Ranking-Based Variable Selection for high-dimensional data

Department of Statistics, London School of Economics and Political Science

Supplementary Material

S1 Details of the implementation of the RBVS algorithm

In this section, we provide a detailed description of our implementation of Algorithm 1, which is available in the R package **rbvs**. First, we recall all necessary notation. By $\mathbf{Z}_i = (Y_i, X_{i1}, X_{ip})$, $i = 1, \dots n$ we denote a random sample we observe, where Y_i is a response and $\{X_{i1,}, \dots, X_{ip}\}$ is the set of the covariates. A chosen (empirical) measure of dependence between the response and j-th covariates is denoted by $\hat{\omega}_j$, positive integer $m < n$ is a subsample size (parameter of our method), B is a positive integer (typically from 50 to 500).

The RBVS algorithm aims to identify the set of covariates which non-spuriously appears at the top of the variable ranking based on the empirical measure $\hat{\omega}_j$. It consists of four steps Implementation of Step 1 is straightforward. It is worth noting that in Step 2 we do not actually need to evaluate complete rankings for each subsample, it is sufficient to find only a partial ranking, i.e. indices of the $k_{\max}$ top ranked variables, as only those are used in 3. The computational complexity of finding a full ranking is $O(p \log(p))$. For the partial ranking, it takes (on average) just $O(p + k_{\max} \log(k_{\max}))$ operations. The gain can be substantial when $p \gg k_{\max}$.

Recall that $\hat{\mathcal{A}}_{k,m} = \text{argmax}_{\mathcal{A} \in \Omega_k} \hat{\pi}_{m,n}(\mathcal{A})$, where Ω_k is the set of all k-element subsets of $\{1, \dots, p\}$. Despite the fact that the definition involves searching of the maximum empirical probability over a set the size of which grows extremely fast, finding $\hat{\mathcal{A}}_{k,m}$ is actually quick. This is because the number of the subsets which could have appeared at the top of the ranking at least once is limited by the total number of evaluated rankings. In Step 3, we apply procedure outlined in Algorithm S1.

The computational complexity of Step 1 is of order $O(k_{\max} B r)$ (for each k we use the fact that at the previous step $k-1$ elements are already in increasing order; we do not need to sort $R_{1,l}, \dots, R_{k,l}$ from scratch). The second part is relatively quick - we need to find the most frequent element among k-element sequences. For each $k = 1, \dots, k_{\max}$,

Algorithm S1 Finding $\hat{\mathcal{A}}_{k,m}$ and computing $\hat{\pi}_{m,n}\big(\hat{\mathcal{A}}_{k,m}\big)$

Input: Variable rankings $(R_{l1},\dots,R_{lk_{\max}})$, $l=1,\dots,Br$.
Output: Estimates $\hat{\mathcal{A}}_{k,m}$ and $\hat{\pi}_{m,n}\big(\hat{\mathcal{A}}_{k,m}\big)$ for $k=1,\dots,k_{\max}$.
procedure kTopRankedSets($\{(R_{l1},\dots,R_{lk_{\max}}\}_{l=1}^{Br}$)
for $k=1,\dots,k_{\max}$ **do**
Step 1 for each l, insert R_{lk} into $S_{l,k-1}$ s.t. resulting sequence $S_{l,k}$ is in increasing order
Step 2 find S_k^* the most frequently occurring among $S_{1,k},\dots,S_{Br,k}$
Step 3 set $\hat{\mathcal{A}}_{k,m}=S_k^*$ and $\hat{\pi}_{m,n}\big(\hat{\mathcal{A}}_{k,m}\big)=\frac{no.\ l\ s.t.\ S_{l,k}=S_k^*}{Br}$
end for
end procedure

the computational complexity is $O(kBr)$. Therefore in total the algorithm we use to find $\hat{\mathcal{A}}_{k,m}$ is of order $O(k_{\max}^2 rB)$. Algorithm S1 can be easily run on multiple CPUs (which is supported by the **rbvs** package) or a GPU, which makes it feasible for extremely large data sets. In practice, Step 3 of the RBVS algorithm (Algorithm 1) takes much less computational time than Step 2. Moreover, the **rbvs** package provides optimised, C-implemented routines performing Algorithm 1 (which includes Algorithm S1).

S2 Real data examples

In this section, we present applications to two real datasets: the Boston housing data and the prostate cancer data.

S2.1 Boston housing data

We apply our methodology to the Boston housing data set (Harrison and Rubinfeld, 1978) which has been frequently adopted to illustrate performance of various variable selection and estimation techniques (see e.g. Radchenko and James (2010), Cho and Fryzlewicz (2012) or Fan et al. (2014). We use Boston Housing data available in the R package **mlbench** (Leisch and Dimitriadou, 2010) containing 15 numerical covariates which may have influence over the median price recorded in $n=506$ locations. As in Cho and Fryzlewicz (2012), we additionally consider interaction terms between the explanatory variables so the final data set has $p=120$ covariates.

Harrison and Rubinfeld (1978) used the linear model to analyse the price, thus we apply RBVS combined with the linear measures introduced Section 3.2.

Figure S1 shows a "RBVS path", i.e. probabilities corresponding to the k-element subsets of covariates the most frequently occurring as the most influential ones (defined by (4)). The probability path for RBVS PC declines much slower than those corresponding to RBVS Lasso and RBVS MC+. This results in a different numbers of selected

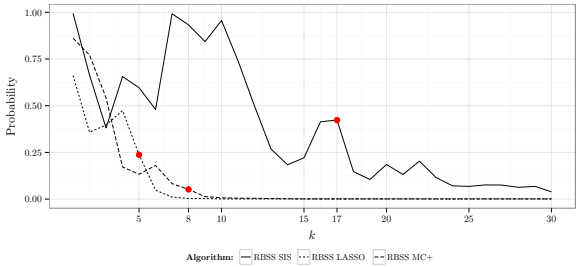

Figure S1: The Boston housing data: the estimated probabilities corresponding to the k-element subsets top-ranked the most frequently. The dots indicate the probability at $k = \hat{s}$, which is the number of elements selected according to the suggested approach. The subsample size $m = \frac{n}{2} = 253$ and $B = 250$.

variables; RBVS PC chooses 17 covariates, while RBVS MC+ selects 8 and RBVS Lasso MC+ selects only 5. We argue that in this example RBVS PC, as based on a marginal measure, includes some variables that are not useful in a predictive model. Intuitively, if two or more variables were highly correlated to the response, then interactions formed of any two of those would be highly correlated to Y.

To investigate predictive usefulness of RBVS based methods, we split the data randomly, assembling approximately 50%, 25% and 25% observations to the train, validation and test sets, respectively. On the training set, we select variables and obtain OLS estimates of the regression coefficients (for Lasso and MC+ we consider all set candidates on their solution paths, for RBVS based methods we take the subsample size equal to $m = \left\{\frac{1}{8}, \frac{2}{8}, \ldots, \frac{7}{8}\right\} n_{train}$). Next, we evaluate the average prediction error on the validation set and choose the covariates minimising the error. Finally, we find the average prediction error, R squared coefficient (R^2) and adjusted R squared (R^2_{adj}) on the test set.

Table S1 reports the results averaged over 500 random splits of the data; PG in this summary corresponds to the linear model studied in Section 2.2 of Pace and Gilley (1997). RBVS PC, RBVS Lasso and RBVS MC+ perform similar to PG in terms of prediction accuracy, which can be seen from the corresponding values of the test error and R^2. On the other hand, RBVS Lasso and RBVS MC+ choose on average only 9 variables and consequently perform best in terms of R^2_{adj}. Lasso and MC+ achieve the best test error; however, they select about 50 variables on average. By contrast, IRBVS Lasso and IRBVS MC+ choose no more than 27 covariates, yet they achieve similar prediction accuracy as Lasso and MC+ respectively. Both RBVS PC and IRBVS PC perform reasonably well in terms of prediction accuracy, however, they select more variables than the remaining RBVS and IRBVS based techniques. This is probably caused by the strong correlations between covariates, which is due to the way the data set has been produced.

				RBVS			IRBVS		
	PG	Lasso	MC+	PC	Lasso	MC+	PC	Lasso	MC+
test error	0.037	0.032	0.032	0.038	0.038	0.038	0.036	0.033	0.033
R^2	0.773	0.803	0.805	0.769	0.766	0.765	0.780	0.798	0.801
R^2_{adj}	0.735	0.638	0.609	0.708	0.748	0.747	0.571	0.739	0.745
# selected variables	18.0	49.3	55.0	25.4	9.2	9.1	44.7	27.6	26.5

Table S1: Boston housing data: test error, R squared, adjusted R squared and the number of selected variables, averaged over 500 test sets.

S2.2 Prostate cancer data set

We analyse the Prostate cancer data (Singh et al., 2002) which is frequently used to evaluate the performance of various classification methods (Pochet et al. (2004), Fan and Fan (2008), Hall and Xue (2014)). It consists of expression levels of $p = 12600$ genes from 52 tumour and 50 normal prostate samples in the training set, and 9 tumour and 25 normal samples in the test set coming from an independent experiment. The response variable Y is binary (1 for tumour samples, 0 for normal samples) and X_j, the expression of the j'th gene, is a continuous variable.

We compare performance of RBVS against its two competitors, StabSel (Meinshausen and Bühlmann, 2010) and the approach of Hall and Miller (2009) (HM). Due to a very huge number of variables, we take the marginal correlation (i.e. PC) as a base learner for both RBVS and StabSel, as it is least computationally demanding across measures studied in the paper. This choice was previously used in this and similar classification problems; see Fan and Lv (2008) and Hall and Xue (2014).

To provide a fair comparison, we apply these three methods with the same subsamples taken from the data, drawn as in Definition 2.4. Besides the number of subsamples and their size, we need to specify the threshold π and the bound for the expected number of false positives EV for StabSel, the significance level α and the cut-off level c for HM. We try several values for each pair of these parameters.

We use RBVS, HM and StabSel on the training set to identify the important genes. Still on the training set, we fit the logistic regression model, using the selected covariates only. Subsequently, we use the fitted model to classify samples in the test set. Finally, we record the number of correctly classified samples. The entire experiment is repeated 50 times, to minimise the impact of a particular random draw, and the medians are reported.

The median correct classification rate on the test set for the RBVS algorithm is 31 out of 34 and this is always achieved using from 3 to 6 genes only, both for subsamples of size $m = \lfloor \frac{n}{2} \rfloor = 51$ and $m = \lfloor \frac{3n}{4} \rfloor = 76$. For some random draws, RBVS selects exactly 4 genes, which result in the classification rate of 33. Figure S2 summarises the corresponding numbers for the StabSel and HM algorithms, with various tuning parameters of these methods. For $m = \lfloor \frac{n}{2} \rfloor$, there exists one pair of parameters that leads to a better error control for StabSel and HM (33 correctly

classified samples), however, RBVS is always better when $m = 76$. The parameters which are the best in this example are much different from those recommended for StabSel and HM. Unlike its competitors, RBVS automatically selects an appropriate number of genes, being particularly effective in this example.

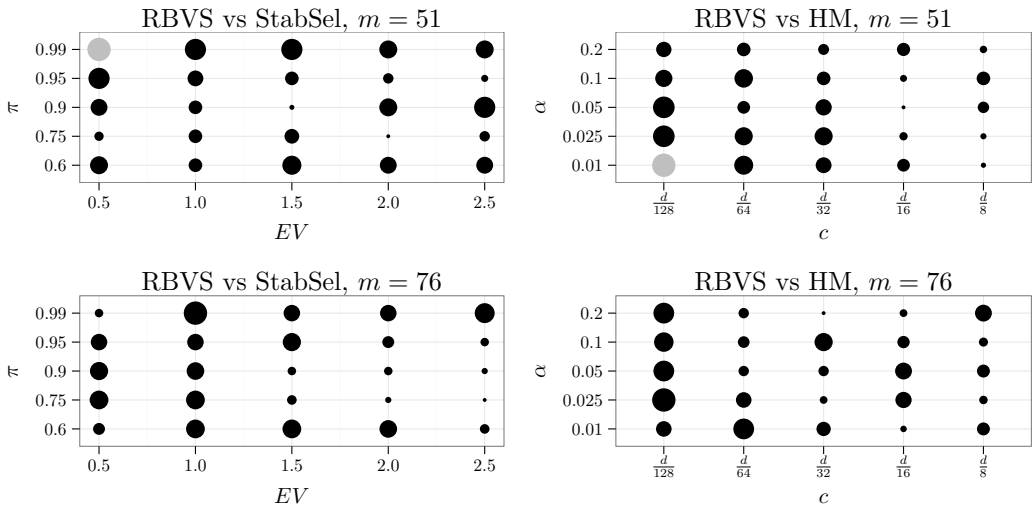

Figure S2: Prostate cancer data set: the median of the number of correctly classified samples on the test set, evaluated over 50 runs of the algorithms studied. The larger a circle, the better classification rate. Grey colour indicates the cases where the median classification rate is no worse than 31, the median classification rate achieved by RBVS PC. The number of subsamples $B = 500$.

S3 Additional high-dimensional simulation study

The aim of the simulation study reported in this section is threefold. First, to provide an extensive comparison of the performance of RBVS and StabSel algorithms. Second, to investigate their utility in the high-dimensional framework. Third, to check how sensitive both approaches are to the choice of the subsample size m.

S3.1 The setting

The data are generated from the following linear model

$$Y_i = \beta_1 X_{i1} + \ldots, \beta_p X_{ip} + \varepsilon_i, \quad i = 1, \ldots, n,$$

where

- X_{ij}'s follow the factor model $X_{ij} = \sum_{l=1}^{K} f_{ijl}\varphi_{il} + \theta_{ij}$, with f_{ijl}, φ_{il}, θ_{ij}, ε_i i.i.d. $\mathcal{N}(0, 1)$ and the number of

factors equal either $K = 0$ (variables independent) or $K = 5$. We choose the factor model, as it provides a non-trivial dependence structure between the covariates and it is relatively easy and quick to simulate. The R package **rbvs** provides a C-implemented routine **gen.factor.model.design** which quickly generates the factor model design matrix.

- The number of non-zero $\beta_j's$ is set to $s = 5, 10$, their indices are drawn uniformly without replacement from $\{1, \ldots, p\}$. Their values are drawn independently and have same distribution as $\beta = \left(|Z| + \frac{\log(n)}{\sqrt{n}}\right) V$, where Z is a standard normal random variable and V is independent of Z with $\mathbb{P}(V = 1) = \mathbb{P}(V = -1) = \frac{1}{2}$.
- The total number of variables $p = 100, 1000, 10000, 100000$.
- The sample size $n = 100, 200, \ldots, 1000$.
- The subsample size is set to $m = 50, 100, \frac{n}{2}$.

Due to a very huge number of variables, we take the marginal correlation as a base learner for both StabSel and (I)RBVS, as it is least computationally demanding across measures studied in the paper. All computations reported in this section are performed with the R package **rbvsGPU** (Baranowski, 2016), which provide a parallel implementation of RBVS PC and IRBVS PC, using to this end the CUDA framework (Luebke, 2008). The number of random splits is set to $B = 500\frac{m}{n}$, such that there always 500 subsamples, each of size m, used in computing the empirical probabilities.

Unlike the RBVS algorithm, StabSel requires specification of the two tuning parameters. From our experience, the values recommended in Meinshausen and Bühlmann (2010) are reasonably "optimal", we decided however to test robustness of the StabSel algorithm against the choice of its parameters. The bound on the error control is set to $EV = 2.5, 5$, while the thresholding probability $\pi = 0.55, 0.6, 0.75, 0.9$.

S3.2 High-dimensional simulation study results

We report results of this high-dimensional simulation study in Tables S2–S13.

$n\backslash p$	10^2	10^3	10^4	10^5
200	1.57	2.38	3.03	3.53
300	1.50	2.27	3.00	3.47
400	1.41	2.33	2.98	3.48
500	1.53	2.32	2.98	3.46
600	1.47	2.29	2.95	3.46
700	1.56	2.34	2.96	3.46
800	1.44	2.27	2.97	3.50
900	1.61	2.34	2.98	3.44
1000	1.48	2.31	2.98	3.45

(a) RBVS PC

$n\backslash p$	10^2	10^3	10^4	10^5
200	**.35**	**.19**	**.41**	**.96**
300	**.16**	**.10**	**.45**	**1.06**
400	**.04**	**.12**	**.49**	**.98**
500	**.03**	**.15**	**.56**	**1.02**
600	**.06**	**.21**	**.62**	**1.18**
700	**.05**	**.26**	**.66**	**1.17**
800	**.04**	**.25**	**.73**	**1.12**
900	**.05**	**.32**	**.72**	**1.31**
1000	**.05**	**.27**	**.74**	**1.28**

(b) IRBVS PC

$n\backslash p$	10^2	10^3	10^4	10^5
200	2.05	2.49	**2.93**	**3.40**
300	2.15	2.57	3.04	**3.46**
400	2.19	2.66	3.11	3.48
500	2.29	2.68	3.11	3.50
600	2.30	2.68	3.11	3.54
700	2.41	2.73	3.14	3.49
800	2.25	2.67	3.14	3.51
900	2.43	2.77	3.19	3.56
1000	2.30	2.70	3.09	3.47

(c) StabSel PC $\pi = 0.55$ $EV = 2.5$

$n\backslash p$	10^2	10^3	10^4	10^5
200	1.91	2.43	**2.94**	**3.41**
300	2.01	2.52	3.05	3.48
400	2.07	2.63	3.11	3.50
500	2.22	2.62	3.10	3.52
600	2.23	2.64	3.12	3.56
700	2.33	2.70	3.16	3.52
800	2.16	2.63	3.15	3.54
900	2.35	2.74	3.18	3.59
1000	2.22	2.67	3.11	3.49

(d) StabSel PC $\pi = 0.6$ $EV = 2.5$

$n\backslash p$	10^2	10^3	10^4	10^5
200	2.07	2.68	3.15	3.62
300	2.23	2.77	3.28	3.73
400	2.27	2.86	3.38	3.81
500	2.42	2.87	3.36	3.76
600	2.40	2.90	3.36	3.77
700	2.50	2.93	3.42	3.77
800	2.37	2.90	3.42	3.77
900	2.54	3.00	3.52	3.81
1000	2.42	2.91	3.34	3.73

(e) StabSel PC $\pi = 0.75$ $EV = 2.5$

$n\backslash p$	10^2	10^3	10^4	10^5
200	1.72	**2.27**	**2.80**	**3.28**
300	1.85	2.35	**2.87**	**3.36**
400	1.92	2.48	**2.97**	**3.38**
500	2.05	2.49	**2.96**	**3.40**
600	2.05	2.48	2.98	**3.40**
700	2.15	2.56	3.02	**3.41**
800	2.02	2.49	3.02	**3.41**
900	2.20	2.59	3.03	3.45
1000	2.09	2.54	**2.98**	**3.38**

(f) StabSel PC $\pi = 0.55$ $EV = 5$

$n\backslash p$	10^2	10^3	10^4	10^5
200	1.68	**2.21**	**2.79**	**3.28**
300	1.83	2.31	**2.86**	**3.37**
400	1.88	2.47	**2.97**	**3.39**
500	2.02	2.47	**2.96**	**3.41**
600	2.02	2.47	2.98	**3.42**
700	2.14	2.54	3.03	**3.42**
800	1.99	2.47	3.02	**3.43**
900	2.17	2.57	3.03	3.46
1000	2.06	2.51	**2.97**	**3.39**

(g) StabSel PC $\pi = 0.6$ $EV = 5$

$n\backslash p$	10^2	10^3	10^4	10^5
200	1.76	2.46	**3.01**	**3.50**
300	1.95	2.58	3.15	3.62
400	2.02	2.68	3.23	3.68
500	2.17	2.70	3.24	3.64
600	2.17	2.73	3.23	3.68
700	2.28	2.78	3.29	3.67
800	2.14	2.73	3.28	3.68
900	2.32	2.83	3.36	3.71
1000	2.19	2.74	3.22	3.63

(h) StabSel PC $\pi = 0.75$ $EV = 5$

Table S2: High-dimensional example: the average number of FP+FN (False Positives and False Negatives) calculated over 500 realisations with $m = 50$ and $B = 500\frac{m}{n}$, number of important variables $s = 5$ and number of factors $K = 0$. Bold: result **better** than the corresponding value for RBVS PC.

$n\backslash p$	10^2	10^3	10^4	10^5
200	1.53	1.82	2.19	2.79
300	1.04	1.40	1.87	2.60
400	.90	1.36	1.89	2.59
500	.85	1.31	1.86	2.55
600	.76	1.34	1.86	2.35
700	.83	1.33	1.90	2.32
800	.73	1.30	1.87	2.31
900	.76	1.32	1.88	2.39
1000	.68	1.30	1.85	2.39

(a) RBVS PC

$n\backslash p$	10^2	10^3	10^4	10^5
200	**1.49**	**.98**	**.66**	**.58**
300	**.60**	**.20**	**.11**	**.40**
400	**.32**	**.09**	**.10**	**.35**
500	**.18**	**.03**	**.09**	**.41**
600	**.12**	**.03**	**.09**	**.32**
700	**.06**	**.01**	**.11**	**.30**
800	**.02**	**.02**	**.15**	**.30**
900	**.01**	**.04**	**.16**	**.36**
1000	**.01**	**.04**	**.18**	**.41**

(b) IRBVS PC

$n\backslash p$	10^2	10^3	10^4	10^5
200	**1.24**	**1.59**	**2.10**	**2.34**
300	1.31	1.49	**1.84**	**2.21**
400	1.33	1.61	1.96	**2.33**
500	1.44	1.61	1.96	**2.28**
600	1.44	1.68	2.01	**2.34**
700	1.55	1.71	2.05	**2.31**
800	1.44	1.69	2.05	2.33
900	1.57	1.74	2.07	2.43
1000	1.50	1.72	2.06	2.41

(c) StabSel PC $\pi = 0.55$ $EV = 2.5$

$n\backslash p$	10^2	10^3	10^4	10^5
200	**1.15**	**1.64**	**2.12**	**2.56**
300	1.17	1.43	**1.82**	**2.48**
400	1.22	1.57	1.96	**2.56**
500	1.29	1.56	1.96	**2.54**
600	1.33	1.63	2.01	**2.35**
700	1.44	1.66	2.06	2.32
800	1.34	1.64	2.06	2.34
900	1.46	1.69	2.08	2.46
1000	1.40	1.70	2.07	2.42

(d) StabSel PC $\pi = 0.6$ $EV = 2.5$

$n\backslash p$	10^2	10^3	10^4	10^5
200	**1.19**	**1.61**	**2.06**	**2.63**
300	1.24	1.60	2.01	2.72
400	1.30	1.73	2.18	2.79
500	1.41	1.75	2.16	2.75
600	1.45	1.82	2.23	2.57
700	1.54	1.87	2.30	2.55
800	1.45	1.82	2.27	2.58
900	1.58	1.88	2.31	2.68
1000	1.51	1.87	2.25	2.63

(e) StabSel PC $\pi = 0.75$ $EV = 2.5$

$n\backslash p$	10^2	10^3	10^4	10^5
200	**1.13**	**1.74**	2.31	**2.51**
300	1.07	**1.33**	**1.72**	**2.14**
400	1.12	1.47	**1.84**	**2.24**
500	1.17	1.45	**1.83**	**2.19**
600	1.20	1.52	1.89	**2.23**
700	1.29	1.54	1.94	**2.23**
800	1.21	1.49	1.94	**2.23**
900	1.36	1.59	1.96	**2.36**
1000	1.26	1.57	1.96	**2.30**

(f) StabSel PC $\pi = 0.55$ $EV = 5$

$n\backslash p$	10^2	10^3	10^4	10^5
200	**1.17**	1.92	2.40	**2.54**
300	**1.02**	**1.29**	**1.72**	**2.13**
400	1.07	1.43	**1.83**	**2.23**
500	1.12	1.42	**1.83**	**2.19**
600	1.15	1.50	1.88	**2.24**
700	1.25	1.53	1.94	**2.23**
800	1.18	1.48	1.95	**2.24**
900	1.33	1.56	1.96	**2.36**
1000	1.23	1.55	1.96	**2.31**

(g) StabSel PC $\pi = 0.6$ $EV = 5$

$n\backslash p$	10^2	10^3	10^4	10^5
200	**1.21**	**1.69**	**2.10**	**2.31**
300	1.05	1.43	1.88	**2.30**
400	1.12	1.59	2.03	**2.43**
500	1.20	1.60	2.02	**2.38**
600	1.23	1.67	2.11	2.48
700	1.34	1.72	2.16	2.45
800	1.27	1.68	2.16	2.47
900	1.41	1.74	2.19	2.58
1000	1.32	1.74	2.15	2.55

(h) StabSel PC $\pi = 0.75$ $EV = 5$

Table S3: High-dimensional example: the average number of FP+FN (False Positives and False Negatives) calculated over 500 realisations with $m = 100$ and $B = 500\frac{m}{n}$, number of important variables $s = 5$ and number of factors $K = 0$. Bold: result **better** than the corresponding value for RBVS PC.

$n\backslash p$	10^2	10^3	10^4	10^5
200	1.57	1.79	2.22	2.59
300	1.18	1.31	1.64	1.98
400	1.07	1.10	1.33	1.61
500	.95	1.00	1.13	1.40
600	.94	.88	1.03	1.15
700	.96	.77	.90	1.00
800	.85	.77	.84	.92
900	.73	.67	.74	.87
1000	.80	.62	.78	.84

(a) RBVS PC

$n\backslash p$	10^2	10^3	10^4	10^5
200	**1.54**	**.90**	**.64**	**.44**
300	1.35	**.80**	**.58**	**.33**
400	1.23	**.86**	**.53**	**.25**
500	1.26	**.87**	**.55**	**.27**
600	1.39	**.80**	**.51**	**.24**
700	1.32	.78	**.46**	**.23**
800	1.28	.82	**.41**	**.24**
900	1.19	.76	**.43**	**.22**
1000	1.21	.75	**.45**	**.29**

(b) IRBVS PC

$n\backslash p$	10^2	10^3	10^4	10^5
200	**1.23**	**1.58**	**2.10**	**2.35**
300	**.88**	**1.18**	**1.54**	**1.81**
400	**.75**	**.96**	**1.31**	**1.56**
500	**.62**	**.83**	1.18	**1.35**
600	**.49**	**.76**	1.08	1.19
700	**.47**	**.62**	.96	1.12
800	**.41**	**.60**	**.82**	1.02
900	**.35**	**.51**	.75	.96
1000	**.32**	**.44**	.82	.92

(c) StabSel PC $\pi = 0.55$ $EV = 2.5$

$n\backslash p$	10^2	10^3	10^4	10^5
200	**1.16**	**1.65**	**2.12**	**2.38**
300	**.81**	**1.22**	**1.57**	**1.87**
400	**.67**	**1.01**	1.33	1.62
500	**.58**	**.91**	1.20	1.41
600	**.48**	**.84**	1.13	1.22
700	**.47**	**.68**	.98	1.13
800	**.43**	**.68**	.88	1.05
900	**.34**	**.59**	.77	.99
1000	**.33**	**.52**	.86	.94

(d) StabSel PC $\pi = 0.6$ $EV = 2.5$

$n\backslash p$	10^2	10^3	10^4	10^5
200	**1.18**	**1.59**	**2.03**	**2.43**
300	**.83**	**1.19**	**1.47**	**1.90**
400	**.68**	**.94**	**1.26**	1.69
500	**.59**	**.87**	**1.09**	1.41
600	**.49**	**.78**	**.98**	**1.11**
700	**.49**	**.64**	**.86**	**1.00**
800	**.45**	**.61**	**.77**	**.89**
900	**.37**	**.51**	**.70**	**.85**
1000	**.35**	**.47**	**.74**	**.82**

(e) StabSel PC $\pi = 0.75$ $EV = 2.5$

$n\backslash p$	10^2	10^3	10^4	10^5
200	**1.15**	**1.77**	2.29	**2.52**
300	**.80**	1.33	1.78	2.06
400	**.66**	1.14	1.52	1.77
500	**.59**	1.05	1.40	1.61
600	**.50**	.97	1.34	1.48
700	**.49**	.82	1.17	1.36
800	**.45**	.82	1.13	1.29
900	**.36**	.71	1.00	1.22
1000	**.34**	.70	1.07	1.21

(f) StabSel PC $\pi = 0.55$ $EV = 5$

$n\backslash p$	10^2	10^3	10^4	10^5
200	**1.15**	1.91	2.41	**2.56**
300	**.80**	1.48	1.86	2.10
400	**.72**	1.33	1.62	1.81
500	**.66**	1.22	1.48	1.68
600	**.56**	1.13	1.46	1.52
700	**.55**	.94	1.28	1.42
800	**.51**	1.01	1.26	1.35
900	**.43**	.91	1.12	1.27
1000	**.42**	.89	1.16	1.29

(g) StabSel PC $\pi = 0.6$ $EV = 5$

$n\backslash p$	10^2	10^3	10^4	10^5
200	**1.20**	**1.72**	**2.12**	**2.33**
300	**.85**	**1.29**	**1.59**	**1.77**
400	**.76**	1.10	1.33	**1.54**
500	**.70**	**.99**	1.19	**1.30**
600	**.63**	.94	1.11	1.17
700	**.63**	**.77**	.99	1.07
800	**.55**	**.76**	.85	.98
900	**.49**	**.66**	.78	.92
1000	**.50**	.65	.86	.87

(h) StabSel PC $\pi = 0.75$ $EV = 5$

Table S4: High-dimensional example: the average number of FP+FN (False Positives and False Negatives) calculated over 500 realisations with $m = \frac{n}{2}$ and $B = 500\frac{m}{n}$, number of important variables $s = 5$ and number of factors $K = 0$. Bold: result **better** than the corresponding value for RBVS PC.

$n\backslash p$	10^2	10^3	10^4	10^5
200	1.77	2.45	3.20	3.70
300	1.66	2.44	3.17	3.68
400	1.62	2.38	3.18	3.66
500	1.63	2.39	3.15	3.63
600	1.50	2.31	3.16	3.61
700	1.61	2.38	3.12	3.72
800	1.54	2.35	3.15	3.67
900	1.54	2.37	3.09	3.81
1000	1.56	2.33	3.10	3.79

(a) RBVS PC

$n\backslash p$	10^2	10^3	10^4	10^5
200	**.28**	**.11**	**.09**	**.48**
300	**.12**	**.03**	**.04**	**.25**
400	**.04**	**.00**	**.05**	**.21**
500	**.02**	**.01**	**.03**	**.15**
600	**.01**	**.00**	**.03**	**.13**
700	**.00**	**.01**	**.04**	**.19**
800	**.00**	**.00**	**.05**	**.17**
900	**.00**	**.00**	**.01**	**.29**
1000	**.00**	**.00**	**.04**	**.15**

(b) IRBVS PC

$n\backslash p$	10^2	10^3	10^4	10^5
200	2.21	2.62	**3.10**	**3.53**
300	2.29	2.66	3.18	**3.63**
400	2.34	2.74	3.20	**3.62**
500	2.39	2.71	3.21	**3.57**
600	2.37	2.75	3.27	**3.57**
700	2.43	2.83	3.26	**3.67**
800	2.37	2.84	3.31	**3.67**
900	2.41	2.87	3.31	**3.78**
1000	2.40	2.73	3.20	**3.72**

(c) StabSel PC $\pi = 0.55$ $EV = 2.5$

$n\backslash p$	10^2	10^3	10^4	10^5
200	2.09	2.57	**3.09**	**3.54**
300	2.20	2.62	3.19	**3.65**
400	2.23	2.71	3.23	**3.64**
500	2.29	2.69	3.21	**3.59**
600	2.28	2.70	3.29	**3.58**
700	2.35	2.80	3.26	**3.68**
800	2.28	2.80	3.32	3.69
900	2.30	2.84	3.32	3.82
1000	2.35	2.71	3.20	**3.74**

(d) StabSel PC $\pi = 0.6$ $EV = 2.5$

$n\backslash p$	10^2	10^3	10^4	10^5
200	2.24	2.77	3.33	3.76
300	2.40	2.92	3.43	3.89
400	2.47	2.99	3.46	3.91
500	2.47	2.91	3.46	3.84
600	2.50	3.01	3.59	3.86
700	2.58	3.04	3.49	3.94
800	2.53	3.05	3.56	3.92
900	2.56	3.11	3.60	4.04
1000	2.52	2.98	3.51	3.95

(e) StabSel PC $\pi = 0.75$ $EV = 2.5$

$n\backslash p$	10^2	10^3	10^4	10^5
200	1.94	**2.43**	**2.98**	**3.42**
300	2.05	2.47	**3.07**	**3.52**
400	2.06	2.55	**3.11**	**3.51**
500	2.11	2.55	**3.09**	**3.47**
600	2.11	2.55	**3.15**	**3.48**
700	2.18	2.65	3.15	**3.56**
800	2.10	2.62	3.20	**3.58**
900	2.13	2.68	3.17	**3.68**
1000	2.20	2.56	**3.07**	**3.62**

(f) StabSel PC $\pi = 0.55$ $EV = 5$

$n\backslash p$	10^2	10^3	10^4	10^5
200	1.88	**2.39**	**2.97**	**3.42**
300	2.01	**2.43**	**3.06**	**3.52**
400	2.03	2.53	**3.10**	**3.52**
500	2.09	2.53	**3.07**	**3.48**
600	2.08	2.53	3.16	**3.49**
700	2.16	2.64	3.15	**3.57**
800	2.08	2.61	3.20	**3.58**
900	2.11	2.66	3.18	**3.70**
1000	2.18	2.53	**3.07**	**3.62**

(g) StabSel PC $\pi = 0.6$ $EV = 5$

$n\backslash p$	10^2	10^3	10^4	10^5
200	2.00	2.60	**3.18**	**3.66**
300	2.14	2.72	3.30	3.78
400	2.19	2.79	3.34	3.81
500	2.23	2.76	3.34	3.71
600	2.26	2.82	3.44	3.76
700	2.32	2.88	3.38	3.84
800	2.26	2.91	3.42	3.82
900	2.28	2.95	3.47	3.95
1000	2.32	2.80	3.34	3.84

(h) StabSel PC $\pi = 0.75$ $EV = 5$

Table S5: High-dimensional example: the average number of FP+FN (False Positives and False Negatives) calculated over 500 realisations with $m = 50$ and $B = 500\frac{m}{n}$, number of important variables $s = 5$ and number of factors $K = 5$. Bold: result **better** than the corresponding value for RBVS PC.

$n\backslash p$	10^2	10^3	10^4	10^5
200	1.58	1.90	2.39	2.82
300	1.21	1.48	2.15	2.58
400	.97	1.48	2.03	2.52
500	.88	1.39	2.01	2.48
600	.90	1.30	2.01	2.50
700	.83	1.41	2.04	2.49
800	.83	1.42	1.97	2.54
900	.76	1.42	1.98	2.59
1000	.77	1.36	2.01	2.63

(a) RBVS PC

$n\backslash p$	10^2	10^3	10^4	10^5
200	**1.51**	**.88**	**.59**	**.31**
300	**.65**	**.23**	**.08**	**.01**
400	**.30**	**.05**	**.01**	**.00**
500	**.16**	**.02**	**.01**	**.00**
600	**.10**	**.00**	**.00**	**.00**
700	**.05**	**.00**	**.00**	**.00**
800	**.03**	**.00**	**.00**	**.00**
900	**.01**	**.00**	**.00**	**.00**
1000	**.02**	**.00**	**.00**	**.01**

(b) IRBVS PC

$n\backslash p$	10^2	10^3	10^4	10^5
200	**1.33**	**1.72**	**2.17**	**2.56**
300	1.45	1.61	**2.07**	**2.37**
400	1.44	1.70	2.05	**2.38**
500	1.50	1.70	2.09	**2.41**
600	1.53	1.69	2.10	2.51
700	1.54	1.78	2.20	**2.43**
800	1.54	1.83	2.15	**2.54**
900	1.55	1.86	2.15	2.61
1000	1.60	1.80	2.19	**2.62**

(c) StabSel PC $\pi = 0.55$ $EV = 2.5$

$n\backslash p$	10^2	10^3	10^4	10^5
200	**1.21**	**1.73**	**2.21**	**2.56**
300	1.33	1.56	**2.06**	**2.37**
400	1.30	1.65	2.04	**2.39**
500	1.36	1.65	2.09	**2.41**
600	1.41	1.66	2.10	2.52
700	1.46	1.75	2.20	**2.44**
800	1.43	1.82	2.15	2.55
900	1.43	1.83	2.17	2.62
1000	1.48	1.77	2.20	2.63

(d) StabSel PC $\pi = 0.6$ $EV = 2.5$

$n\backslash p$	10^2	10^3	10^4	10^5
200	**1.24**	**1.72**	**2.18**	**2.56**
300	1.40	1.71	2.25	2.59
400	1.42	1.82	2.28	2.62
500	1.48	1.81	2.30	2.64
600	1.54	1.87	2.31	2.75
700	1.58	1.96	2.41	2.67
800	1.57	1.97	2.40	2.77
900	1.57	2.03	2.41	2.88
1000	1.64	1.98	2.41	2.85

(e) StabSel PC $\pi = 0.75$ $EV = 2.5$

$n\backslash p$	10^2	10^3	10^4	10^5
200	**1.21**	**1.86**	2.41	**2.74**
300	1.21	**1.46**	**1.97**	**2.25**
400	1.21	1.52	**1.95**	**2.30**
500	1.24	1.53	**1.97**	**2.32**
600	1.29	1.53	**2.00**	**2.41**
700	1.31	1.65	2.06	**2.33**
800	1.27	1.69	2.02	**2.43**
900	1.32	1.70	2.05	**2.52**
1000	1.36	1.63	2.09	**2.54**

(f) StabSel PC $\pi = 0.55$ $EV = 5$

$n\backslash p$	10^2	10^3	10^4	10^5
200	**1.23**	1.96	2.49	**2.75**
300	**1.16**	**1.41**	**1.96**	**2.24**
400	1.17	1.50	**1.95**	**2.30**
500	1.21	1.50	**1.97**	**2.33**
600	1.24	1.49	**2.00**	**2.42**
700	1.25	1.64	2.06	**2.34**
800	1.24	1.65	2.01	**2.44**
900	1.28	1.68	2.05	**2.53**
1000	1.32	1.61	2.09	**2.54**

(g) StabSel PC $\pi = 0.6$ $EV = 5$

$n\backslash p$	10^2	10^3	10^4	10^5
200	**1.27**	**1.74**	**2.24**	**2.53**
300	1.21	1.57	**2.11**	**2.49**
400	1.21	1.67	2.15	**2.51**
500	1.27	1.68	2.19	2.54
600	1.35	1.70	2.19	2.65
700	1.36	1.79	2.31	2.57
800	1.36	1.85	2.26	2.67
900	1.36	1.89	2.29	2.74
1000	1.41	1.82	2.29	2.75

(h) StabSel PC $\pi = 0.75$ $EV = 5$

Table S6: High-dimensional example: the average number of FP+FN (False Positives and False Negatives) calculated over 500 realisations with $m = 100$ and $B = 500\frac{m}{n}$, number of important variables $s = 5$ and number of factors $K = 5$. Bold: result **better** than the corresponding value for RBVS PC.

$n\backslash p$	10^2	10^3	10^4	10^5
200	1.59	1.88	2.38	2.79
300	1.37	1.41	1.83	2.12
400	1.10	1.17	1.45	1.70
500	.92	1.08	1.24	1.48
600	.91	.89	1.13	1.29
700	.82	.87	1.01	1.14
800	.80	.84	.88	1.10
900	.83	.75	.80	.93
1000	.72	.73	.86	.91

(a) RBVS PC

$n\backslash p$	10^2	10^3	10^4	10^5
200	**1.35**	**.85**	**.56**	**.30**
300	1.45	**.78**	**.58**	**.29**
400	1.44	**.78**	**.48**	**.24**
500	1.23	**.84**	**.52**	**.29**
600	1.29	**.81**	**.51**	**.24**
700	1.17	**.80**	**.50**	**.20**
800	1.23	**.83**	**.48**	**.25**
900	1.34	.82	**.46**	**.21**
1000	1.19	.79	**.51**	**.19**

(b) IRBVS PC

$n\backslash p$	10^2	10^3	10^4	10^5
200	**1.34**	**1.75**	**2.19**	**2.59**
300	**1.07**	**1.26**	**1.65**	**2.06**
400	**.81**	**1.05**	**1.37**	**1.69**
500	**.63**	**.94**	**1.23**	1.48
600	**.58**	**.74**	1.14	1.34
700	**.48**	**.75**	1.09	1.19
800	**.43**	**.67**	.89	1.17
900	**.42**	**.60**	.82	1.01
1000	**.37**	**.56**	.87	.99

(c) StabSel PC $\pi = 0.55$ $EV = 2.5$

$n\backslash p$	10^2	10^3	10^4	10^5
200	**1.23**	**1.75**	**2.22**	**2.59**
300	**1.00**	**1.29**	**1.71**	**2.08**
400	**.74**	**1.12**	**1.42**	**1.69**
500	**.59**	**.99**	1.27	1.50
600	**.55**	**.81**	1.17	1.37
700	**.46**	**.82**	1.14	1.23
800	**.39**	**.75**	.93	1.21
900	**.38**	**.65**	.86	1.02
1000	**.35**	**.63**	.90	1.01

(d) StabSel PC $\pi = 0.6$ $EV = 2.5$

$n\backslash p$	10^2	10^3	10^4	10^5
200	**1.25**	**1.73**	**2.19**	**2.56**
300	**1.01**	**1.26**	**1.68**	**1.95**
400	**.77**	**1.05**	**1.32**	**1.60**
500	**.62**	**.93**	**1.16**	**1.40**
600	**.58**	**.74**	**1.03**	**1.18**
700	**.47**	**.75**	**.99**	**1.10**
800	**.41**	**.68**	**.80**	**1.03**
900	**.40**	**.59**	**.76**	**.93**
1000	**.38**	**.58**	**.81**	**.85**

(e) StabSel PC $\pi = 0.75$ $EV = 2.5$

$n\backslash p$	10^2	10^3	10^4	10^5
200	**1.20**	**1.84**	2.40	**2.74**
300	**.97**	**1.35**	1.86	2.29
400	**.74**	1.22	1.63	1.85
500	**.59**	1.09	1.46	1.74
600	**.58**	.92	1.34	1.58
700	**.45**	.92	1.33	1.44
800	**.41**	.86	1.10	1.43
900	**.39**	.77	1.07	1.25
1000	**.37**	.76	1.07	1.26

(f) StabSel PC $\pi = 0.55$ $EV = 5$

$n\backslash p$	10^2	10^3	10^4	10^5
200	**1.23**	1.93	2.48	**2.78**
300	**1.01**	1.48	1.96	2.33
400	**.79**	1.32	1.72	1.92
500	**.65**	1.21	1.56	1.80
600	**.60**	1.05	1.43	1.65
700	**.50**	1.10	1.42	1.49
800	**.47**	.99	1.21	1.49
900	**.45**	.92	1.18	1.29
1000	**.42**	.93	1.17	1.31

(g) StabSel PC $\pi = 0.6$ $EV = 5$

$n\backslash p$	10^2	10^3	10^4	10^5
200	**1.27**	**1.79**	**2.23**	**2.56**
300	**1.02**	**1.32**	**1.69**	**2.03**
400	**.82**	**1.17**	**1.40**	**1.66**
500	**.69**	**1.05**	1.25	**1.44**
600	**.64**	**.88**	1.15	1.30
700	**.55**	**.87**	1.14	**1.13**
800	**.52**	**.80**	.91	1.13
900	**.51**	**.72**	.84	.96
1000	**.49**	**.70**	.90	.94

(h) StabSel PC $\pi = 0.75$ $EV = 5$

Table S7: High-dimensional example: the average number of FP+FN (False Positives and False Negatives) calculated over 500 realisations with $m = \frac{n}{2}$ and $B = 500\frac{m}{n}$, number of important variables $s = 5$ and number of factors $K = 5$. Bold: result **better** than the corresponding value for RBVS PC.

$n\backslash p$	10^2	10^3	10^4	10^5
200	6.46	7.50	8.32	8.93
300	6.27	7.48	8.33	8.88
400	6.39	7.44	8.31	8.81
500	6.31	7.38	8.18	8.82
600	6.35	7.41	8.31	8.85
700	6.29	7.47	8.22	8.85
800	6.34	7.43	8.17	8.82
900	6.41	7.46	8.24	8.87
1000	6.30	7.44	8.25	8.81

(a) RBVS PC

$n\backslash p$	10^2	10^3	10^4	10^5
200	**1.82**	**1.52**	**3.01**	**6.38**
300	**1.41**	**1.51**	**2.94**	**5.97**
400	**1.49**	**1.59**	**3.08**	**5.61**
500	**1.20**	**1.54**	**2.87**	**5.37**
600	**1.33**	**1.67**	**3.33**	**5.69**
700	**1.57**	**2.02**	**3.05**	**5.83**
800	**1.46**	**1.76**	**3.08**	**5.35**
900	**1.66**	**2.17**	**3.52**	**6.08**
1000	**1.29**	**1.91**	**3.04**	**5.36**

(b) IRBVS PC

$n\backslash p$	10^2	10^3	10^4	10^5
200	6.97	**7.49**	**8.17**	**8.82**
300	6.96	7.60	8.35	8.92
400	7.14	7.69	8.32	8.84
500	6.98	7.57	8.21	8.83
600	7.06	7.72	8.37	8.92
700	7.18	7.73	8.39	8.89
800	7.15	7.75	8.35	8.85
900	7.23	7.73	8.41	9.02
1000	7.12	7.64	8.33	8.89

(c) StabSel PC $\pi = 0.55$ $EV = 2.5$

$n\backslash p$	10^2	10^3	10^4	10^5
200	6.67	**7.39**	**8.17**	**8.84**
300	6.71	7.51	8.34	8.94
400	6.90	7.59	8.33	8.86
500	6.74	7.50	8.23	8.83
600	6.87	7.64	8.40	8.94
700	6.97	7.67	8.41	8.93
800	6.94	7.68	8.37	8.88
900	7.01	7.67	8.43	9.05
1000	6.89	7.58	8.33	8.91

(d) StabSel PC $\pi = 0.6$ $EV = 2.5$

$n\backslash p$	10^2	10^3	10^4	10^5
200	6.92	7.81	8.54	9.11
300	6.99	7.96	8.71	9.24
400	7.25	8.10	8.76	9.17
500	7.09	7.96	8.61	9.14
600	7.21	8.08	8.86	9.25
700	7.34	8.13	8.79	9.25
800	7.26	8.11	8.74	9.20
900	7.39	8.20	8.78	9.37
1000	7.25	8.02	8.71	9.21

(e) StabSel PC $\pi = 0.75$ $EV = 2.5$

$n\backslash p$	10^2	10^3	10^4	10^5
200	**6.29**	**7.08**	**7.95**	**8.70**
300	6.39	**7.15**	**8.12**	**8.74**
400	6.53	**7.33**	**8.11**	**8.71**
500	6.40	**7.17**	**7.98**	**8.66**
600	6.53	**7.36**	**8.19**	**8.77**
700	6.56	**7.34**	**8.18**	**8.75**
800	6.61	**7.38**	**8.12**	**8.71**
900	6.68	**7.40**	**8.22**	**8.87**
1000	6.60	**7.31**	**8.12**	**8.75**

(f) StabSel PC $\pi = 0.55$ $EV = 5$

$n\backslash p$	10^2	10^3	10^4	10^5
200	**6.13**	**7.00**	**7.92**	**8.70**
300	6.27	**7.08**	**8.09**	**8.76**
400	6.42	**7.27**	**8.11**	**8.72**
500	**6.30**	**7.11**	**7.98**	**8.67**
600	6.45	**7.33**	**8.19**	**8.80**
700	6.46	**7.30**	**8.18**	**8.77**
800	6.51	**7.32**	**8.12**	**8.75**
900	6.57	**7.34**	**8.22**	8.90
1000	6.51	**7.26**	**8.11**	**8.76**

(g) StabSel PC $\pi = 0.6$ $EV = 5$

$n\backslash p$	10^2	10^3	10^4	10^5
200	**6.26**	**7.45**	**8.29**	8.98
300	6.46	7.62	8.53	9.08
400	6.70	7.73	8.54	9.05
500	6.50	7.63	8.42	9.02
600	6.69	7.77	8.61	9.13
700	6.73	7.84	8.62	9.13
800	6.80	7.81	8.53	9.10
900	6.87	7.84	8.61	9.24
1000	6.74	7.71	8.52	9.10

(h) StabSel PC $\pi = 0.75$ $EV = 5$

Table S8: High-dimensional example: the average number of FP+FN (False Positives and False Negatives) calculated over 500 realisations with $m = 50$ and $B = 500\frac{m}{n}$, number of important variables $s = 10$ and number of factors $K = 0$. Bold: result **better** than the corresponding value for RBVS PC.

$n\backslash p$	10^2	10^3	10^4	10^5
200	4.69	5.79	6.75	7.61
300	4.21	5.42	6.53	7.38
400	3.97	5.31	6.37	7.31
500	3.77	5.30	6.40	7.22
600	3.85	5.36	6.37	7.24
700	3.95	5.35	6.42	7.24
800	4.01	5.31	6.40	7.24
900	4.01	5.37	6.41	7.24
1000	3.98	5.21	6.44	7.25

(a) RBVS PC

$n\backslash p$	10^2	10^3	10^4	10^5
200	**2.09**	**1.22**	**.93**	**1.38**
300	**.90**	**.46**	**.40**	**.70**
400	**.51**	**.17**	**.25**	**.75**
500	**.32**	**.07**	**.38**	**.84**
600	**.16**	**.15**	**.36**	**.86**
700	**.24**	**.18**	**.39**	**1.13**
800	**.11**	**.15**	**.47**	**1.09**
900	**.12**	**.16**	**.63**	**1.10**
1000	**.08**	**.13**	**.55**	**1.18**

(b) IRBVS PC

$n\backslash p$	10^2	10^3	10^4	10^5
200	5.29	**5.31**	**6.05**	**6.91**
300	5.43	**5.32**	**6.05**	**6.79**
400	5.58	5.42	**6.02**	**6.88**
500	5.49	5.47	**6.11**	**6.85**
600	5.62	5.64	**6.15**	**6.79**
700	5.60	5.62	**6.20**	**6.89**
800	5.68	5.62	**6.35**	**6.89**
900	5.69	5.66	**6.34**	**6.98**
1000	5.65	5.60	**6.35**	**6.94**

(c) StabSel PC $\pi = 0.55$ $EV = 2.5$

$n\backslash p$	10^2	10^3	10^4	10^5
200	4.73	**5.20**	**6.06**	**6.92**
300	4.93	**5.13**	**6.03**	**6.80**
400	5.02	**5.29**	**6.03**	**6.89**
500	4.94	5.31	**6.11**	**6.87**
600	5.14	5.52	**6.14**	**6.82**
700	5.17	5.50	**6.22**	**6.90**
800	5.25	5.48	**6.37**	**6.92**
900	5.28	5.54	**6.35**	**7.02**
1000	5.23	5.49	**6.36**	**6.96**

(d) StabSel PC $\pi = 0.6$ $EV = 2.5$

$n\backslash p$	10^2	10^3	10^4	10^5
200	**4.58**	**5.34**	**6.21**	**7.01**
300	4.84	5.54	**6.43**	**7.25**
400	5.04	5.65	6.41	7.32
500	5.00	5.74	6.54	7.26
600	5.21	5.93	6.60	7.34
700	5.24	5.90	6.73	7.34
800	5.34	5.95	6.85	7.35
900	5.39	5.98	6.79	7.49
1000	5.33	5.89	6.78	7.37

(e) StabSel PC $\pi = 0.75$ $EV = 2.5$

$n\backslash p$	10^2	10^3	10^4	10^5
200	**4.48**	**5.08**	**6.06**	**6.98**
300	4.60	**4.91**	**5.77**	**6.59**
400	4.71	**5.02**	**5.80**	**6.68**
500	4.64	**5.03**	**5.87**	**6.62**
600	4.84	**5.24**	**5.87**	**6.58**
700	4.91	**5.21**	**5.93**	**6.69**
800	4.95	**5.21**	**6.10**	**6.68**
900	5.02	**5.25**	**6.12**	**6.81**
1000	4.96	**5.20**	**6.08**	**6.74**

(f) StabSel PC $\pi = 0.55$ $EV = 5$

$n\backslash p$	10^2	10^3	10^4	10^5
200	**4.22**	**5.07**	**6.11**	**6.97**
300	4.35	**4.82**	**5.74**	**6.58**
400	4.41	**4.91**	**5.78**	**6.68**
500	4.39	**4.93**	**5.82**	**6.64**
600	4.60	**5.13**	**5.85**	**6.60**
700	4.70	**5.14**	**5.93**	**6.72**
800	4.71	**5.14**	**6.08**	**6.72**
900	4.83	**5.19**	**6.12**	**6.85**
1000	4.74	**5.14**	**6.08**	**6.75**

(g) StabSel PC $\pi = 0.6$ $EV = 5$

$n\backslash p$	10^2	10^3	10^4	10^5
200	**4.06**	**5.14**	**6.07**	**6.96**
300	4.22	**5.12**	**6.16**	**7.02**
400	4.38	5.31	**6.17**	**7.10**
500	4.37	5.34	**6.29**	**7.07**
600	4.60	5.60	**6.33**	**7.12**
700	4.71	5.57	6.42	**7.13**
800	4.78	5.57	6.61	**7.17**
900	4.85	5.62	6.55	7.29
1000	4.79	5.57	6.57	**7.20**

(h) StabSel PC $\pi = 0.75$ $EV = 5$

Table S9: High-dimensional example: the average number of FP+FN (False Positives and False Negatives) calculated over 500 realisations with $m = 100$ and $B = 500\frac{m}{n}$, number of important variables $s = 10$ and number of factors $K = 0$. Bold: result **better** than the corresponding value for RBVS PC.

$n\backslash p$	10^2	10^3	10^4	10^5
200	4.66	5.84	6.78	7.54
300	3.52	4.52	5.46	6.29
400	2.80	3.62	4.51	5.46
500	2.38	3.19	3.98	4.72
600	2.20	2.77	3.44	4.17
700	2.13	2.58	3.21	3.81
800	1.99	2.35	3.04	3.53
900	1.81	2.14	2.81	3.27
1000	1.70	1.96	2.51	3.05

(a) RBVS PC

$n\backslash p$	10^2	10^3	10^4	10^5
200	**1.96**	**1.34**	**1.09**	**1.25**
300	**1.62**	**1.01**	**.63**	**.46**
400	**1.61**	**.97**	**.54**	**.35**
500	**1.65**	**.95**	**.51**	**.31**
600	**1.58**	**.90**	**.58**	**.26**
700	**1.61**	**.89**	**.51**	**.26**
800	**1.36**	**.90**	**.48**	**.26**
900	**1.44**	**.80**	**.54**	**.22**
1000	**1.39**	**.90**	**.56**	**.30**

(b) IRBVS PC

$n\backslash p$	10^2	10^3	10^4	10^5
200	5.28	**5.31**	**6.05**	**6.93**
300	4.82	**4.22**	**4.94**	**5.63**
400	4.51	**3.46**	**4.06**	**4.88**
500	4.35	**3.02**	**3.64**	**4.29**
600	4.33	**2.70**	**3.15**	**3.82**
700	4.29	**2.48**	**2.90**	**3.54**
800	4.26	**2.29**	**2.86**	**3.26**
900	4.23	**2.09**	**2.60**	**3.02**
1000	4.21	**1.87**	**2.29**	**2.91**

(c) StabSel PC $\pi = 0.55$ $EV = 2.5$

$n\backslash p$	10^2	10^3	10^4	10^5
200	4.70	**5.19**	**6.04**	**6.93**
300	4.00	**4.09**	**4.95**	**5.64**
400	3.52	**3.34**	**4.05**	**4.88**
500	3.11	**2.87**	**3.64**	**4.32**
600	2.96	**2.62**	**3.17**	**3.82**
700	2.87	**2.40**	**2.93**	**3.56**
800	2.76	**2.20**	**2.87**	**3.25**
900	2.61	**2.01**	**2.63**	**3.04**
1000	2.68	**1.76**	**2.30**	**2.93**

(d) StabSel PC $\pi = 0.6$ $EV = 2.5$

$n\backslash p$	10^2	10^3	10^4	10^5
200	**4.51**	**5.31**	**6.24**	**7.01**
300	3.64	**4.23**	**5.04**	**5.80**
400	2.99	**3.44**	**4.15**	**4.99**
500	2.54	**2.95**	**3.71**	**4.33**
600	2.36	**2.69**	**3.19**	**3.88**
700	**2.11**	**2.43**	**2.98**	**3.53**
800	2.04	**2.27**	**2.94**	**3.29**
900	**1.80**	**2.06**	**2.63**	**3.04**
1000	**1.69**	**1.83**	**2.36**	**2.88**

(e) StabSel PC $\pi = 0.75$ $EV = 2.5$

$n\backslash p$	10^2	10^3	10^4	10^5
200	**4.47**	**5.09**	**6.06**	**6.97**
300	3.78	**4.02**	**4.95**	**5.77**
400	3.29	**3.27**	**4.14**	**4.95**
500	2.96	**2.86**	**3.70**	**4.49**
600	2.83	**2.62**	**3.22**	**3.94**
700	2.72	**2.41**	**3.00**	**3.67**
800	2.63	**2.20**	**2.97**	**3.43**
900	2.48	**2.02**	**2.72**	**3.22**
1000	2.51	**1.75**	**2.43**	3.10

(f) StabSel PC $\pi = 0.55$ $EV = 5$

$n\backslash p$	10^2	10^3	10^4	10^5
200	**4.18**	**5.05**	**6.12**	**7.02**
300	**3.39**	**4.04**	**5.00**	**5.82**
400	2.81	**3.30**	**4.18**	**4.98**
500	2.41	**2.88**	**3.77**	**4.53**
600	2.25	**2.63**	**3.30**	**3.98**
700	**2.02**	**2.43**	**3.08**	**3.75**
800	**1.98**	**2.26**	**3.02**	**3.46**
900	**1.79**	**2.06**	**2.78**	**3.26**
1000	**1.67**	**1.84**	**2.48**	3.18

(g) StabSel PC $\pi = 0.6$ $EV = 5$

$n\backslash p$	10^2	10^3	10^4	10^5
200	**4.01**	**5.13**	**6.05**	**6.92**
300	**3.16**	**4.03**	**4.96**	**5.66**
400	**2.55**	**3.27**	**4.08**	**4.92**
500	**2.10**	**2.84**	**3.63**	**4.29**
600	**2.00**	**2.66**	**3.18**	**3.83**
700	**1.74**	**2.42**	**2.97**	**3.54**
800	**1.69**	**2.22**	**2.89**	**3.26**
900	**1.49**	**2.03**	**2.62**	**3.02**
1000	**1.37**	**1.78**	**2.32**	**2.88**

(h) StabSel PC $\pi = 0.75$ $EV = 5$

Table S10: High-dimensional example: the average number of FP+FN (False Positives and False Negatives) calculated over 500 realisations with $m = \frac{n}{2}$ and $B = 500\frac{m}{n}$, number of important variables $s = 10$ and number of factors $K = 0$. Bold: result **better** than the corresponding value for RBVS PC.

$n\backslash p$	10^2	10^3	10^4	10^5
200	7.23	8.05	8.77	9.35
300	7.04	8.02	8.74	9.27
400	7.02	7.90	8.68	9.26
500	6.83	7.88	8.62	9.21
600	6.96	7.94	8.69	9.18
700	7.14	7.90	8.63	9.11
800	6.96	7.97	8.63	9.13
900	7.01	7.94	8.68	9.22
1000	6.94	7.90	8.55	9.08

(a) RBVS PC

$n\backslash p$	10^2	10^3	10^4	10^5
200	**2.49**	**1.94**	**4.04**	**8.38**
300	**1.55**	**1.73**	**3.26**	**7.48**
400	**1.93**	**1.52**	**2.76**	**7.14**
500	**1.19**	**1.01**	**2.51**	**6.21**
600	**1.68**	**1.38**	**2.94**	**6.12**
700	**2.09**	**1.45**	**2.65**	**5.50**
800	**1.57**	**1.38**	**2.41**	**5.19**
900	**1.44**	**1.44**	**2.84**	**6.09**
1000	**1.66**	**1.41**	**2.13**	**4.89**

(b) IRBVS PC

$n\backslash p$	10^2	10^3	10^4	10^5
200	7.37	**7.97**	**8.67**	**9.24**
300	7.41	8.17	8.79	9.33
400	7.48	8.15	8.81	9.39
500	7.44	8.09	8.74	9.34
600	7.55	8.17	8.83	9.31
700	7.66	8.25	8.83	9.31
800	7.51	8.20	8.80	9.31
900	7.67	8.26	8.89	9.38
1000	7.55	8.13	8.73	9.23

(c) StabSel PC $\pi = 0.55$ $EV = 2.5$

$n\backslash p$	10^2	10^3	10^4	10^5
200	**7.13**	**7.88**	**8.68**	**9.26**
300	7.24	8.09	8.80	9.33
400	7.26	8.08	8.83	9.39
500	7.25	8.02	8.75	9.35
600	7.39	8.09	8.85	9.34
700	7.50	8.16	8.85	9.33
800	7.29	8.12	8.82	9.34
900	7.47	8.20	8.90	9.41
1000	7.33	8.06	8.76	9.26

(d) StabSel PC $\pi = 0.6$ $EV = 2.5$

$n\backslash p$	10^2	10^3	10^4	10^5
200	7.42	8.26	9.00	9.51
300	7.50	8.47	9.12	9.54
400	7.61	8.51	9.14	9.62
500	7.59	8.45	9.09	9.56
600	7.74	8.61	9.17	9.59
700	7.84	8.55	9.18	9.56
800	7.71	8.59	9.16	9.55
900	7.89	8.61	9.21	9.67
1000	7.75	8.51	9.10	9.48

(e) StabSel PC $\pi = 0.75$ $EV = 2.5$

$n\backslash p$	10^2	10^3	10^4	10^5
200	**6.75**	**7.63**	**8.50**	**9.13**
300	**6.92**	**7.83**	**8.59**	**9.23**
400	**6.91**	**7.76**	**8.61**	**9.25**
500	6.95	**7.77**	**8.53**	**9.20**
600	7.01	**7.78**	**8.64**	9.22
700	**7.11**	**7.86**	**8.60**	9.19
800	6.96	**7.82**	**8.60**	9.18
900	7.13	**7.90**	8.71	9.30
1000	6.97	**7.76**	**8.55**	9.12

(f) StabSel PC $\pi = 0.55$ $EV = 5$

$n\backslash p$	10^2	10^3	10^4	10^5
200	**6.63**	**7.56**	**8.48**	**9.13**
300	**6.83**	**7.76**	**8.58**	**9.23**
400	**6.84**	**7.71**	**8.62**	9.27
500	6.85	**7.71**	**8.53**	**9.21**
600	**6.95**	**7.74**	**8.65**	9.24
700	**7.05**	**7.84**	**8.61**	9.23
800	**6.91**	**7.78**	**8.59**	9.20
900	7.06	**7.87**	8.73	9.30
1000	**6.92**	**7.73**	8.55	9.13

(g) StabSel PC $\pi = 0.6$ $EV = 5$

$n\backslash p$	10^2	10^3	10^4	10^5
200	**6.83**	**7.94**	8.81	9.38
300	7.04	8.19	8.95	9.43
400	7.10	8.19	8.99	9.53
500	7.10	8.15	8.89	9.48
600	7.27	8.24	9.01	9.48
700	7.34	8.32	9.01	9.49
800	7.20	8.28	8.97	9.48
900	7.38	8.34	9.08	9.59
1000	7.20	8.22	8.91	9.41

(h) StabSel PC $\pi = 0.75$ $EV = 5$

Table S11: High-dimensional example: the average number of FP+FN (False Positives and False Negatives) calculated over 500 realisations with $m = 50$ and $B = 500\frac{m}{n}$, number of important variables $s = 10$ and number of factors $K = 5$. Bold: result **better** than the corresponding value for RBVS PC.

$n\backslash p$	10^2	10^3	10^4	10^5
200	5.43	6.53	7.47	8.34
300	5.07	6.17	7.23	8.05
400	4.67	5.92	7.06	7.95
500	4.65	6.05	7.05	7.80
600	4.55	5.84	6.99	7.87
700	4.54	5.96	6.99	7.83
800	4.48	5.98	6.97	7.80
900	4.55	6.03	7.10	7.87
1000	4.56	5.89	7.03	7.71

(a) RBVS PC

$n\backslash p$	10^2	10^3	10^4	10^5
200	**2.13**	**1.57**	**1.13**	**1.89**
300	**1.01**	**.56**	**.30**	**.60**
400	**.56**	**.13**	**.21**	**.31**
500	**.44**	**.14**	**.23**	**.28**
600	**.31**	**.13**	**.16**	**.60**
700	**.28**	**.15**	**.16**	**.40**
800	**.25**	**.23**	**.20**	**.26**
900	**.12**	**.21**	**.24**	**.53**
1000	**.20**	**.19**	**.26**	**.15**

(b) IRBVS PC

$n\backslash p$	10^2	10^3	10^4	10^5
200	5.64	**5.88**	**6.80**	**7.60**
300	5.84	**5.91**	**6.66**	**7.45**
400	5.93	5.95	**6.72**	**7.50**
500	5.93	6.11	**6.77**	**7.41**
600	5.96	5.99	**6.71**	**7.47**
700	5.95	6.13	**6.84**	**7.51**
800	5.93	6.16	**6.89**	**7.57**
900	6.03	6.21	**6.92**	**7.58**
1000	5.95	6.06	**6.87**	**7.41**

(c) StabSel PC $\pi = 0.55$ $EV = 2.5$

$n\backslash p$	10^2	10^3	10^4	10^5
200	**5.21**	**5.78**	**6.76**	**7.59**
300	5.38	**5.76**	**6.62**	**7.45**
400	5.50	**5.83**	**6.72**	**7.51**
500	5.49	**5.97**	**6.78**	**7.45**
600	5.50	5.87	**6.72**	**7.50**
700	5.55	6.02	**6.83**	**7.54**
800	5.54	6.06	**6.91**	**7.61**
900	5.67	6.08	**6.95**	**7.63**
1000	5.55	5.95	**6.88**	**7.46**

(d) StabSel PC $\pi = 0.6$ $EV = 2.5$

$n\backslash p$	10^2	10^3	10^4	10^5
200	**5.09**	**5.93**	**6.91**	**7.67**
300	5.42	**6.17**	**7.12**	**7.83**
400	5.58	6.20	7.19	**7.89**
500	5.59	6.39	7.20	7.86
600	5.64	6.35	7.24	7.94
700	5.70	6.47	7.27	7.96
800	5.72	6.45	7.35	8.04
900	5.85	6.55	7.41	8.02
1000	5.76	6.41	7.31	7.91

(e) StabSel PC $\pi = 0.75$ $EV = 2.5$

$n\backslash p$	10^2	10^3	10^4	10^5
200	**4.99**	**5.70**	**6.80**	**7.71**
300	**5.05**	**5.50**	**6.37**	**7.21**
400	5.16	**5.55**	**6.39**	**7.28**
500	5.16	**5.68**	**6.54**	**7.25**
600	5.17	**5.58**	**6.46**	**7.25**
700	5.25	**5.72**	**6.54**	**7.29**
800	5.24	**5.74**	**6.62**	**7.38**
900	5.36	**5.76**	**6.61**	**7.38**
1000	5.24	**5.66**	**6.58**	**7.23**

(f) StabSel PC $\pi = 0.55$ $EV = 5$

$n\backslash p$	10^2	10^3	10^4	10^5
200	**4.72**	**5.66**	**6.80**	**7.72**
300	**4.80**	**5.41**	**6.34**	**7.20**
400	4.96	**5.42**	**6.38**	**7.29**
500	4.94	**5.58**	**6.52**	**7.26**
600	4.97	**5.49**	**6.43**	**7.26**
700	5.01	**5.65**	**6.54**	**7.31**
800	5.05	**5.68**	**6.62**	**7.41**
900	5.20	**5.71**	**6.63**	**7.41**
1000	5.06	**5.60**	**6.58**	**7.24**

(g) StabSel PC $\pi = 0.6$ $EV = 5$

$n\backslash p$	10^2	10^3	10^4	10^5
200	**4.55**	**5.73**	**6.75**	**7.53**
300	**4.74**	**5.79**	**6.76**	**7.65**
400	4.97	**5.85**	**6.88**	**7.72**
500	4.97	**6.03**	**6.92**	**7.65**
600	5.02	5.92	**6.94**	**7.75**
700	5.13	6.11	7.05	**7.76**
800	5.14	6.14	7.11	7.86
900	5.29	6.21	7.14	**7.84**
1000	5.15	6.06	7.09	7.71

(h) StabSel PC $\pi = 0.75$ $EV = 5$

Table S12: High-dimensional example: the average number of FP+FN (False Positives and False Negatives) calculated over 500 realisations with $m = 100$ and $B = 500\frac{m}{n}$, number of important variables $s = 10$ and number of factors $K = 5$. Bold: result **better** than the corresponding value for RBVS PC.

$n\backslash p$	10^2	10^3	10^4	10^5
200	5.47	6.51	7.54	8.35
300	4.33	5.22	6.24	7.08
400	3.48	4.29	5.34	6.09
500	2.96	3.80	4.65	5.30
600	2.59	3.35	4.08	4.85
700	2.27	2.96	3.74	4.22
800	2.16	2.72	3.46	3.95
900	1.98	2.45	3.13	3.70
1000	1.83	2.32	2.90	3.44

(a) RBVS PC

$n\backslash p$	10^2	10^3	10^4	10^5
200	**2.21**	**1.48**	**1.11**	**1.83**
300	**1.95**	**1.13**	**.67**	**.45**
400	**1.59**	**1.00**	**.57**	**.30**
500	**1.73**	**.99**	**.56**	**.32**
600	**1.70**	**.97**	**.56**	**.30**
700	**1.70**	**.90**	**.49**	**.28**
800	**1.53**	**.91**	**.50**	**.29**
900	**1.52**	**.91**	**.50**	**.27**
1000	**1.46**	**.90**	**.51**	**.26**

(b) IRBVS PC

$n\backslash p$	10^2	10^3	10^4	10^5
200	5.65	**5.86**	**6.81**	**7.61**
300	5.02	**4.74**	**5.54**	**6.29**
400	4.71	**4.08**	**4.71**	**5.33**
500	4.49	**3.57**	**4.13**	**4.76**
600	4.38	**3.15**	**3.68**	**4.36**
700	4.29	**2.86**	**3.39**	**3.76**
800	4.23	**2.59**	**3.17**	**3.57**
900	4.29	**2.35**	**2.90**	**3.38**
1000	4.24	**2.17**	**2.71**	**3.17**

(c) StabSel PC $\pi = 0.55$ $EV = 2.5$

$n\backslash p$	10^2	10^3	10^4	10^5
200	**5.22**	**5.74**	**6.78**	**7.60**
300	4.35	**4.63**	**5.52**	**6.29**
400	3.88	**3.99**	**4.72**	**5.33**
500	3.50	**3.44**	**4.14**	**4.78**
600	3.26	**2.97**	**3.67**	**4.36**
700	3.07	**2.74**	**3.39**	**3.76**
800	2.96	**2.52**	**3.18**	**3.58**
900	2.80	**2.26**	**2.92**	**3.38**
1000	2.71	**2.09**	**2.72**	**3.18**

(d) StabSel PC $\pi = 0.6$ $EV = 2.5$

$n\backslash p$	10^2	10^3	10^4	10^5
200	**5.10**	**5.89**	**6.90**	**7.66**
300	**4.11**	**4.78**	**5.70**	**6.42**
400	3.50	**4.10**	**4.81**	**5.42**
500	3.05	**3.55**	**4.25**	**4.80**
600	2.70	**3.08**	**3.76**	**4.39**
700	2.49	**2.84**	**3.46**	**3.83**
800	2.30	**2.58**	**3.24**	**3.65**
900	2.11	**2.32**	**2.94**	**3.44**
1000	1.92	**2.14**	**2.73**	**3.19**

(e) StabSel PC $\pi = 0.75$ $EV = 2.5$

$n\backslash p$	10^2	10^3	10^4	10^5
200	**4.96**	**5.69**	**6.82**	**7.72**
300	**4.13**	**4.52**	**5.52**	**6.39**
400	3.65	**3.91**	**4.78**	**5.46**
500	3.32	**3.40**	**4.24**	**4.92**
600	3.09	**2.94**	**3.70**	**4.52**
700	2.93	**2.68**	**3.46**	**3.88**
800	2.80	**2.55**	**3.25**	**3.69**
900	2.69	**2.24**	**3.03**	**3.52**
1000	2.57	**2.09**	**2.82**	**3.32**

(f) StabSel PC $\pi = 0.55$ $EV = 5$

$n\backslash p$	10^2	10^3	10^4	10^5
200	**4.70**	**5.70**	**6.87**	**7.72**
300	**3.78**	**4.51**	**5.60**	**6.39**
400	**3.29**	**3.89**	**4.84**	**5.49**
500	**2.89**	**3.43**	**4.26**	**4.99**
600	**2.55**	**2.95**	**3.75**	**4.57**
700	2.40	**2.73**	**3.51**	**3.92**
800	2.23	**2.59**	**3.31**	**3.74**
900	2.03	**2.28**	**3.08**	**3.57**
1000	1.89	**2.17**	2.92	**3.36**

(g) StabSel PC $\pi = 0.6$ $EV = 5$

$n\backslash p$	10^2	10^3	10^4	10^5
200	**4.53**	**5.71**	**6.77**	**7.57**
300	**3.54**	**4.56**	**5.52**	**6.31**
400	**3.04**	**3.91**	**4.72**	**5.33**
500	**2.67**	**3.39**	**4.13**	**4.76**
600	**2.28**	**2.96**	**3.65**	**4.36**
700	**2.08**	**2.67**	**3.39**	**3.75**
800	**1.96**	**2.54**	**3.19**	**3.60**
900	**1.75**	**2.23**	**2.94**	**3.39**
1000	**1.62**	**2.10**	**2.73**	**3.15**

(h) StabSel PC $\pi = 0.75$ $EV = 5$

Table S13: High-dimensional example: the average number of FP+FN (False Positives and False Negatives) calculated over 500 realisations with $m = \frac{n}{2}$ and $B = 500\frac{m}{n}$, number of important variables $s = 10$ and number of factors $K = 5$. Bold: result **better** than the corresponding value for RBVS PC.

S3.3 Some comments

We address each issue brought up in the introduction of this section in the comments below.

1. Comparison of StabSel to RBVS:

   - In the fixed m cases, RBVS typically outperforms StabSel. Moreover, for a moderate value of $m = 100$ and p fixed, the average number of false positives and false negatives decreases with n, which does not hold for StabSel.
   - When the subsample size is set to $\frac{m}{2}$, there typically exists a set of parameters for StabSel such that it slightly outperforms RBVS. We have checked that RBVS in this setting selects slightly more false positives.
   - Overall, performance of StabSel is sensitive to the choice of its parameter.
   - "Optimal" parameters for StabSel in one example are not necessarily best in another case. For instance, in the $s = 5$, $K = 0$ and $m = \frac{n}{2}$ case $\pi = 0.75$ and $EV = 2.5$ results in the best error control, while for $s = 5$, $K = 0$ and $m = 50$ setting $EV = 5$ and $\pi = 0.6$ yields best $FP + FN$ rate.
   - IRBVS almost uniformly outperforms both RBVS and StabSel, which demonstrates that the iterative extension of our methodology significantly improves its vanilla variant.

2. General comments on the impact of high-dimensionality:

   - Perhaps a bit unexpectedly, performance of the IRBVS algorithm improves with dimensionality p growing. This phenomenon can be explained by the fact that a single irrelevant covariate is the less likely to appear at the top of the ranking, the more covariates with similar (spurious) impact on the response there are. We note that this surprising "blessing of dimensionality" has been observed in Fan et al. (2009).
   - IRBVS performs very well even for small/moderate values of n and m, even when p is very large.

3. Comments on the choice of the subsample size m:

   - For the IRBVS algorithm, $m = 100$ yields best $FP + FN$ in this example, often close to 0. On the other hand, choosing $\frac{m}{2}$ results in IRBVS occasionally picking some irrelevant covariates. We emphasise again, however, that IRBVS seems to outperform RBVS and StabSel.

- For the RBVS and StabSel algorithms, $m = \frac{n}{2}$ leads to best performance.
- The subsample size set to a small number ($m = 50$) results in a worse selection of the important variables.

Bibliography

R. Baranowski. **rbvsGPU**: Ranking-Based Variable Selection on GPU, 2016. URL `https://github.com/rbaranowski/rbvsGPU`. R package version 1.0.

H. Cho and P. Fryzlewicz. High dimensional variable selection via tilting. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 74:593–622, 2012.

J. Fan and Y. Fan. High dimensional classification using features annealed independence rules. *Ann. Statist.*, 36: 2605–2637, 2008.

J. Fan and J. Lv. Sure independence screening for ultrahigh dimensional feature space. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 70(5):849–911, 2008.

J. Fan, R. Samworth, and Y. Wu. Ultrahigh dimensional feature selection: beyond the linear model. *J. Mach. Learn. Res.*, 10:2013–2038, 2009.

J. Fan, Y. Ma, and W. Dai. Nonparametric independence screening in sparse ultra-high dimensional varying coefficient models. *J. Amer. Statist. Assoc.*, 109:1270–1284, 2014.

P. Hall and H. Miller. Using generalized correlation to effect variable selection in very high dimensional problems. *J. Comput. Graph. Statist.*, 18, 2009.

P. Hall and J.-H. Xue. On selecting interacting features from high-dimensional data. *Comput. Statist. Data Anal.*, 71:694–708, 2014.

D. Harrison and D. L. Rubinfeld. Hedonic housing prices and the demand for clean air. *J. Environ. Econ. Manage.*, 5:81–102, 1978.

F. Leisch and E. Dimitriadou. *mlbench: Machine Learning Benchmark Problems*, 2010. R package version 2.1-1.

D. Luebke. CUDA: Scalable parallel programming for high-performance scientific computing. In *5th IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, pages 836–838. IEEE, 2008.

N. Meinshausen and P. Bühlmann. Stability selection. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 72:417–473, 2010.

R. K. Pace and O. W. Gilley. Using the spatial configuration of the data to improve estimation. *J. Real Estate. Financ.*, 14:333–340, 1997.

N. Pochet, F. De Smet, J. A. K. Suykens, and B. L. R. De Moor. Systematic benchmarking of microarray data classification: assessing the role of non-linearity and dimensionality reduction. *Bioinformatics*, 20:3185–3195, 2004.

P. Radchenko and G. M. James. Forward-lasso with adaptive shrinkage. *Ann. Appl. Stat.*, 5:427–448, 2010.

D. Singh, P. G. Febbo, K. Ross, D. G. Jackson, J. Manola, C. Ladd, P. Tamayo, A. A. Renshaw, A. V. D'Amico, and J. P. Richie. Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell*, 1:203–209, 2002.