SUPPLEMENTARY MATERIAL

6 Proofs of theorems

6.1 Regularity conditions

This section is a review of the formal regularity conditions required to specify the distribution of the SPVIM values (Williamson and Feng, 2020). We define the linear space $\mathcal{R} := \{c(P_1 - P_2) : c \in \mathbb{R}, P_1, P_2 \in \mathcal{M}\}$ of finite signed measures generated by $\mathcal{M}$. For any $R \in \mathcal{R}$, we consider the supremum norm $\|R\|_\infty := |c| \sup_z |F_1(z) - F_2(z)|$, where F_1 and F_2 are the distribution functions corresponding to P_1 and P_2, respectively, and we have used the representation $R = c(P_1 - P_2)$. For distribution $P_{0,\epsilon} := P_0 + \epsilon h$ with $\epsilon \in \mathbb{R}$ and $h \in \mathcal{R}$, we define $f_{0,\epsilon,s} = f_{P_{0,\epsilon},s}$ to be the oracle prediction function with respect to each subset $s \in \{1, \ldots, p\}$. Let $\dot{V}(f, P_0; h)$ denote the Gâteaux derivative of $P \mapsto V(f, P)$ at P_0 in the direction $h \in \mathcal{R}$. The Gâteaux derivatives for several common choices of V are provided in Williamson et al. (2021). Next, we define the random function $g_{n,s} : z \mapsto \dot{V}(f_{n,s}, P_0; \delta_z - P_0) - \dot{V}(f_{0,s}, P_0; \delta_z - P_0)$, where δ_z is the degenerate distribution on $\{z\}$. For each $s \subseteq \{1, \ldots, p\}$, we require the following conditions to hold:

(A1) *(optimality)* there is some $C > 0$ such that for each sequence $f_1, f_2, \cdots \in \mathcal{F}_s$ with $\|f_j - f_{0,s}\|_{\mathcal{F}_s} \to 0$, there is a J such that for all $j > J$, $|V(f_j, P_0) - V(f_{0,s}, P_0)| \leq C\|f_j - f_{0,s}\|_{\mathcal{F}_s}^2$;

(A2) there is some $\delta > 0$ such that for each sequence $\epsilon_1, \epsilon_2, \ldots \in \mathbb{R}$ and $h, h_1, h_2, \ldots \in \mathcal{R}$ satisfying that $\epsilon_j \to 0$ and $\|h_j - h\|_\infty \to 0$, it holds that

$$\sup_{f \in \mathcal{F}_s : \|f - f_{0,s}\|_{\mathcal{F}_s} < \delta} \left| \frac{V(f, P_0 + \epsilon_j h_j) - V(f, P_0)}{\epsilon_j} - \dot{V}(f, P_0; h_j) \right| \to 0;$$

(A3) $\|f_{0,\epsilon,s} - f_{0,s}\|_{\mathcal{F}_s} = o(\epsilon)$ for each $h \in \mathcal{R}$;

(A4) $f \mapsto \dot{V}(f, P_0; h)$ is continuous at $f_{0,s}$ relative to $\mathcal{F}_s$ for each $h \in \mathcal{R}$;

(A5) $\|f_{n,s} - f_{0,s}\|_{\mathcal{F}_s} = o_P(n^{-1/4})$;

(A6) $E_{P_0}[\int \{g_{n,s}(z)\}^2 dP_0(z)] = o_P(1)$;

(A7) for $\gamma > 0$ and sequence $\gamma_1, \gamma_2, \ldots \in \mathbb{R}^+$ satisfying that $|\gamma_j - \gamma| \to 0$, $c = \gamma_n n$.

In settings with missing data, a modified version of (A5) and (A6) must hold for on average over the imputed datasets:

(A5) *(in missing data settings)* $M^{-1} \sum_{m=1}^{M} \|f_{m,n,s} - f_{0,s}\|_{\mathcal{F}_s} = o_P(n^{-1/4})$;

(A6) *(in missing data settings)* $M^{-1} \sum_{m=1}^{M} E_{P_0}[\int \{g_{m,n,s}(z)\}^2 dP_0(z)] = o_P(1)$,

where $f_{m,n,s}$ is a prediction function estimated using the mth imputed dataset, and $g_{m,n,s}$ is defined as above but replacing all instances of $f_{n,s}$ with $f_{m,n,s}$, and replacing the ideal-data unit z with the observed-data unit o.

6.2 Proof of Lemma 1

The result follows under conditions (A1)–(A8) and an application of results in Chapter 4 of Rubin (1987). Using this result, we can write that

$$\sqrt{n}(\psi_{M,c,n} - \psi_0) \to_d W \sim N(0, \sigma^2),$$

where a consistent estimator of σ^2 is given by $\sigma^2_{M,n} + \frac{m+1}{m}\tau^2_{M,n}$. Recall that (A8) requires consistency of the imputation-based estimators as $M \to \infty$.

6.3 Proof of Theorem 1

Before proving the theorem, we state and prove a lemma that will be useful.

Lemma S3. *For any $\alpha \in (0,1)$, $k \in \{0, \ldots, p - R_n(\alpha)\}$ and $q \in (0,1)$, if conditions (A1)–(A6) hold for each $s \subseteq \{1, \ldots, p\}$ and (A7) holds, then the procedure $S_n(\alpha)$ satisfies the following: (a) when based on Holm-adjusted p-values, $FWER \leq \alpha$ both in finite samples and asymptotically; and (b) when based on a step-down maxT or minP procedure, $FWER \leq \alpha$ asymptotically.*

Proof. Under the collection of conditions (A1)–(A7), $\sqrt{n}(\psi_{c,n} - \psi_0) \rightarrow_d Z \sim N(0, \Sigma_0)$ by Theorem 1 in Williamson and Feng (2020), where $\Sigma_0 = E_0\{\phi_0(O)\phi_0(O)^\top\}$ and ϕ_0 is the vector of efficient influence function values provided in Williamson and Feng (2020) for each j. Therefore, the centered and scaled test statistics T_n follow a multivariate Gaussian distribution.

Thus, by Proposition 3.8 in Dudoit and van der Laan (2008), when $S_n(\alpha)$ is based on Holm-adjusted p-values the procedure has finite-sample and asymptotic control of the FWER. When $S_n(\alpha)$ is based on a step-down maxT or minP procedure, the procedure has asymptotic control of the FWER as a result of Theorems 5.2 and 5.7 in Dudoit and van der Laan (2008), respectively. □

Under conditions (A1)–(A7) and (B1)–(B2), an application of Lemma S3 and Theorem 6.3 in Dudoit and van der Laan (2008) to the procedure $S_n^+(k, \alpha)$ yields that

$$Pr_{P_0}(V_n^+(k, \alpha) > k) = \alpha_n \text{ and } Pr_{P_0}(V_n^+(k, \alpha)/R_n^+(k, \alpha) > q) = \alpha_n \text{ for all } n,$$

i.e., the gFWER(k) and PFP(q) are controlled in finite samples at level α_n.

If additionally conditions (B3)–(B4) hold, then an application of Lemma S3 and Theorem 6.5 in Dudoit and van der Laan (2008) to the procedure $S_n^+(k, \alpha)$ yields that

$$\limsup_{n\to\infty} Pr_{P_0}(V_n^+(k, \alpha) > k) \leq \alpha \text{ and } \limsup_{n\to\infty} Pr_{P_0}(V_n^+(k, \alpha)/R_n^+(k, \alpha) > q) \leq \alpha,$$

i.e., the gFWER(k) and PFP(q) are controlled asymptotically at level α.

Finally, under the above conditions, an application of Lemma S3 and Theorem 6.6 in Dudoit and van der Laan (2008) to the procedure $S_n^+(k, \alpha)$ yields that the FDR is controlled asymptotically.

In missing-data settings, we simply require that condition (A8) additionally hold, and modify the above displays to use $S_{M,n}^+(\alpha)$, $V_{M,n}^+(\alpha)$, and $R_{M,n}^+(\alpha)$ in place of $S_n^+(\alpha)$, $V_n^+(\alpha)$, and $R_n^+(\alpha)$.

6.4 Proof of Lemma 2

Suppose that we are in a complete-data setting. Without loss of generality, suppose that we use Holm-adjusted p-values to construct the initial set of selected variables and that the augmented set is chosen so as to control the gFWER(k). For a fixed sample size n and constant k_n, this results in selected set $S_n := S_n^+(k_n, \alpha)$, where $|S_n| = k_n$. The claim of persistence is equivalent to showing that

$$V(f_{n,S_n}, P_0) - V(f_*, P_0) \to_P 0.$$

We can decompose the left-hand side of the above expression into two terms:

$$V(f_{n,S_n}, P_0) - V(f_*, P_0) = \{V(f_{n,S_n}, P_0) - V(f_{0,S_n}, P_0)\} - \{V(f_{0,S_n}, P_0) - V(f_*, P_0)\}. \quad \text{(S6)}$$

The first term in (S6) is the contribution to the limiting behavior of $V(f_{n,S_n}, P_0) - V(f_*, P_0)$ from estimating f_0 for a fixed S_n; by condition (A1),

$$|V(f_{n,S_n}, P_0) - V(f_{0,S_n}, P_0)| \leq C\|f_{n,S_n} - f_{0,S_n}\|^2_{\mathcal{F}_{S_n}} \to_P 0.$$

The second term in (S6) is the contribution to the limiting behavior of $V(f_{n,S_n}, P_0) - V(f_*, P_0)$ from selecting S_n compared to the population-optimal set. To study this term, recall that for a fixed p, we have under conditions (A1), (A2), (A5), (A6), and (A7) that $\psi_{c,n,j} \to_P \psi_{0,j}$ for each $j \in \{1, \ldots, p\}$. Thus, for each $j \in S_0$, the p-value $p_{n,j}$ associated with testing the null hypothesis $H_{0,j} : \psi_{0,j} = 0$ converges to 0. This implies that as $n \to \infty$, $S_n(\alpha) \to_P S_0$. Moreover, by condition (B3), $S_0 \subseteq S_n^+(k_n, \alpha)$ as $n \to \infty$. By definition, $\psi_{0,j} > 0$ if and only if $V(f_{0,s\bigcup\{j\}}, P_0) - V(f_{0,s}, P_0) > 0$ for some $s \subseteq \{1, \ldots, p\}$. This implies that for $j \in S_0^c$, $V(f_{0,s\bigcup\{j\}}, P_0) - V(f_{0,s}, P_0) = 0$ for all $s \subseteq \{1, \ldots, p\}$. In particular, for $j \in S_0^c$,

$$V(f_{0,S_0\bigcup\{j\}}, P_0) - V(f_{0,S_0}, P_0) = 0.$$

This implies that $S_n^+(k_n, \alpha) \to_P S_0$, which further implies that $\{V(f_{0,S_n}, P_0) - V(f_*, P_0)\} \to_P 0$, proving the claim with

$$V(f_{n,S_n}, P_0) - V(f_*, P_0) = o_P(n^{-1/2}).$$

In a setting with missing data, we consider the imputation-based analogue of the above result. Suppose that we have a selected set $S_{M,n} := S_{M,n}^+(\alpha)$. Then

$$\frac{1}{M}\sum_{m=1}^{M} V(f_{m,n,S_{M,n}}, P_0) - V(f_*, P_0) = \frac{1}{M}\sum_{m=1}^{M}\{V(f_{m,n,S_{M,n}}, P_0) - V(f_{0,S_{M,n}})\} - \{V(f_{0,S_{M,n}} - V(f_*, P_0))\}.$$

Under conditions (A1), (A2), and (A5)–(A8), the same logic applies to the second term in the above display as applied to the second term in Equation (S6), so $\{V(f_{0,S_{M,n}}, P_0) - V(f_*, P_0)\} \to_P 0$. For the first term in the display, an application of (A1) to each of the m terms in the average yields the desired convergence in probability.

7 Additional numerical experiments

7.1 Replicating all numerical experiments

All numerical experiments presented here and in the main manuscript can be replicated using code available on GitHub.

In all cases, our simulated dataset consisted of independent replicates of (X, Y), where $X = (X_1, \ldots, X_p)$ and Y followed a Bernoulli distribution with success probability $\Phi\{\beta_{00} + f(\beta_0, x)\}$ conditional on $X = x$, where Φ denotes the cumulative distribution function of the standard normal distribution. Under this specification, Y followed a probit model. A summary of the eight scenarios is provided in Table S1.

In Scenarios 3–5, we investigate the effect of departures from a multivariate normal feature distribution and a linear outcome regression model under a similar setup to Scenario 1. We set $\beta_{00} = 0.5$ and $\beta_0 = (-1, 1, -0.5, 0.5, 1/3, -1/3, \mathbf{0}_{p-6})^\top$, where $\mathbf{0}_k$ denotes a zero-vector of

Scenario	Outcome regression	Feature distribution	Importance	p
1	Linear	Independent normal	Mix	$\{30, 500\}$
2	Nonlinear	Correlated normal	Weak	6
3	Linear	Independent nonnormal	Mix	$\{30, 500\}$
4	Nonlinear	Independent normal	Mix	$\{30, 500\}$
5	Nonlinear	Independent nonnormal	Mix	$\{30, 500\}$
6	Linear	Independent normal	Weak	6
7	Linear	Correlated normal	Weak	6
8	Nonlinear	Independent normal	Weak	6

Table S1: Summary of the eight data-generating scenarios considered in the numerical experiments.

dimension k. We vary $p \in \{30, 500\}$. In Scenario 3, we set $f(\beta_0, x) = x\beta_0$, but in contrast to Scenario 1, X follows a nonnormal feature distribution specified by

$$X_1 \sim N(0.5, 1);\ X_2 \sim Binomial(0.5);\ X_3 \sim Weibull(1.75, 1.9);\ X_4 \sim Lognormal(0.5, 0.5);$$
$$X_5 \sim Binomial(0.5);\ X_6 \sim N(0.25, 1);\ (X_7, \ldots, X_p) \sim MVN(0, I_{p-6}). \quad \text{(S7)}$$

In Scenarios 4 and 5, the outcome regression follows the same nonlinear specification as in Scenario 2. Specifically, using a centering and scaling function c_j for each variable,

$$\begin{aligned} f(\beta_0, x) = 2[&\beta_{0,1} f_1\{c_1(x_1)\} + \beta_{0,2} f_2\{c_2(x_2), c_3(x_3)\} + \beta_{0,3} f_3\{c_3(x_3)\} \\ &+ \beta_{0,4} f_4\{c_4(x_4)\} + \beta_{0,5} f_2\{c_5(x_5), c_1(x_1)\} + \beta_{0,6} f_5\{c_6(x_6)\}], \quad \text{(S8)} \\ f_1(x) = &\sin\left(\frac{\pi}{4}x\right), f_2(x, y) = xy, f_3(x) = \tanh(x), \\ f_4(x) = &\cos\left(\frac{\pi}{4}x\right), f_5(x) = -\tanh(x), \end{aligned}$$

where tanh denotes the hyperbolic tangent. In Scenario 4, $X \sim MVN(0, I_p)$, while in Scenario 5, X follows the distribution specified in Equation (S7). In these scenarios, only the first six features truly influence the outcome; some of the features are strongly important, while others are more weakly important.

In the final scenarios, we investigate the effect of correlated features and departures from a linear outcome regression model in a setting where the features are equally, and weakly,

important; these settings are similar to Scenario 2. In these cases, we set $p = 6$, $\beta_{00} = 0.5$, $\beta_0 = (0, 1, 0, 0, 0, 1)^\top$, and $X \sim MVN(0, \Sigma)$, where $\Sigma_{i,j} = \rho_1^{|i-j|}$ for i, j not in the active set, and $\Sigma_{i,j} = I_p + \rho_2(J_p - I_p)$ for i, j in the active set, where J_p is a $p \times p$ matrix of ones. In Scenarios 6 and 7 we set $f(\beta_0, x) = x\beta_0$, while in Scenario 8 f is specified as in Equation (S8). In Scenarios 6 and 8 we set $\rho_1 = \rho_2 = 0$, while in Scenario 7 we set $\rho_1 = 0.3$ and $\rho_2 = 0.95$.

7.2 Tuning parameters for variable selection

The tuning parameters that specify each variable selection procedure are as follows. For the intrinsic selection algorithm, we determined k and q for error control using a target specificity at $n = 3000$ of 75% for $p = 6$, 85% for $p = 30$, and 95% for $p = 500$. For target specificity denoted by s_p and $s_0 = \sum_{j=1}^{p} I(\beta_{0j} > 0)$, we set $k = \lceil (1 - s_p)(p - s_0) \rceil$, where $\lceil \cdot \rceil$ denotes the ceiling; and set $q = k\{p^{-1}(p - s_0)(n/200)^{1/2} + k\}^{-1}$. The exact values of k (for $gFWER(k)$ control) and q (for $PFP(q)$ control) are provided in Table S2. For stability selection, we specified stability selection threshold equal to 0.9 and target per-comparison type I error rate of 0.04. For the lasso with knockoffs, we set target FDR equal to 0.2.

For cases with missing data, the methods compared are: stability selection within bootstrap imputation, lasso + SS (LJ); bootstrap imputation with bolasso, lasso + SS (BI-BL); SPVIM + gFWER, intrinsic selection to control the generalized familywise error rate; SPVIM + PFP, intrinsic selection to control the proportion of false positives among the rejected variables; and SPVIM + FDR, intrinsic selection to control the false discovery rate.

For cases with complete data the methods compared are: lasso; lasso + SS, lasso with stability selection; lasso + KF, lasso with knockoffs; SPVIM + gFWER, intrinsic selection to control the generalized familywise error rate; SPVIM + PFP, intrinsic selection to control the proportion of false positives among the rejected variables; and SPVIM + FDR, intrinsic selection to control the false discovery rate.

n	p	SS_q	Target specificity	k	q
200	30	23	0.762	6	0.882
500	30	23	0.774	6	0.826
1500	30	23	0.809	5	0.695
3000	30	23	0.854	4	0.564
200	500	91	0.812	94	0.990
500	500	91	0.824	88	0.983
1500	500	91	0.861	69	0.962
3000	500	91	0.904	48	0.926

Table S2: Values of: the number of variables selected in each bootstrap run of stability selection (SS_q), target specificity for $gFWER(k)$ and $PFP(q)$ control, and k and q used for $gFWER$ and PFP control, respectively, in the numerical experiments.

7.3 Super Learner specification

The specific candidate learners and their corresponding tuning parameters for our Super Learner library are provided in Tables S3 (Scenarios 1, 3–5) and S4 (Scenarios 2, 6–8). In both cases, we used a wide variety of algorithms, each with several tuning parameter values, in an effort to be robust to model misspecification. It is possible that with a different library of learners, different results could be obtained.

For the internal library in our intrinsic selection procedure in Scenarios 1 and 3–5, we first pre-screened variables based on their univariate rank correlation with the outcome, and then fit boosted trees with maximum depth equal to three and shrinkage equal to 0.1. In Scenarios 2 and 6–8, we again first pre-screened variables based on their univariate rank correlation with the outcome, and then fit a logistic regression or boosted trees with maximum depth equal to four, shrinkage equal to 0.1, and number of rounds equal to 100. Recall that within the intrinsic selection procedure, we estimate the optimal prediction function for each subset s of the p features. The univariate rank correlation screen operated as follows: if $|s| \leq 2$, we did no screening; if $2 < |s| < 100$, we picked the top two variables ranked by univariate correlation with the outcome; and if $|s| \geq 100$, we picked the top ten variables ranked by univariate correlation with the outcome. This screening substantially reduced the computation time for the intrinsic selection procedure, and reflects the type of aggressive screen that is used in some cases (Neidich et al., 2019). Also, the univariate comparisons of each feature to the null model

Candidate Learner	R Implementation	Tuning Parameter and possible values	Tuning parameter description
Random forests	`ranger` (Wright and Ziegler, 2017)	`mtry` $\in \{1/2, 1, 2\}\sqrt{p}$ †	Number of variables to possibly split at in each node
Gradient boosted trees	`xgboost` (Chen et al., 2019)	`max.depth` $\in \{1, 3\}$	Maximum tree depth
Support vector machines	`ksvm` (Karatzoglou et al., 2004)		
Lasso	`glmnet` (Friedman et al., 2010)	λ chosen via 10-fold CV	ℓ_1 regularization parameter

Table S3: Candidate learners in the Super Learner ensemble for Scenarios 1 and 3–5 along with their R implementation, tuning parameter values, and description of the tuning parameters. All tuning parameters besides those listed here are set to their default values. In particular, the random forests are grown with 500 trees, a minimum node size of 5 for continuous outcomes and 1 for binary outcomes, and a subsampling fraction of 1; the boosted trees are grown with a maximum of 1000 trees, shrinkage rate of 0.1, and a minimum of 10 observations per node; and the SVMs are fit with radial basis kernel, cost of constraints violation equal to 1, upper bound on training error (`nu`) equal to 0.2, `epsilon` equal to 0.1, and three-fold cross-validation with a sigmoid for calculating class probabilities.
†: p denotes the total number of predictors.

(with no features) are given high weight in the intrinsic importance measure, so screening should not have much impact on the final intrinsic importance estimate.

7.4 Additional results from Scenarios 1 and 2 with missing data

In the main manuscript, we presented results with a maximum of 40% missing data in some variables in Scenarios 1 and 2. In Figure S1 we present results in an intermediate setting with a maximum of 20% missing data in some variables; the results in this setting tend to be similar to the results with maximum 40% missing data.

In In Figure S2 we present results in an intermediate setting with a maximum of 20% missing data in some variables, which again tend to be similar to the results with maximum 40% missing data.

In Figures S3–S4, we display the empirical selection probability for each active-set variable under each selection algorithm in Scenario 1. All active-set variables are selected with high

Candidate Learner	R Implementation	Tuning Parameter and possible values	Tuning parameter description
Random forests	`ranger`	`min.node.size` $\in$ $\{1, 20, 50, 100, 250, 500\}$	Minimum node size
Gradient boosted trees	`xgboost`	`shrinkage` $\in \{1 \times 10^{-2}, 1 \times 10^{-1}\}$ `ntrees` $\in \{100, 1000\}$	Shrinkage Number of trees
Support vector machines	`ksvm`		
Lasso	`glmnet`	λ chosen via 10-fold CV	ℓ_1 regularization parameter

Table S4: Candidate learners in the Super Learner ensemble for Scenarios 2 and 6–8 along with their R implementation, tuning parameter values, and description of the tuning parameters. All tuning parameters besides those listed here are set to their default values. In particular, the random forests are grown with 500 trees and a subsampling fraction of 1; the boosted trees are grown with a minimum of 10 observations per node; and the SVMs are fit with radial basis kernel, cost of constraints violation equal to 1, upper bound on training error (`nu`) equal to 0.2, `epsilon` equal to 0.1, and three-fold cross-validation with a sigmoid for calculating class probabilities.

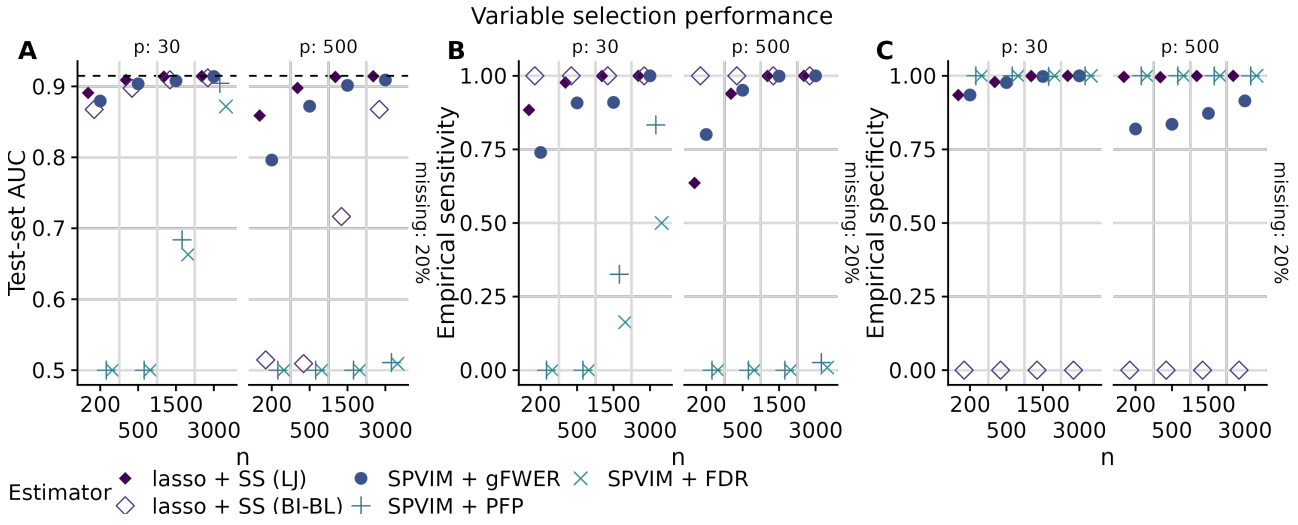

Figure S1: Test-set AUC (panel A) and empirical variable selection sensitivity (panel B) and specificity (panel C) vs n for each estimator and missing data proportion equal to 0.2, in Scenario 1 (a linear model for the outcome and multivariate normal features). The dotted line in panel A shows the true (optimal) test-set AUC.

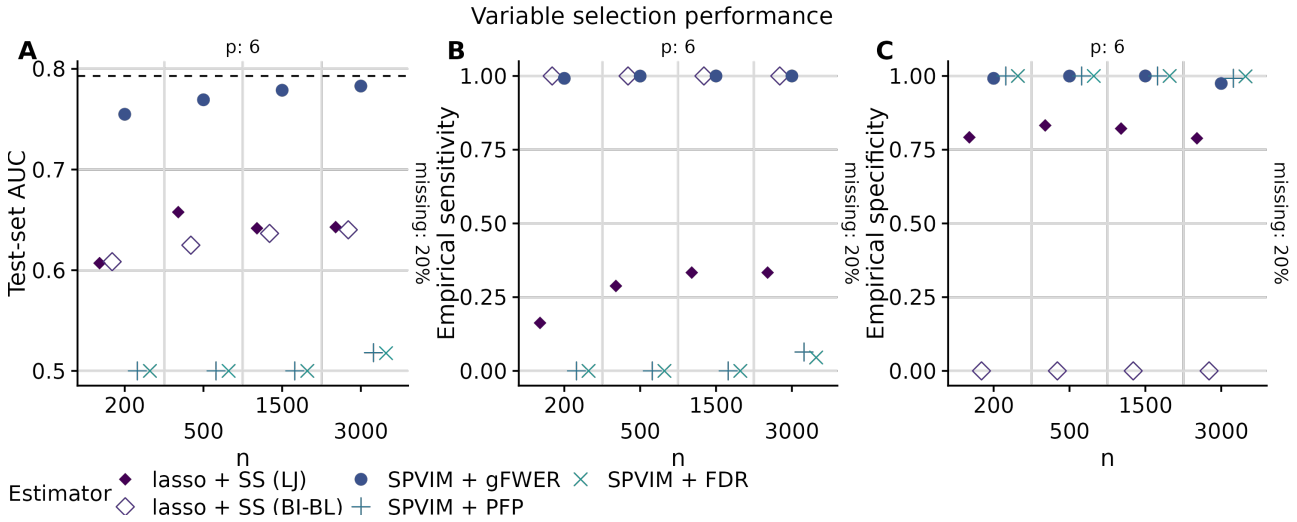

Figure S2: Test-set AUC (panel A) and empirical variable selection sensitivity (panel B) and specificity (panel C) vs n for each estimator and missing data proportion equal to 0.2, in Scenario 2 (a nonlinear model for the outcome and correlated multivariate normal features). The dotted line in panel A shows the true (optimal) test-set AUC.

probability by all procedures, with the exception of SPVIM + FDR and SPVIM + PFP. In small samples, all estimators besides lasso + SS (BI-BL) sometimes fail to select variables 5 and 6, the variables with smallest intrinsic importance; these variables are selected with low probability by SPVIM + PFP and SPVIM + FDR at all sample sizes considered here. In the higher dimensional case, SPVIM + gFWER selects these variables in cases where lasso + SS (LJ) does not. This reflects the low true importance of these variables combined with tuning parameters that provide strict PFP and FDR control. As the proportion of missing data increases, the selection probabilities tend to decrease slightly.

In Figures S5–S6, we display the empirical selection probability for each active-set variable under each selection algorithm in Scenario 2. In this scenario, as expected, the selection probability is low for lasso + SS (LJ) and high for SPVIM + gFWER (as reflected in the empirical sensitivity presented in the main manuscript). Variables 2 and 3, which are highly correlated and include an interaction term not modelled by the lasso, have the lowest selection probability for lasso + SS (LJ), as expected (though lasso + SS (BI-BL) has perfect sensitivity, it also has zero specificity).

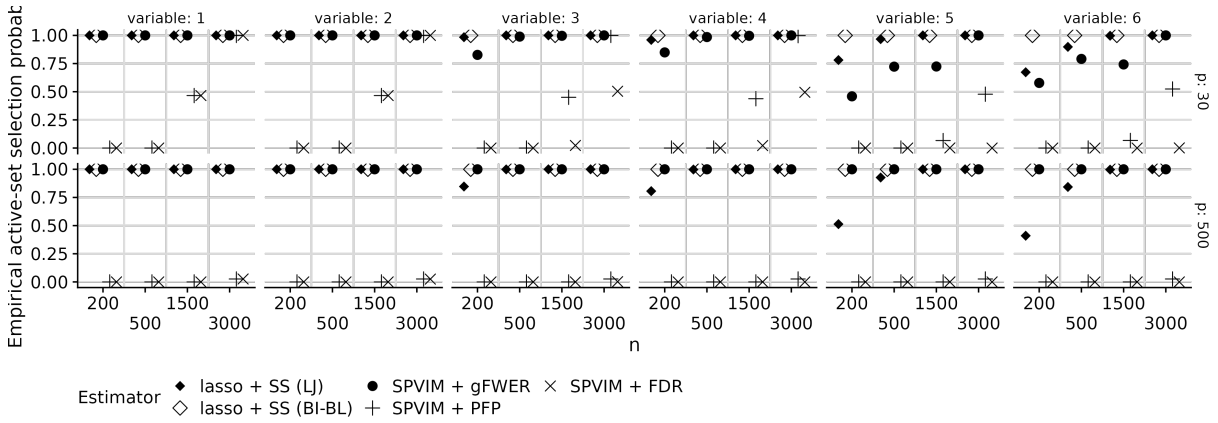

Figure S3: Empirical selection probability for each active-set variable vs n for each estimator and dimension with missing data proportion equal to 0.2, in Scenario 1 (a linear model for the outcome and multivariate normal features).

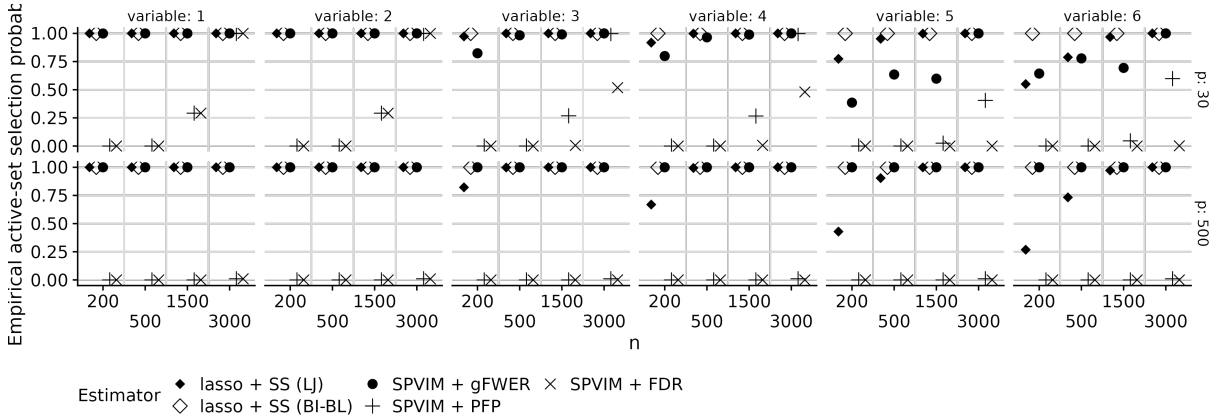

Figure S4: Empirical selection probability for each active-set variable vs n for each estimator and dimension with missing data proportion equal to 0.4, in Scenario 1 (a linear model for the outcome and multivariate normal features).

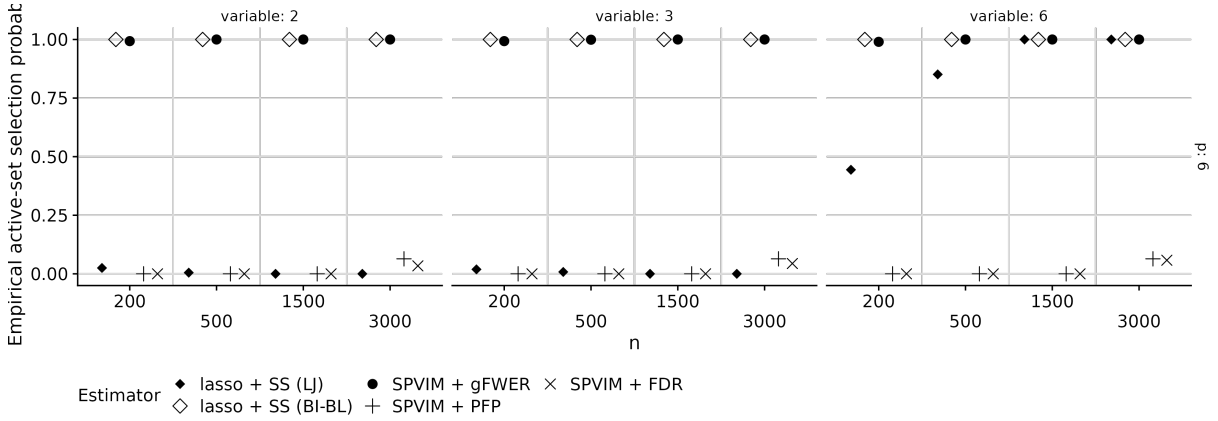

Figure S5: Empirical selection probability for each active-set variable vs n for each estimator, in Scenario 2 (a nonlinear model for the outcome and correlated multivariate normal features).

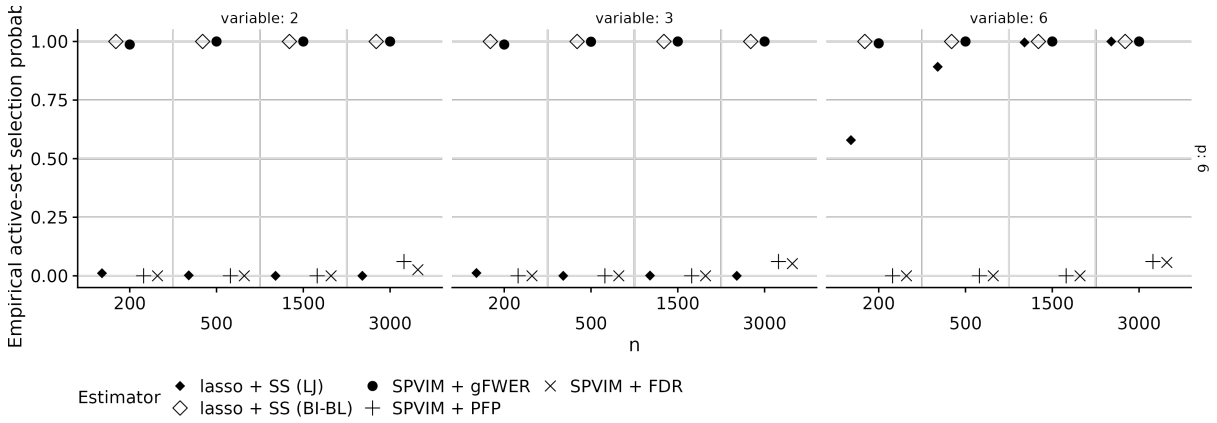

Figure S6: Empirical selection probability for each active-set variable vs n for each estimator, in Scenario 2 (a nonlinear model for the outcome and correlated multivariate normal features).

7.5 Results from Scenarios 3–8 with missing data

In Scenario 3, we generate features from a nonnormal joint distribution and the outcome is a linear combination of these features. We display the results of this experiment in Figure S7. We observe similar performance in this scenario to the performance we observed in Scenario 1: test-set AUC increases towards the optimal value with increasing sample size for all estimators, though slowest for SPVIM + FDR and SPVIM + PFP; empirical sensitivity and specificity tend to both increase, with the exception of the lasso + SS (BI-BL) algorithm, which has near-zero specificity at all sample sizes considered here.

In Scenario 4, we generate features from a multivariate normal distribution and the outcome is a nonlinear combination of these features. In this case, lasso-based methods follow a misspecified mean model. We display the results of this experiment in Figure S8. We observe that test-set AUC tends to increase quickly towards the optimal AUC with increasing sample size for the SPVIM + gFWER procedure, but increases more slowly for lasso-based procedures; empirical sensitivity and specificity tend to both increase, with the exception of the lasso + SS (BI-BL) algorithm, which again has near-zero specificity at all sample sizes considered here. In this case, among the algorithms with non-zero specificity, SPVIM + gFWER has the highest sensitivity at all sample sizes considered here.

In Figure S9, we display the results of the experiment conducted under Scenario 5, in which

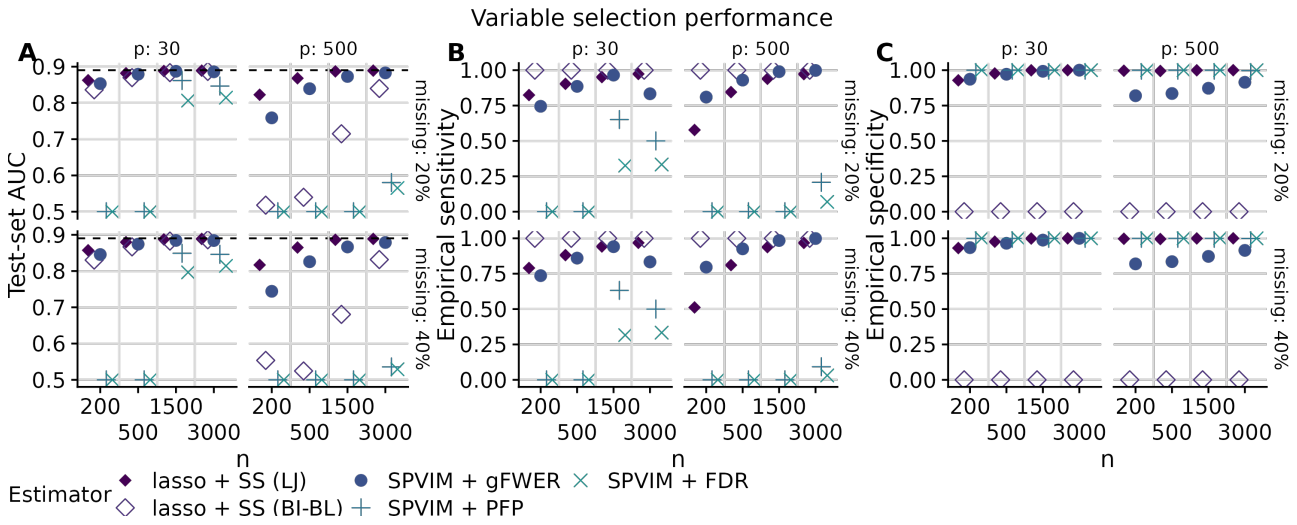

Figure S7: Test-set AUC (panel A) and empirical variable selection sensitivity (panel B) and specificity (panel C) vs n for each estimator and missing data proportion, in Scenario 3 (a linear model for the outcome and nonnormal features). The dotted line in panel A shows the true (optimal) test-set AUC.

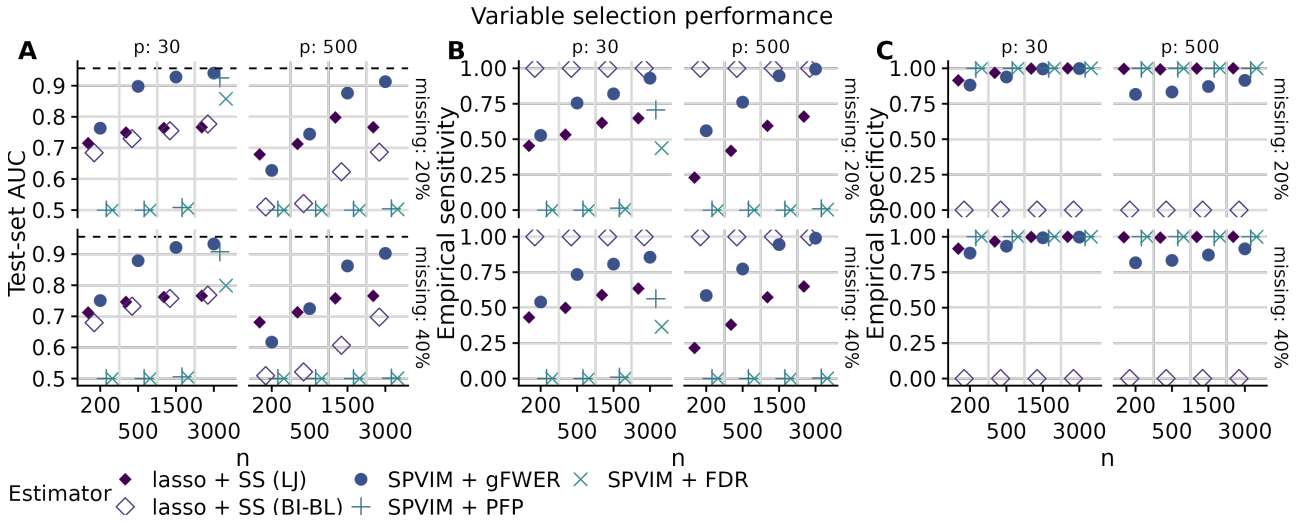

Figure S8: Test-set AUC (panel A) and empirical variable selection sensitivity (panel B) and specificity (panel C) vs n for each estimator and missing data proportion, in Scenario 4 (a nonlinear model for the outcome and normal features). The dotted line in panel A shows the true (optimal) test-set AUC.

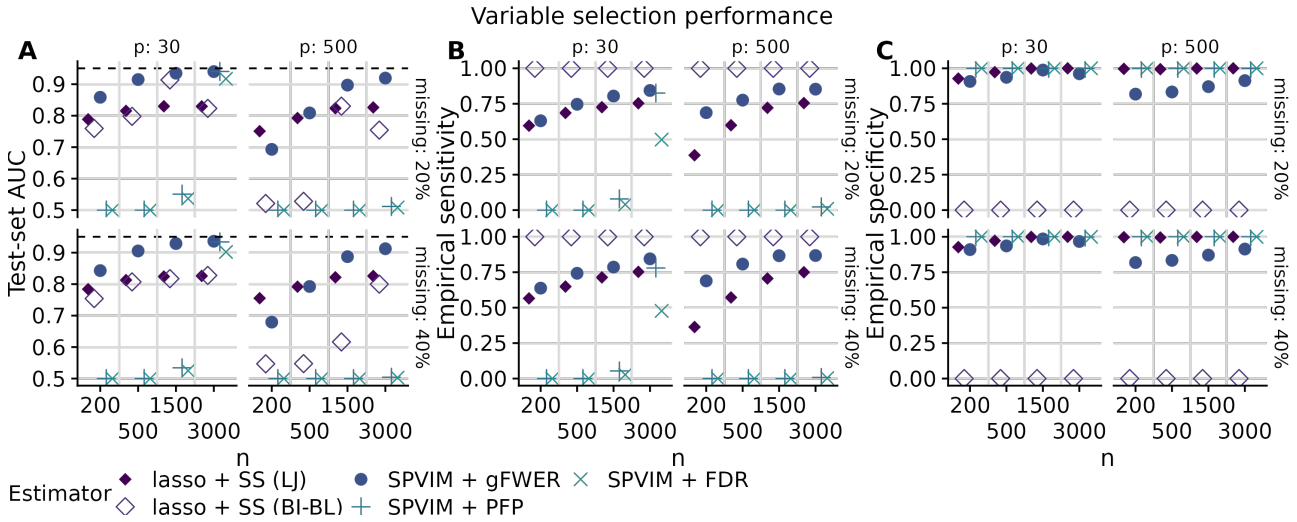

Figure S9: Test-set AUC (panel A) and empirical variable selection sensitivity (panel B) and specificity (panel C) vs n for each estimator and missing data proportion, in Scenario 5 (a nonlinear model for the outcome and nonnormal features). The dotted line in panel A shows the true (optimal) test-set AUC.

the features are nonnormal and the outcome-feature relationship is nonlinear. In this case, the lasso-based methods are misspecified. In panel A, we observe that lasso-based methods have test-set AUC increasing slowly with n, while SPVIM + gFWER has test-set AUC approaching the optimal value more quickly. In panels B and C, we see that sensitivity tends to be lower than in Scenario 1 for all procedures, though still increasing towards one; and that specificity trends are similar to those in Scenario 1. In all cases considered here, SPVIM + gFWER has higher empirical sensitivity than lasso + SS (LJ), and often has comparable specificity, particularly in the lower-dimensional setting.

In Scenarios 6–8, the features are more weakly important. We present the results of the experiments under these scenarios in Figures S10–S12. In Scenario 7, we observe reduced variable selection performance for the lasso-based procedures compared to Scenario 6. In Scenario 8, we observe similar trends to Scenario 2, though performance for the lasso-based methods tends to be better than the performance we observed in Scenario 2, reflecting that this scenario does not involve correlation among the features. These experiments suggest that correlation makes variable selection more difficult, particularly in combination with a misspecified outcome regression model.

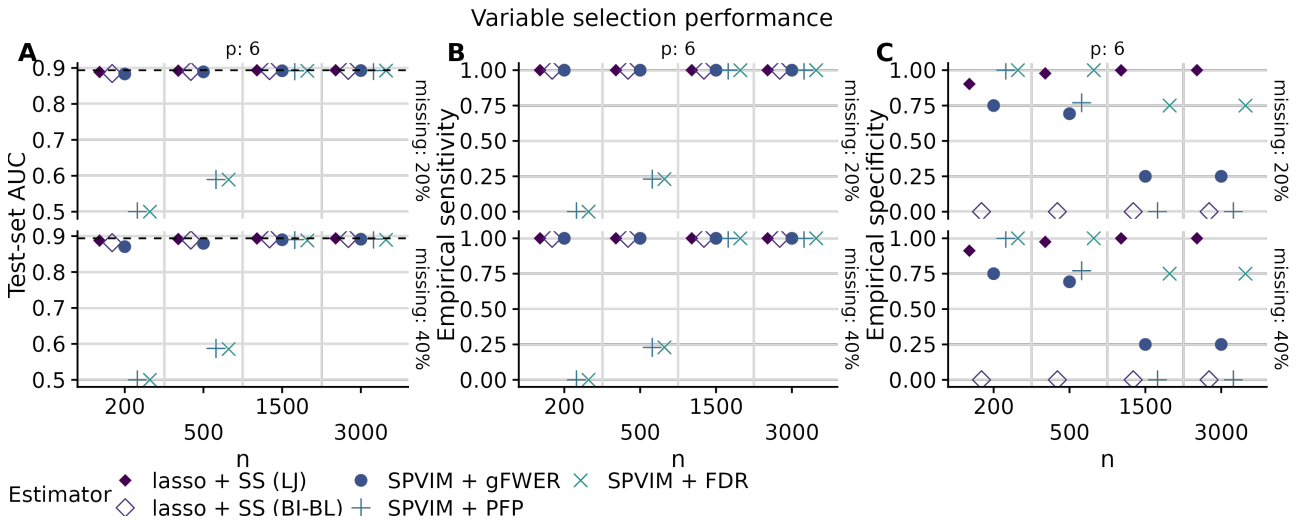

Figure S10: Test-set AUC (panel A) and empirical variable selection sensitivity (panel B) and specificity (panel C) vs n for each estimator and missing data proportion, in Scenario 6 (a weak linear model for the outcome and normal features). The dotted line in panel A shows the true (optimal) test-set AUC.

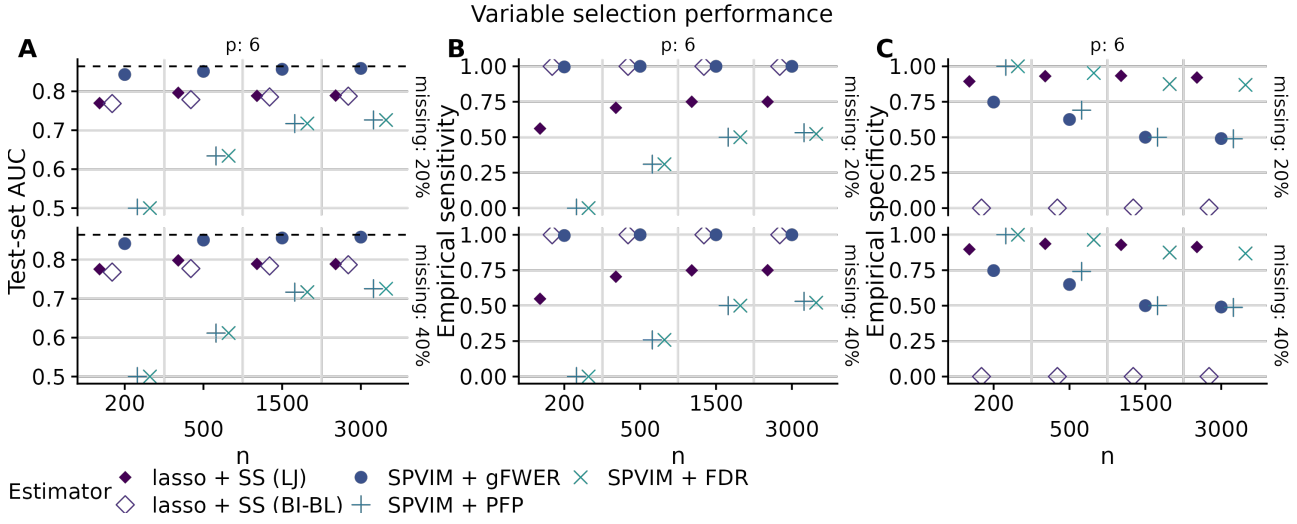

Figure S11: Test-set AUC (panel A) and empirical variable selection sensitivity (panel B) and specificity (panel C) vs n for each estimator and missing data proportion, in Scenario 7 (a weak nonlinear model for the outcome and correlated normal features). The dotted line in panel A shows the true (optimal) test-set AUC.

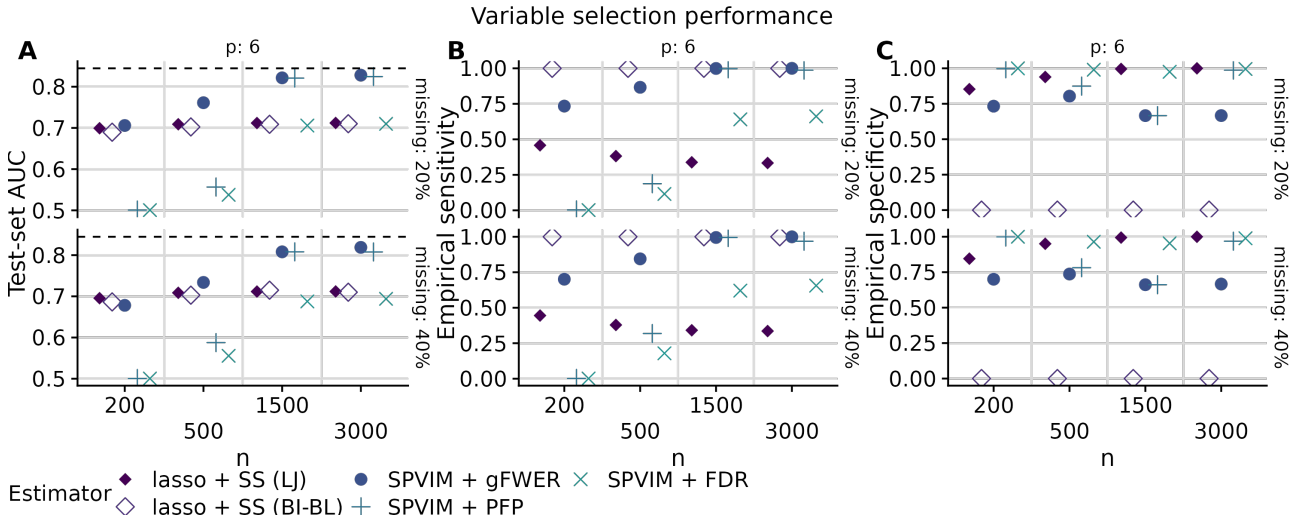

Figure S12: Test-set AUC (panel A) and empirical variable selection sensitivity (panel B) and specificity (panel C) vs n for each estimator and missing data proportion, in Scenario 8 (a weak nonlinear model for the outcome and normal features). The dotted line in panel A shows the true (optimal) test-set AUC.

In Figures S13–S24, we display the empirical selection probability for each active-set variable under each selection algorithm in Scenarios 3–8. We observe similar performance in Scenario 3 as in Scenario 1. In Scenarios 3 and 4, we observe that most procedures select variables 1, 2, 3, 4, and 6 with high probability as sample size increases. However, in the higher-dimensional case lasso-based procedures select variable 5 with lower probability than our proposed intrinsic selection procedure. Variable 5 is moderately important (its coefficient is 1, compared to a maximum coefficient of 2), but the function relating this variable to the outcome is highly nonlinear over its support. In Scenario 6–8, we observe similar patterns to Scenario 5: variables 2 and 3 tend to be selected infrequently by the lasso-based procedures, but with high frequency by the intrinsic selection procedure.

7.6 Results with completely-observed data

Here, we consider Scenarios 1–8 with completely-observed data. We compare our intrinsic selection algorithm to the lasso, the lasso with stability selection, and the lasso with knockoffs; these latter three algorithms are often used in variable selection analyses with fully-observed data. In Figures S25–S32, we present the results of these experiments. The results tend to be

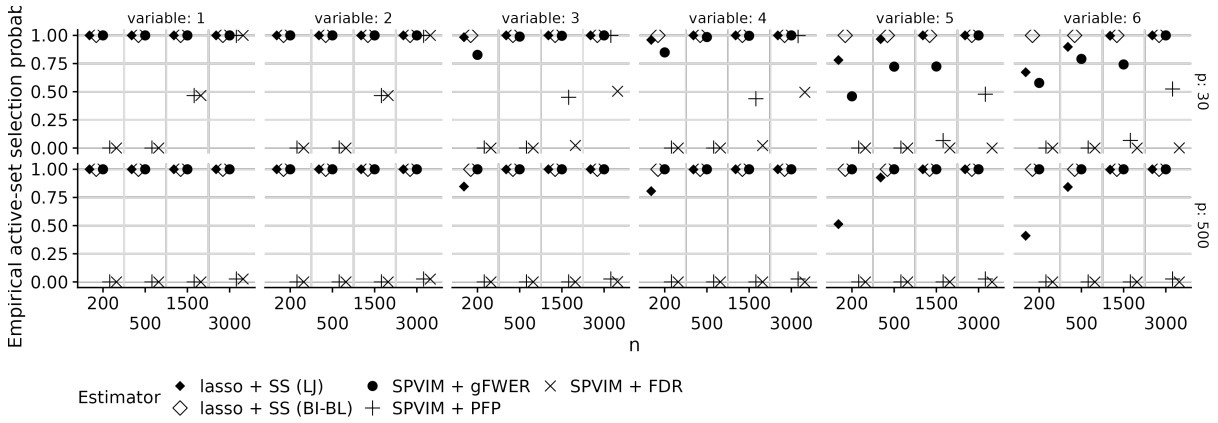

Figure S13: Empirical selection probability for each active-set variable vs n for each estimator and dimension with missing data proportion equal to 0.2, in Scenario 3 (a linear model for the outcome and nonnormal features).

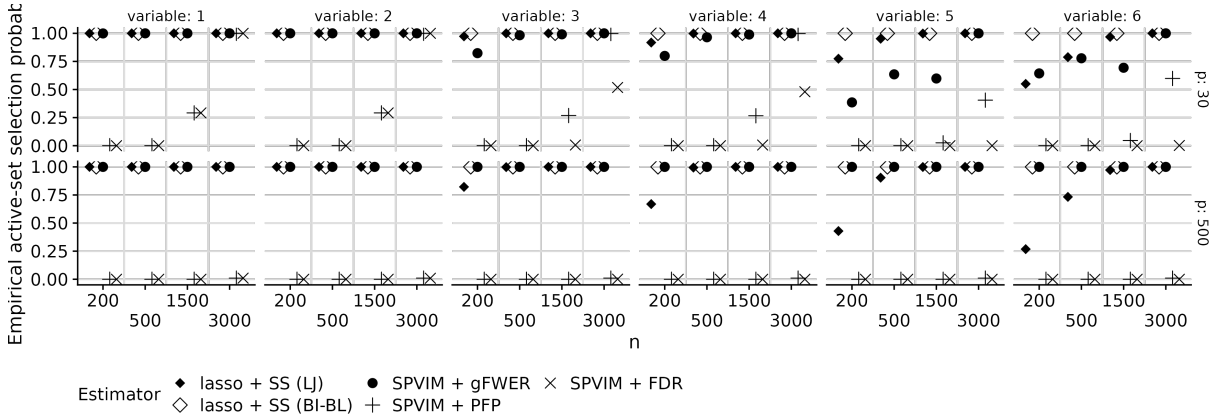

Figure S14: Empirical selection probability for each active-set variable vs n for each estimator and dimension with missing data proportion equal to 0.4, in Scenario 3 (a linear model for the outcome and nonnormal features).

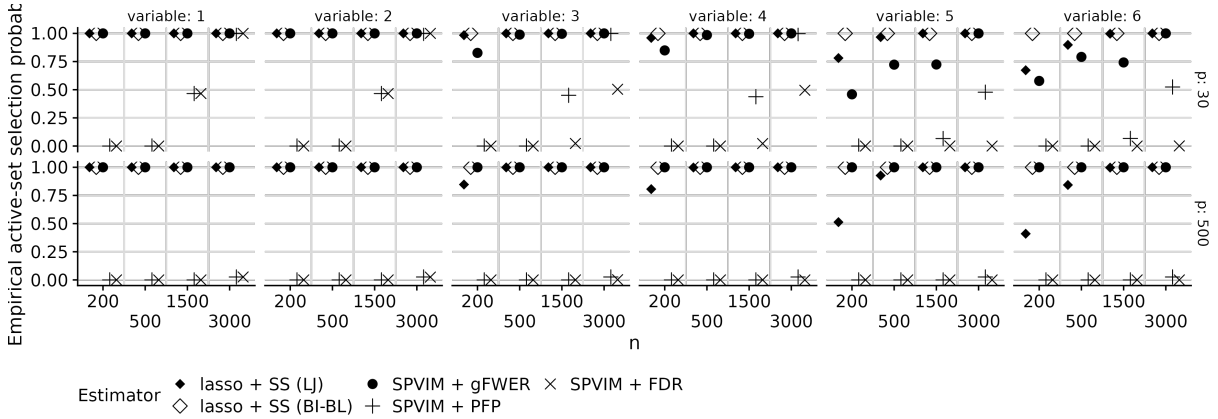

Figure S15: Empirical selection probability for each active-set variable vs n for each estimator and dimension with missing data proportion equal to 0.2, in Scenario 4 (a nonlinear model for the outcome and multivariate normal features).

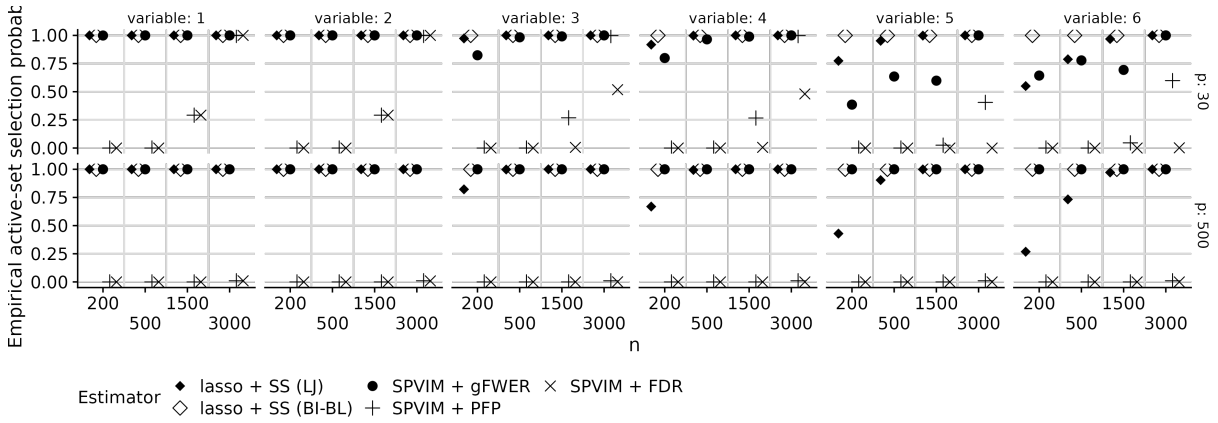

Figure S16: Empirical selection probability for each active-set variable vs n for each estimator and dimension with missing data proportion equal to 0.4, in Scenario 4 (a nonlinear model for the outcome and multivariate normal features).

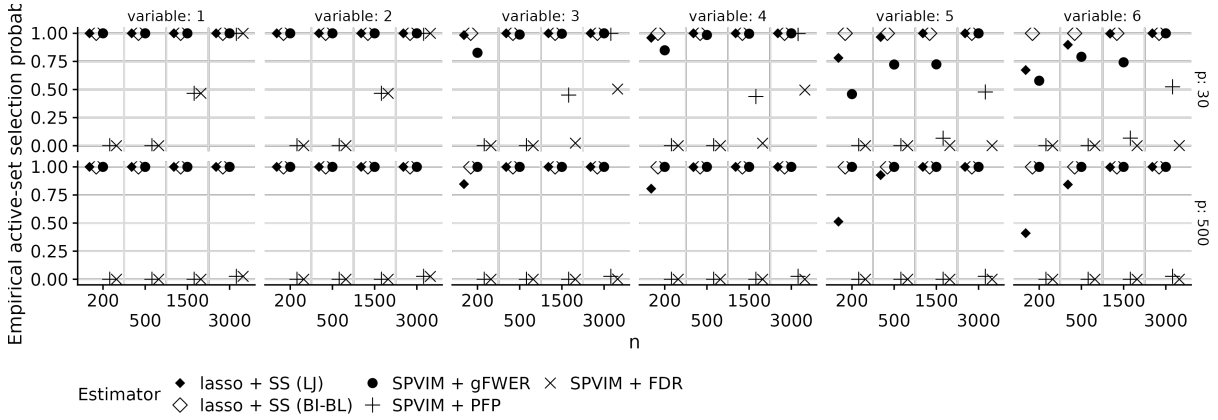

Figure S17: Empirical selection probability for each active-set variable vs n for each estimator and dimension with missing data proportion equal to 0.2, in Scenario 5 (a nonlinear model for the outcome and nonnormal features).

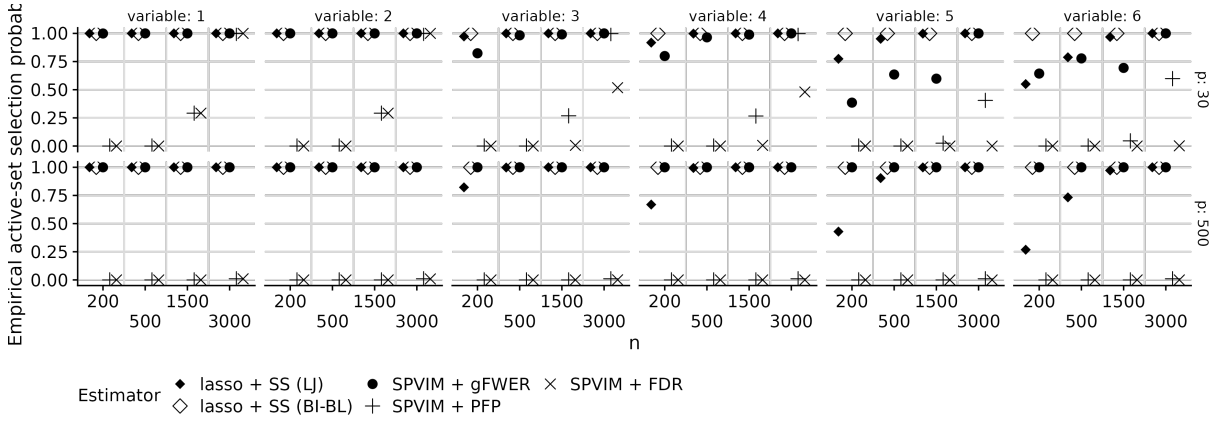

Figure S18: Empirical selection probability for each active-set variable vs n for each estimator and dimension with missing data proportion equal to 0.4, in Scenario 5 (a nonlinear model for the outcome and nonnormal features).

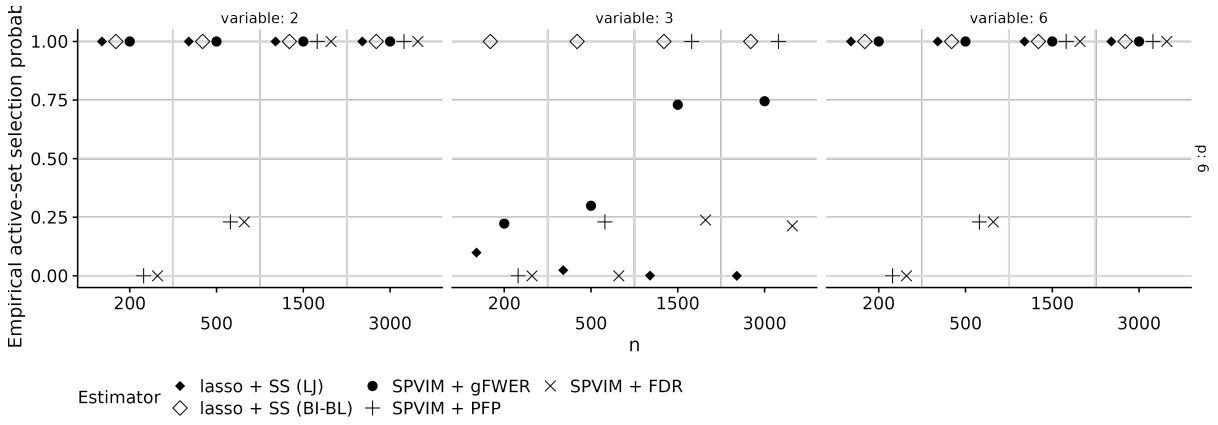

Figure S19: Empirical selection probability for each active-set variable vs n for each estimator, in Scenario 6 (a weak linear model for the outcome and normal features).

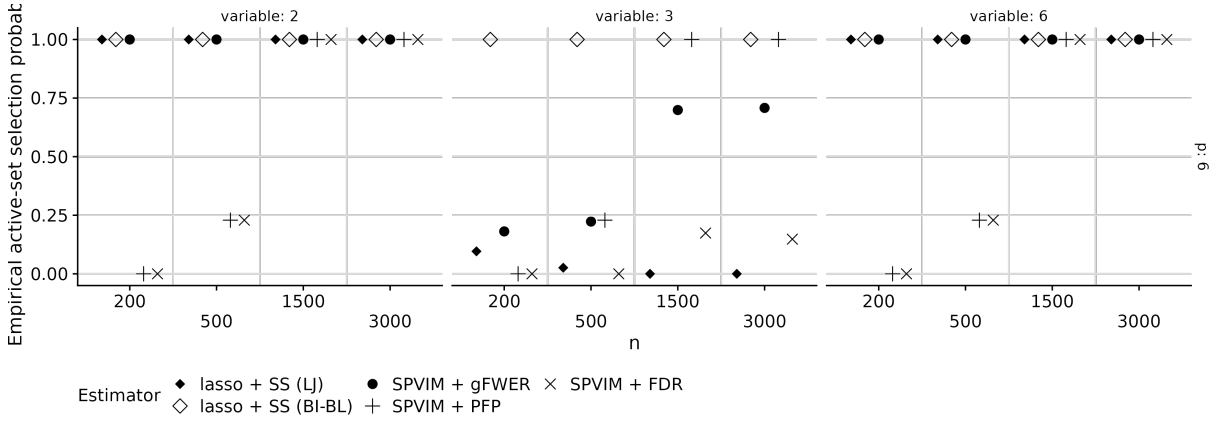

Figure S20: Empirical selection probability for each active-set variable vs n for each estimator, in Scenario 6 (a weak linear model for the outcome and normal features).

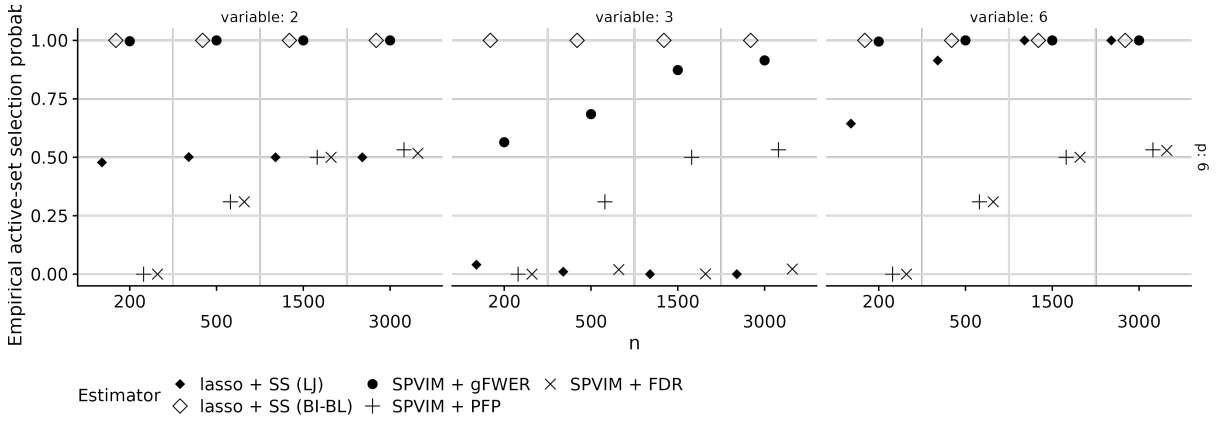

Figure S21: Empirical selection probability for each active-set variable vs n for each estimator, in Scenario 7 (a weak linear model for the outcome and correlated normal features).

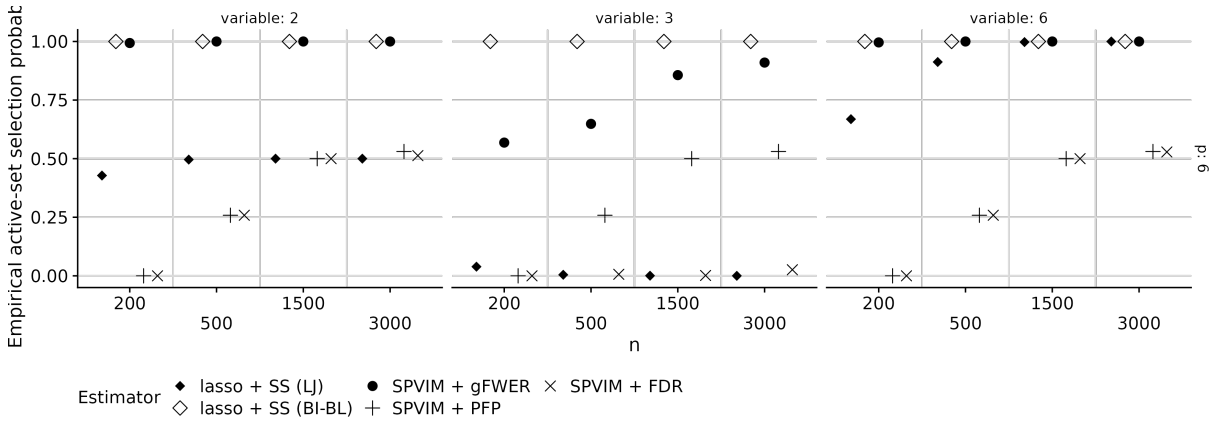

Figure S22: Empirical selection probability for each active-set variable vs n for each estimator, in Scenario 7 (a weak linear model for the outcome and correlated normal features).

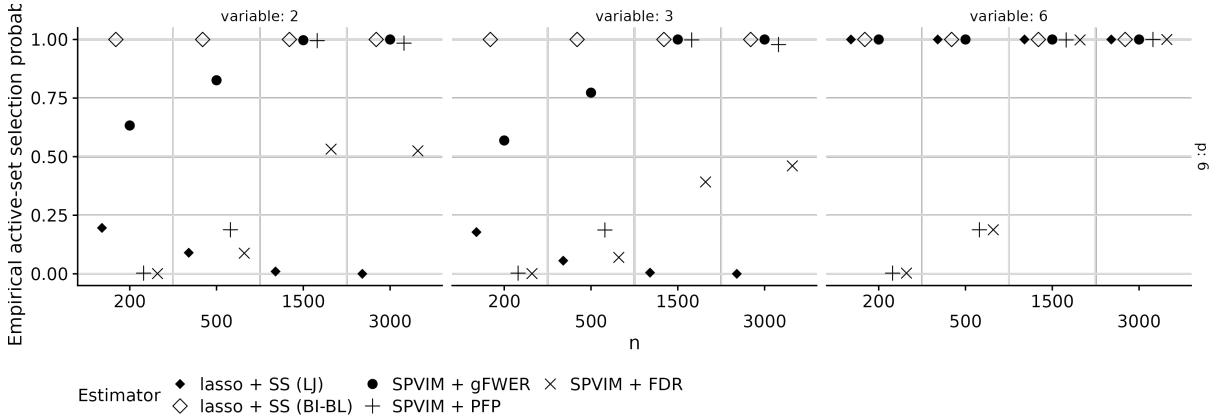

Figure S23: Empirical selection probability for each active-set variable vs n for each estimator, in Scenario 8 (a weak nonlinear model for the outcome and normal features).

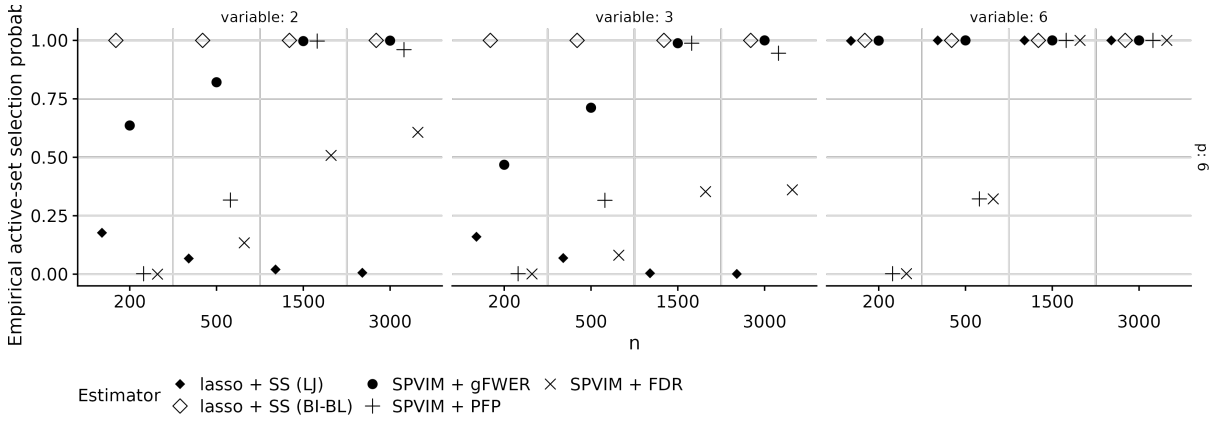

Figure S24: Empirical selection probability for each active-set variable vs n for each estimator, in Scenario 8 (a weak nonlinear model for the outcome and normal features).

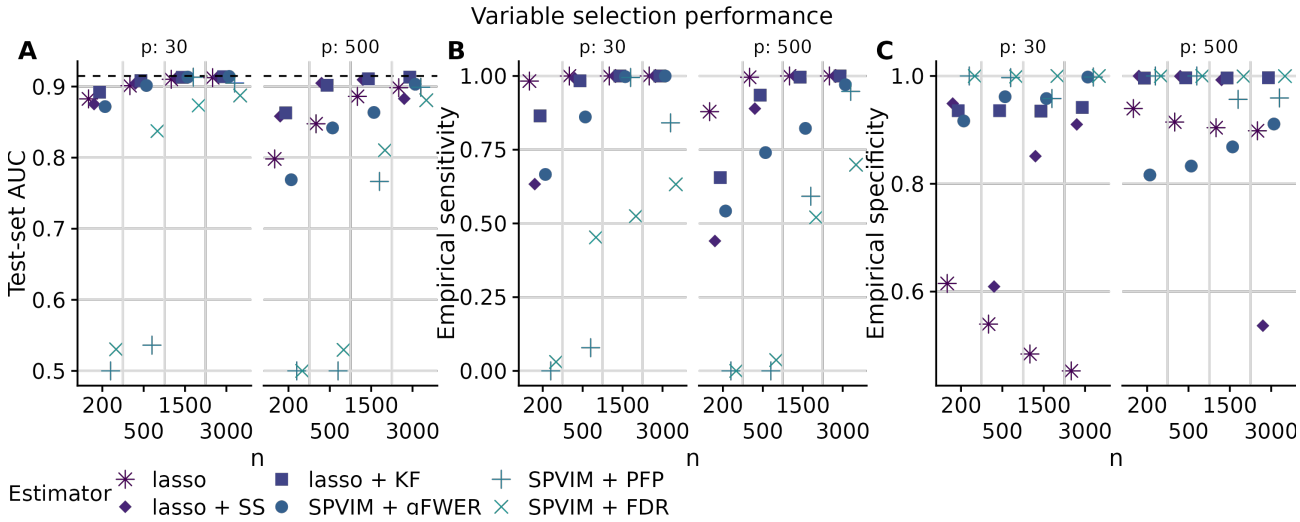

Figure S25: Test-set AUC (panel A) and empirical variable selection sensitivity (panel B) and specificity (panel C) vs n for each estimator and missing data proportion equal to 0, in Scenario 1 (a linear model for the outcome and multivariate normal features). The dotted line in panel A shows the true (optimal) test-set AUC.

similar to the results with missing data: when a linear outcome regression model is correctly specified, our intrinsic procedure tends to perform as well as the lasso-based procedures; when the linear outcome regression model is misspecified, our gFWER-controlling procedure tends to perform better than the lasso-based procedures. In settings with more weakly important variables, our intrinsic procedures continue to perform well. We present the proportion of replications where each variable was selected in Figures S33–S40, again observing similar trends to the missing-data cases.

7.7 Summary of results from Scenarios 1–8

Taken together, these results suggest that (a) as the missing data proportion increases, performance of all procedures tends to degrade; (b) the outcome distribution (linear vs nonlinear) appears to have a larger effect on test-set AUC than the covariate distribution (normal vs non-normal); (c) weakly important variables are less likely to be selected by lasso-based procedures than strongly important variables; and (d) correlation causes further degradation in performance for lasso-based methods. Variable selection performance (sensitivity and specificity) is similar asymptotically across Scenarios 1 and 3–5. This last finding is surprising, since the

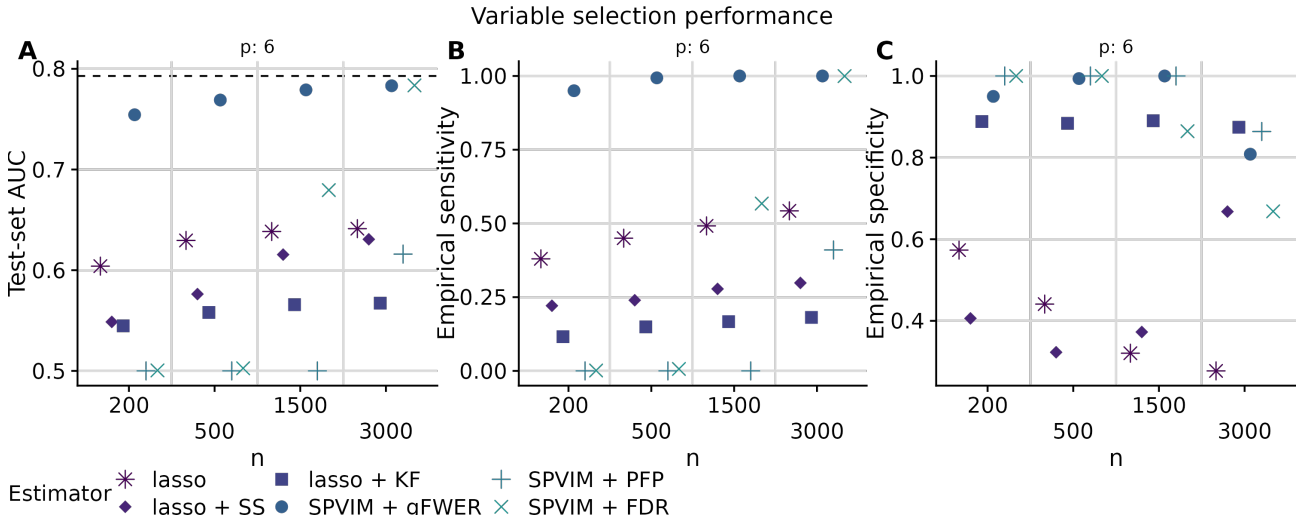

Figure S26: Test-set AUC (panel A) and empirical variable selection sensitivity (panel B) and specificity (panel C) vs n for each estimator and missing data proportion equal to 0, in Scenario 2 (a nonlinear model for the outcome and correlated multivariate normal features), when the data are completely observed. The dotted line in panel A shows the true (optimal) test-set AUC.

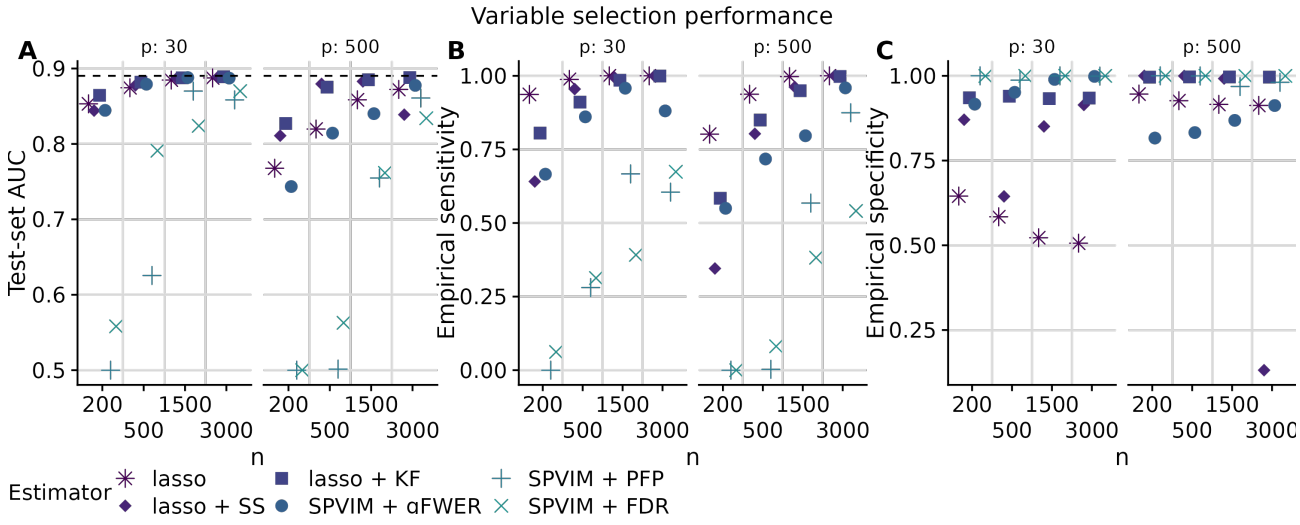

Figure S27: Test-set AUC (panel A) and empirical variable selection sensitivity (panel B) and specificity (panel C) vs n for each estimator and missing data proportion, in Scenario 3 (a linear model for the outcome and nonnormal features), when the data are completely observed. The dotted line in panel A shows the true (optimal) test-set AUC.

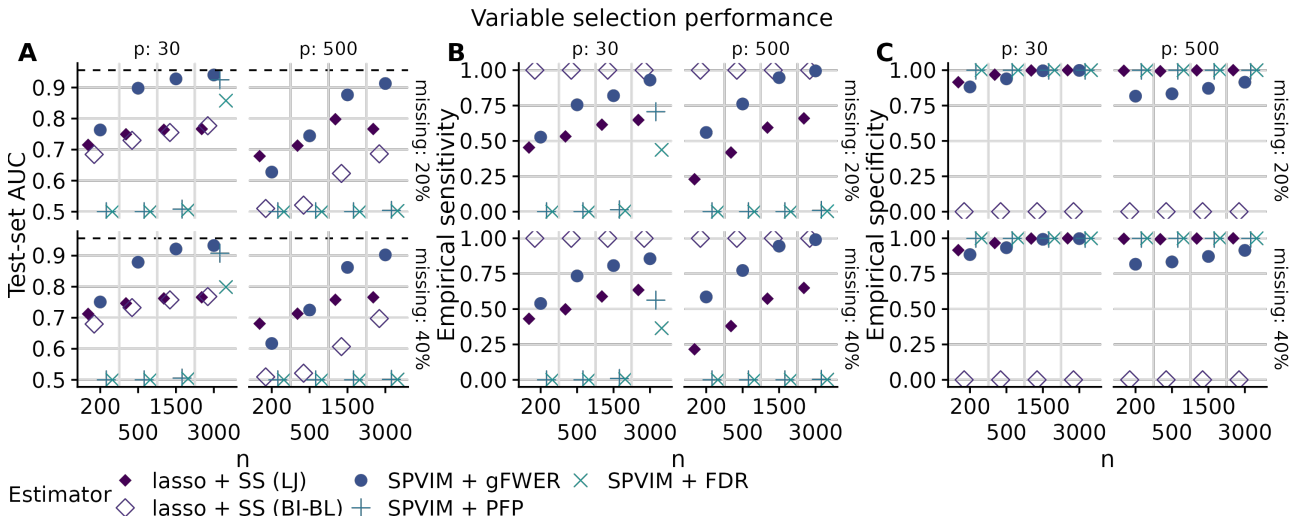

Figure S28: Test-set AUC (panel A) and empirical variable selection sensitivity (panel B) and specificity (panel C) vs n for each estimator and missing data proportion, in Scenario 4 (a nonlinear model for the outcome and normal features), when the data are completely observed. The dotted line in panel A shows the true (optimal) test-set AUC.

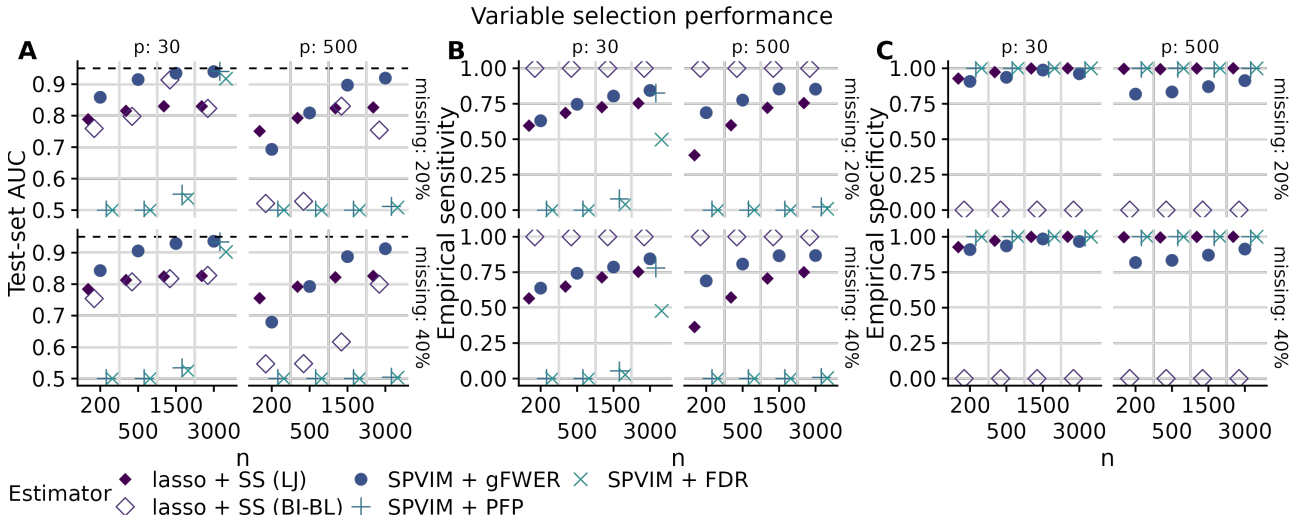

Figure S29: Test-set AUC (panel A) and empirical variable selection sensitivity (panel B) and specificity (panel C) vs n for each estimator and missing data proportion, in Scenario 5 (a nonlinear model for the outcome and nonnormal features), when the data are completely observed. The dotted line in panel A shows the true (optimal) test-set AUC.

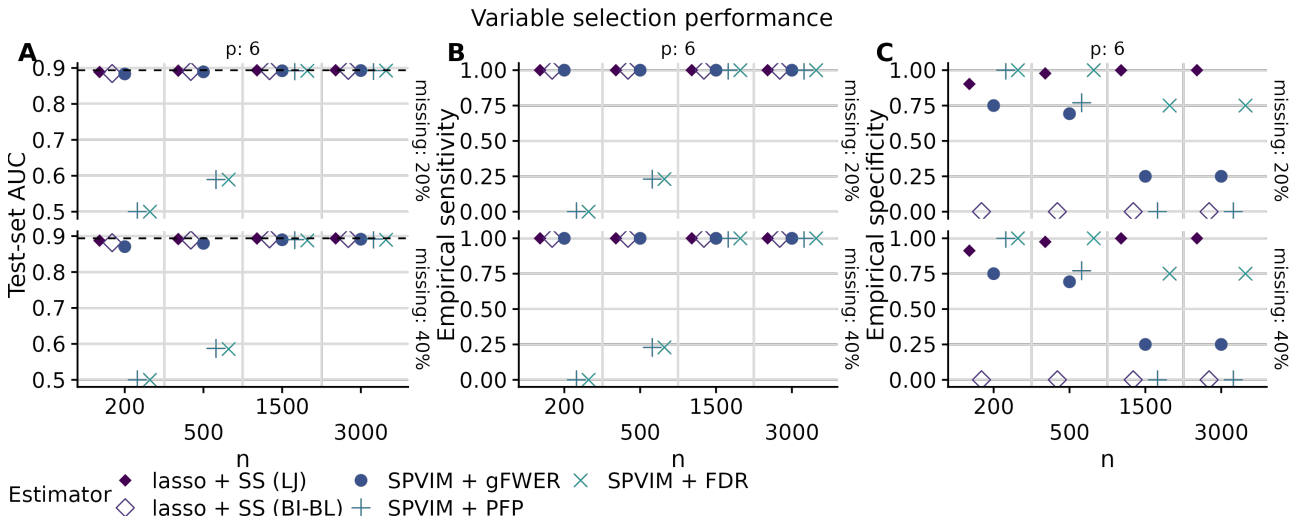

Figure S30: Test-set AUC (panel A) and empirical variable selection sensitivity (panel B) and specificity (panel C) vs n for each estimator and missing data proportion, in Scenario 6 (a weak linear model for the outcome and normal features), when the data are completely observed. The dotted line in panel A shows the true (optimal) test-set AUC.

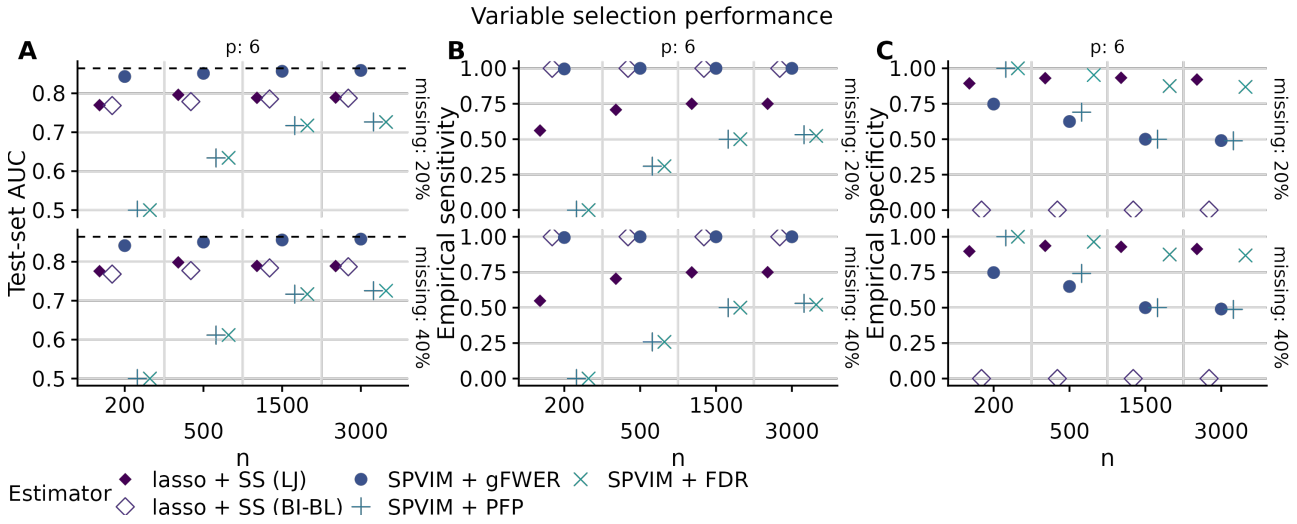

Figure S31: Test-set AUC (panel A) and empirical variable selection sensitivity (panel B) and specificity (panel C) vs n for each estimator and missing data proportion, in Scenario 7 (a weak nonlinear model for the outcome and correlated normal features), when the data are completely observed. The dotted line in panel A shows the true (optimal) test-set AUC.

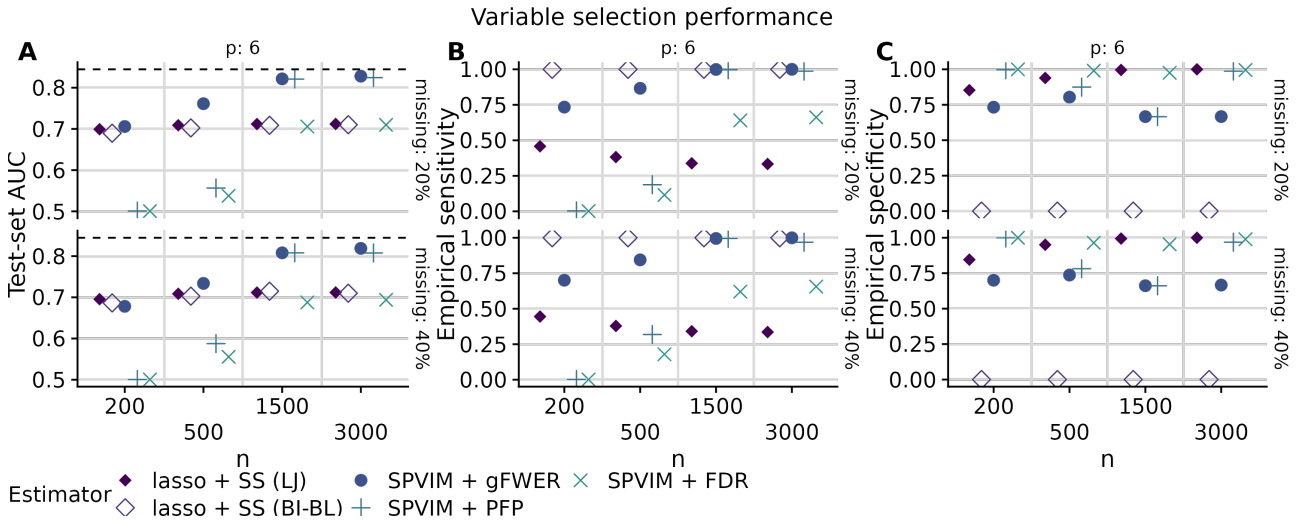

Figure S32: Test-set AUC (panel A) and empirical variable selection sensitivity (panel B) and specificity (panel C) vs n for each estimator and missing data proportion, in Scenario 8 (a weak nonlinear model for the outcome and normal features), when the data are completely observed. The dotted line in panel A shows the true (optimal) test-set AUC.

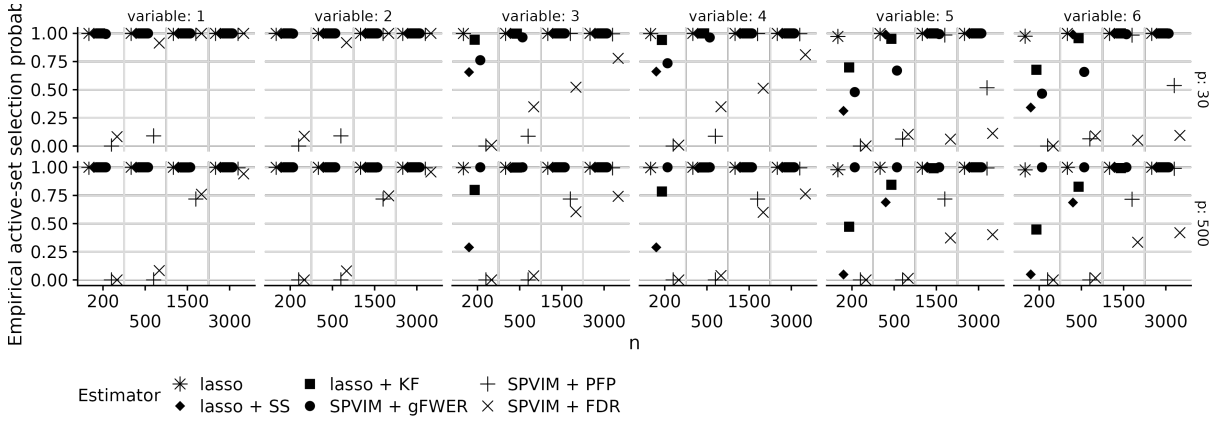

Figure S33: Empirical selection probability for each active-set variable vs n for each estimator and dimension with missing data proportion equal to 0, in Scenario 1 (a linear model for the outcome and multivariate normal features), when the data are completely observed.

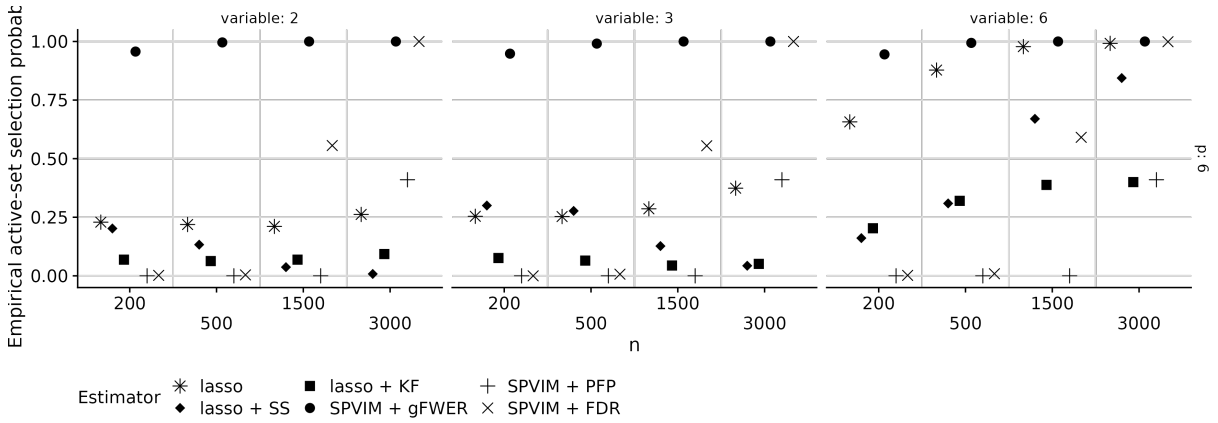

Figure S34: Empirical selection probability for each active-set variable vs n for each estimator, in Scenario 2 (a nonlinear model for the outcome and correlated multivariate normal features), when the data are completely observed.

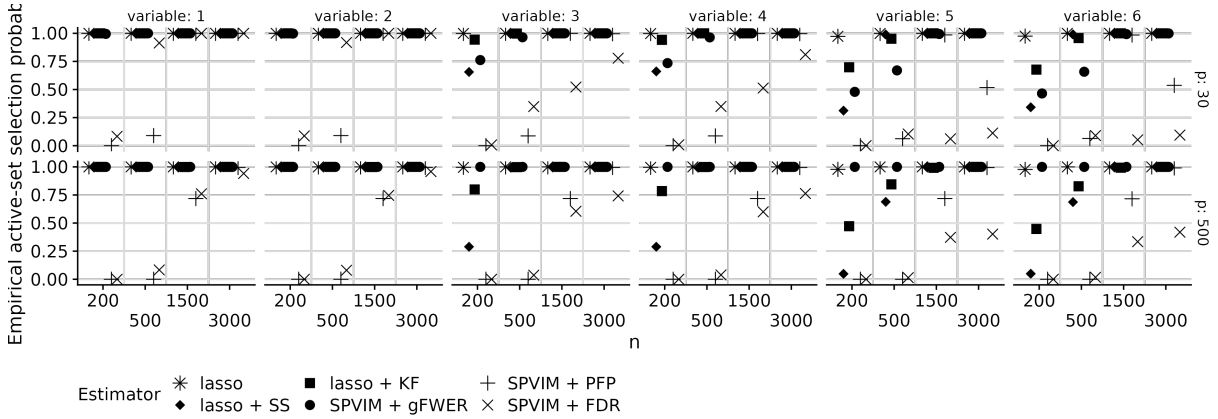

Figure S35: Empirical selection probability for each active-set variable vs n for each estimator and dimension with missing data proportion equal to 0, in Scenario 3 (a linear model for the outcome and nonnormal features), when the data are completely observed.

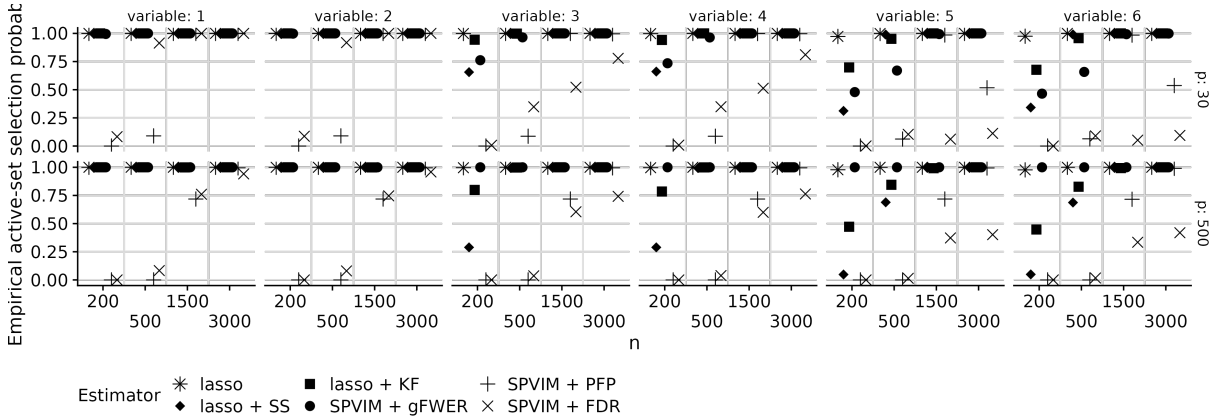

Figure S36: Empirical selection probability for each active-set variable vs n for each estimator and dimension with missing data proportion equal to 0, in Scenario 4 (a nonlinear model for the outcome and multivariate normal features), when the data are completely observed.

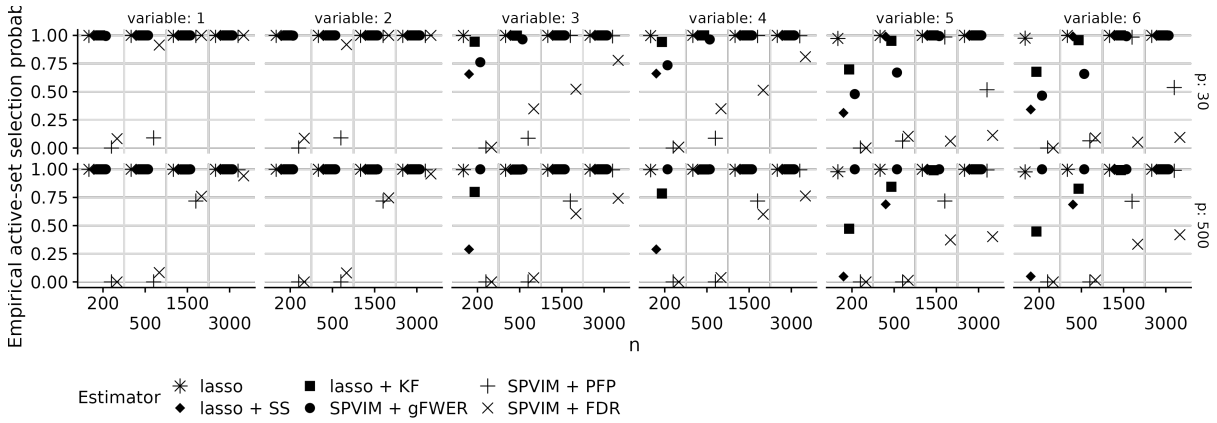

Figure S37: Empirical selection probability for each active-set variable vs n for each estimator and dimension with missing data proportion equal to 0, in Scenario 5 (a nonlinear model for the outcome and nonnormal features), when the data are completely observed.

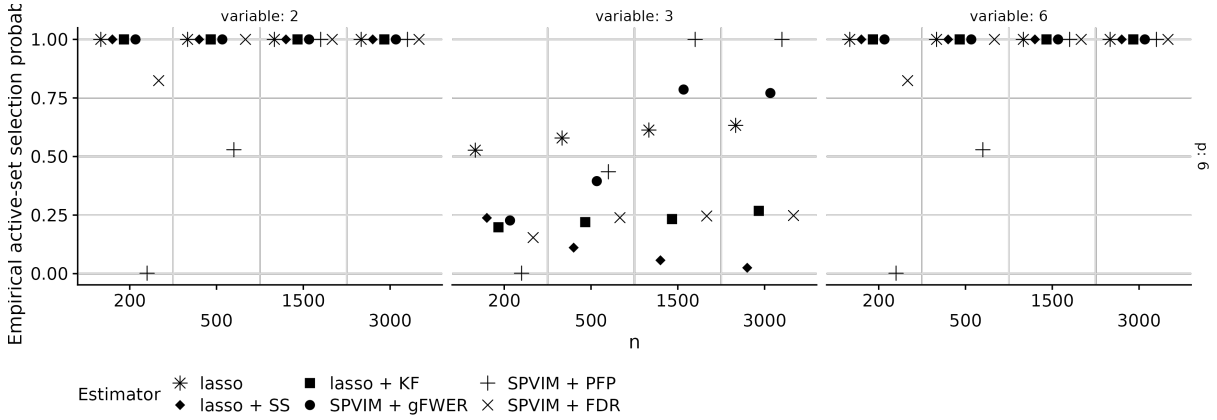

Figure S38: Empirical selection probability for each active-set variable vs n for each estimator, in Scenario 6 (a weak linear model for the outcome and normal features), when the data are completely observed.

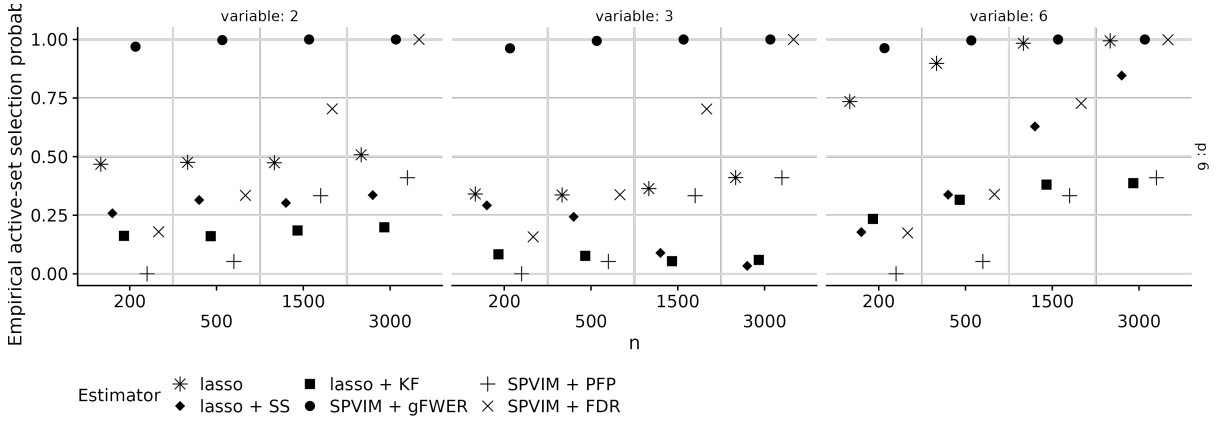

Figure S39: Empirical selection probability for each active-set variable vs n for each estimator, in Scenario 7 (a weak linear model for the outcome and correlated normal features), when the data are completely observed.

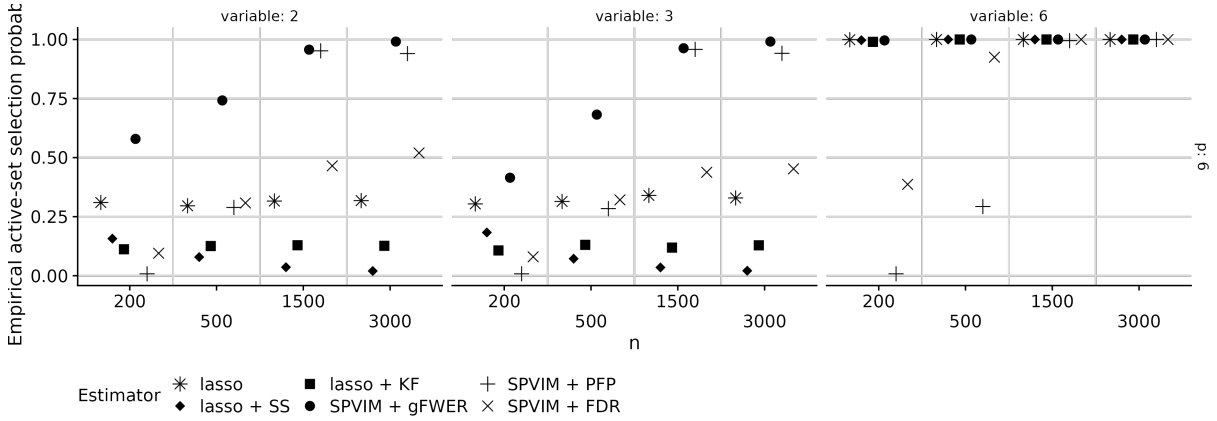

Figure S40: Empirical selection probability for each active-set variable vs n for each estimator, in Scenario 8 (a weak nonlinear model for the outcome and normal features), when the data are completely observed.

variable selection performance of the lasso is not guaranteed in misspecified settings. However, as we saw in Scenarios 2, 7, and 8, in adversarial cases the lasso-based estimators can have poor variable selection performance, as suggested by theory. Additionally, in the plots describing empirical selection probability for lasso-based estimators, we saw that while lasso-based procedures may have good overall selection performance, some important variables may still be missed, even in the non-adversarial settings. In contrast, our intrinsic variable selection procedure is more robust to model misspecification. Finally, we saw that our proposal performs comparably to commonly used variable selection procedures in settings both with and without missing data when lasso-based estimators are correctly specified.

8 Additional details for the pancreatic cancer analysis

We had two overall objectives:

1. separate mucinous cysts from non-mucinous cysts, where a mucinous cyst is thought to have some malignant potential; and

2. separate cysts with high maglinant potential from cysts with low or no malignant potential.

Table S5: All biomarkers of interest for the pancreatic cancer analysis.

Biomarker	Description
CEA	Carcinoembryonic antigen. Serum levels may be elevated in some types of cancer (e.g., colorectal cancer, pancreatic cancer).
CEA mucinous call	Binary indicator of whether CEA > 192.
ACTB	Actin Beta (Hata et al., 2017)
Molecules score	Methylated DNA levels of selected genes (Hata et al., 2017)
Molecules neoplasia call	Binary indicator of whether molecules score > 25
Telomerase score	Telomerase activity measured using telomere repeat amplification protocol (Hata et al., 2016)
Telomerase neoplasia call	Binary indicator of whether telomerase score > 730
AREG score	Amphiregulin (AREG) overexpression (Tun et al., 2012)
AREG mucinous call	Binary indicator of whether AREG score > 112
Glucose score	Glucometer glucose level (Zikos et al., 2015)
Glucose mucinous call	Binary indicator of whether glucose score < 50
Combined mucinuous call	Binary indicator of whether AREG score > 112 and glucose score < 50
Fluorescence score	Fluorescent protease activity (Ivry et al., 2017)
Fluorescence mucinuous call	Binary indicator of whether fluorescence score > 1.23
DNA mucinous call	Presence of mutations in a DNA sequencing panel (Singhi et al., 2018)
DNA neoplasia call (v1)	Binary indicator of methylated DNA levels of selected genes being above a threshold (Majumder et al., 2019)
DNA neoplasia call (v2)	Binary indicator of methylated DNA levels of selected genes being above a threshold (Majumder et al., 2019)
MUC3AC score	Expression of protein Mucin 3AC
MUC5AC score	Expression of protein Mucin 5AC (Cao et al., 2013)
Ab score	Monoclonal antibody reactivity (Das et al., 2014)
Ab neoplasia call	Binary indicator of whether Ab score > 0.104

To meet these objectives, we want to assess both individual biomarkers and panels of biomarkers, both using continuous markers and binary calls.

8.1 Data preprocessing

To create analysis data from the raw data, we selected the following variables: participant ID, institution, the entire set of continuous biomarkers and binary calls (listed in Table S5). The proportion of missing data in the biomarkers ranged from a minimum of 24.5% to a maximum of 68.3%; the median proportion of missing data was 31%.

8.2 Imputing missing data

Our analyses are all based on multiple imputation via chained equations (MICE, implemented in the R package `mice`; van Buuren, 2007; van Buuren and Groothuis-Oudshoorn, 2010). For $i = 1, \ldots, n$ and $j = 1, \ldots, r$ (where $n = 321$ is the sample size and $r = 21$ denotes the total number of biomarkers), we denote the ith measurement of biomarker j by X_{ij} and the outcome of interest by Y_i. We used the following model to impute missing biomarker values:

$$X_{i,j,\text{mis}} \sim Y_i + X_{i,j,\text{obs}} + \text{Institution}_i.$$

These models allow us to relate observed biomarker values (and the institution at which each specimen was collected) to the unobserved biomarker values. All imputations were performed using a maximum of 20 iterations and predictive mean matching (PMM; van Buuren and Groothuis-Oudshoorn, 2010) to create 10 fully-imputed datasets. In some cases, the PMM algorithm failed to converge; in these cases, we used tree-based imputation.

8.3 Variable selection procedures

We use the same variable selection procedures as in the main manuscript: stability selection within bootstrap imputation (denoted by lasso + SS (LJ)) or bootstrap imputation with bolasso for variable selection (denoted by lasso + SS (BI-BL)), with final predictions made using logistic regression; and intrinsic selection designed to control the gFWER, PFP, and FDR, both with and without using Rubin's Rules via Lemma 1 (denoted SPVIM + {gFWER, PFP, FDR}, respectively), with final predictions made using the Super Learner, with library described in Table S6. We based tuning parameter selection on a similar setting from the simulations: in this case, the sample size is 321 and there are 21 biomarkers, so we set $k = 5$, $q = 0.8$, the number of variables selected in each bootstrap run of stability selection equal to 9 (based on a target per-family error rate of $p(0.04)$ and threshold of 0.9).

8.4 Assessing prediction performance

Assessing prediction performance is complicated by both the imputation step and the initial variable selection step. To address this, we performed imputation within cross-fitting within Monte-Carlo sampling; this provides an unbiased assessment of the entire procedure, from imputation to variable selection to prediction. More specifically, for each of 100 replicates and each outcome, we performed the procedure outlined in Algorithm 2.

Algorithm 2 Imputation and pooled variable selection within cross-fitting and Monte-Carlo sampling

for $b = 1, \ldots, 50$ **do**
  generate a random vector $B_n \in \{1, \ldots, 5\}^n$ by sampling uniformly from $\{1, \ldots, 5\}$ with replacement, and for each $v \in \{1, \ldots, 5\}$, denote by D_v the data with index in $\{i : B_{n,i} = v\}$;
  for $v = 1, \ldots, 5$ **do**
    if using a bootstrap imputation-based procedure, create 100 bootstrap datasets based on the data in $\cup_{j \neq v} D_j$ and a single imputed dataset for each;
    create 10 imputed datasets $\{Z_{k,-v}\}_{k=1}^{10}$ based on the data in $\cup_{j \neq v} D_j$ using MICE;
    create 10 imputed datasets $\{Z_{k,v}\}_{k=1}^{10}$ based on the data in D_v using MICE;
    apply the chosen variable selection procedure on the training data, resulting in a final set of selected variables S_v;
    for $k = 1, \ldots, 10$ **do**
      train the chosen prediction algorithm on the training data $Z_{k,-v}$ using only variables in S_v;
      obtain $\text{AUC}_{k,v}$ and its associated variance $\text{var(AUC)}_{k,v}$ by predicting on the withheld test data $Z_{k,v}$ and measure prediction performance using AUC;
    end for
    combine the AUCs and associated variance estimators into AUC_v and var(AUC)_v using Rubin's rules;
  end for
  compute $\text{CV-AUC}_b = \frac{1}{5} \sum_{v=1}^{v} \text{AUC}_v$ and $\text{var(CV-AUC)}_b = \frac{1}{5} \sum_{v=1}^{v} \text{var(AUC)}_v$;
end for
compute overall performance by averaging over the Monte-Carlo iterations.

8.5 Obtaining a final set of selected biomarkers

We obtain a final set of selected biomarkers by applying the variable selection procedure to the full set of observations for each imputed dataset.

Candidate Learner	R Implementation	Tuning Parameter and possible values	Tuning parameter description
Random forests	`ranger`	`max.depth` $\in \{1, 10, 20, 30, 100, \infty\}$	Maximum tree depth
Gradient boosted trees	`xgboost`	`max.depth` $= \{4\}$ `nrounds` $\in \{100, 500, 2000\}$	Maximum tree depth Number of boosting iterations
Elastic net	`glmnet`	mixing parameter α $\in \{0, \frac{1}{4}, \frac{1}{2}, \frac{3}{4}, 1\}$	Trade-off between ℓ_1 and ℓ_2 regularization‡

Table S6: Candidate learners in the Super Learner ensemble for the pancreatic cyst data analysis along with their R implementation, tuning parameter values, and description of the tuning parameters. All tuning parameters besides those listed here are set to their default values. In particular, the random forests are grown with `mtry` $= \sqrt{p}^\dagger$, a minimum node size of 5 for continuous outcomes and 1 for binary outcomes, and a subsampling fraction of 1; the boosted trees are grown with shrinkage rate of 0.1 and a minimum of 10 observations per node; and the ℓ_1 tuning parameter for the elastic net is determined via 10-fold cross-validation.
†: p denotes the total number of predictors.

8.6 Super Learner specification

As in the simulations, we used a different specification for the internal Super Learner in the intrinsic selection procedure (max. depth 4 boosted trees (all tuning parameter values in Table S6) with pre-screening via univariate rank correlation with the outcome) and all other Super Learners (Table S6). In all cases, the final Super Learner fit for prediction performance of the selected set of variables used the candidate learners in Table S6.

9 Additional results from the pancreatic cyst analysis

In the main manuscript, we performed an analysis with goal of predicting whether a cyst was mucinous, using Algorithm 2 to assess prediction performance. In Table S7, we present the biomarkers selected using each procedure. Here, we show results using this same algorithm for the outcome of whether a cyst has high malignancy potential.

We present the results of our analysis in Figure S41 and Table S8. In Figure S41, we see that the PFP- and FDR-controlling intrinsic selection procedures again select no variables, on average, as we saw in the analysis of the mucinous outcome in the main manuscript. Prediction

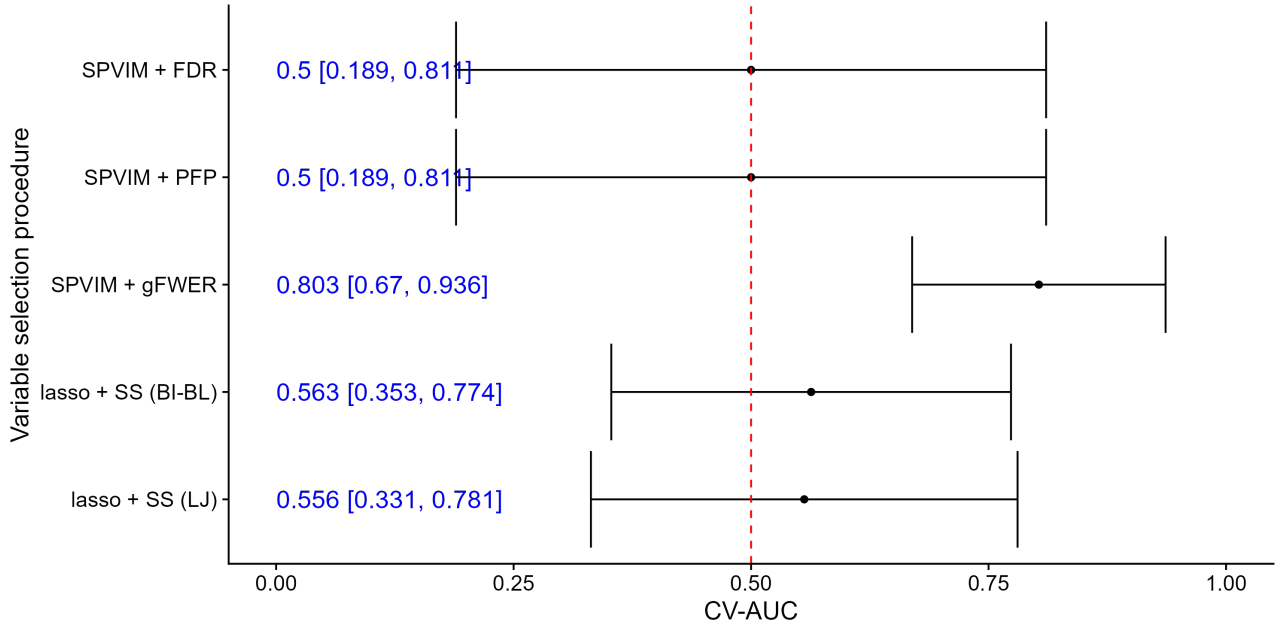

Figure S41: Cross-validated area under the receiver operating characteristic curve (CV-AUC) for predicting whether a cyst has high malignancy potential averaged over 100 replicates of the imputation-within-cross-validated procedure (Algorithm 2) for each variable selection algorithm. Prediction performance for lasso-based methods is based on logistic regression on the selected variables, while performance for Super Learner-based methods is based on a Super Learner. Error bars denote 95% confidence intervals based on the average variance over the 100 replications.

performance is also poor for the lasso-based estimators. Compared to the mucinous outcome, we observe reduced prediction performance for the gFWER-controlling intrinsic selection procedure, with an estimated cross-validated AUC of 0.803 (95% confidence interval [0.67, 0.936]). In Table S8, we display the final set of biomarkers selected by each procedure. Several biomarkers are selected across all two or more procedures that selected any variables on the full dataset. An antibody score was selected across all three procedures. Variables appearing in two or more procedures included an ACTB score, four neoplasia calls (binary variables), a glucose score, a combined amphiregulin- and glucose-based mucinous call, a fluorescence score and its associated mucinous call, and an antibody-based neoplasia call. Selection across the majority of procedures suggests that these variables may useful for predicting whether a cyst has high malignancy potential.

Table S7: Biomarkers selected by each selection procedure for predicting whether a cyst is mucinous on the full imputed dataset. Full definitions of each variable are provided in the Supplementary Material.

Biomarker	lasso + SS (LJ)	lasso + SS (BI-BL)	SPVIM + gFWER	Number of procedures
CEA	No	Yes	No	1
CEA mucinous call	No	Yes	No	1
ACTB	No	Yes	No	1
Molecules (M) score	No	Yes	No	1
M neoplasia call	No	Yes	Yes	2
Telomerase (T) score	No	Yes	No	1
T neoplasia call	No	Yes	No	1
AREG (A) score	Yes	Yes	Yes	3
A mucinous call	No	Yes	No	1
Glucose (G) score	No	Yes	Yes	2
G mucinous call	Yes	Yes	Yes	3
A and G mucinous call	Yes	Yes	Yes	3
Fluorescence (F) score	Yes	Yes	Yes	3
F mucinous call	No	Yes	Yes	2
DNA mucinous call	No	Yes	No	1
DNA neoplasia call (v1)	No	Yes	No	1
DNA neoplasia call (v2)	No	Yes	Yes	2
MUC3AC score	Yes	Yes	Yes	3
MUC5AC score	No	Yes	No	1
Ab score	No	Yes	No	1
Ab neoplasia call	No	Yes	Yes	2

Table S8: Biomarkers selected by each selection procedure for predicting whether a cyst has high malignancy potential on the full imputed dataset. Full definitions of each variable are provided in Table S5.

Biomarker	lasso + SS (LJ)	lasso + SS (BI-BL)	SPVIM + gFWER	Number of procedures
CEA	No	Yes	No	1
CEA mucinous call	No	Yes	No	1
ACTB	No	Yes	Yes	2
Molecules (M) score	No	Yes	No	1
M neoplasia call	No	Yes	Yes	2
Telomerase (T) score	No	Yes	No	1
T neoplasia call	Yes	Yes	No	2
AREG (A) score	No	Yes	No	1
A mucinous call	No	Yes	No	1
Glucose (G) score	No	Yes	Yes	2
G mucinous call	No	Yes	No	1
A and G mucinous call	No	Yes	Yes	2
Fluorescence (F) score	No	Yes	Yes	2
F mucinous call	No	Yes	Yes	2
DNA mucinous call	No	Yes	No	1
DNA neoplasia call (v1)	No	Yes	Yes	2
DNA neoplasia call (v2)	No	Yes	Yes	2
MUC3AC score	No	Yes	No	1
MUC5AC score	No	Yes	No	1
Ab score	Yes	Yes	Yes	3
Ab neoplasia call	No	Yes	Yes	2