Supplementary materials for *Variable selection with multiply-imputed datasets: choosing between stacked and grouped methods*

Jiacong Du[1], Jonathan Boss[1], Peisong Han[1], Lauren J Beesley[1], Michael Kleinsasser[1]
Stephen A Goutman[2], Stuart Batterman[3], Eva L Feldman[2], Bhramar Mukherjee[1]
[1]Department of Biostatistics, University of Michigan, Ann Arbor, MI
[2]Department of Neurology, University of Michigan, Ann Arbor, MI
[3]Department of Environmental Health Science, University of Michigan, Ann Arbor, MI
Corresponding author Jiacong Du jiacong@umich.edu

January 3, 2022

Table S1: Summary of methods for handling missing data and variable selection in the literature.

Class	Method	Key reference First author, year	Method description	Limitations
Sequential handling missing data by multiple imputation and variable selection	Ad-hoc thresholding	Wood et al. (2008)	First apply the existing variable selection method, e.g., stepwise selection, on each imputed dataset. A variable is considered selected if it is selected in at least πD imputed datasets, $0 < \pi \leq 1$.	No clear guideline on how to choose the thresholding parameter π.
	Bootstrapped imputation and variable selection	Long and Johnson (2015)	This method imputes missing values in the bootstrapped samples from the original data and applies penalized regression on each bootstrapped imputed dataset. Proportions of each covariate selected over all bootstrap imputed datasets are provided and the final active set is determined by a thresholding.	1.Coefficient estimates can only be obtained after the final active set is determined. 2.Still requires ad-hoc thresholding.
	Elastic net regularization on the stacked imputed datasets	Wan et al. (2015)	Simultaneously performs variable selection and regression coefficient estimation by maximizing the data likelihood after stacking all imputed datasets subject to elastic net penalization. Each data point is weighted by an observational weight which quantifies the proportion of observed covariates out of the total number of covariates.	1.Only continuous outcomes are considered in the paper. 2.Using the proposed observational weights can introduce bias if the missing data mechanism is dependent on the outcome.
	Group LASSO regularization on the pooled likelihood for imputed datasets	Chen and Wang (2013)	Group LASSO penalized optimization over all imputed datasets, where the groups are defined by regression coefficients corresponding to the same variable across all imputed datasets. Variable selection and point estimation are obtained simultaneously.	1. Existing optimization algorithm uses local quadratic approximation, which is reliant on a subjective hard thresholding rule to select variables. 2.The existing implementation is limited to continuous outcomes and a LASSO penalty function.
Simultaneously handling missing data and variable selection	Bayesian variable selection by data augmentation strategy	Yang et al. (2005)	The Simultaneously Impute and Select method jointly models the missing data mechanism, model parameters, and selection indicators in a Bayesian framework, and samples joint posterior draws using Gibbs sampler.	Can be computationally intensive for a large amount of missingness.
	Penalized maximum likelihood estimators combined with EM algorithm	Garcia et al. (2010)	Variable selection and coefficient estimates are simultaneously obtained by maximizing the penalized observed data likelihood. SCAD and ALASSO penalty functions have been studied. Additional information criteria involving penalized observed data likelihood have been developed for choosing tuning parameters.	Computational challenges for integrating the augmented likelihood with no closed forms over the missing data distribution to find the observed data likelihood.
	Penalized estimating equations combined with IPW/AIPW	Johnson et al. (2008)	Missing data is handled by IPW for each complete-observed data point, and variable selection is achieved by solving the penalized estimating equations.	Assumes that the missing data pattern is monotone.

Table S2: Missing data proportion for 30 variables in the ALS data. The data in total contains 266 observations with 167 cases and 99 controls.

Variables	**Proportion**	**Variables**	**Proportion**
PCB 174	0.508	PCB 153	0.237
PCB 110	0.429	PCB 202	0.199
cis-chlordane	0.398	cis-nonachlor	0.147
trans-nonachlor	0.380	beta-HCH	0.132
PCB 175	0.368	PBDE 99	0.086
trans-chlordane	0.365	p,p'-DDE	0.083
PCB 118	0.361	PeCB	0.060
PCB 180	0.350	PBDE 100	0.056
PBDE 28	0.305	Education1	0.038
PCB 138	0.278	Education2	0.038
PBDE 154	0.267	PCB 151	0.034
PBDE 153	0.256	PBDE 47	0.015
BMI	0.244	Age	0.011
BMI_slope	0.244	Sex	0.000
PBDE 85	0.241	ALS	0.000

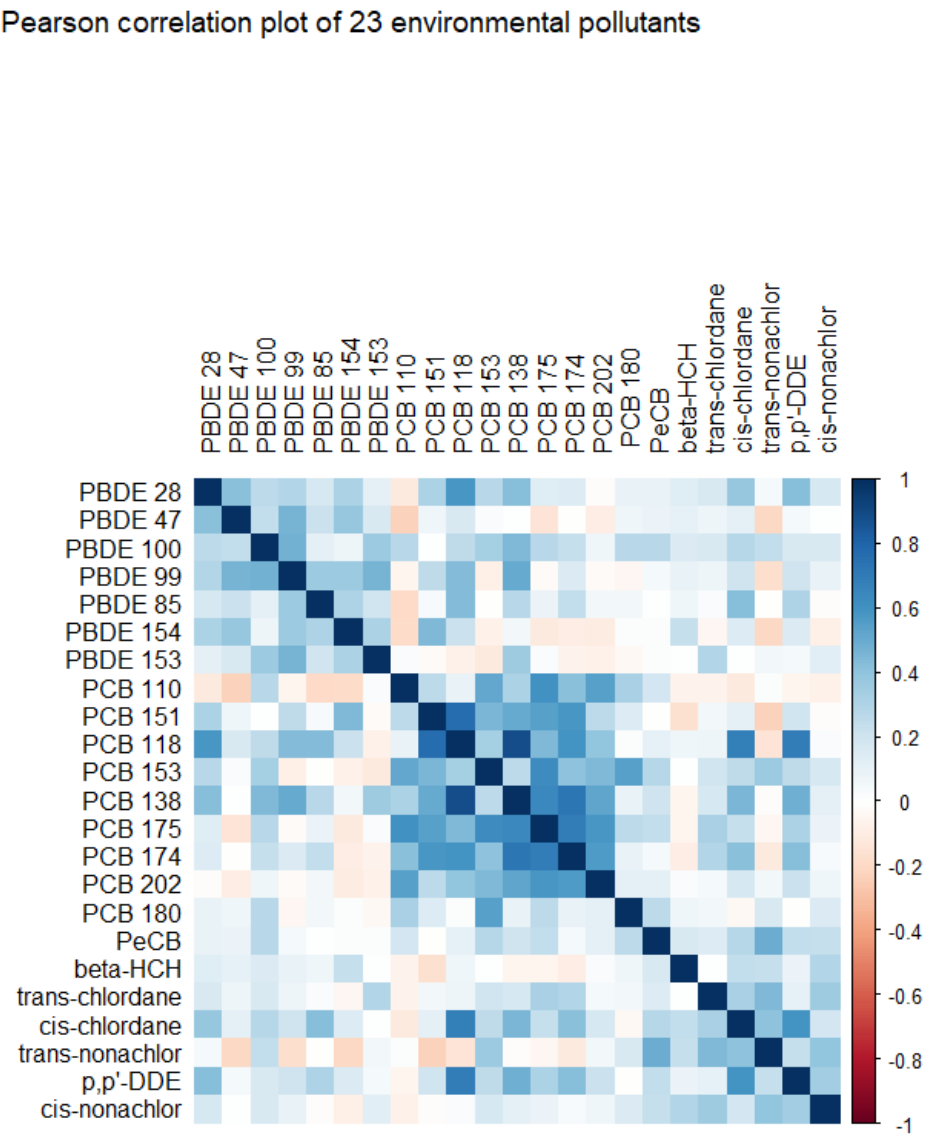

Figure S1: Pairwise Pearson correlation matrix for 23 POPs collected through the University of Michigan ALS Patient Biorepository. The dataset contains 266 observations (167 cases and 99 controls).

Table S3: Regression coefficient estimates for five POPs collected as part of the University of Michigan ALS Patients Biorepository case-control study (167 ALS cases and 99 healthy controls). Results are based on 10 imputed datasets. Only five of the 23 POPs are displayed because the other 18 POPs were not selected by any method.

POPs	SLASSO	SaLASSO	SENET	SaENET	SLASSO (w)	SaLASSO (w)	SENET (w)	SaENET (w)	GLASSO	GaLASSO
PBDE 153	0.087	-	0.095	-	0.090	0.075	0.099	0.067	0.081	-
PeCB	0.316	0.751	0.354	0.743	0.268	0.645	0.307	0.655	0.295	0.969
trans-chlordane	0.067	0.001	0.104	0.003	0.068	0.081	0.105	0.068	0.047	-
cis-nonachlor	0.220	0.578	0.278	0.572	0.225	0.524	0.284	0.530	0.185	0.194
PCB151	0.060	0.173	0.109	0.171	0.035	-	0.090	-	0.031	-
# selected	**5**	**4**	**5**	**4**	**5**	**4**	**5**	**4**	**5**	**2**
# removed	**18**	**19**	**18**	**19**	**18**	**19**	**18**	**19**	**18**	**21**

Table S4: Sensitivity (SENS), specificity (SPEC), MSE for non-null coefficients ($MSE_{non-null}$), and MSE for null coefficients (MSE_{null}) for simulation Cases 1-4 with 50 imputed datasets. Cases 1 and 2 have 500 observations with 20 covariates, while Cases 3 and 4 have 1000 observations with 100 covariates. Cases 1 and 3 correspond to concentrated signals and Cases 2 and 4 correspond to distributed signals.

	SLASSO	SaLASSO	SENET	SaENET	SLASSO (w)	SaLASSO (w)	SENET (w)	SaENET (w)	GLASSO	GaLASSO
					Case 1					
$SENS$	1.00	1.00	1.00	1.00	1.00	0.99	1.00	0.99	0.98	0.94
$SPEC$	0.53	0.93	0.53	0.92	0.55	0.93	0.54	0.93	0.49	0.98
$MSE_{non-null}$	1.51	0.86	1.50	0.84	1.58	0.86	1.57	0.83	2.26	1.51
MSE_{null}	0.20	0.12	0.20	0.12	0.22	0.12	0.22	0.13	0.14	0.06
					Case 2					
$SENS$	0.98	0.92	0.98	0.92	0.98	0.91	0.98	0.92	0.95	0.80
$SPEC$	0.57	0.94	0.56	0.93	0.59	0.94	0.58	0.94	0.56	0.98
$MSE_{non-null}$	1.75	1.47	1.74	1.44	1.81	1.53	1.79	1.46	2.66	2.43
MSE_{null}	0.18	0.12	0.18	0.13	0.19	0.12	0.20	0.13	0.12	0.08
					Case 3					
$SENS$	1.00	0.98	1.00	0.98	1.00	0.98	1.00	0.98	0.98	0.77
$SPEC$	0.34	0.73	0.34	0.72	0.43	0.80	0.42	0.79	0.10	0.86
$MSE_{non-null}$	1.04	0.52	1.03	0.55	1.21	0.51	1.19	0.55	3.30	2.02
MSE_{null}	1.08	1.47	1.11	1.44	0.87	1.23	0.91	1.18	29.70	1.09
					Case 4					
$SENS$	0.95	0.90	0.95	0.90	0.95	0.89	0.95	0.90	0.94	0.77
$SPEC$	0.39	0.75	0.38	0.75	0.48	0.81	0.47	0.81	0.17	0.95
$MSE_{non-null}$	1.21	0.93	1.19	0.90	1.31	0.97	1.29	0.93	2.28	2.46
MSE_{null}	0.95	1.31	0.99	1.25	0.78	1.16	0.82	1.08	6.75	0.39

Table S5: Sensitivity (SENS), specificity (SPEC), MSE for non-null coefficients ($MSE_{non-null}$), and MSE for null coefficients (MSE_{null}) for simulation Cases 1-4 with 50 imputed datasets. Cases 1 and 2 have 500 observations with 20 covariates, while Cases 3 and 4 have 1000 observations with 100 covariates. Cases 1 and 3 correspond to concentrated signals and Cases 2 and 4 correspond to distributed signals. For each measure, SLASSO is used as the benchmark and the ratio to SLASSO is presented.

	SLASSO	SaLASSO	SENET	SaENET	SLASSO (w)	SaLASSO (w)	SENET (w)	SaENET (w)	GLASSO	GaLASSO
					Case 1					
$SENS$	1.00	1.00	1.00	1.00	1.00	0.99	1.00	0.99	0.98	0.94
$SPEC$	1.00	1.74	0.99	1.74	1.02	1.75	1.02	1.74	0.91	1.83
$MSE_{non-null}$	1.00	0.57	0.99	0.56	1.05	0.57	1.04	0.55	1.50	1.00
MSE_{null}	1.00	0.58	1.01	0.61	1.10	0.60	1.11	0.66	0.70	0.32
					Case 2					
$SENS$	1.00	0.93	1.00	0.93	1.00	0.92	1.00	0.93	0.96	0.82
$SPEC$	1.00	1.65	0.99	1.64	1.03	1.66	1.02	1.65	0.99	1.72
$MSE_{non-null}$	1.00	0.84	0.99	0.82	1.04	0.87	1.02	0.83	1.52	1.39
MSE_{null}	1.00	0.69	1.02	0.71	1.06	0.67	1.08	0.74	0.66	0.45
					Case 3					
$SENS$	1.00	0.99	1.00	0.99	1.00	0.98	1.00	0.98	0.98	0.77
$SPEC$	1.00	2.14	0.99	2.12	1.26	2.33	1.24	2.32	0.30	2.51
$MSE_{non-null}$	1.00	0.50	0.99	0.53	1.16	0.48	1.14	0.52	3.16	1.94
MSE_{null}	1.00	1.36	1.02	1.33	0.81	1.13	0.84	1.09	27.42	1.01
					Case 4					
$SENS$	1.00	0.95	1.00	0.96	1.00	0.94	1.00	0.95	1.00	0.81
$SPEC$	1.00	1.95	0.98	1.95	1.24	2.11	1.21	2.11	0.43	2.46
$MSE_{non-null}$	1.00	0.77	0.99	0.74	1.08	0.80	1.06	0.77	1.89	2.03
MSE_{null}	1.00	1.38	1.04	1.32	0.82	1.22	0.87	1.13	7.11	0.42

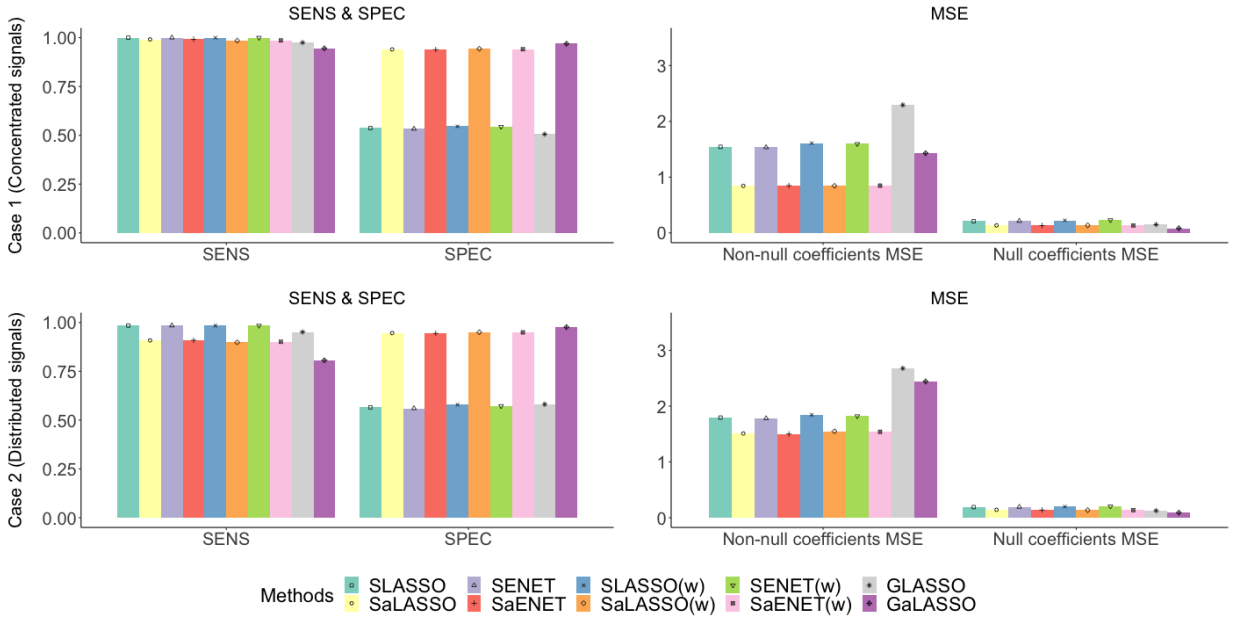

Figure S2: Simulation results for Case 1 (top panel) and Case 2 (bottom panel) where n=500 and p=20 for 10 imputed datasets. Sensitivity (SENS) and specificity (SPEC) are on the left and MSE for non-null and null coefficients are on the right.

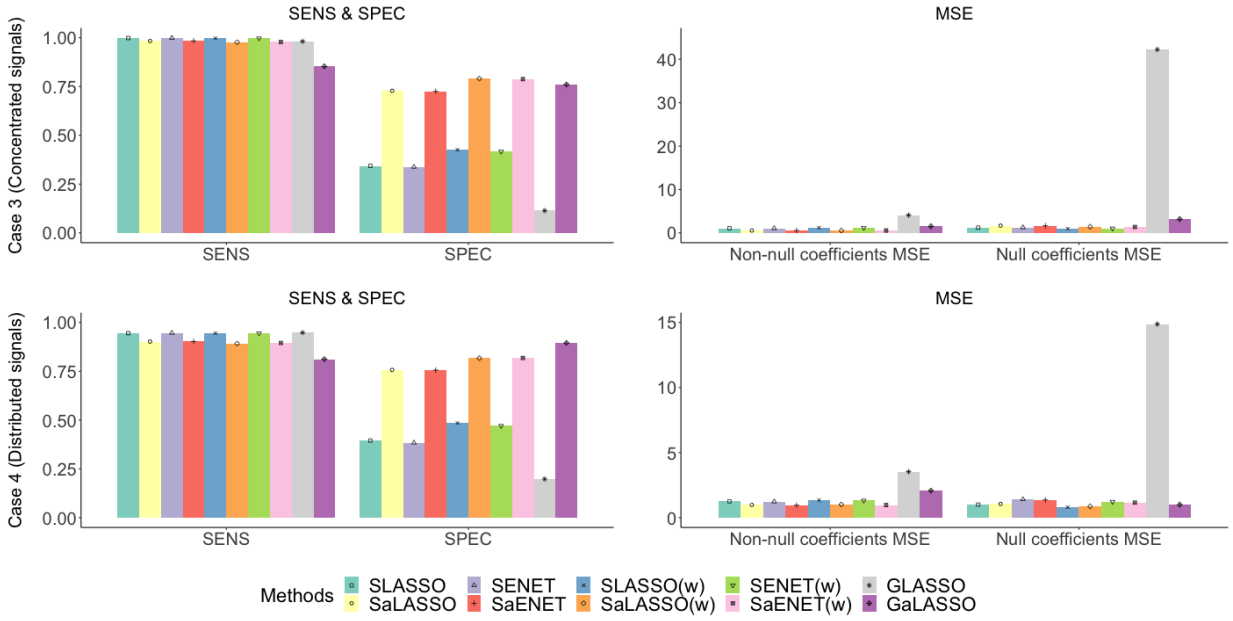

Figure S3: Simulation results for Case 3 (top panel) and Case 4 (bottom panel) where n=1000 and p=100 for 10 imputed datasets. Sensitivity (SENS) and specificity (SPEC) are on the left and MSE for non-null and null coefficients are on the right.

Table S6: Sensitivity (SENS), specificity (SPEC), MSE for non-null coefficients ($MSE_{non-null}$), and MSE for null coefficients (MSE_{null}) for simulation Cases 1-4 with 10 imputed datasets. Cases 1 and 2 have 500 observations with 20 covariates, while Cases 3 and 4 have 1000 observations with 100 covariates. Cases 1 and 3 correspond to concentrated signals and Cases 2 and 4 correspond to distributed signals.

	SLASSO	SaLASSO	SENET	SaENET	SLASSO (w)	SaLASSO (w)	SENET (w)	SaENET (w)	GLASSO	GaLASSO
					Case 1					
$SENS$	1.00	0.99	1.00	0.99	1.00	0.98	1.00	0.99	0.98	0.94
$SPEC$	0.54	0.94	0.53	0.94	0.55	0.94	0.54	0.94	0.51	0.97
$MSE_{non-null}$	1.54	0.84	1.53	0.85	1.61	0.84	1.60	0.85	2.29	1.43
MSE_{null}	0.21	0.14	0.21	0.13	0.22	0.14	0.23	0.13	0.15	0.08
					Case 2					
$SENS$	0.98	0.91	0.98	0.91	0.98	0.90	0.98	0.90	0.95	0.81
$SPEC$	0.56	0.95	0.56	0.94	0.58	0.95	0.57	0.95	0.58	0.98
$MSE_{non-null}$	1.79	1.51	1.78	1.50	1.84	1.55	1.82	1.54	2.68	2.44
MSE_{null}	0.19	0.14	0.19	0.14	0.20	0.14	0.20	0.14	0.13	0.09
					Case 3					
$SENS$	1.00	0.98	1.00	0.98	1.00	0.98	1.00	0.98	0.98	0.85
$SPEC$	0.34	0.73	0.34	0.72	0.43	0.79	0.42	0.79	0.11	0.76
$MSE_{non-null}$	1.07	0.55	1.05	0.56	1.22	0.54	1.20	0.55	4.09	1.54
MSE_{null}	1.22	1.68	1.26	1.64	0.99	1.41	1.05	1.37	42.27	3.19
					Case 4					
$SENS$	0.94	0.90	0.95	0.90	0.94	0.89	0.94	0.90	0.95	0.81
$SPEC$	0.40	0.76	0.38	0.76	0.48	0.82	0.47	0.82	0.20	0.90
$MSE_{non-null}$	1.26	0.98	1.24	0.96	1.35	1.02	1.32	0.97	3.53	2.08
MSE_{null}	1.00	1.42	1.06	1.36	0.80	1.23	0.87	1.16	14.87	1.01

Table S7: Sensitivity (SENS), specificity (SPEC), MSE for non-null coefficients ($MSE_{non-null}$), and MSE for null coefficients (MSE_{null}) for simulation Cases 1-4 with 10 imputed datasets. Cases 1 and 2 have 500 observations with 20 covariates, while Cases 3 and 4 have 1000 observations with 100 covariates. Cases 1 and 3 correspond to concentrated signals and Cases 2 and 4 correspond to distributed signals. For each measure, SLASSO is used as the benchmark and the ratio to SLASSO is presented.

	SLASSO	SaLASSO	SENET	SaENET	SLASSO (w)	SaLASSO (w)	SENET (w)	SaENET (w)	GLASSO	GaLASSO
					Case 1					
$SENS$	1.00	0.99	1.00	0.99	1.00	0.98	1.00	0.99	0.98	0.94
$SPEC$	1.00	1.75	0.99	1.75	1.02	1.76	1.01	1.75	0.94	1.80
$MSE_{non-null}$	1.00	0.55	0.99	0.55	1.04	0.55	1.03	0.55	1.49	0.93
MSE_{null}	1.00	0.64	1.01	0.64	1.07	0.65	1.10	0.64	0.71	0.40
					Case 2					
$SENS$	1.00	0.92	1.00	0.92	1.00	0.91	1.00	0.92	0.97	0.82
$SPEC$	1.00	1.67	0.99	1.67	1.02	1.68	1.01	1.68	1.03	1.73
$MSE_{non-null}$	1.00	0.84	0.99	0.84	1.03	0.86	1.01	0.86	1.49	1.36
MSE_{null}	1.00	0.75	1.02	0.73	1.05	0.71	1.07	0.72	0.66	0.48
					Case 3					
$SENS$	1.00	0.98	1.00	0.99	1.00	0.98	1.00	0.98	0.98	0.85
$SPEC$	1.00	2.12	0.99	2.11	1.24	2.30	1.22	2.30	0.33	2.22
$MSE_{nonnull}$	1.00	0.52	0.99	0.52	1.14	0.50	1.12	0.52	3.83	1.44
MSE_{null}	1.00	1.38	1.03	1.34	0.81	1.16	0.86	1.12	34.62	2.61
					Case 4					
$SENS$	1.00	0.95	1.00	0.96	1.00	0.94	1.00	0.95	1.00	0.86
$SPEC$	1.00	1.92	0.97	1.91	1.23	2.07	1.19	2.07	0.50	2.27
$MSE_{nonnull}$	1.00	0.78	0.98	0.76	1.08	0.81	1.05	0.77	2.81	1.66
MSE_{null}	1.00	1.41	1.05	1.36	0.80	1.22	0.86	1.15	14.80	1.00

References

Chen, Q. and S. Wang (2013). Variable selection for multiply-imputed data with application to dioxin exposure study. *Statistics in Medicine 32*(21), 3646–3659.

Garcia, R. I., J. G. Ibrahim, and H. Zhu (2010). Variable selection in the cox regression model with covariates missing at random. *Biometrics 66*(1), 97–104.

Johnson, B. A., D. Lin, and D. Zeng (2008). Penalized estimating functions and variable selection in semiparametric regression models. *Journal of the American Statistical Association 103*(482), 672–680.

Long, Q. and B. A. Johnson (2015). Variable selection in the presence of missing data: resampling and imputation. *Biostatistics 16*(3), 596–610.

Wan, Y., S. Datta, D. J. Conklin, and M. Kong (2015). Variable selection models based on multiple imputation with an application for predicting median effective dose and maximum effect. *Journal of Statistical Computation and Simulation 85*(9), 1902–1916.

Wood, A. M., I. R. White, and P. Royston (2008). How should variable selection be performed with multiply imputed data? *Statistics in Medicine 27*(17), 3227–3246.

Yang, X., T. R. Belin, and W. J. Boscardin (2005). Imputation and variable selection in linear regression models with missing covariates. *Biometrics 61*(2), 498–506.