized in Figures 1 through Figure 4 and further detailed in the Supplementary Tables (note that, unless noted otherwise, the elastic net had comparable performance with that of LASSO and for simplicity it was omitted from the figures, but full results can be found in the Supplementary Tables). The STOPES method outperforms or has similar performance to the other methods for all the simulations investigated, while the STOPES-min method follows closely behind. All methods have similar model error across simulations (Figures 1) except for Lasso which displays a much larger model error likely due to the larger bias in the Lasso estimates as well as selecting some of the most parsimonious models (on average Lasso and STOPES included in the model the least amount of noisy predictors). BERDS and the univariate screening at 0.05 level result in the most number of noisy predictors selected in the final models (Figures 2), regardless of the configuration, with the univariate screening tripling the number of selected noisy predictors as the number of total noisy predictors increases from 10 to 30 (about 0.5 noisy predictors for 10 noisy predictors, increasing to about 1 noisy predictor selected out of 20 noisy predictors and tripling to about 1.5 noisy predictors out of 30 predictors).

Except for LASSO, all methods have satisfactory performance in terms of including the true predictors in the final models (Figures 3), with BERDS and the univariate screening having the highest proportion of true predictors included (but as seen above at the expense of selecting more noisy predictors as well) while LASSO's ability to include all the true predictors diminishes with the increased sample size (as it moves further away from the small n large p scenario for which LASSO performs best).

The most challenging criterion is the proportion of times the different techniques selected the true model as the final model, i.e. included all the true predictors and none of the noisy predictors (Figures 4). In all the scenarios studied, the STOPES had the highest power for this criterion except for a few scenarios in which the STOPES-min was superior. BERDS consistently exhibited lowest power among most settings. As the number of true predictors increased, the power of including the exact model diminished (for more details see the Supplementary Tables).

Introducing correlation among predictors did not change substantially the operating characteristics of STOPES and STOPES-min whereas for the other methods the performance deteriorated particularly for BERDS and the univariate screening and especially for increasing numbers of noisy predictors. As expected, for scenarios with highest correlated true predictors (Models 7 and 10), the elastic net is more powerful than LASSO in selecting exactly the true model (see Supplementary Tables 7 and 10).

All methods except for LASSO and the elastic net exhibited small bias of the estimated coefficients (see Supplementary Table 11). The empirically chosen cutpoints for the two proposed methods tend to be closer together than those produced by BERDS for most scenarios investigated, especially as the sample size n increases (Supplementary Table 12).

Encouragingly, as seen in Figure 5, the proposed methods showed excellent asymptotic behaviour showing model error decreasing, power increasing towards 1, and number of noisy predictors dropping towards zero with increasing sample size, and converging towards an asymptote at a much faster rate than the other competitors: by sample sizes of $n = 300$, STOPES and STOPES-min have stabilized to an asymptote while for the other methods the power was still below 1 and the number of noisy predictors above 0 with sample sizes as large as $n = 500$. Similar convergence patterns were observed for the other configurations studied (data not shown).

4 Data Analysis

We applied the proposed methods to a dataset consisting of all adult patients that underwent partial hepatectomy at Memorial Sloan Kettering Cancer Center between 2002 and 2003 (n=314). This is a subset of a larger study population reported in Sima et al. (2009) which included all patients undergoing partial hepatectomy between 1995 and 2003; we only focused on a subset of the original data in order to illustrate the methods on a dataset that resembles the sample sizes used in the simulation study. The outcome of interest for this analysis was the amount of blood loss experienced during surgery. To achieve normality we have used the log-transformed blood loss measurements. There were 22 available preoperative factors that we investigated as potential predictors.

Since all the investigated methods (except the Univariate $p < 0.05$) involve random sampling from the data (and thus a re-run of the analysis can potentially result in a different model), we applied the methods 100 times and summarized the final models selected by the different methods and the number of times they were selected. Furthermore, for STOPES, STOPES-min, and BERDS we varied m, the number of splits, to be 20, 100, and 1000 to study whether the model converges to a single final model as expected when $m \to \infty$. These results are reported in Table 4.

As we increased the number of splits, the STOPES and STOPES-min methods converged to a single final model (the same model for both methods when $m = 1000$ splits were employed), while LASSO and elastic net were divided between four different models. Out of the 100 runs, BERDS selected about three quarters of the time a two-variable model, while the remaining times it selected much bigger models. Except for STOPES and STOPES-min, the models selected predominantly by the other methods were all different from each other but they all included the number of segments resected indicating the importance of this predictor.

Summaries (median and IQR) of the cutpoints selected by the different methods are reported in Table 5. STOPES-min and STOPES display very stable behavior with similar distributions for the optimal cutpoints selected with tight ranges indicating convergence, whereas BERDS results in wider ranges that do not improve as the number of splits is increased.

The models predominantly picked by STOPES and STOPES-min have low residual sums of squares without growing too large in size (as more variables are added into the model, the residual sum of squares becomes smaller at the expense of more complex models). We also report the adjusted R^2 but there was not much separation for this criterion among the different models and there was still a lot of variability unexplained even after accounting for all the predictors in the model (adjusted R^2 of 0.16 for the full model) indicating that blood loss during the surgery is a difficult outcome to predict.

In conclusion, the data analysis confirms the findings from the simulation study, with our new methods resulting in models that provide tight fit to the data (with residual sum of squares close to the plateau values beyond which further lowering the RSS may not warrant increasing the complexity of the model; see the type I/type II errors tradeoff discussed in Section 1). Moreover, as we increased the number of splits performed, these methods converged to a single final model contrary to the other methods for which convergence was not attained. This is in line with the findings in the simulation study where we observed that the proposed methods already converged to an asymptote with sample sizes of 300, while the others were converging at a much slower rate.

5 Discussion

Many variable selection approaches have been proposed over the years, yet model selection continues to pose difficulty to researchers. In this article we introduce a new model selection method which presents a number of advantages over the existing methods. By deriving an empirical p-value threshold via minimizing the prediction error of a sequence of nested models (from an intercept only model to a full model), our method avoids the shortcomings of the univariable screening approach as well as of other techniques that use arbitrary cutoffs to decide which variables to include in the model. It also bypasses the risk of omitting correlated true predictors or predictors that are important only after adjusting for other variables. Furthermore, using repeated data splitting to estimate the objective function to be minimized it avoids overfitting and results in adequate yet parsimonious model selection with good prediction ability. Indeed, we have used extensive simulations to study the finite sample properties of the two new methods and have demonstrated that they have superior performance compared to competitor methods, with low model errors and with highest power of detecting the correct model while including the least amount of noisy predictors in the model. Due to its general nature in the minimization step and to its excellent performance, we recommend STOPES as the preferred choice to STOPES-min.

Encouragingly, with increasing sample size the proposed method had operating characteristics stabilizing to an asymptote and converging to a single final model at smaller sample sizes than the other methods. Note that in our simulations we have used $m = 20$ splits (Thall et al. 1997) to estimate the threshold α^*. However for data analysis we recommend using a larger number of repeated splits (such as $m = 500, 1000$ recognizing that for certain datasets the methods will converge to a final model at a faster or slower pace.

The scenarios considered here were meant to mimic typical scenarios encountered in practice with moderate sample sizes and moderate numbers of predictors, and thus our methods were not specifically examined under other scenarios such as large p small n problems.

The algorithm is straightforward and easy to implement using readily available software; R code is provided in the Supplementary Material and can be obtained from the authors per request, while building an R package is underway. Extensions to other regression types such as Cox regression or logistic regression are possible by changing the optimization objective function (for example using concordance index for Cox models and AUC for logistic regression) and are currently being investigated.

ACKNOWLEDGEMENTS: We thank Patrick Hilden and Venkat Seshan for

their valuable assistance with the implementation of the BERDS method.

References

Auger, I. E. and Lawrence, C. E. (1989) Algorithms for the optimal identification of segment neighborhoods. *Bulletin of Mathematical Biology*, **51**, 39–54.

Dziak, J., Li, R. and Collins, L. (2005) Critical review and comparison of variable selection procedures for linear regression. Technical Report Penn State University.

Freedman, D. A. (1983) A note on screening regression equations. *The American Statistician*, **37**, 152–155.

Freedman, D. A. and Pee, D. (1989) Return to a note on screening regression equations. *The American Statistician*, **43**, 279–282.

Harrell, F. (2001) *Regression Modelling Strategies: With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis.* New York: Springer.

Hastie, T., Tibshirani, R. and Friedman, J. (2009) *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* New York: Springer.

Killick, R., Fearnhead, P., Aston, J. and Eckley, I. A. (2012) Optimal detection of changepoints with a linear computational cost. *Journal of the American Statistical Association*, **107**, 1590–1598.

Miller, A. J. (2002) *Subset selection in regression.* New York: Chapman and Hall.

Scott, A. J. and Knott, M. (1974) A cluster analysis method for grouping means in the analysis of variance. *Biometrics*, **30**, 507–512.

Sima, C. S., Jarnagin, W. R., Fong, Y., Elkin, E., Fischer, M., Wuest, D., D'Angelica, M., DeMatteo, R. P., Blumgart, L. H. and Gönen, M. (2009) Predicting the risk of perioperative transfusion for patients undergoing elective hepatectomy. *Annals of Surgery*, **250**, 914–921.

Thall, P. F., Russell, K. E. and Simon, R. M. (1997) Variable selection in regression via repeated data splitting. *Journal of Computational and Graphical Statistics*, **6**, 416–434.

Thall, P. F., Simon, R. and Grier, D. A. (1992) Test-based variable selection via cross-validation. *Journal of Computational and Graphical Statistics*, **1**, 41–61.

Zou, H. and Hastie, T. (2005) Regularization and variable selection via theelastic net. *J. R. Statist. Soc. B*, **67**, 301–320.

Table 1: Different configurations used in simulations.

Model	Number of True Predictors	Number of Noisy Predictors	Sample size n	Correlation
M0	0	10, 20, or 30	100, 200, or 300	None
M1	1, 2, or 3	10, 20, or 30	100, 200, or 300	None
M2	1, 2, or 3	10, 20, or 30	100, 200, or 300	$\rho_n = 0.3$
M3	1, 2, or 3	10, 20, or 30	100, 200, or 300	$\rho_n = 0.6$
M4	1, 2, or 3	10, 20, or 30	100, 200, or 300	$\rho_n = 0.8$
M5	3	10, 20, or 30	100, 200, or 300	$\rho_t = 0.3$
M6	3	10, 20, or 30	100, 200, or 300	$\rho_t = 0.6$
M7	3	10, 20, or 30	100, 200, or 300	$\rho_t = 0.8$
M8	3	10, 20, or 30	100, 200, or 300	$\rho_t = 0.3, \rho_n = 0.6$
M9	3	10, 20, or 30	100, 200, or 300	$\rho_t = 0.6, \rho_n = 0.6$
M10	3	10, 20, or 30	100, 200, or 300	$\rho_t = 0.8, \rho_n = 0.6$

Table 2: True coefficients β_1, β_2, β_3 under different sample sizes and correlation structures.

Sample size	Correlation			
	$\rho_t = 0$	$\rho_t = 0.3$	$\rho_t = 0.6$	$\rho_t = 0.8$
n=100	0.58	(0.42, 0.42, 0.54)	(0.33, 0.33, 0.54)	(0.29, 0.29, 0.54)
n=200	0.34	(0.26, 0.26, 0.34)	(0.21, 0.21, 0.34)	(0.19, 0.19, 0.34)
n=300	0.27	(0.21, 0.21, 0.27)	(0.17, 0.17, 0.27)	(0.15, 0.15, 0.27)

Table 3: Proportion of noisy variables found significant at the 0.05 level in the final model under the null model, $M0$.

Method	$n = 100$			$n = 200$			$n = 300$		
	$p = 10$	$p = 20$	$p = 30$	$p = 10$	$p = 20$	$p = 30$	$p = 10$	$p = 20$	$p = 30$
STOPES-min	0.051	0.045	0.039	0.049	0.047	0.041	0.048	0.049	0.043
STOPES	0.052	0.045	0.040	0.048	0.048	0.042	0.048	0.049	0.044
Univar $p < 0.05$	0.048	0.044	0.039	0.046	0.045	0.040	0.047	0.046	0.043
BERDS	0.058	0.059	0.062	0.054	0.055	0.053	0.055	0.052	0.051

Table 4: Final models selected by the different methods for the blood loss analysis and their corresponding diagnostics.

			Frequency (out of 100) the models were selected as final models											
			STOPES-min			STOPES			BERDS[d]			Univar		Elastic
Models selected[a]	RSS^b	R^{2c}	$m = 20$	$m = 100$	$m = 1000$	$m = 20$	$m = 100$	$m = 1000$	$m = 20$	$m = 100$	$m = 1000$	$p < 0.05$	LASSO	Net
Intercept	34.60		0	0	0	0	0	0	0	0	0	0	18	9
Segm rx	30.89	0.10	0	0	0	0	0	0	0	0	1	0	34	16
Lobes, Segm rx	30.64	0.11	4	0	0	5	0	0	0	0	0	0	35	56
Segm rx, Exhep rx	29.34	0.15	0	0	0	0	0	0	82	68	75	0	0	0
Diagnosis, Segm rx, Exhep rx	29.15	0.155	0	0	0	0	0	0	1	4	6	0	0	0
Lobes, Segm rx, Exhep rx	29.04	0.15	32	18	0	18	8	0	0	0	0	0	13	19
Lobes, Segm rx, Maj exhep rx, Exhep rx	29.03	0.15	36	77	100	42	76	100	0	0	0	0	0	0
Lobes, Bilateral rx, Segm rx, Maj exhep rx, Exhep rx	28.87	0.14	25	5	0	33	16	0	0	0	0	0	0	0
Lobes, Bilateral rx, Segm rx, Maj exhep rx, Exhep rx, Size largest tumor	28.87	0.15	1	0	0	2	0	0	0	0	0	0	0	0
Diagnosis, Segm rx, Exhep rx BMI, Size largest tumor, Bili	27.81	0.178	0	0	0	0	0	0	12	21	10	0	0	0
Diagnosis, Segm rx, Exhep rx BMI, Size largest tumor, Bili Steatosis	27.58	0.181	0	0	0	0	0	0	4	4	5	0	0	0
Diagnosis, Lobes, Bilateral rx, Segm rx, Maj exhep rx, Exhep rx, Size largest tumor	28.33	0.16	2	0	0	0	0	0	0	0	0	0	0	0
Diagnosis, Lobes, Bilateral rx, Segm rx, Maj exhep rx, Exhep rx, Size largest tumor, Alb preop	28.07	0.16	0	0	0	0	0	0	0	0	0	100	0	0

[a] For simplicity, the names of the variables have been shortened: “rx” stands for “resection”; the full names of the variables are as following: number of segments resected, lobes (0 vs 1), extrahepatic resection (0 vs 1), diagnosis (malignant vs benign), major extrahepatic procedure (0 vs 1), bilateral resection, size largest tumor, BMI (continuous), bilirubin preop (continuous), steatosis (0 vs 1), Albumin preop (continuous).

[b] RSS stands for residual sum of squares.

[c] Adjusted R^2.

[d] BERDS selected two other larger models (10 and 12 variables) for 1, 2, or 3 times (out of 100) and for simplicity these were not reported (consequently frequencies in the BERDS columns do not sum to 100.

Table 5: Median and interquartile range of the p-value cutpoints selected by the different methods over the 100 runs of the blood loss data analysis.

	$m = 20$	$m = 100$	$m = 1000$
Method			
STOPES-min	0.0013 (0.0008, 0.0023)	0.0013 (0.001, 0.0016)	0.0014 (0.0013, 0.0015)
STOPES	0.0016 (0.001, 0.0026)	0.0015 (0.001, 0.0019)	0.0014 (0.0013, 0.0015)
BERDS	0.0078 (0.0029, 0.0384)	0.01 (0.003, 0.07)	0.01 (0.003, 0.047)

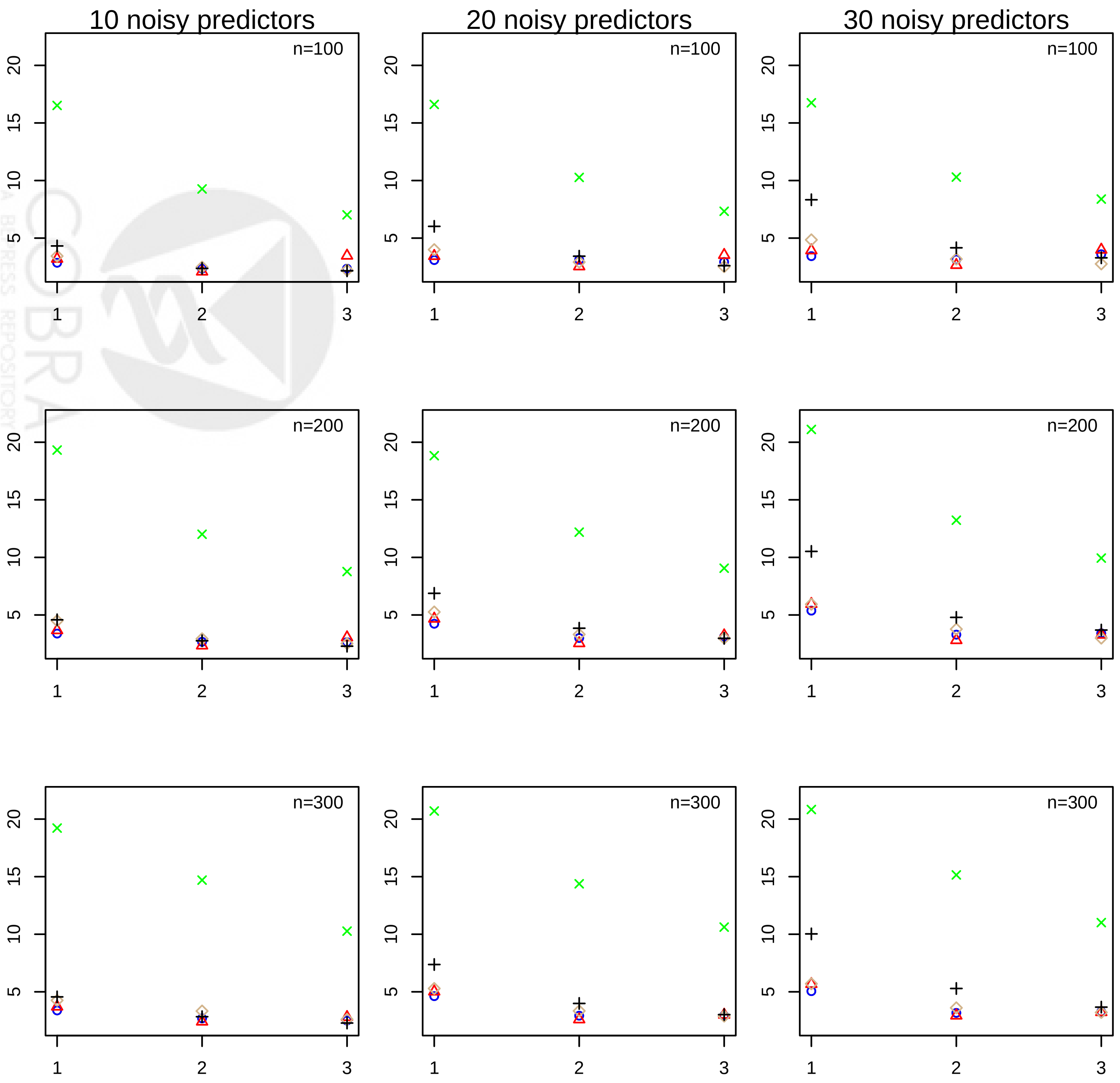

Figure 1: Average Ratios Model Errors (ME) for the models selected by STOPES (△), STOPES-min (○), BERDS (◇), Univariate $p < 0.05$ (+), and LASSO (×) under Model 1 with 1, 2, or 3 true predictors (x-axis).

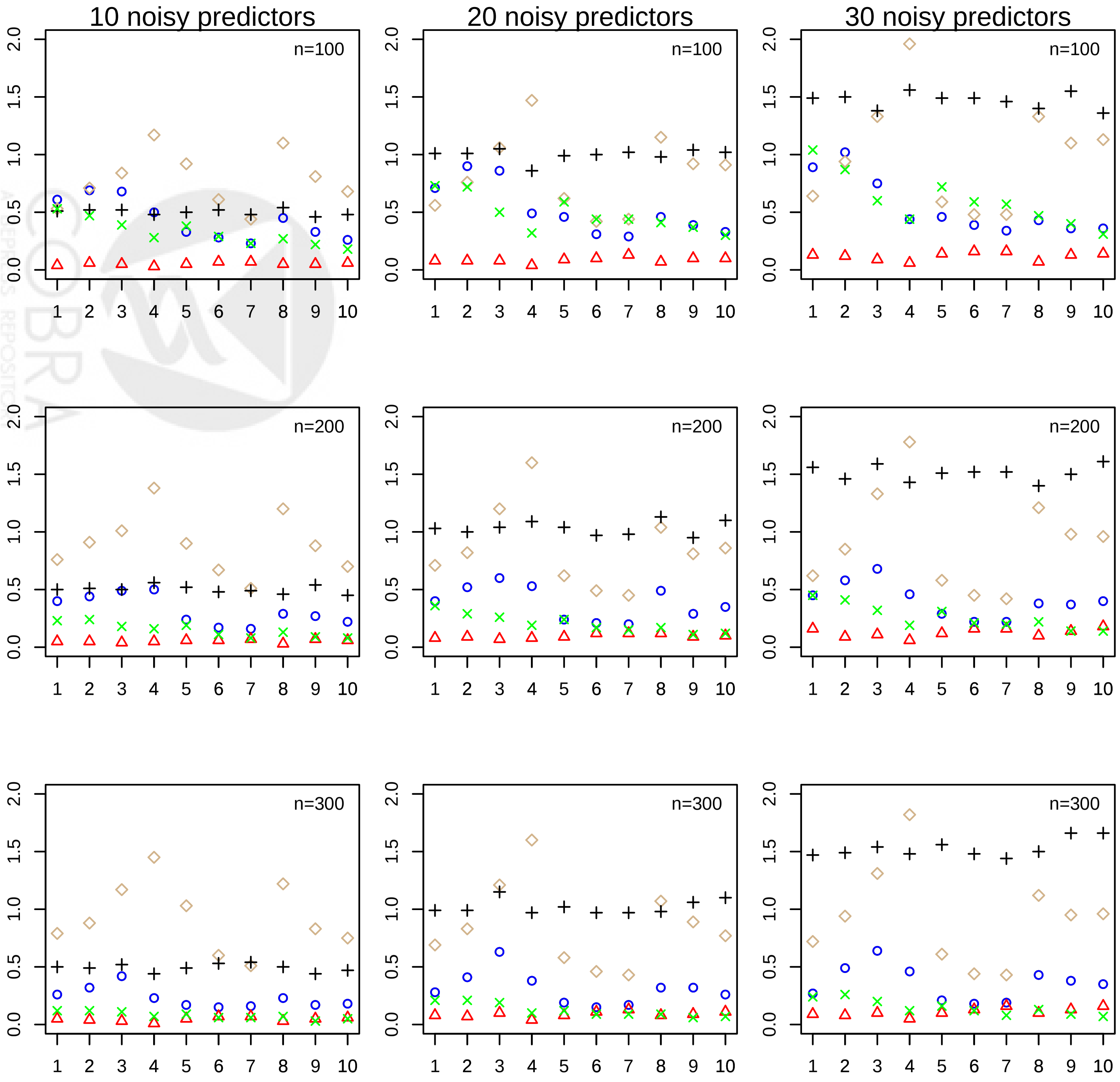

Figure 2: Average number of noisy predictors included in the model by STOPES (△), STOPES-min (○), BERDS (◇), Univariate $p < 0.05$ (+), and LASSO (×) assuming three true predictors in any of the 10 models (M1-M10 on x-axis).

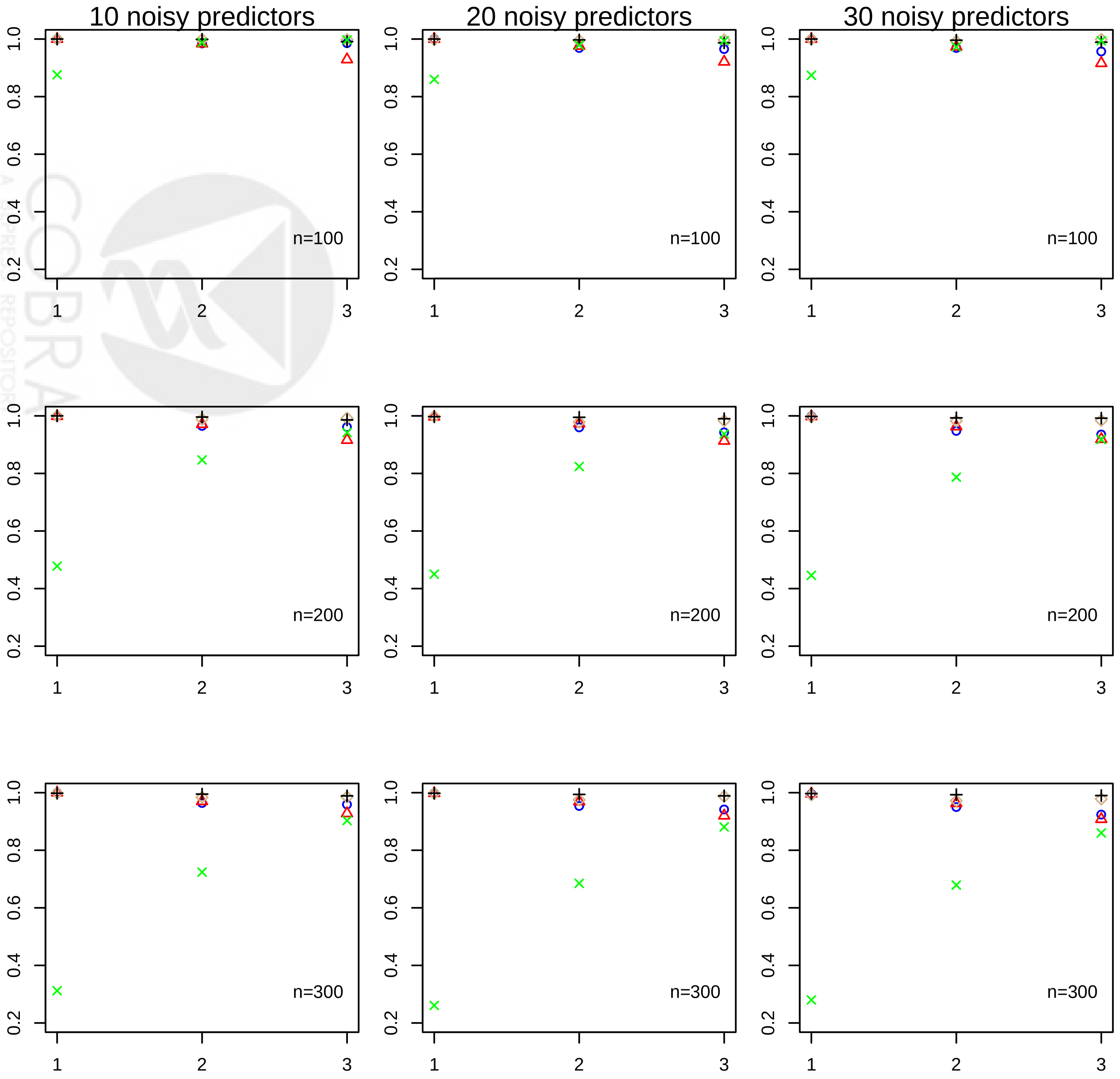

Figure 3: Average proportion of true predictors selected by STOPES (△), STOPES-min (○), BERDS (◇), Univariate $p < 0.05$ (+), and LASSO (×) under Model 1 with 1, 2, or 3 true predictors (x-axis).

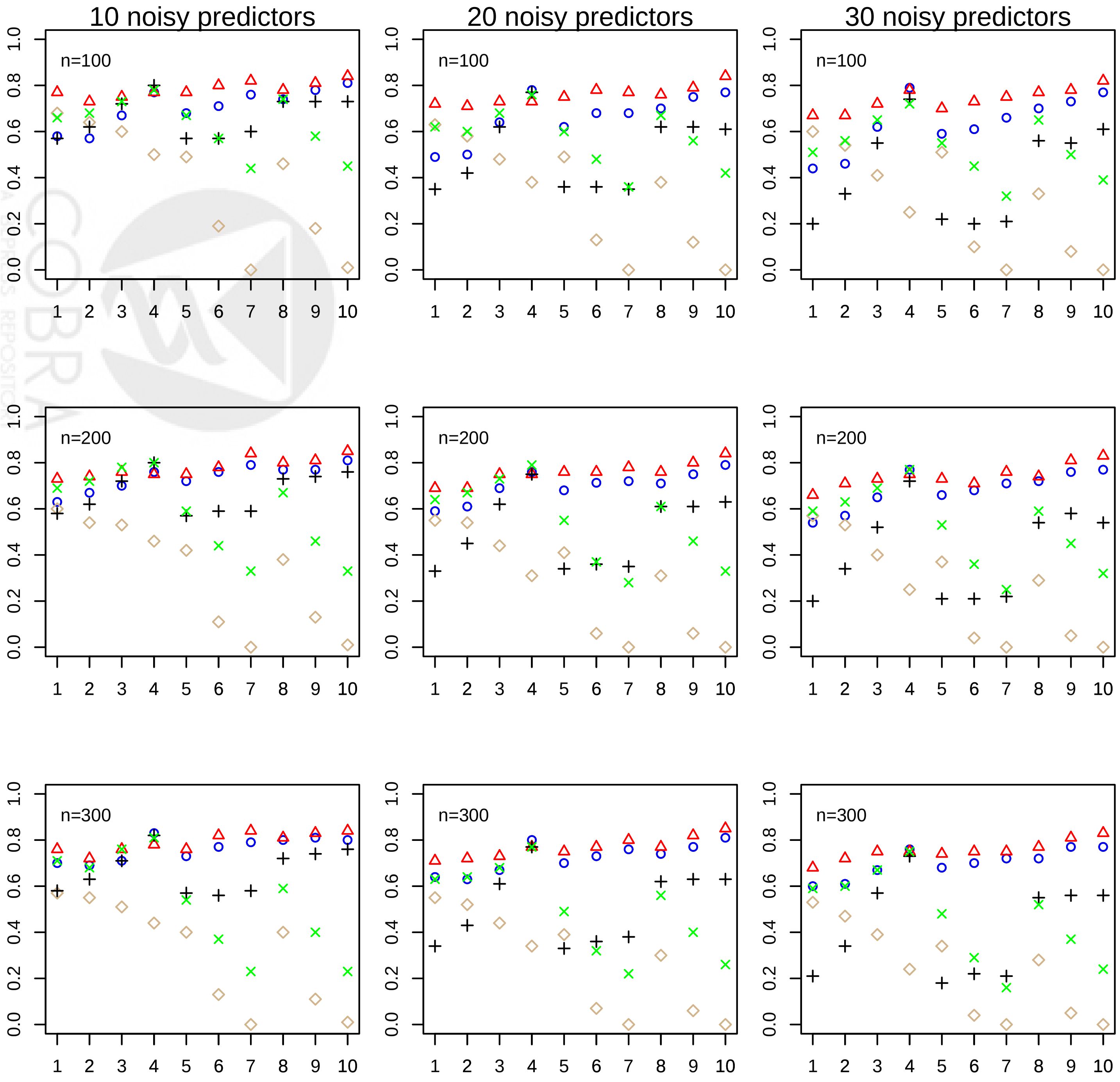

Figure 4: Average number of times the correct model was selected by STOPES (△), STOPES-min (○), BERDS (◇), Univariate $p < 0.05$ (+), and LASSO (×) assuming three true predictors in any of the 10 models (M1-M10 on x-axis).

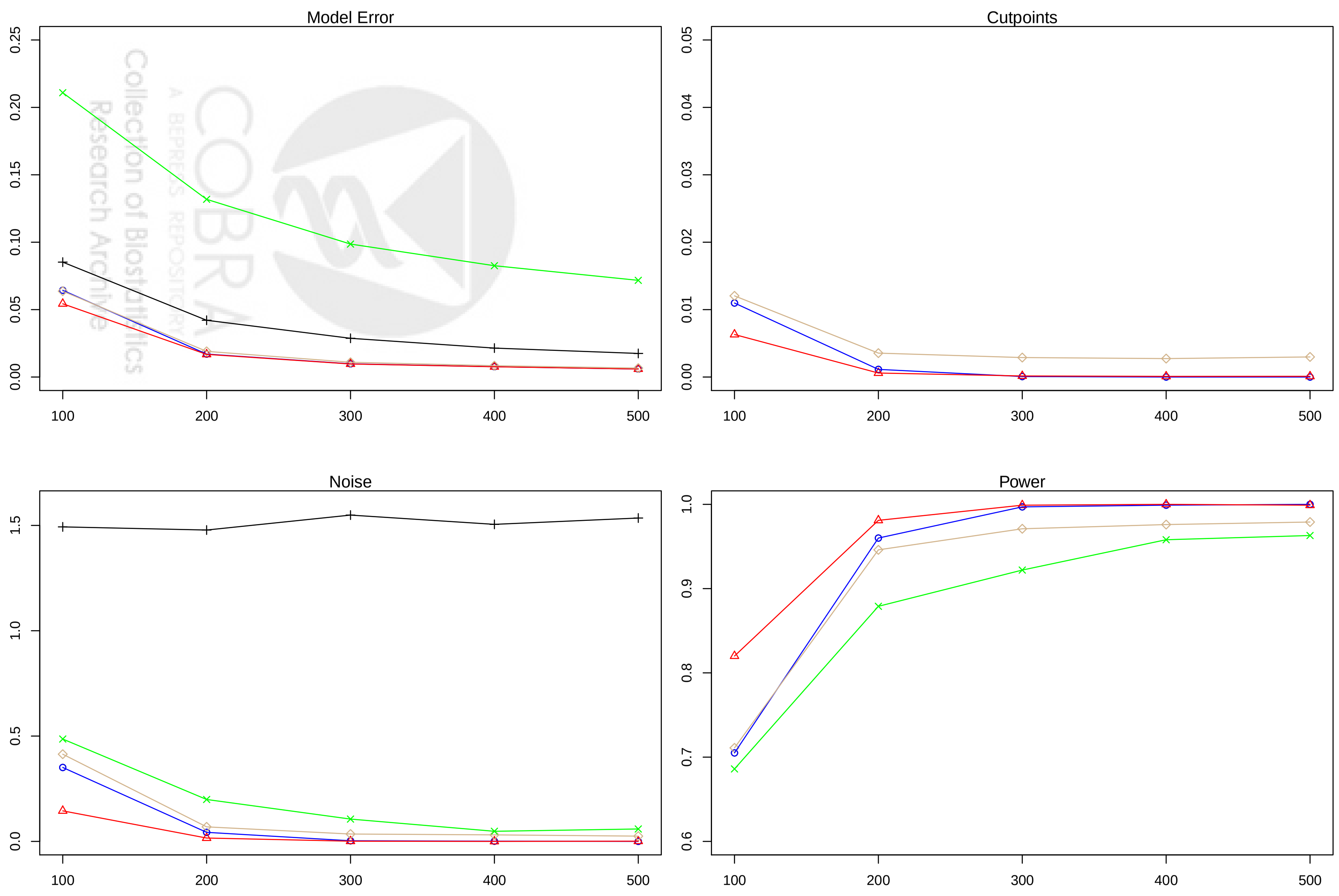

Figure 5: Asymptotic behavior of STOPES (△), STOPES-min (○), BERDS (◇), Univariate $p < 0.05$ (+), and LASSO (×) assuming 30 noisy predictors and 2 true predictors under Model 1, while increasing the sample size $n = 100$ through $n = 500$ (x-axis).