Supplementary Material for "Subdata selection with a large number of variables"

Rakhi Singh
Binghamton University, NY, USA
and
John Stufken*
George Mason University, VA, USA

May 18, 2023

*The authors gratefully acknowledge support through NSF grant DMS-1935729 and DMS-2304767

1 Summary

In the Supplementary Material, we show the results for the following situations:

- Section 2 uses $p = 500$, active effects coefficients equal to 1, and error variance to be 9. We see the effect of changing n.
- Section 3 uses $p = 500$, active effects coefficients come from $N(5, 1)$, and error is $N(0, 1)$. We see the effect of changing n.
- Section 4 uses $p = 500$, active effects coefficients come from $N(5, 1)$, and error is $N(0, 1)$. We see the effect of changing k.
- Section 5 uses $p = 5000$, active effects coefficients come from $N(5, 1)$, and error is $N(0, 1)$. We see the effect of changing n.

The figures are in pairs. Odd numbered figures are for MSEs whereas even numbered figures show the power and error of the corresponding case.

First six figures in Sections 2–5 correspond to $\mathbf{\Sigma} = (0^{I(i \neq j)})$, next six correspond to $\mathbf{\Sigma} = (0.5^{I(i \neq j)})$. Last six figures in Sections 2–4 correspond to $\mathbf{\Sigma}$ is a randomly generated correlation matrix (called 'Random'). We consider p_1 to be 10, 25, and 50 for Sections 2–4 and p_1 to be 25, 50, and 75 for Section 5.

For Sections 2, 3, and 5, the sample size k is fixed at 1000. For Section 4, we vary k but keep n fixed at 10^5.

In all figures, we see that our method performs better in terms of both MSE and screening performance.

In addition, in Section 6, for $p = 500$, active effects coefficients come from $N(5, 1)$, and error is $N(0, 1)$, $\mathbf{\Sigma} = (0.5^{I(i \neq j)})$, we see the effect of changing tuning parameters. We see that as long as $nsample > 100$, all other considered choices of tuning parameters perform similarly.

Finally, in Section 7, we show tables like Tables 5 and 6 in the main manuscript, but now for the lognormal and mixture distributions.

2 For $p = 500$ with error variance equals to 9

Figure 1: MSE for $k = 1000$, $p = 500$, $p_1 = 10$, and $\mathbf{\Sigma} = (0^{I(i \neq j)})$.

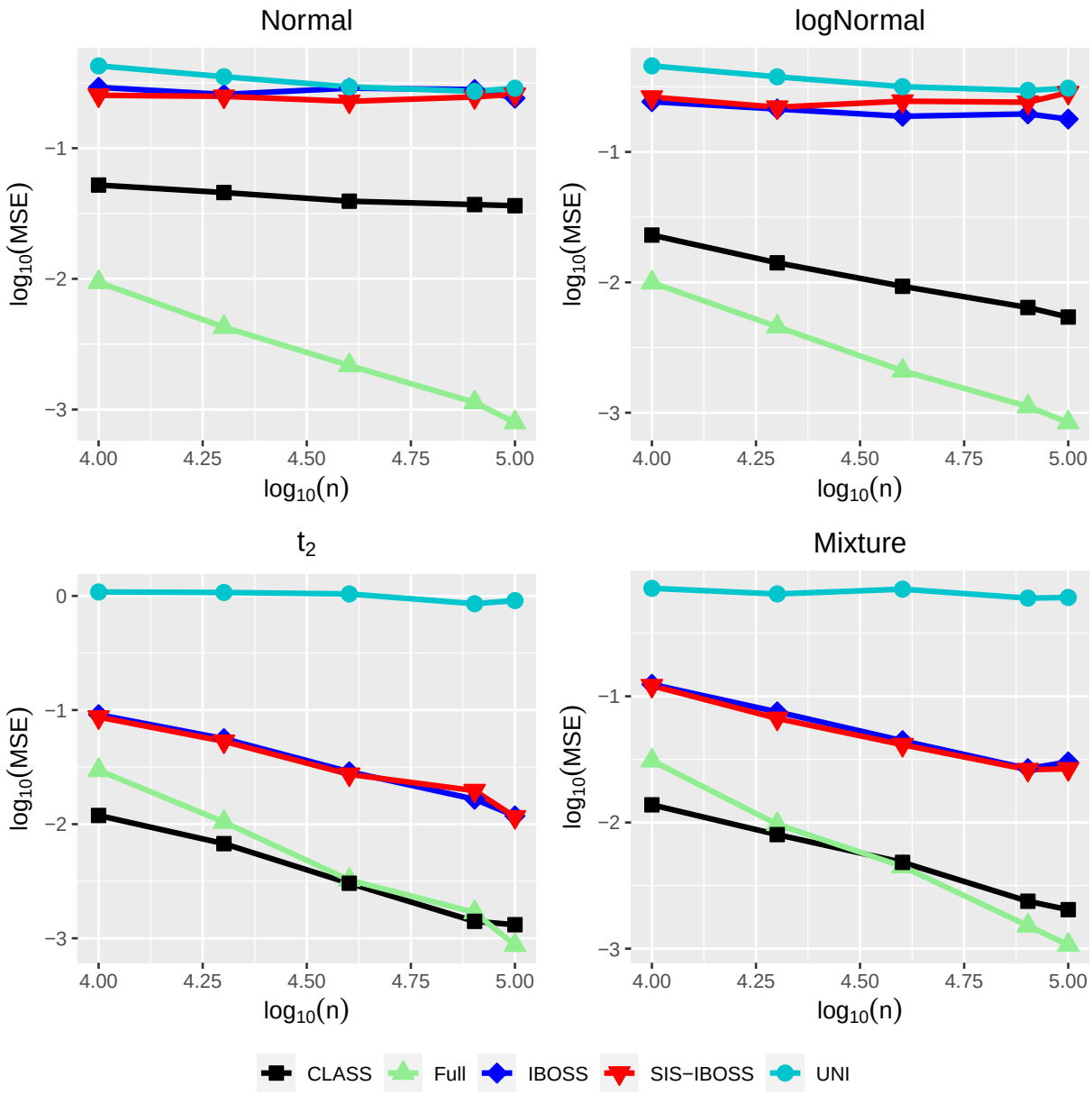


Figure 2: Variable selection performance for $k = 1000$, $p = 500$, $p_1 = 10$, and $\mathbf{\Sigma} = (0^{I(i \neq j)})$. The solid lines represent the power, whereas the dashed lines represent the error.

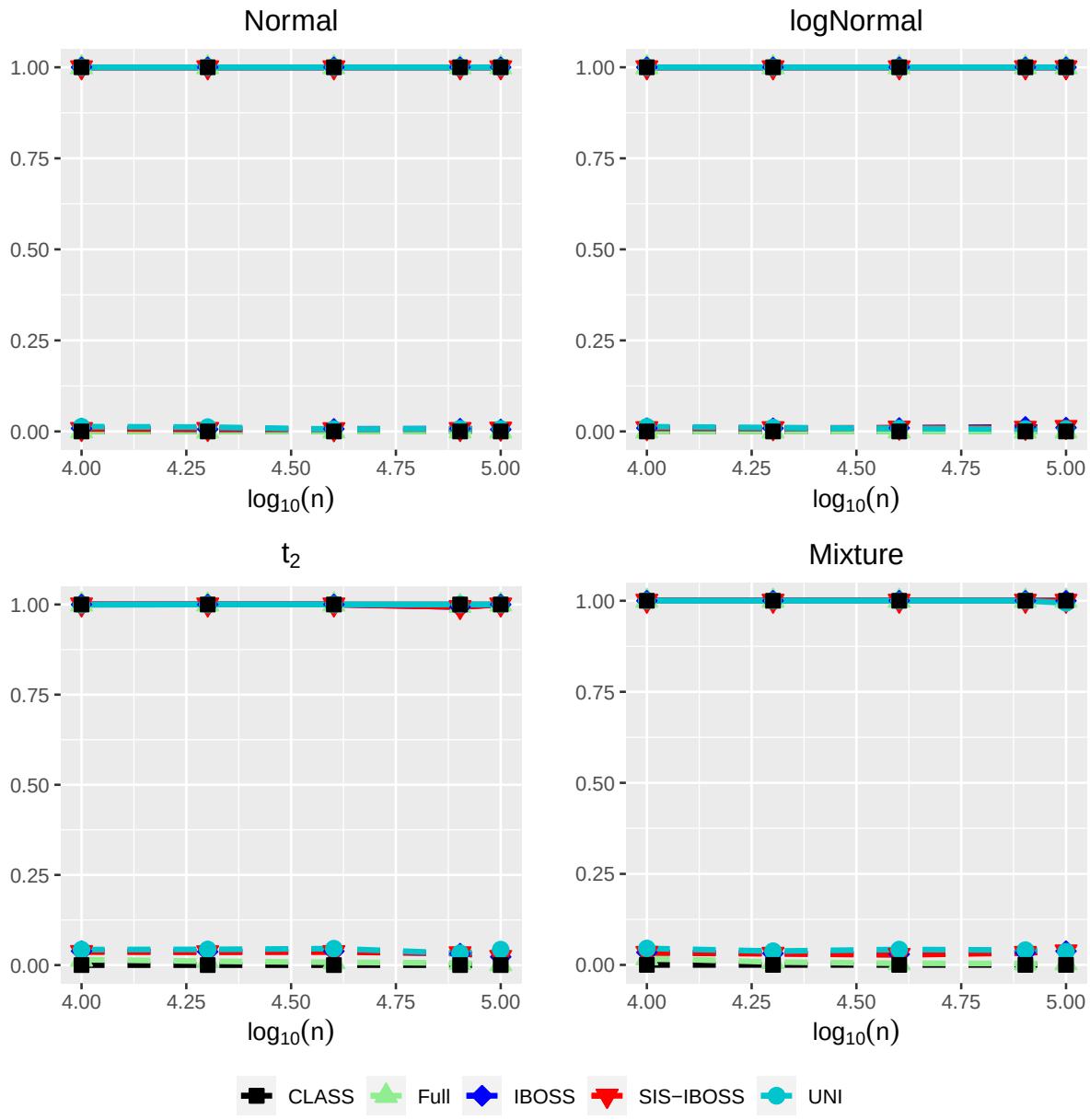


Figure 3: MSE for $k = 1000$, $p = 500$, $p_1 = 25$, and $\mathbf{\Sigma} = (0^{I(i \neq j)})$.

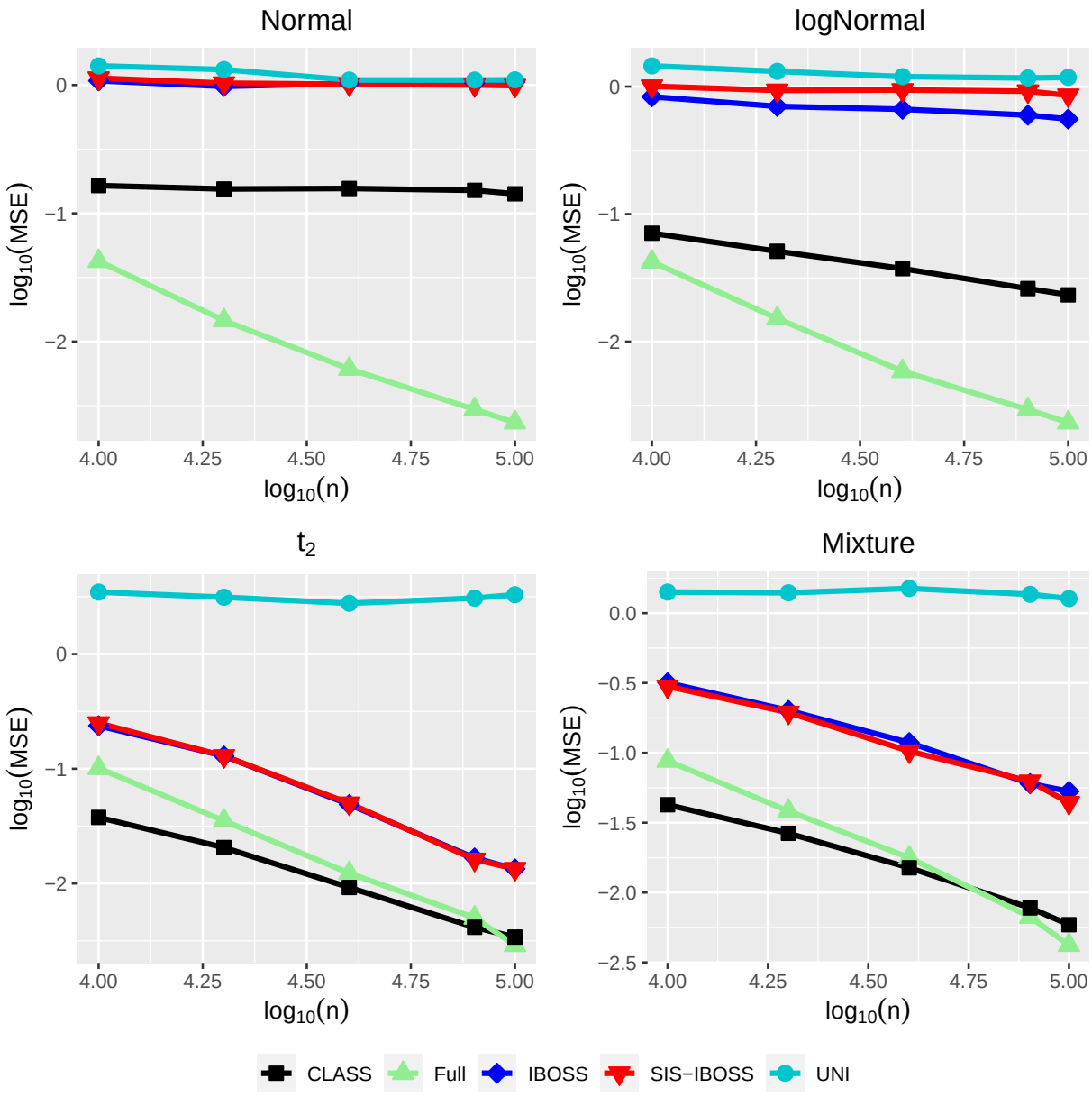


Figure 4: Variable selection performance for $k = 1000$, $p = 500$, $p_1 = 25$, and $\mathbf{\Sigma} = (0^{I(i \neq j)})$. The solid lines represent the power, whereas the dashed lines represent the error.

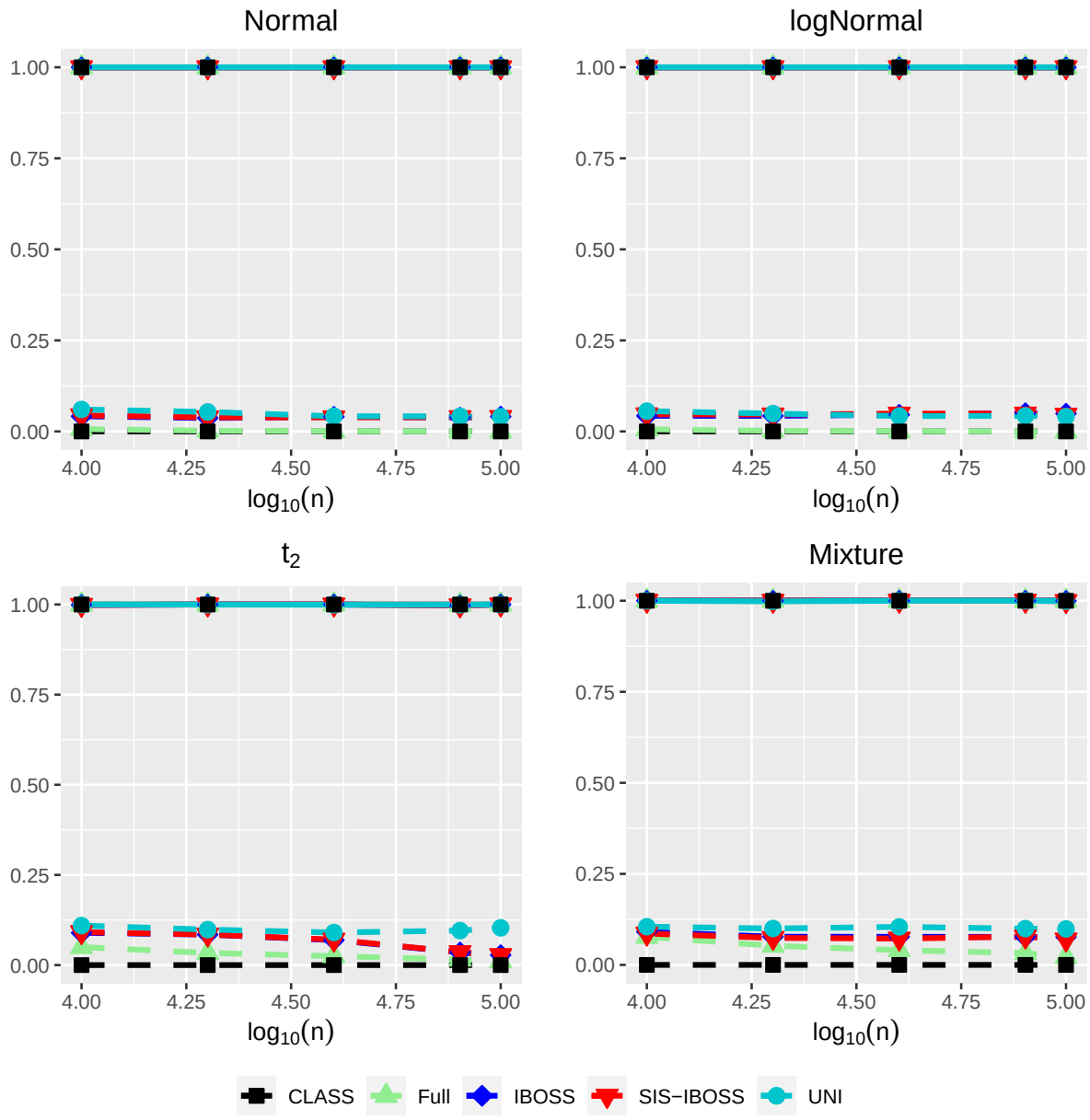


Figure 5: MSE for $k = 1000$, $p = 500$, $p_1 = 50$, and $\mathbf{\Sigma} = (0^{I(i \neq j)})$.

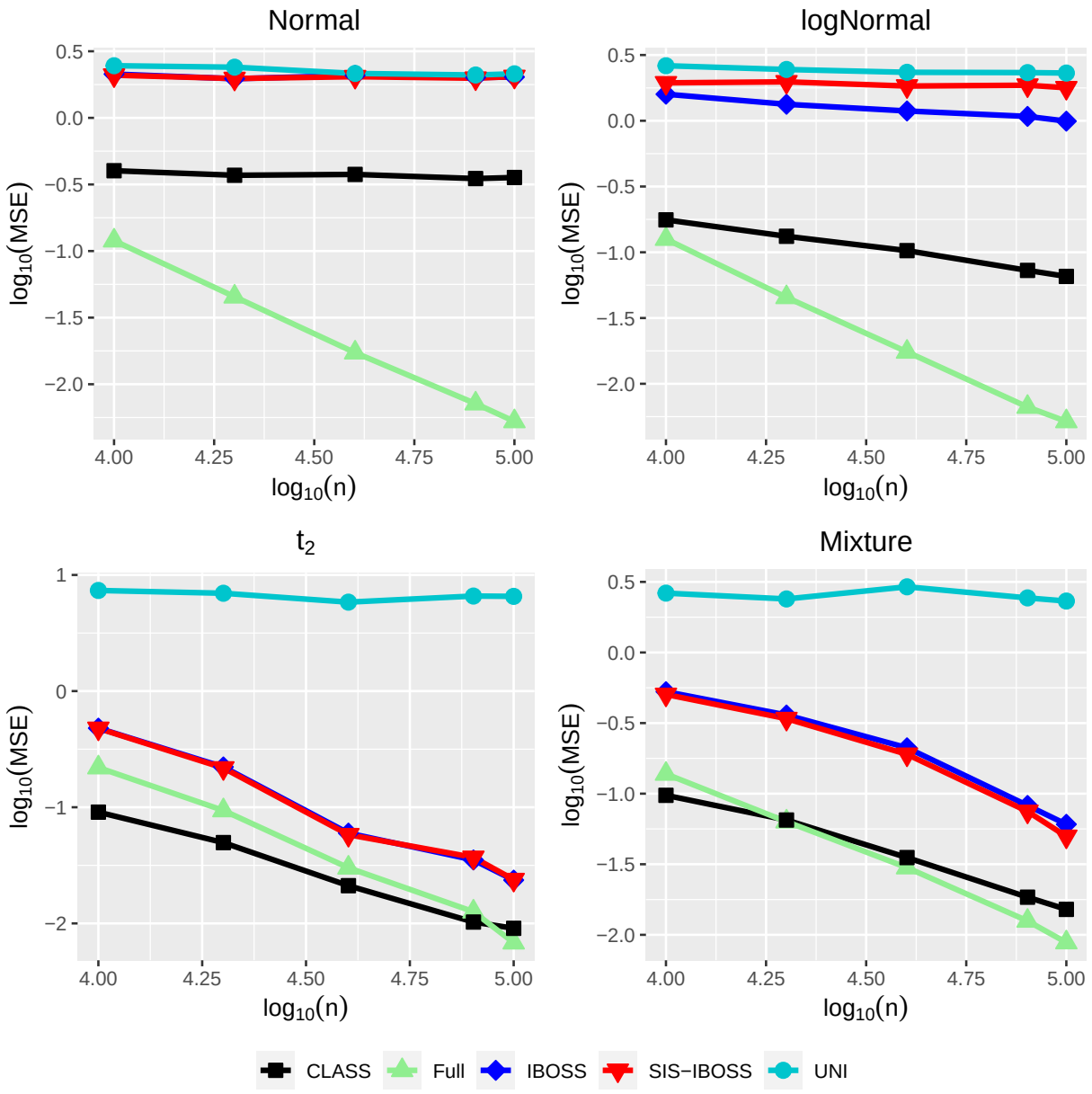


Figure 6: Variable selection performance for $k = 1000$, $p = 500$, $p_1 = 50$, and $\mathbf{\Sigma} = (0^{I(i \neq j)})$. The solid lines represent the power, whereas the dashed lines represent the error.

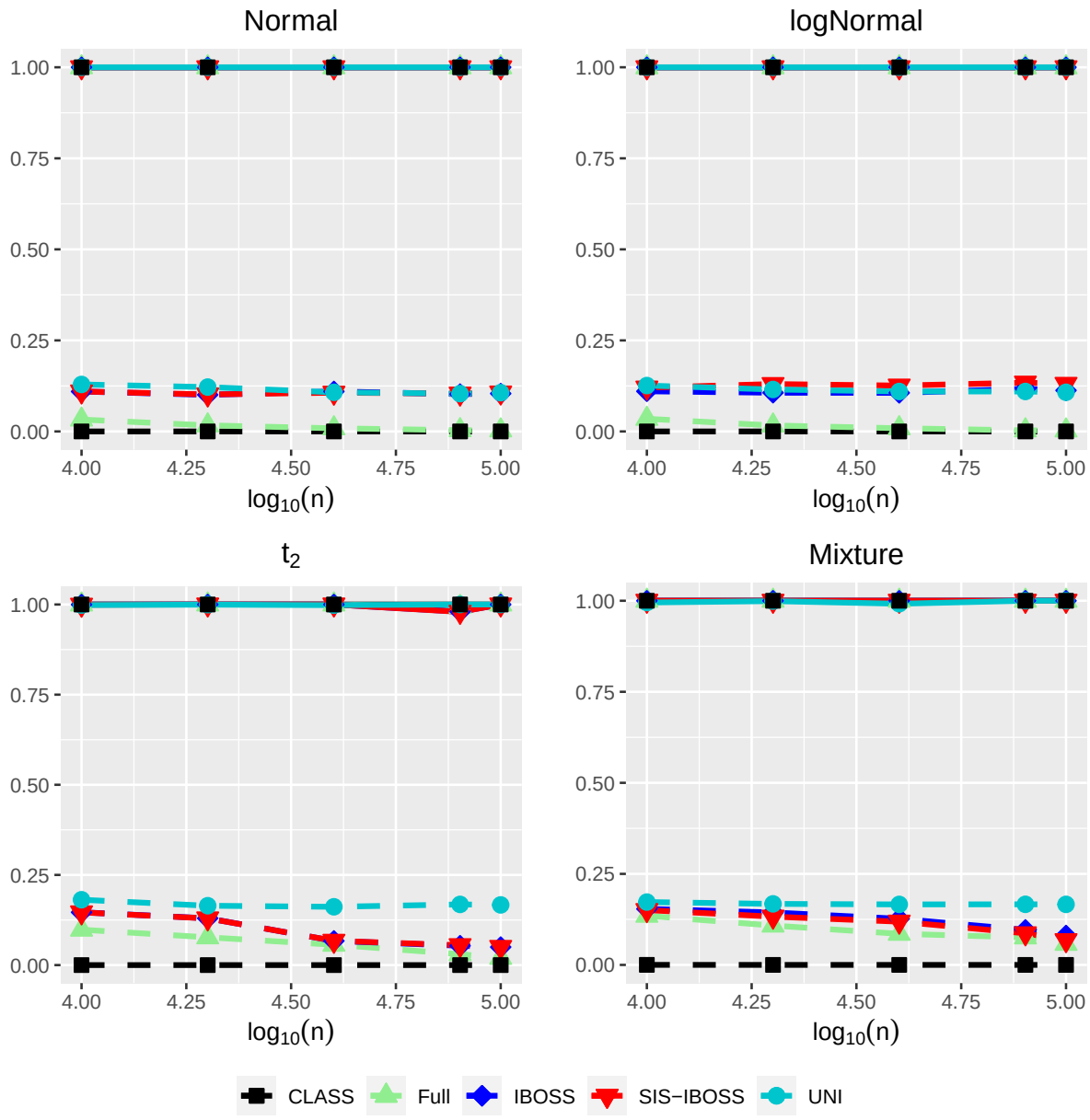


Figure 7: MSE for $k = 1000$, $p = 500$, $p_1 = 10$, and $\mathbf{\Sigma} = (0.5^{I(i \neq j)})$.

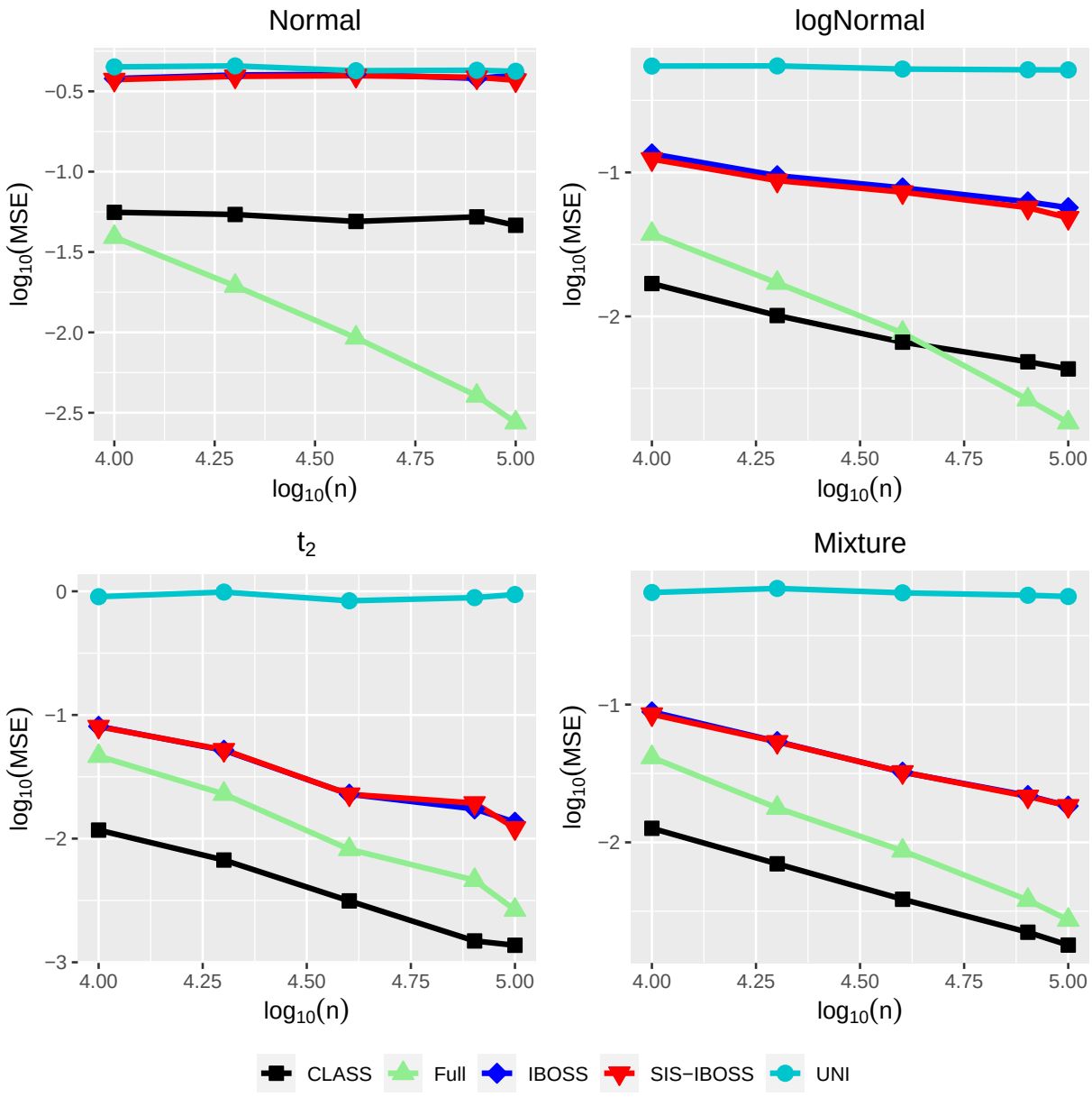


Figure 8: Variable selection performance for $k = 1000$, $p = 500$, $p_1 = 10$, and $\mathbf{\Sigma} = (0.5^{I(i \neq j)})$. The solid lines represent the power, whereas the dashed lines represent the error.

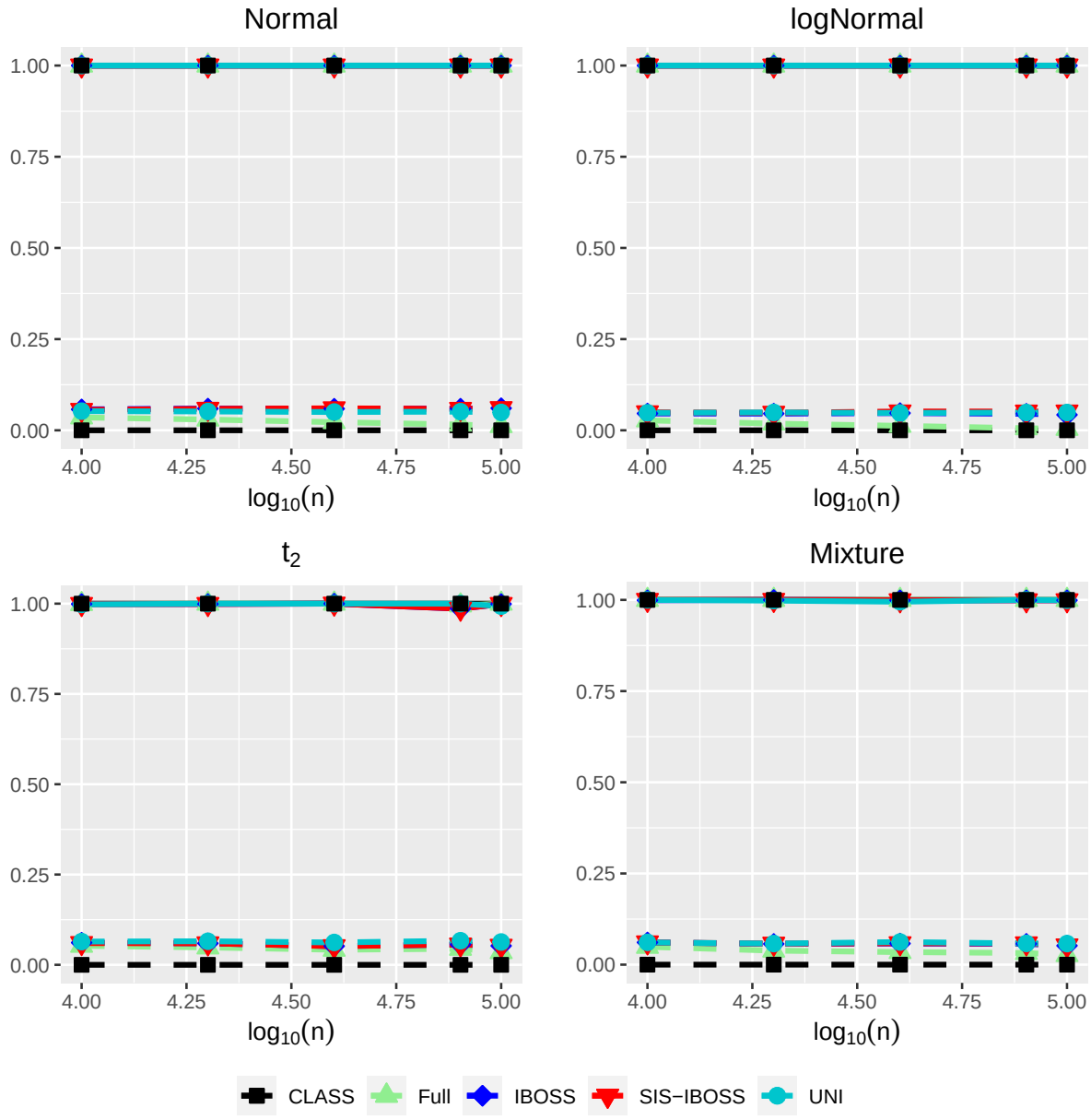


Figure 9: MSE for $k = 1000$, $p = 500$, $p_1 = 25$, and $\mathbf{\Sigma} = (0.5^{I(i \neq j)})$.

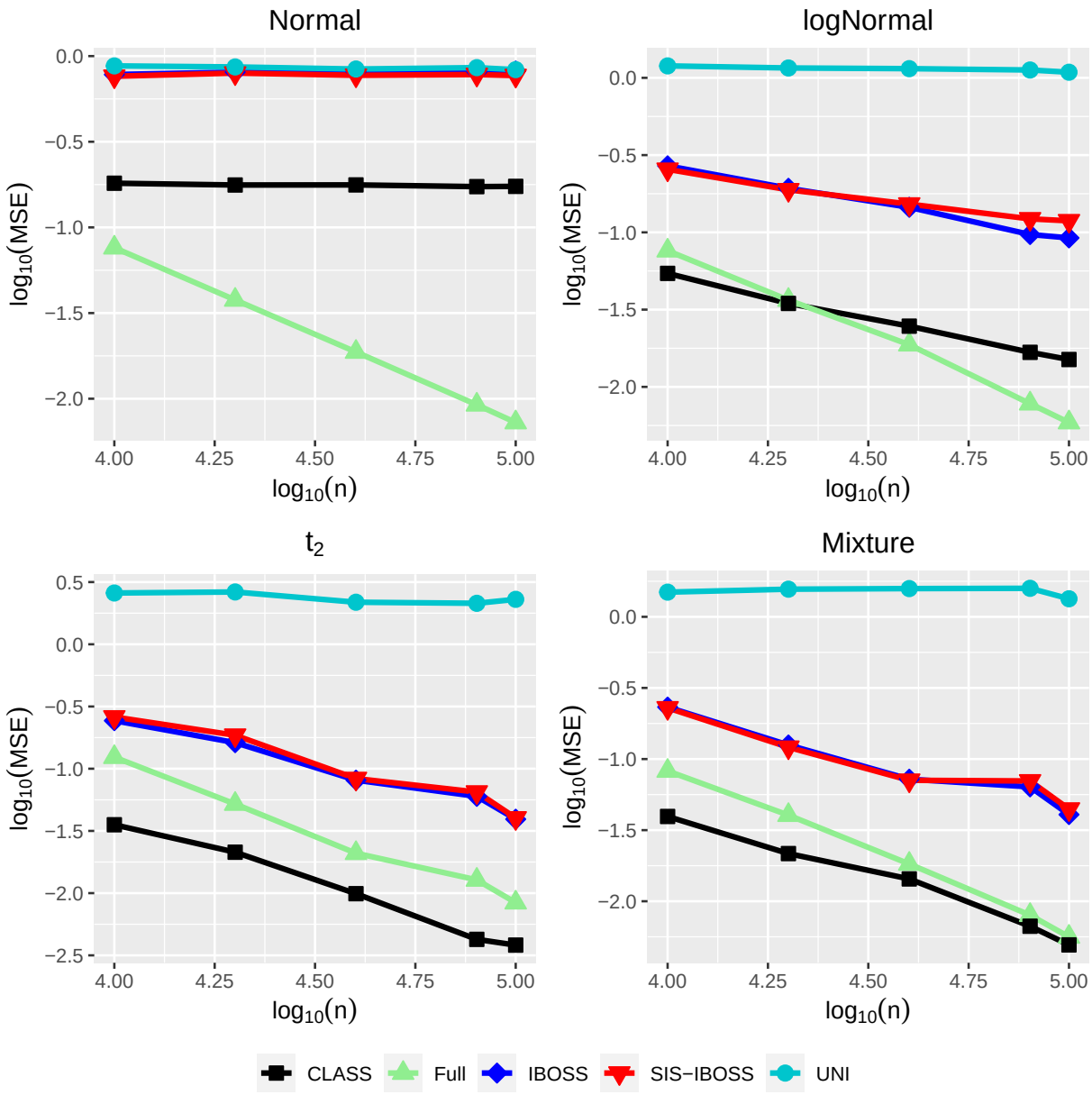


Figure 10: Variable selection performance for $k = 1000$, $p = 500$, $p_1 = 25$, and $\mathbf{\Sigma} = (0.5^{I(i \neq j)})$. The solid lines represent the power, whereas the dashed lines represent the error.

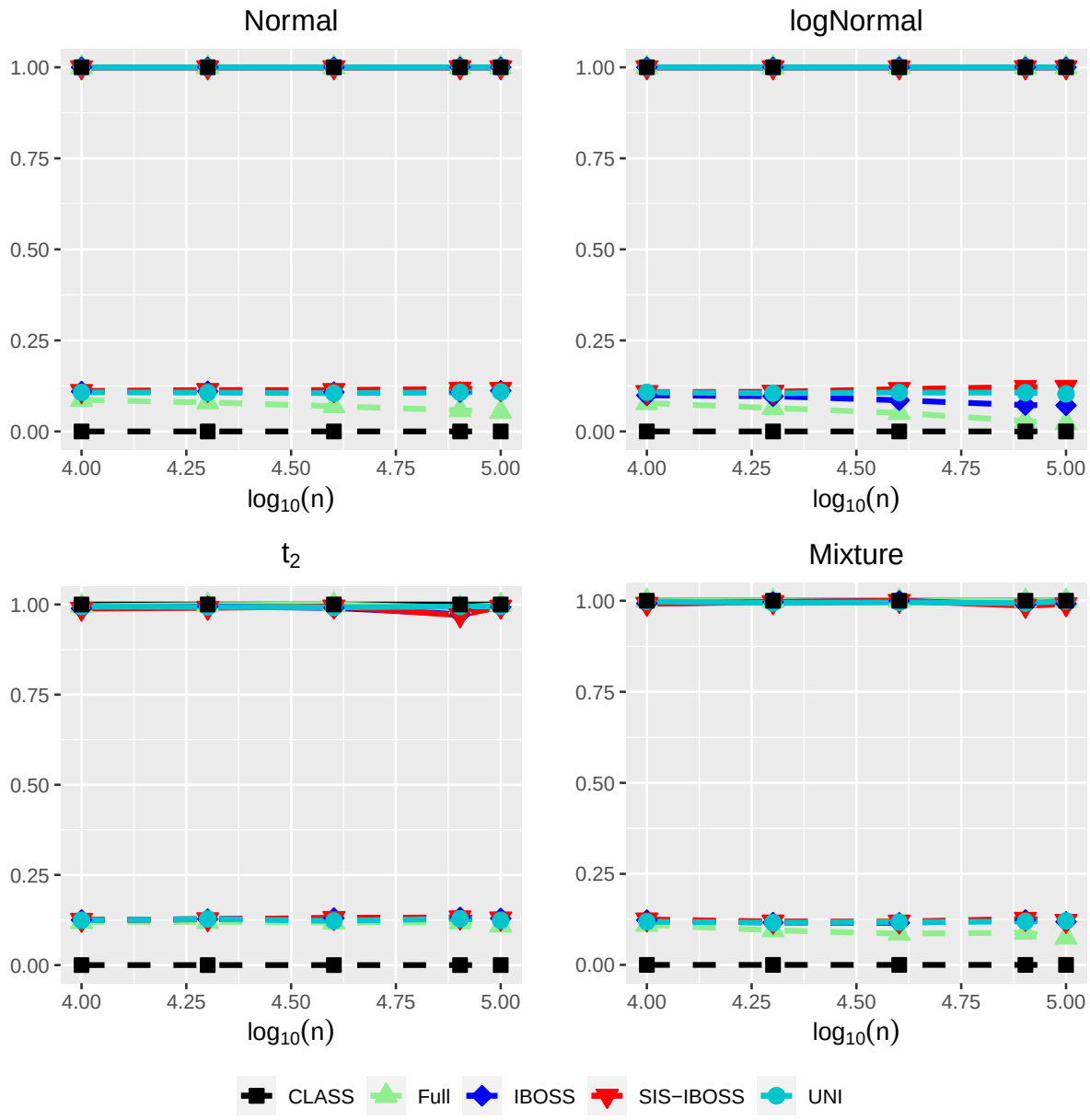


Figure 11: MSE for $k = 1000$, $p = 500$, $p_1 = 50$, and $\mathbf{\Sigma} = (0.5^{I(i \neq j)})$.

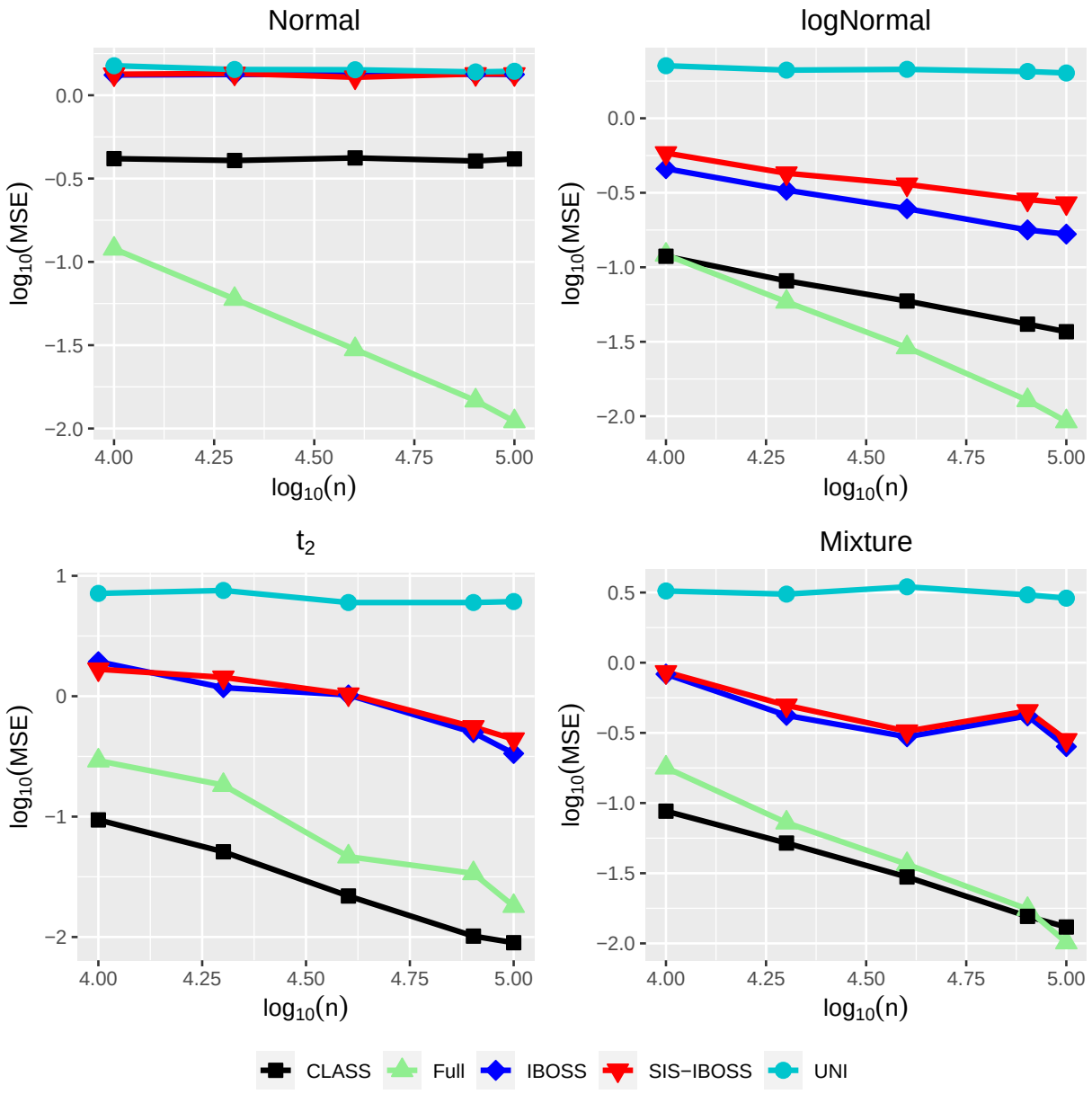


Figure 12: Variable selection performance for $k = 1000$, $p = 500$, $p_1 = 50$, and $\mathbf{\Sigma} = (0.5^{I(i \neq j)})$. The solid lines represent the power, whereas the dashed lines represent the error.

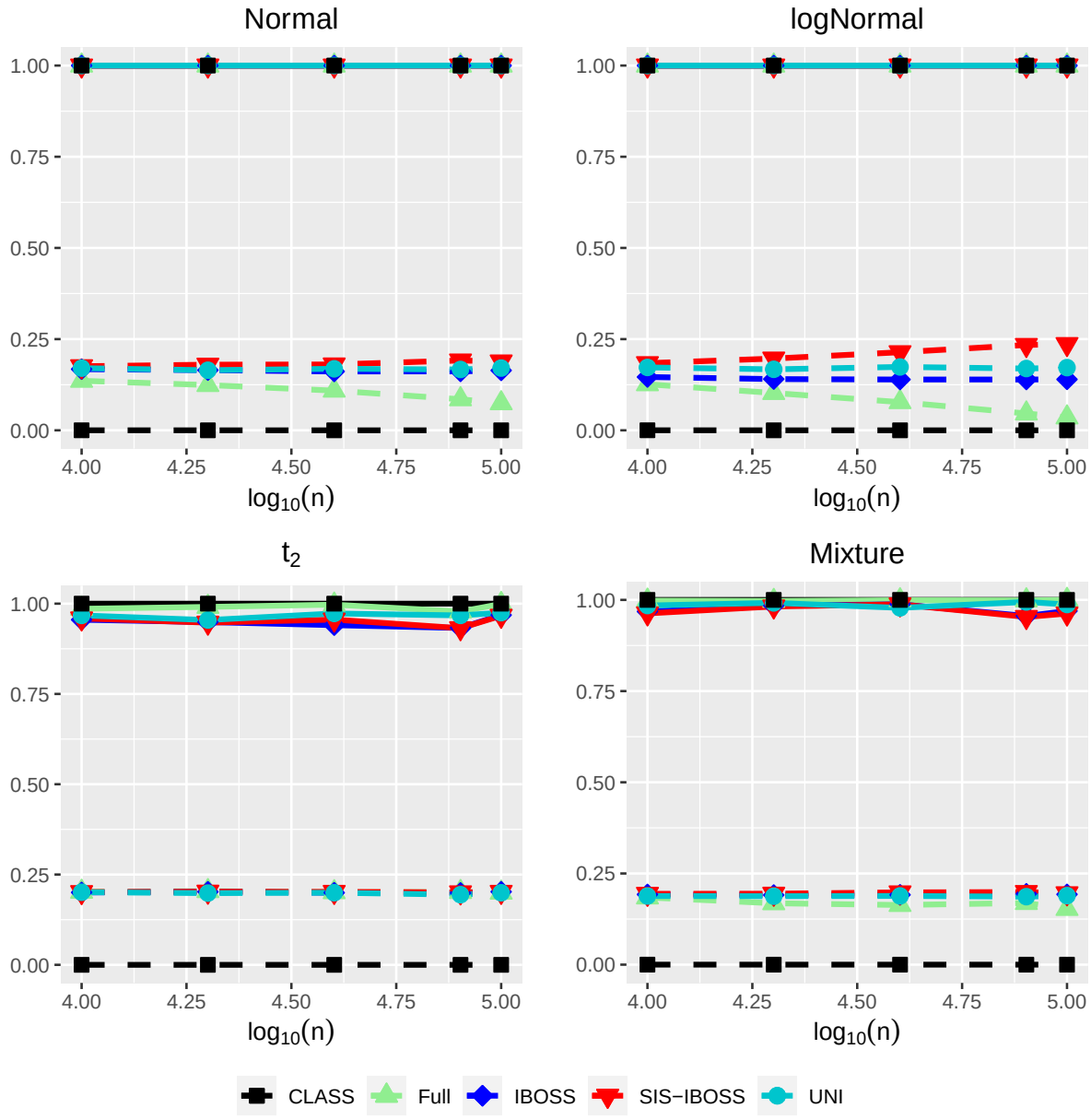


Figure 13: MSE for $k = 1000$, $p = 500$, $p_1 = 10$, and $\mathbf{\Sigma}$ is Random.

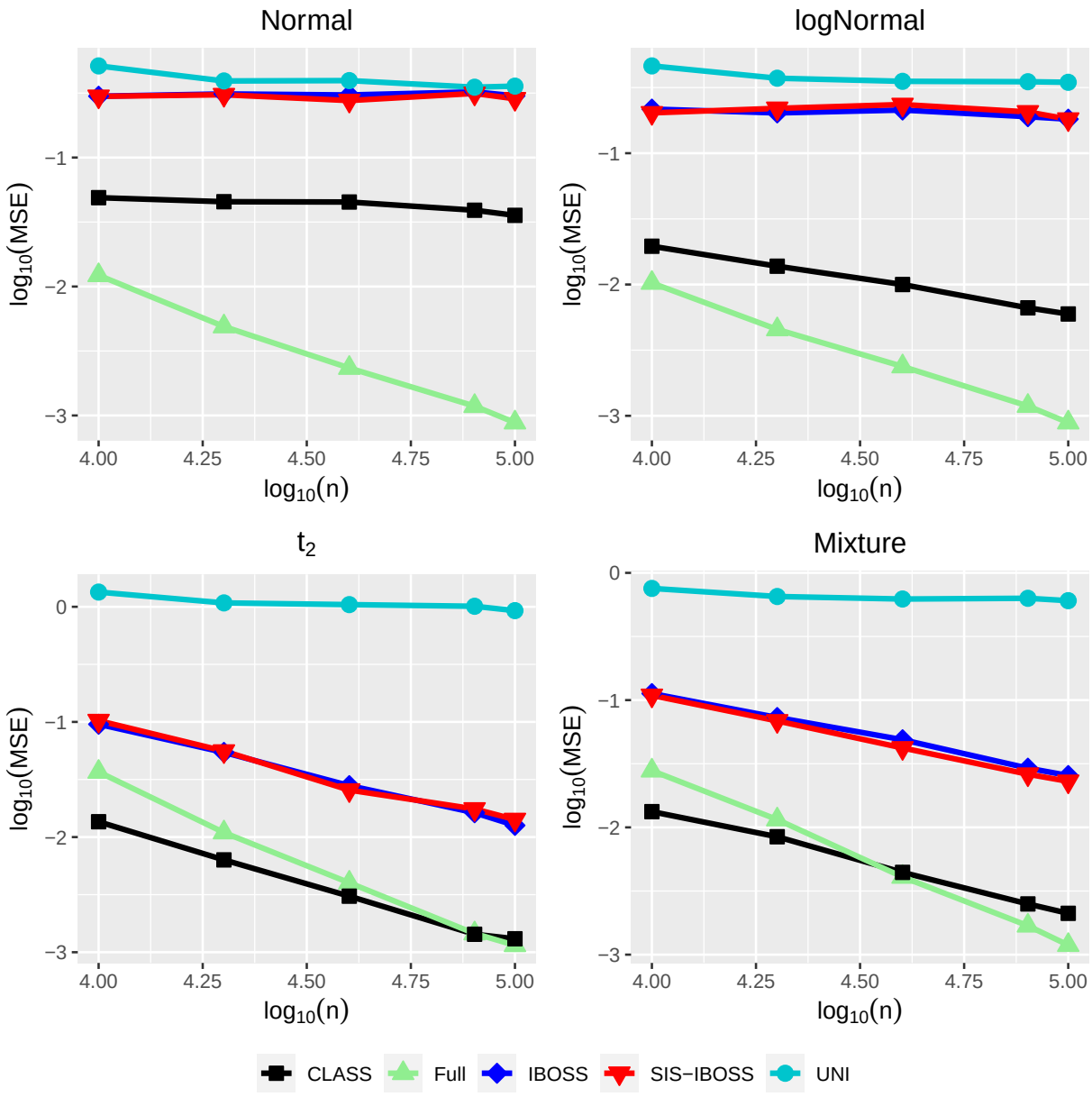


Figure 14: Variable selection performance for $k = 1000$, $p = 500$, $p_1 = 10$, and $\mathbf{\Sigma}$ is Random. The solid lines represent the power, whereas the dashed lines represent the error.

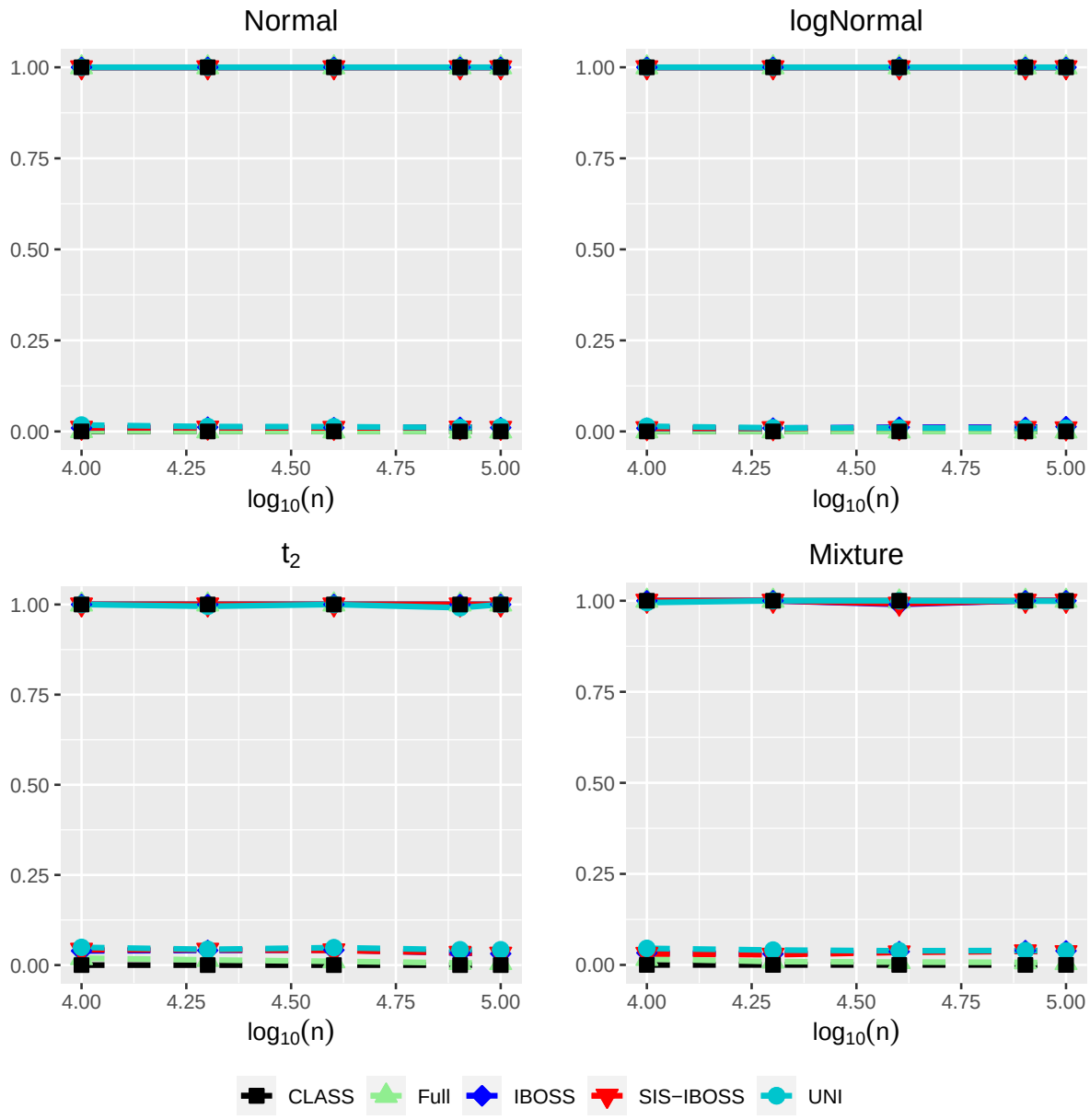


Figure 15: MSE for $k = 1000$, $p = 500$, $p_1 = 25$, and $\mathbf{\Sigma}$ is Random.

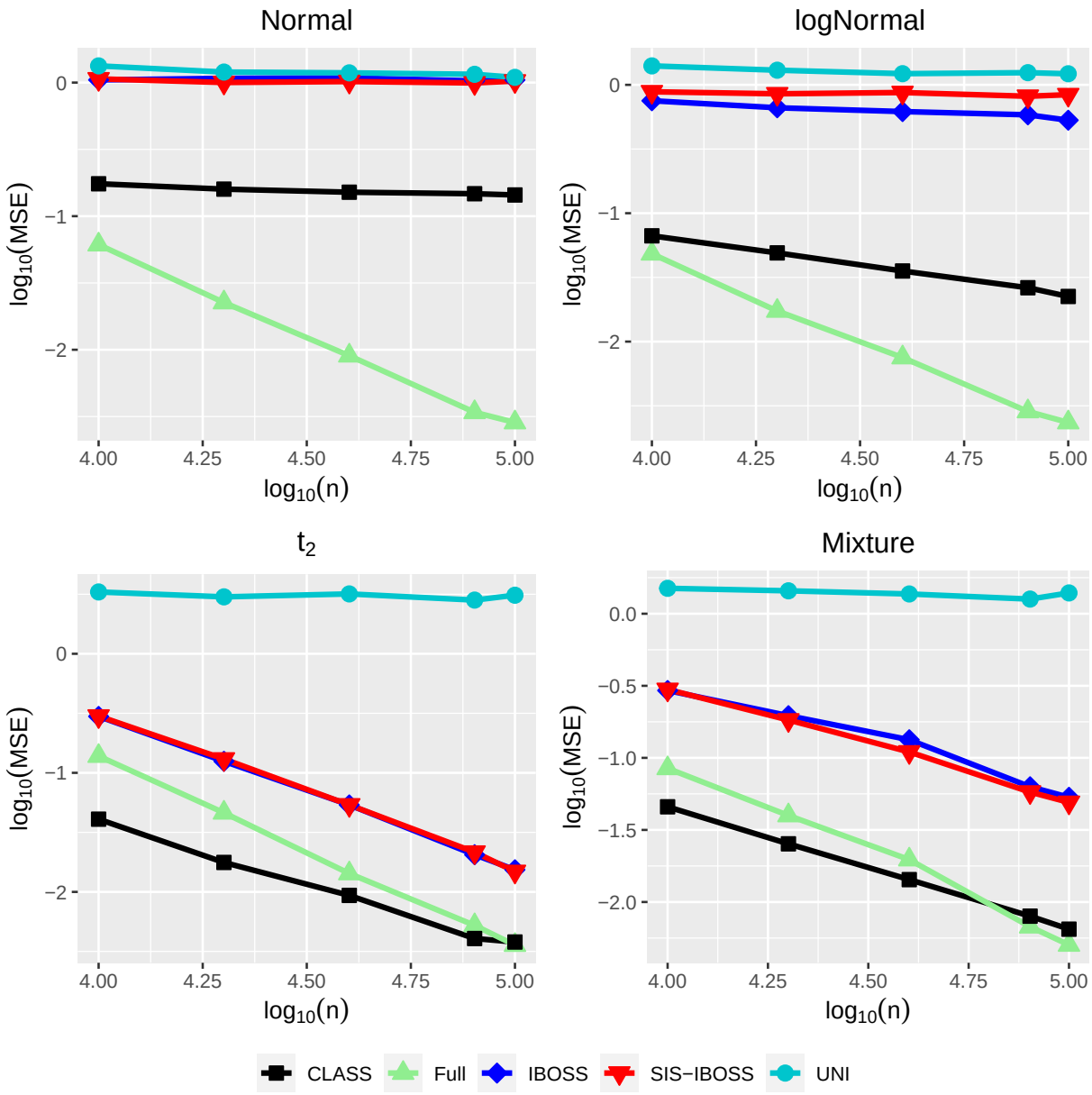


Figure 16: Variable selection performance for $k = 1000$, $p = 500$, $p_1 = 25$, and $\mathbf{\Sigma}$ is Random. The solid lines represent the power, whereas the dashed lines represent the error.

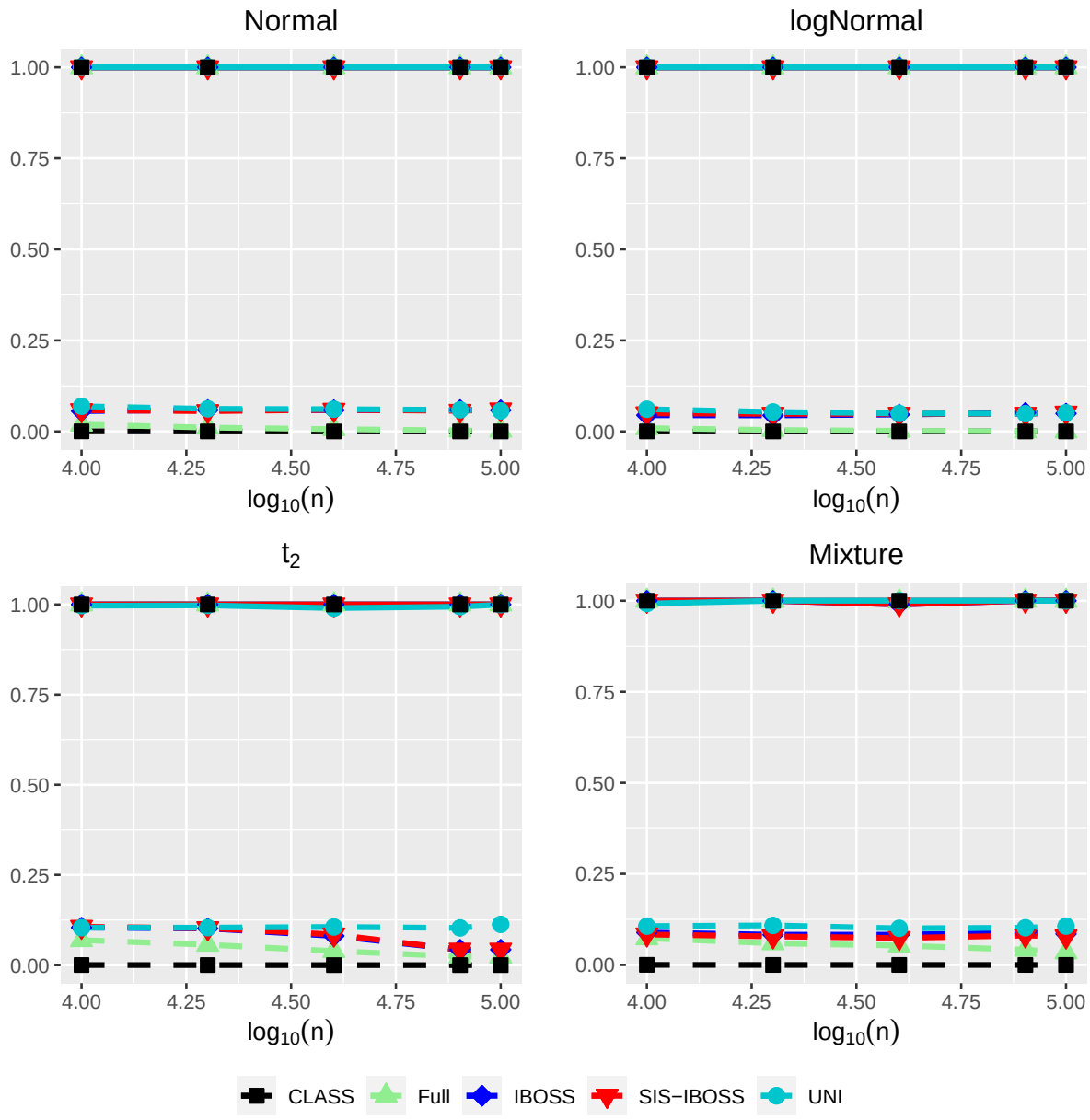


Figure 17: MSE for $k = 1000$, $p = 500$, $p_1 = 50$, and $\mathbf{\Sigma}$ is Random.

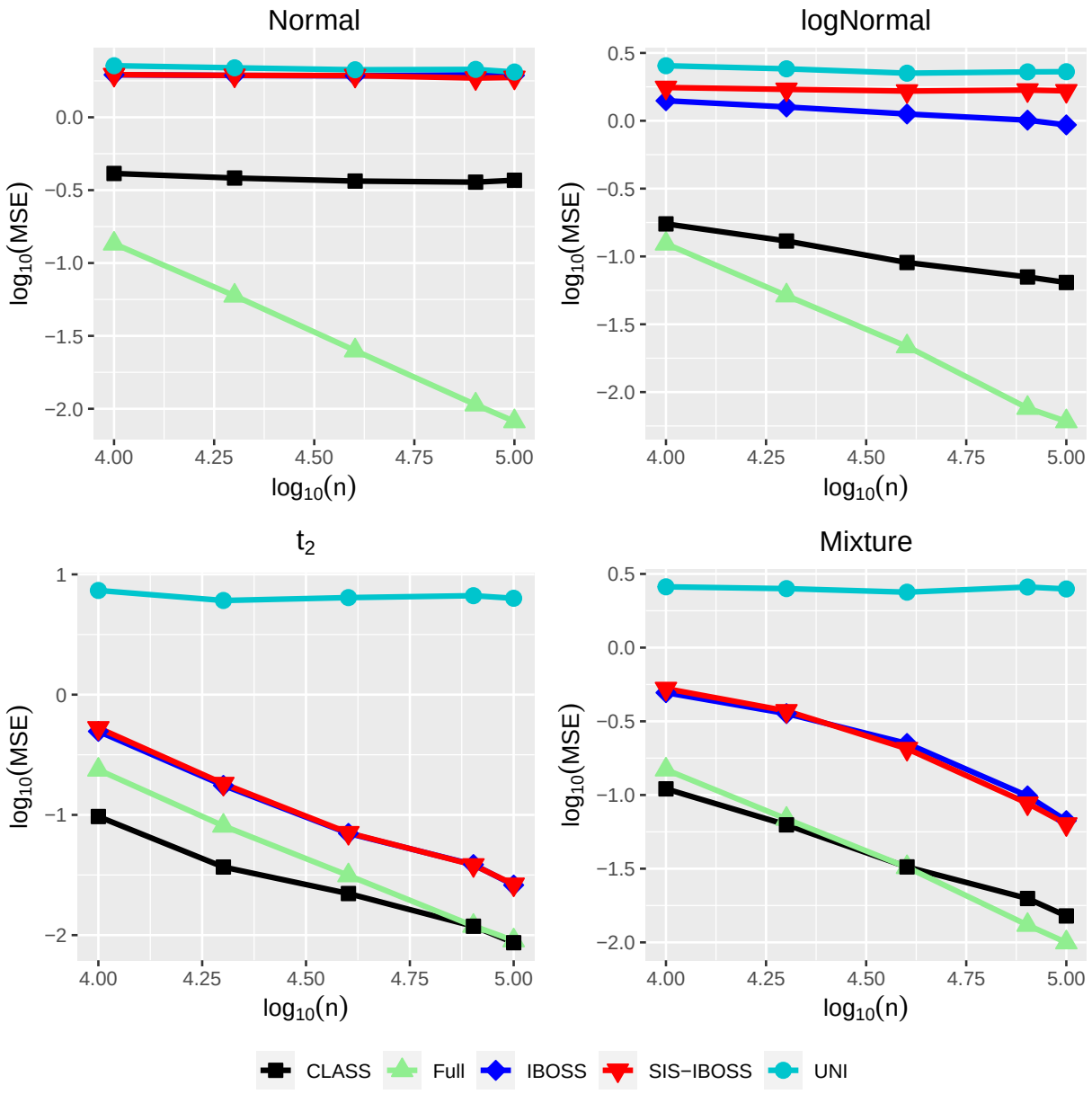
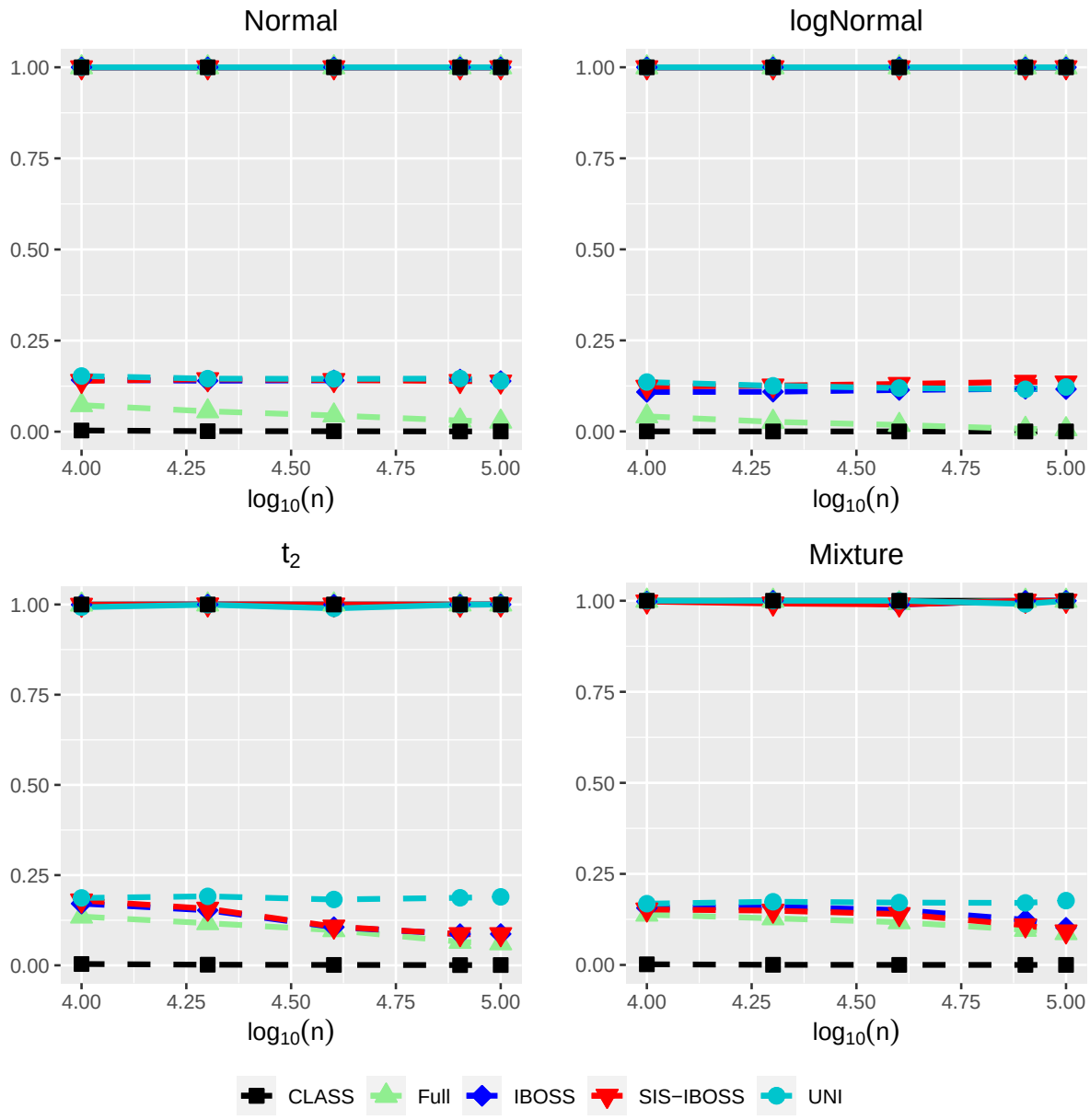


Figure 18: Variable selection performance for $k = 1000$, $p = 500$, $p_1 = 50$, and $\mathbf{\Sigma}$ is Random. The solid lines represent the power, whereas the dashed lines represent the error.



3 For $p = 500$ with error standard deviation equals 1

Figure 19: MSE for $k = 1000$, $p = 500$, $p_1 = 10$, and $\mathbf{\Sigma} = (0^{I(i \neq j)})$.

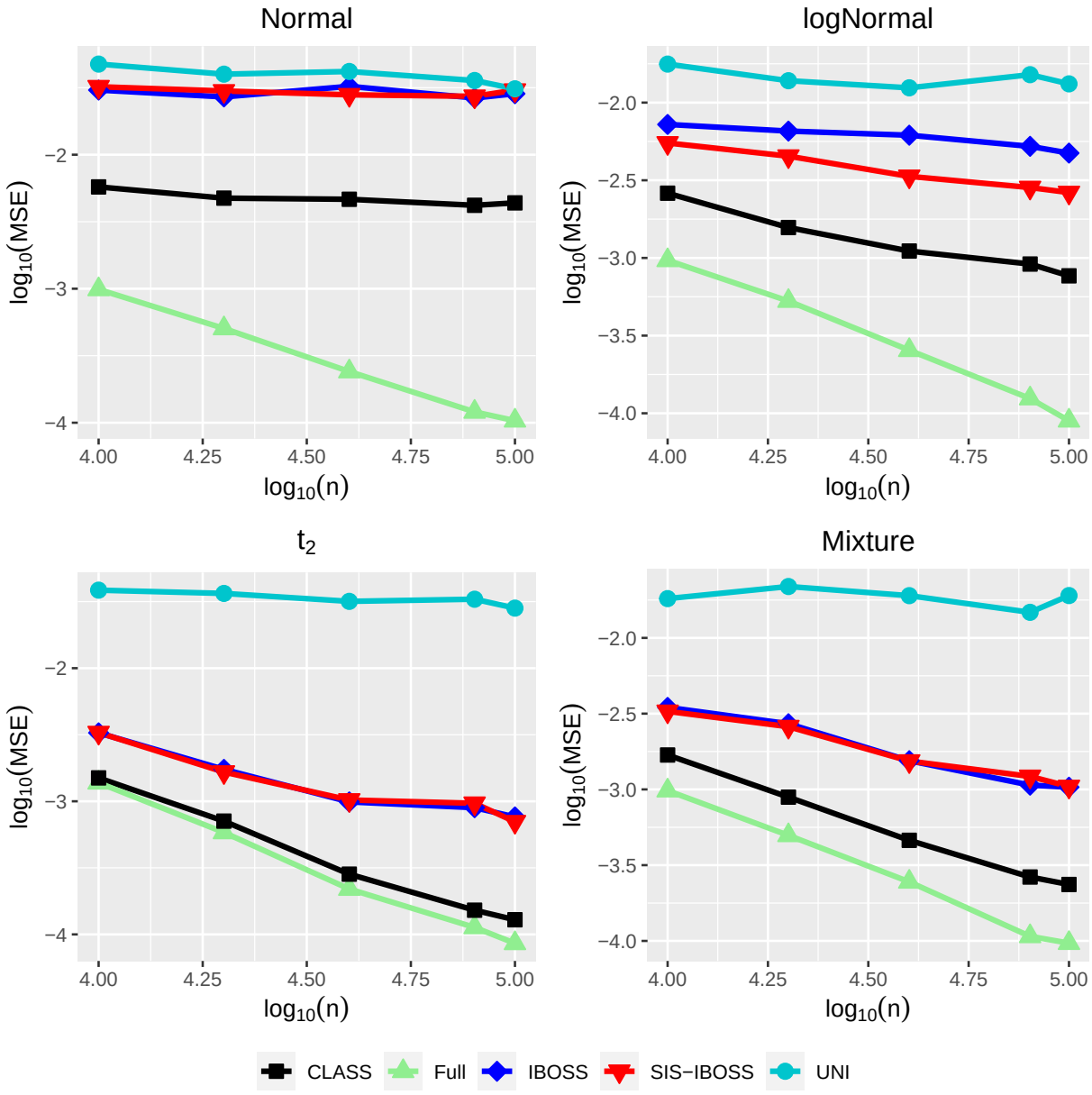


Figure 20: Variable selection performance for $k = 1000$, $p = 500$, $p_1 = 10$, and $\mathbf{\Sigma} = (0^{I(i \neq j)})$. The solid lines represent the power, whereas the dashed lines represent the error.

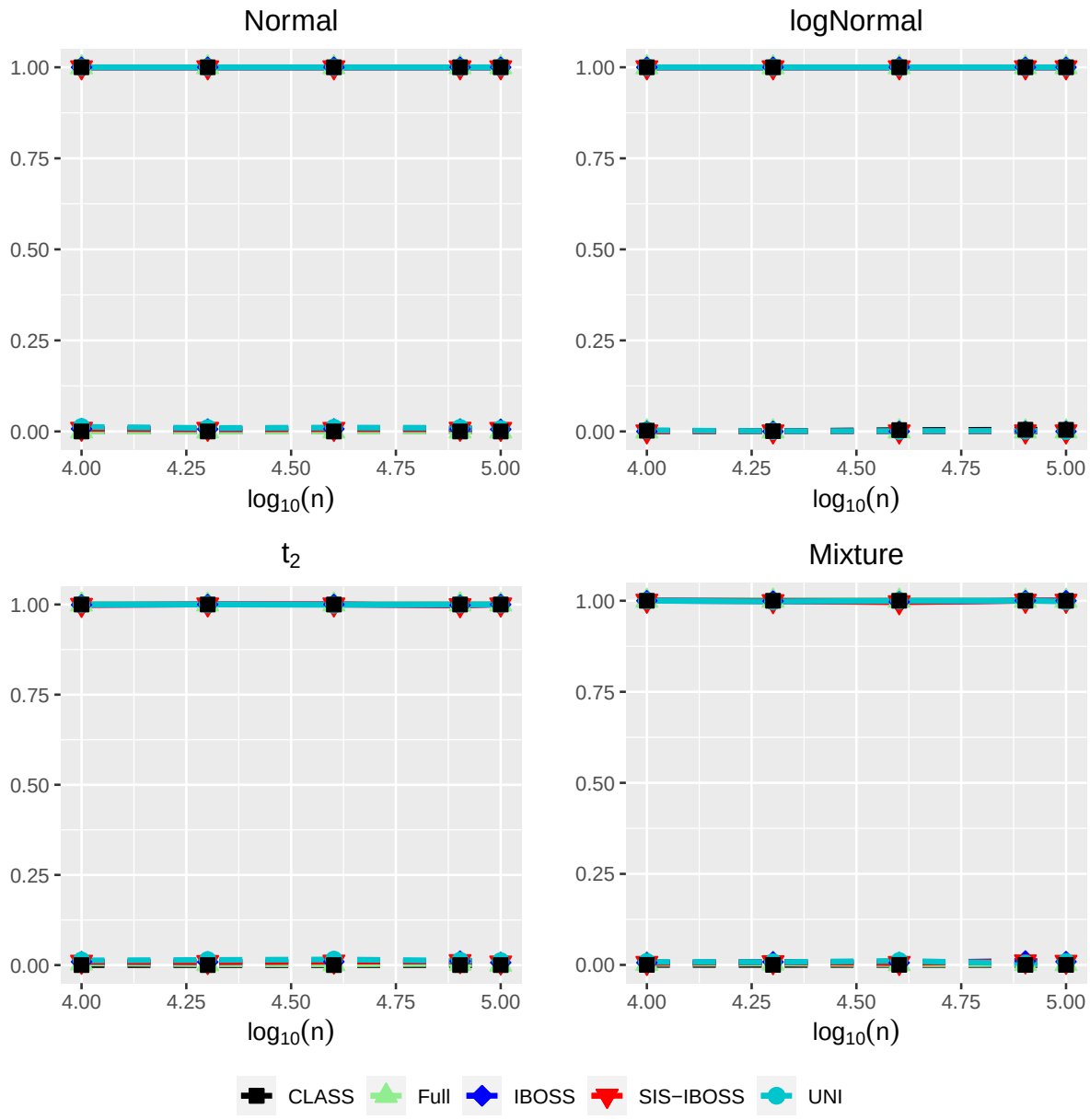


Figure 21: MSE for $k = 1000$, $p = 500$, $p_1 = 25$, and $\mathbf{\Sigma} = (0^{I(i \neq j)})$.

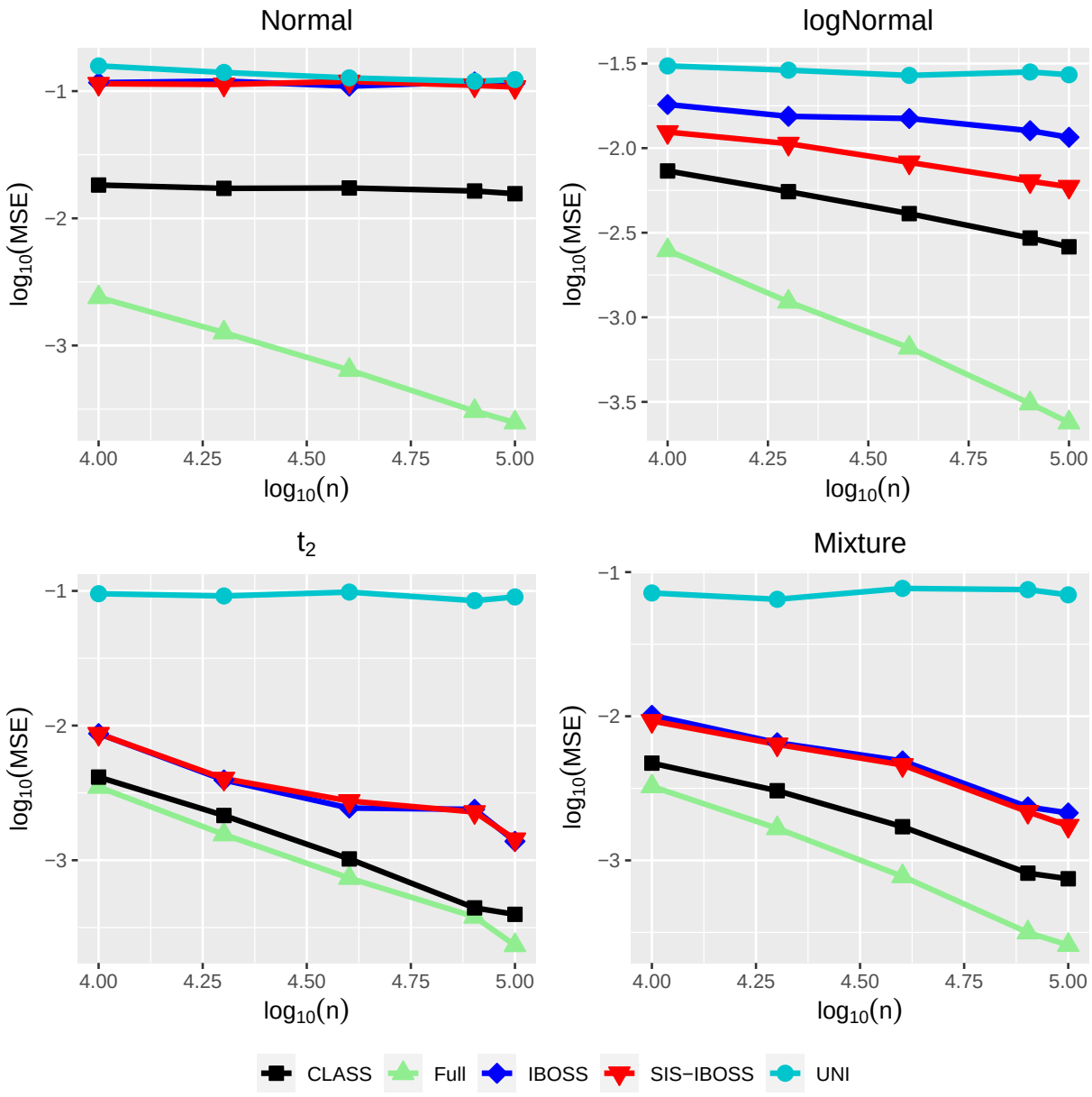


Figure 22: Variable selection performance for $k = 1000$, $p = 500$, $p_1 = 25$, and $\mathbf{\Sigma} = (0^{I(i \neq j)})$. The solid lines represent the power, whereas the dashed lines represent the error.

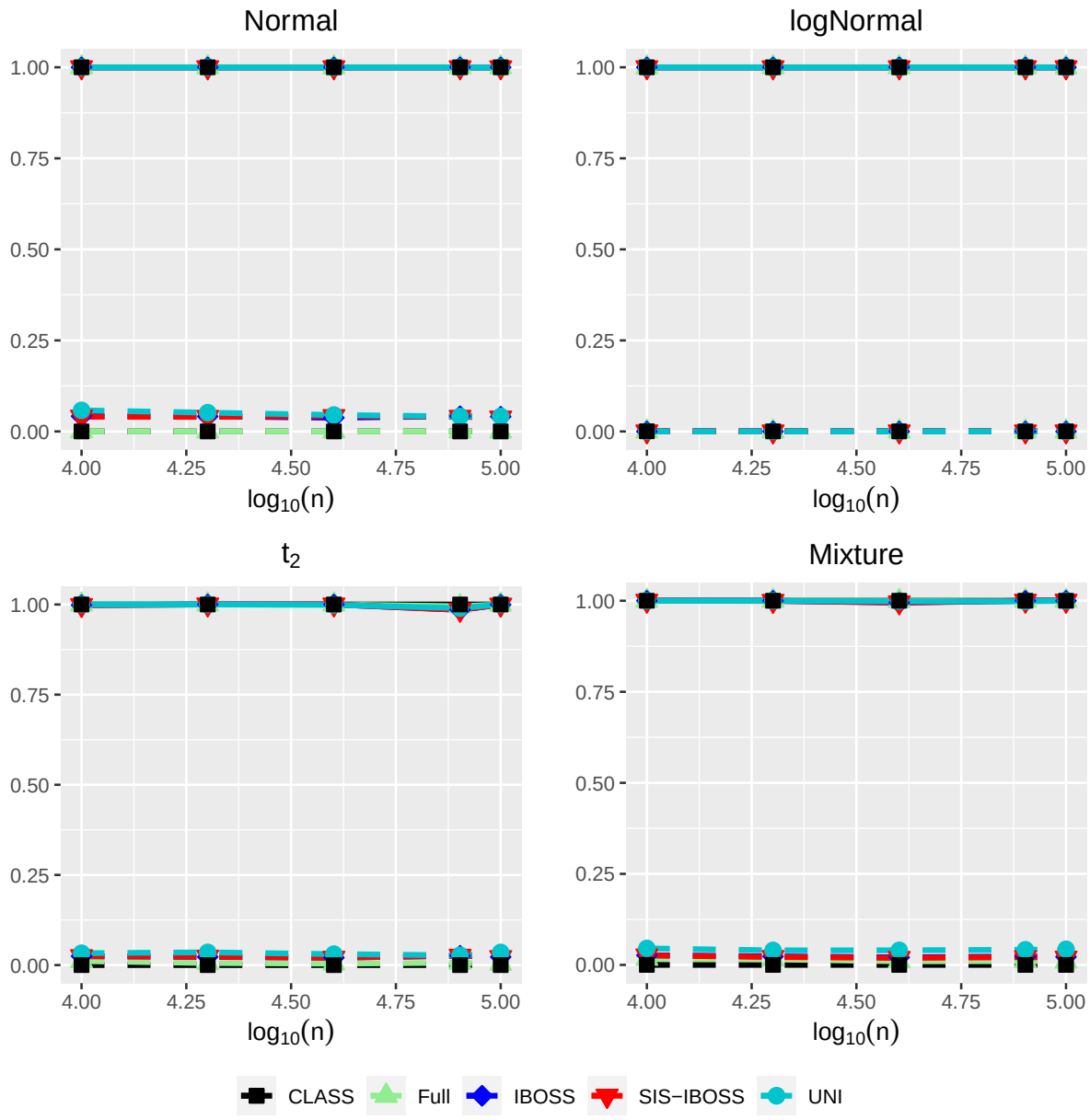


Figure 23: MSE for $k = 1000$, $p = 500$, $p_1 = 50$, and $\mathbf{\Sigma} = (0^{I(i \neq j)})$.

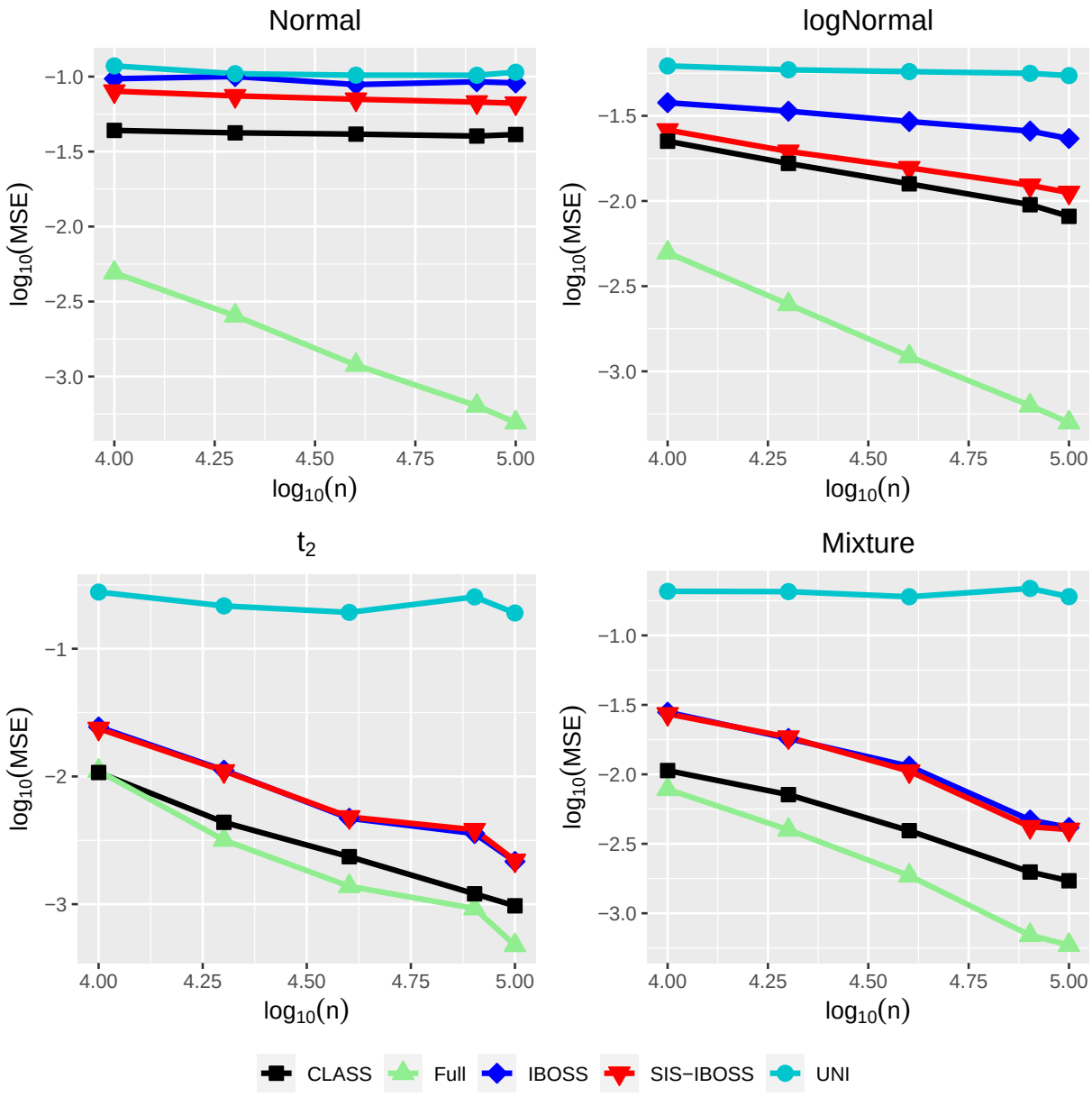


Figure 24: Variable selection performance for $k = 1000$, $p = 500$, $p_1 = 50$, and $\mathbf{\Sigma} = (0^{I(i \neq j)})$. The solid lines represent the power, whereas the dashed lines represent the error.

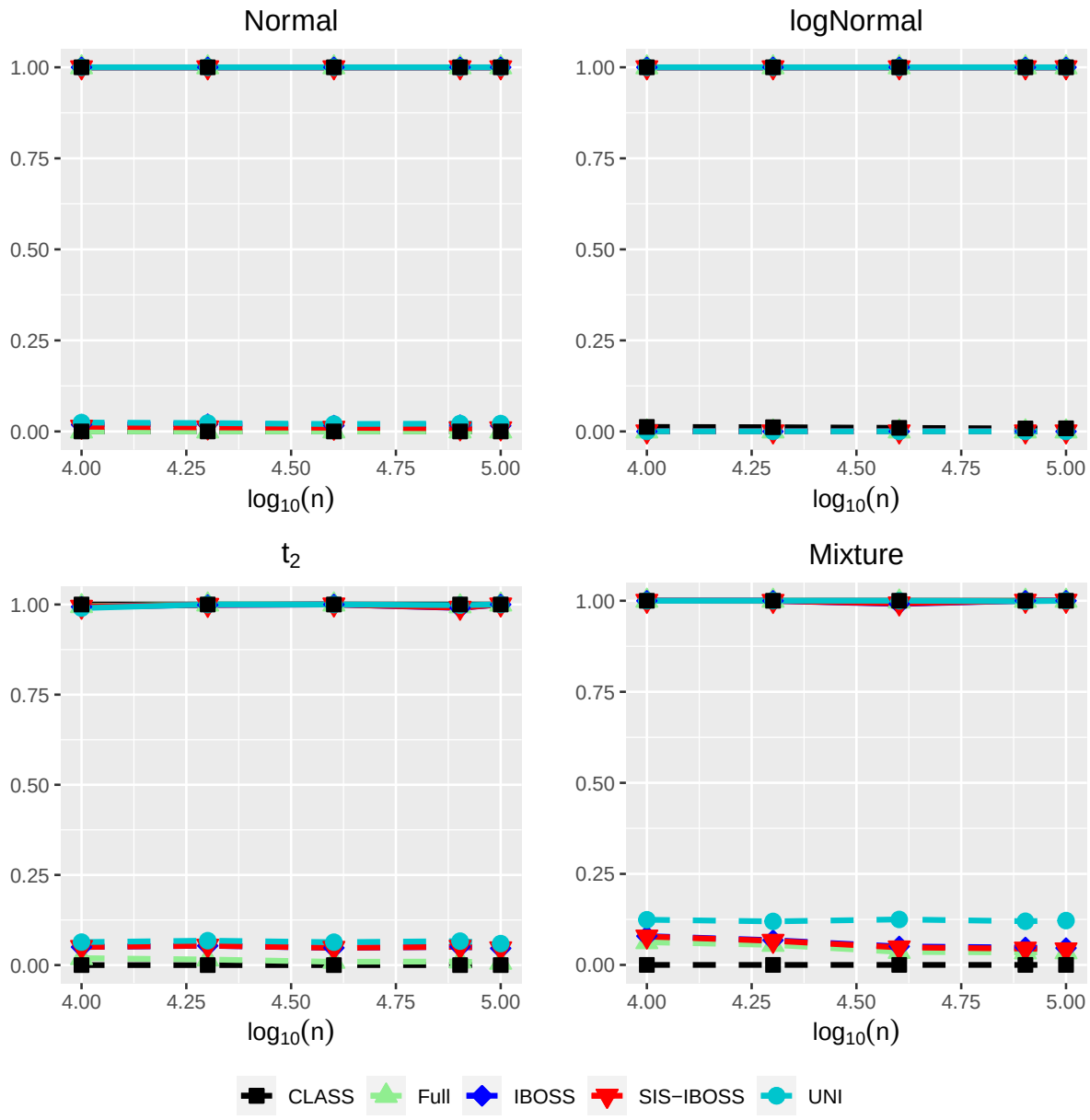


Figure 25: MSE for $k = 1000$, $p = 500$, $p_1 = 10$, and $\mathbf{\Sigma} = (0.5^{I(i \neq j)})$.

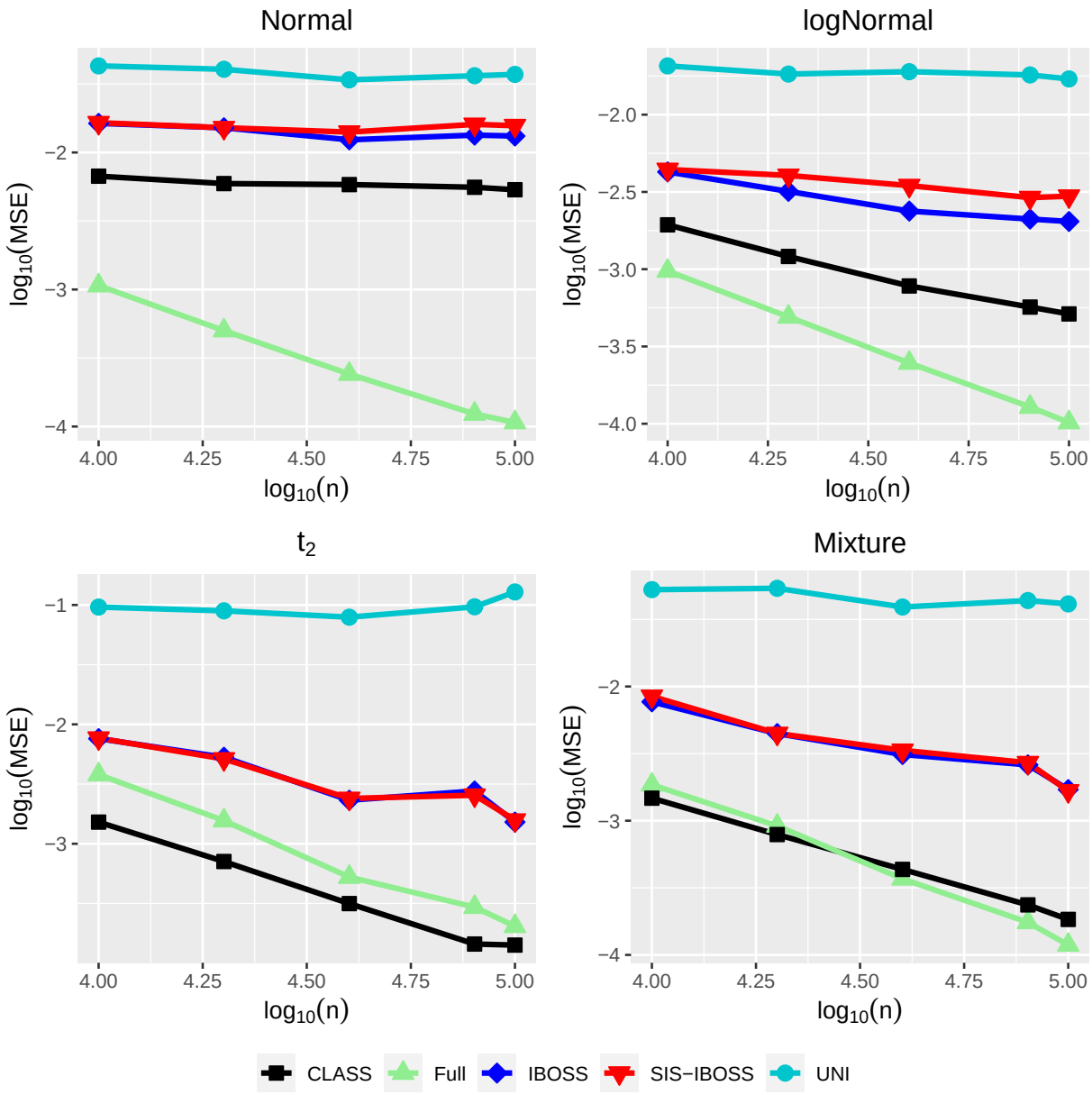


Figure 26: Variable selection performance for $k = 1000$, $p = 500$, $p_1 = 10$, and $\mathbf{\Sigma} = (0.5^{I(i \neq j)})$. The solid lines represent the power, whereas the dashed lines represent the error.

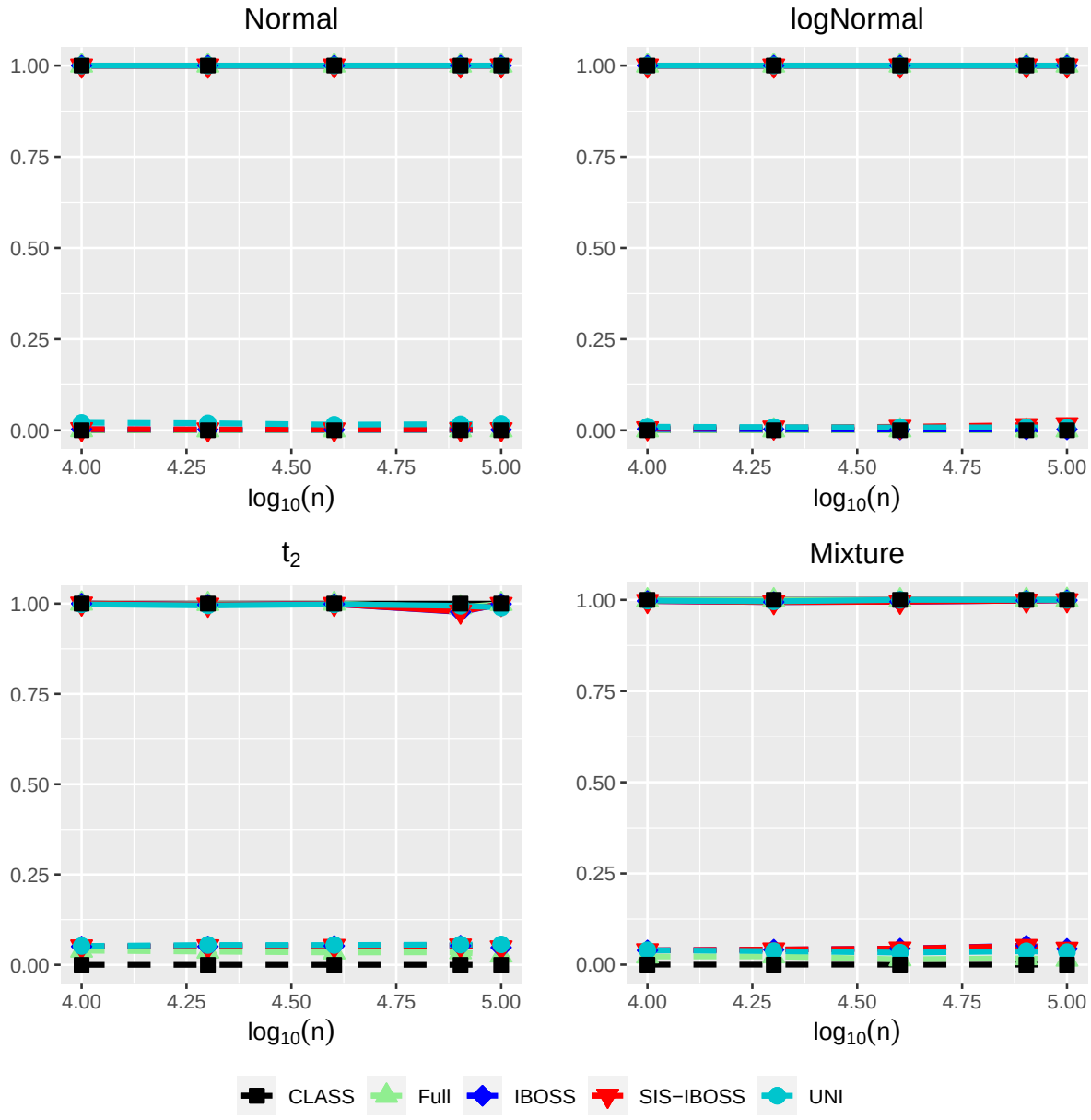


Figure 27: MSE for $k = 1000$, $p = 500$, $p_1 = 25$, and $\mathbf{\Sigma} = (0.5^{I(i \neq j)})$.

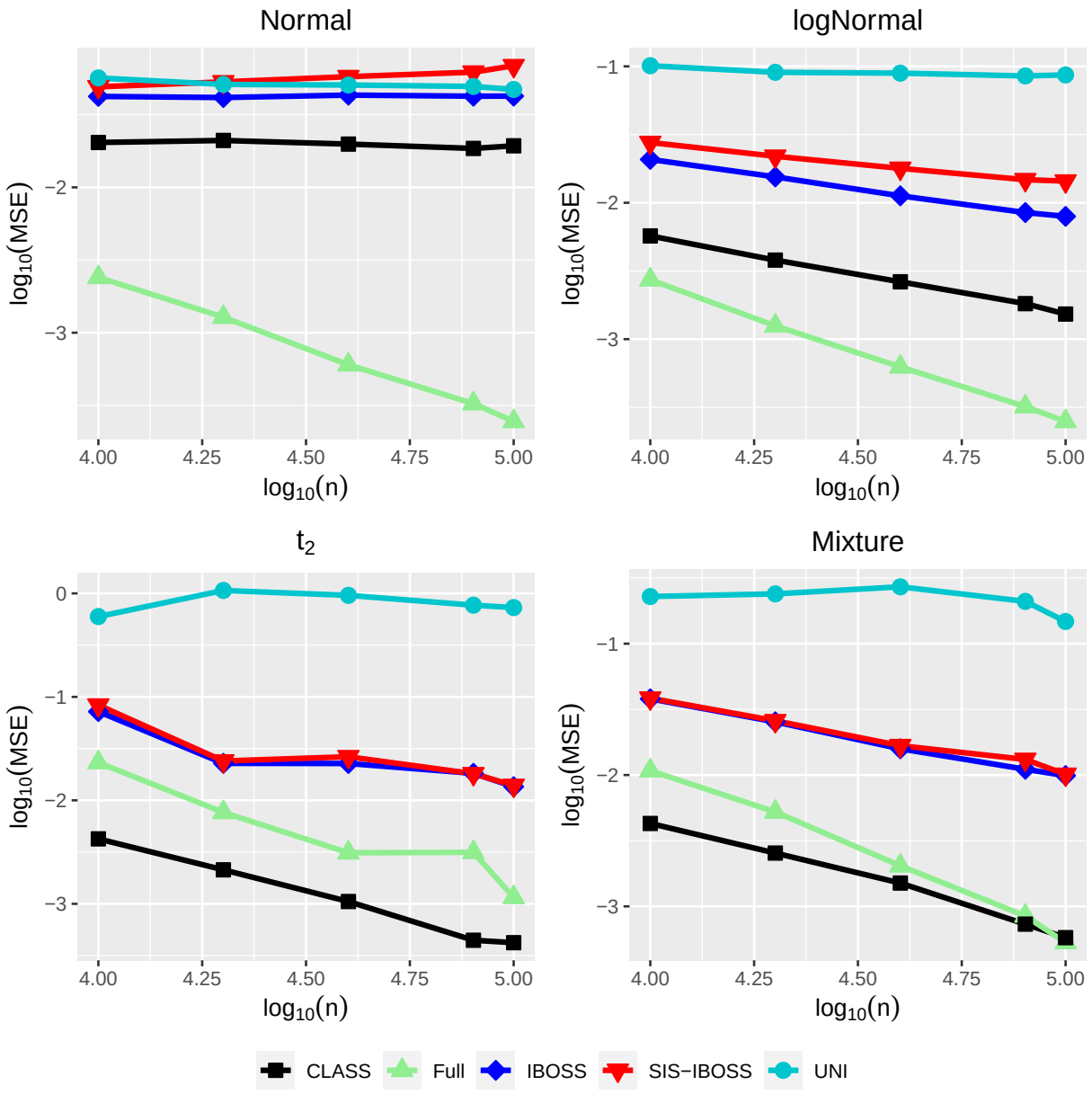


Figure 28: Variable selection performance for $k = 1000$, $p = 500$, $p_1 = 25$, and $\mathbf{\Sigma} = (0.5^{I(i \neq j)})$. The solid lines represent the power, whereas the dashed lines represent the error.

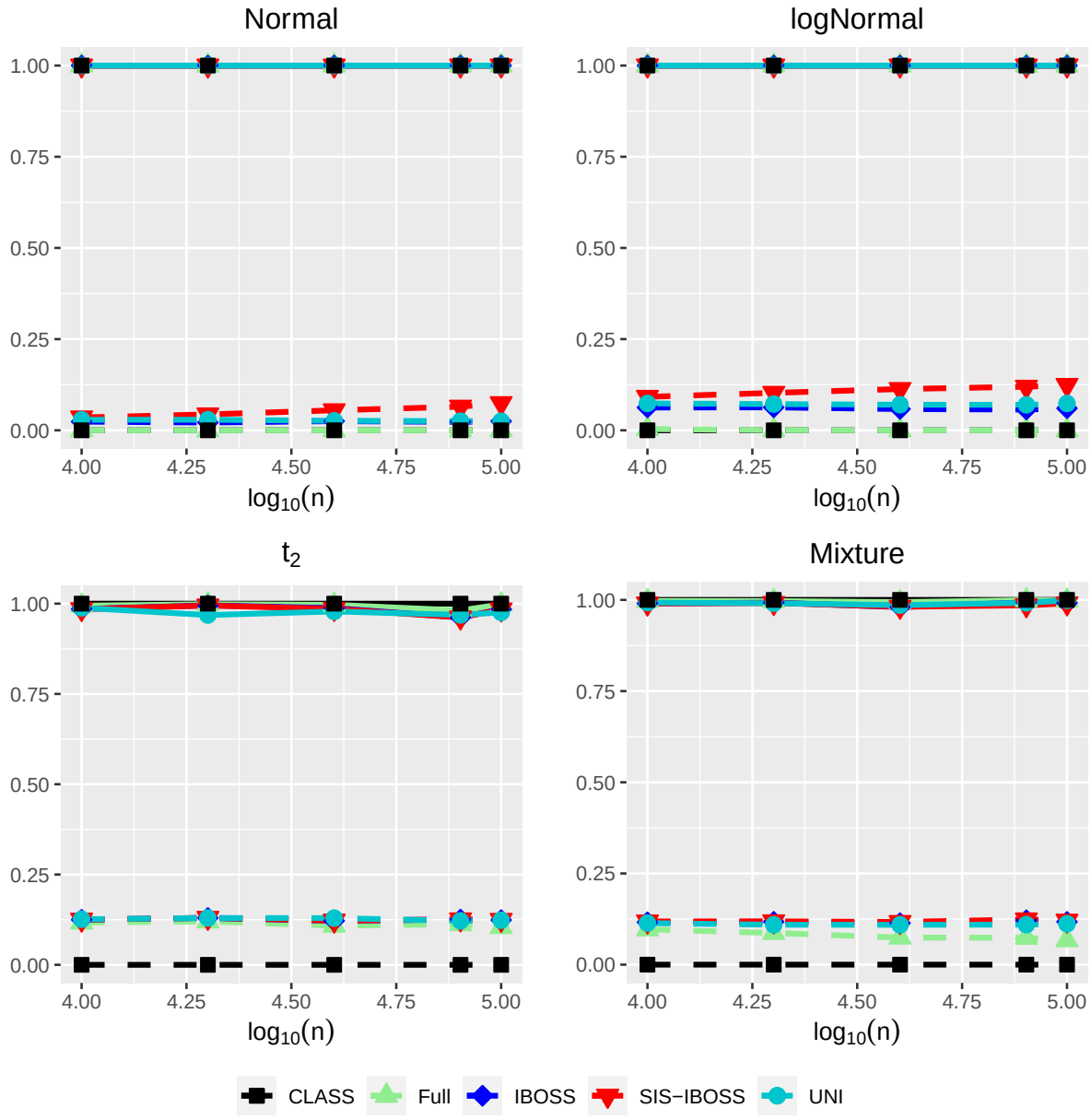


Figure 29: MSE for $k = 1000$, $p = 500$, $p_1 = 50$, and $\mathbf{\Sigma} = (0.5^{I(i \neq j)})$.

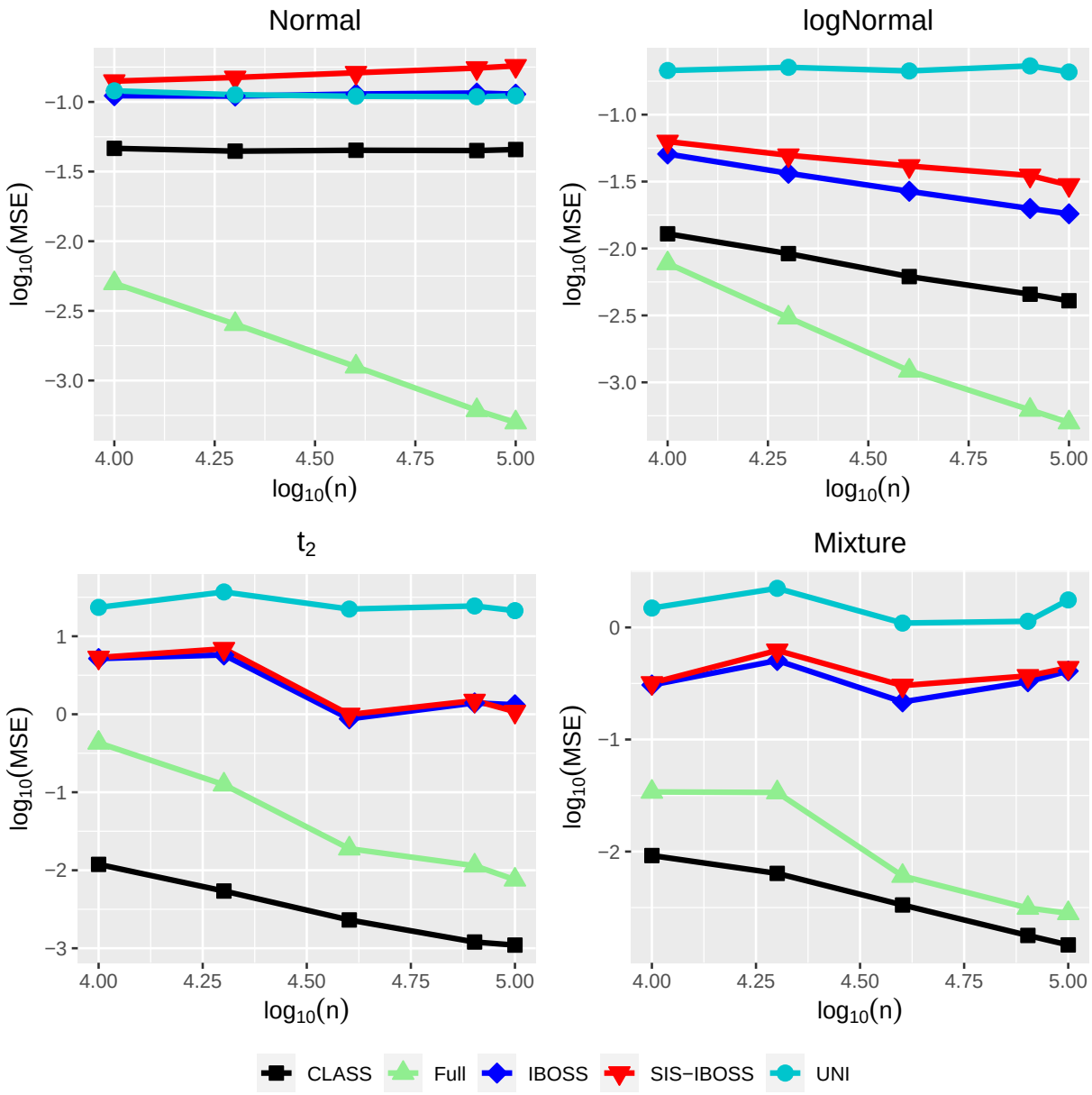


Figure 30: Variable selection performance for $k = 1000$, $p = 500$, $p_1 = 50$, and $\mathbf{\Sigma} = (0.5^{I(i \neq j)})$. The solid lines represent the power, whereas the dashed lines represent the error.

Figure 31: MSE for $k = 1000$, $p = 500$, $p_1 = 10$, and $\mathbf{\Sigma}$ is Random.

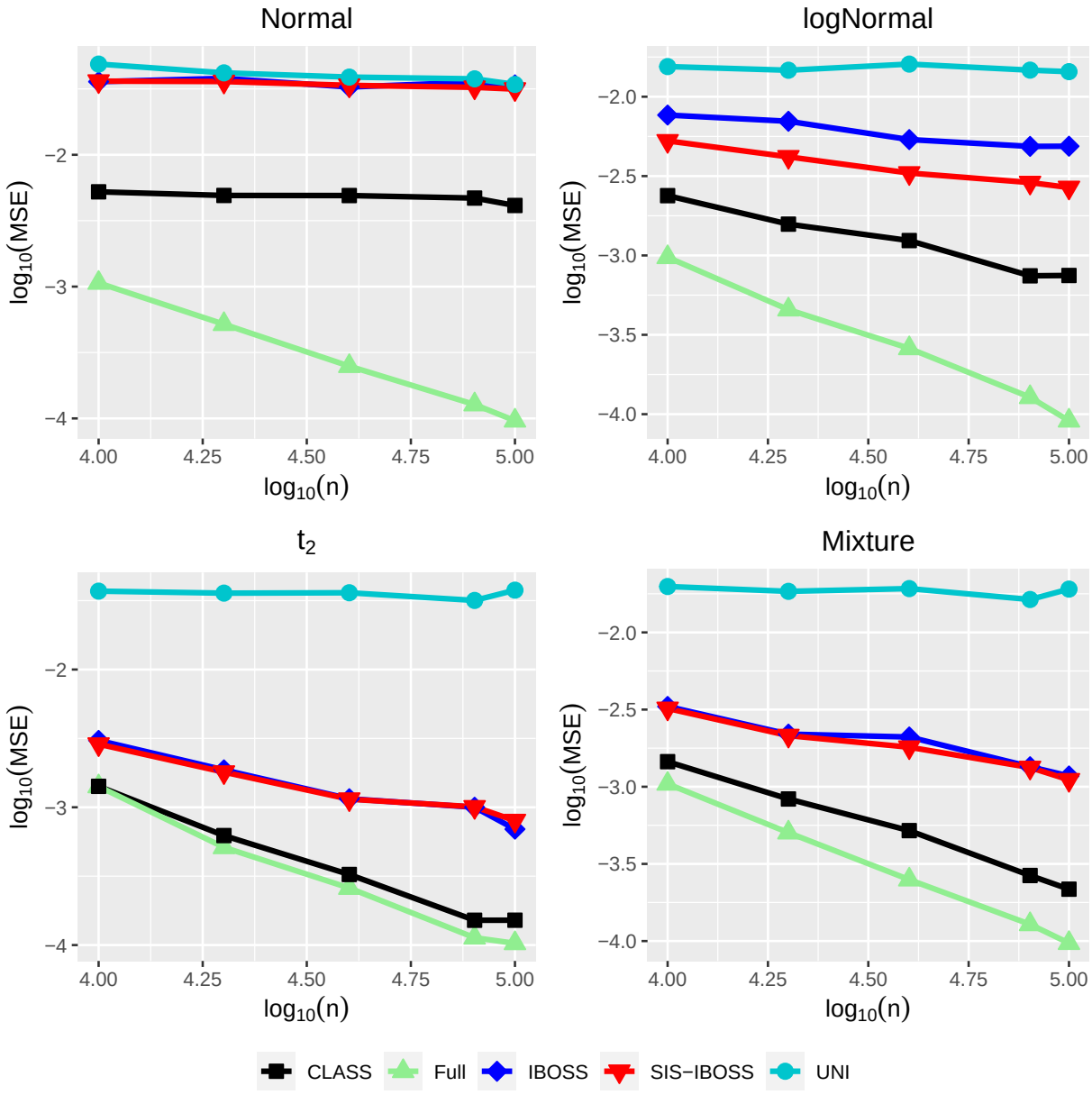


Figure 32: Variable selection performance for $k = 1000$, $p = 500$, $p_1 = 10$, and $\mathbf{\Sigma}$ is Random. The solid lines represent the power, whereas the dashed lines represent the error.

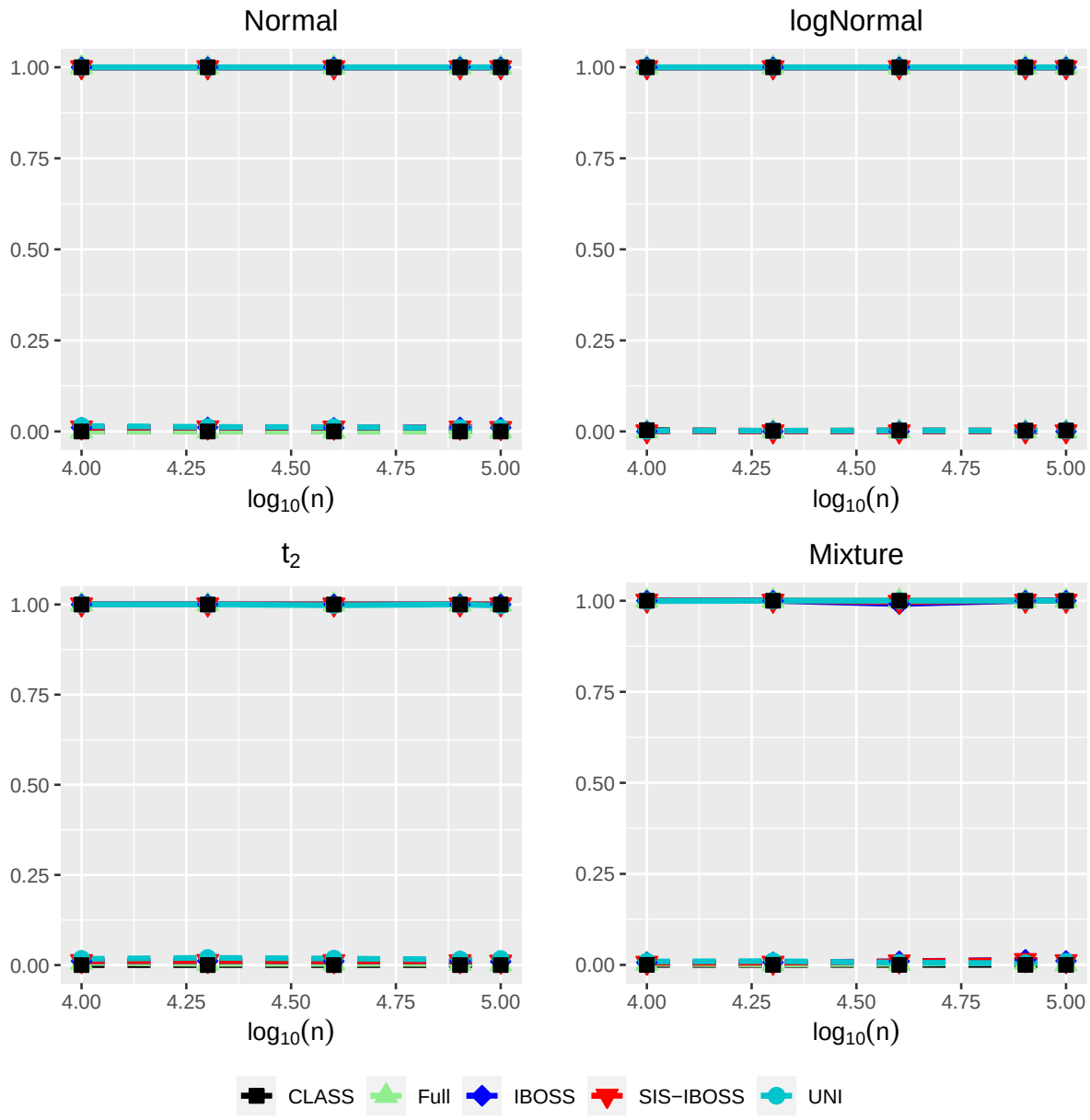


Figure 33: MSE for $k = 1000$, $p = 500$, $p_1 = 25$, and $\mathbf{\Sigma}$ is Random.

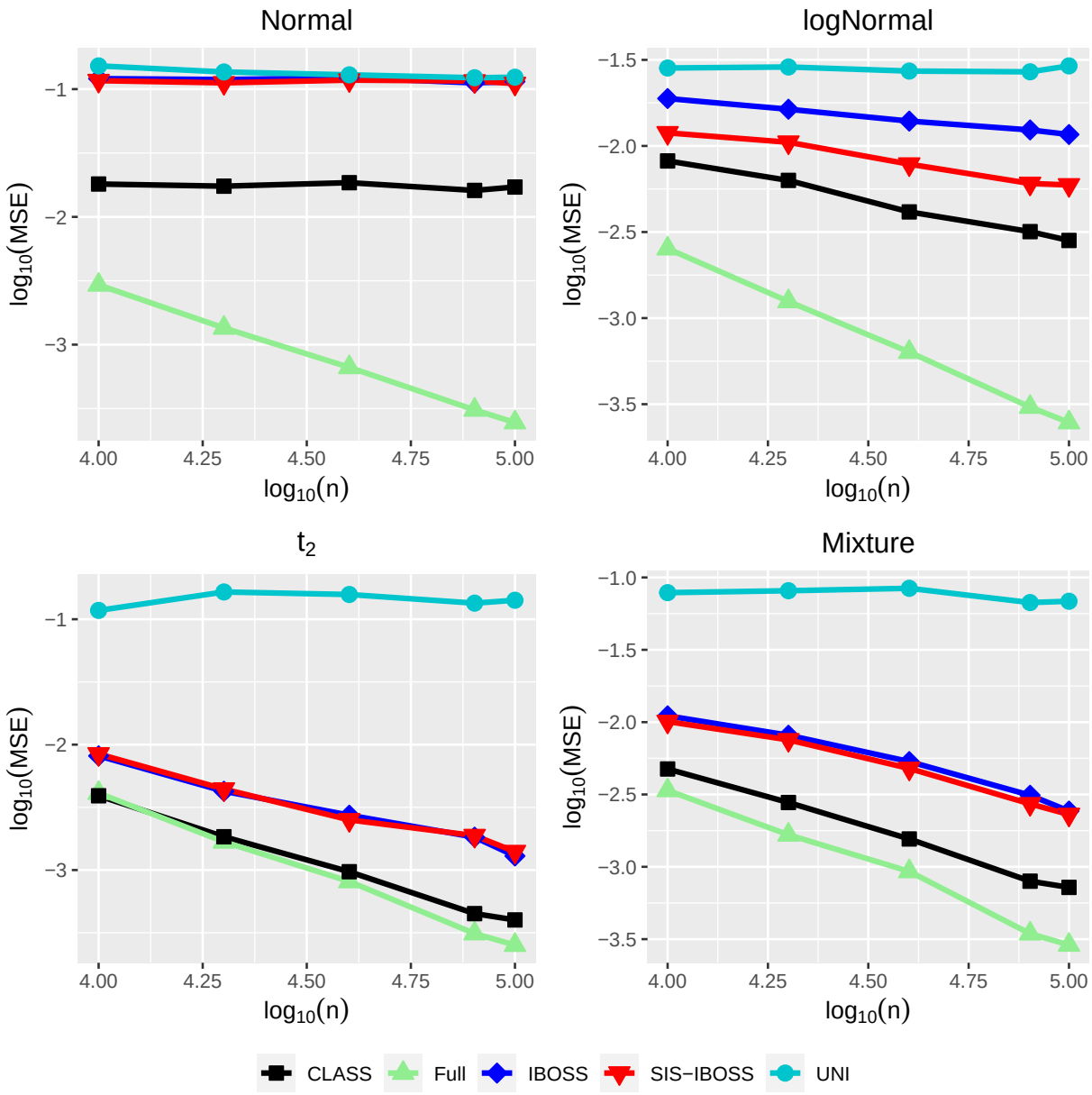


Figure 34: Variable selection performance for $k = 1000$, $p = 500$, $p_1 = 25$, and $\mathbf{\Sigma}$ is Random. The solid lines represent the power, whereas the dashed lines represent the error.

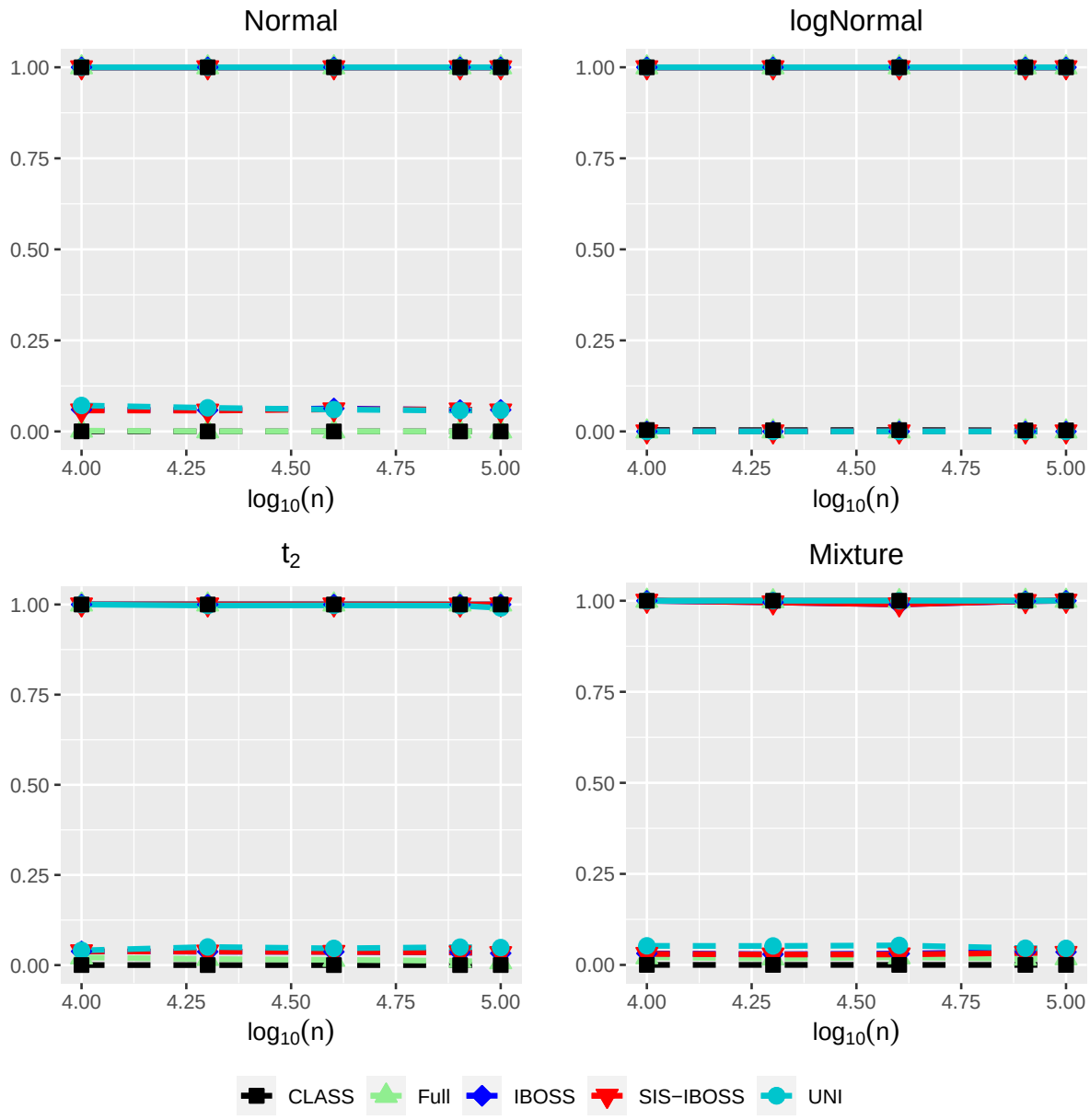


Figure 35: MSE for $k = 1000$, $p = 500$, $p_1 = 50$, and $\mathbf{\Sigma}$ is Random.

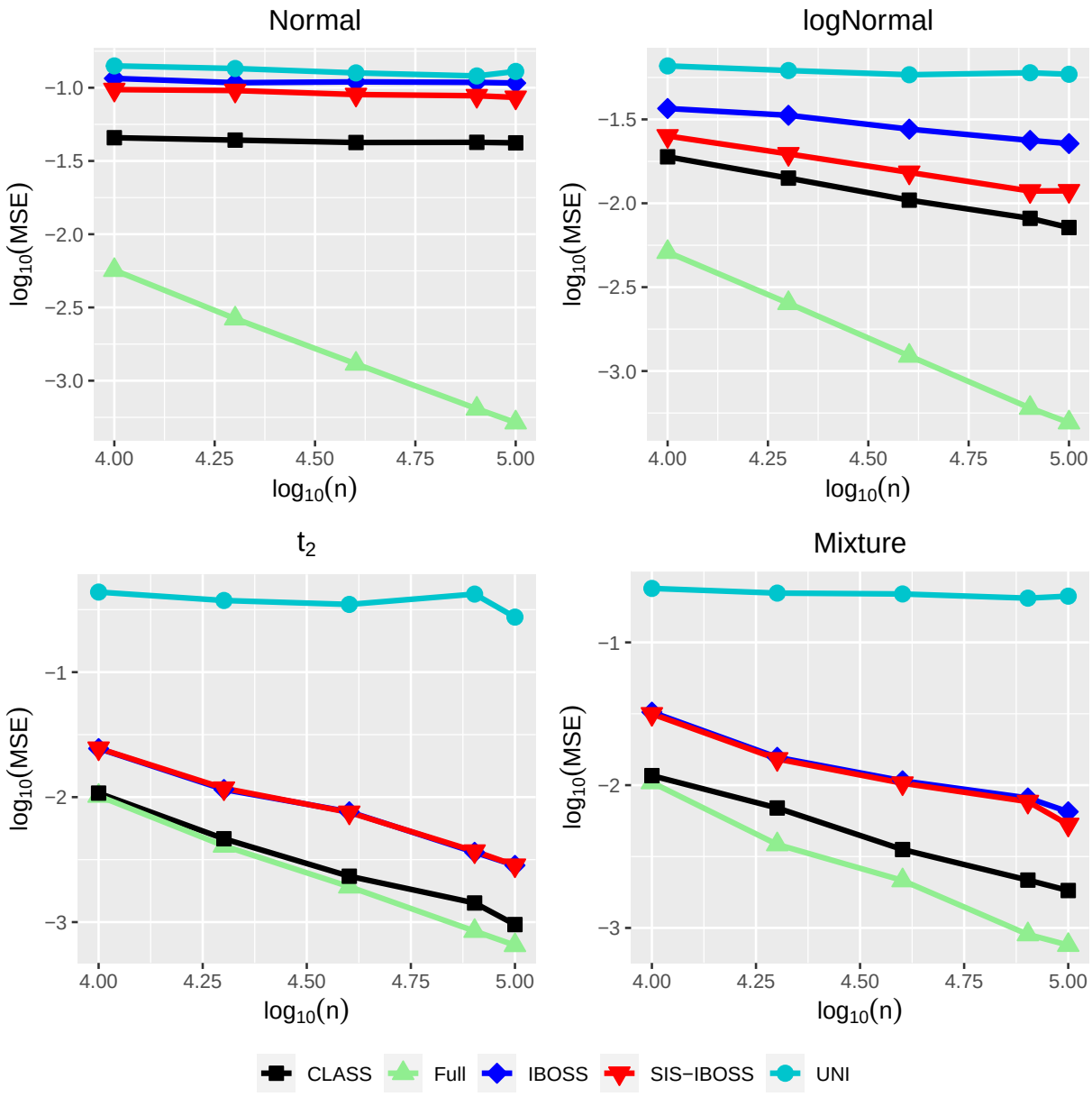
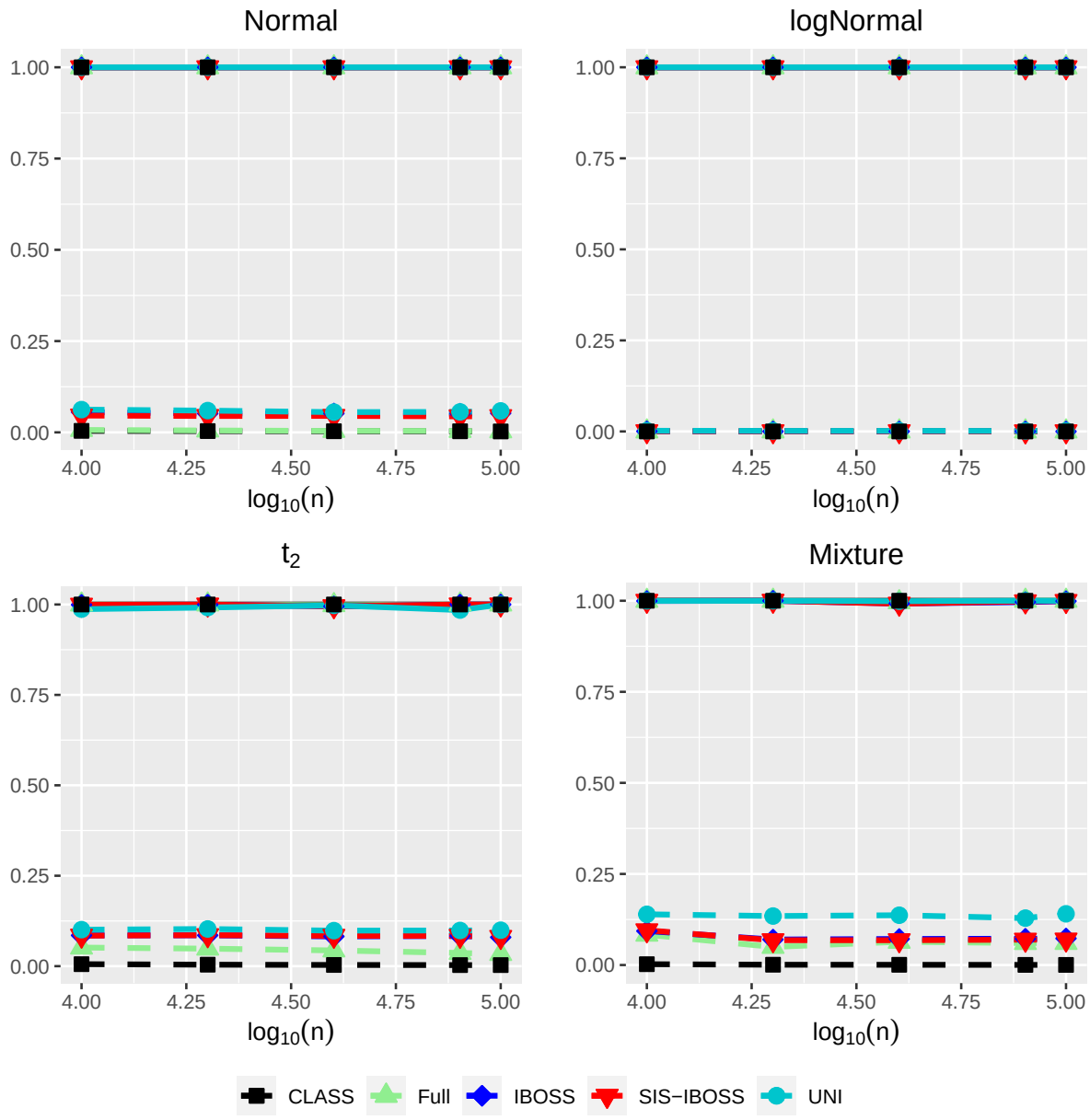


Figure 36: Variable selection performance for $k = 1000$, $p = 500$, $p_1 = 50$, and $\mathbf{\Sigma}$ is Random. The solid lines represent the power, whereas the dashed lines represent the error.



4 Changing k for $p = 500$ and $n = 10^5$ with error standard deviation equals 1

Figure 37: MSE for $n = 10^5$, $p = 500$, $p_1 = 10$, and $\mathbf{\Sigma} = (0^{I(i \neq j)})$.

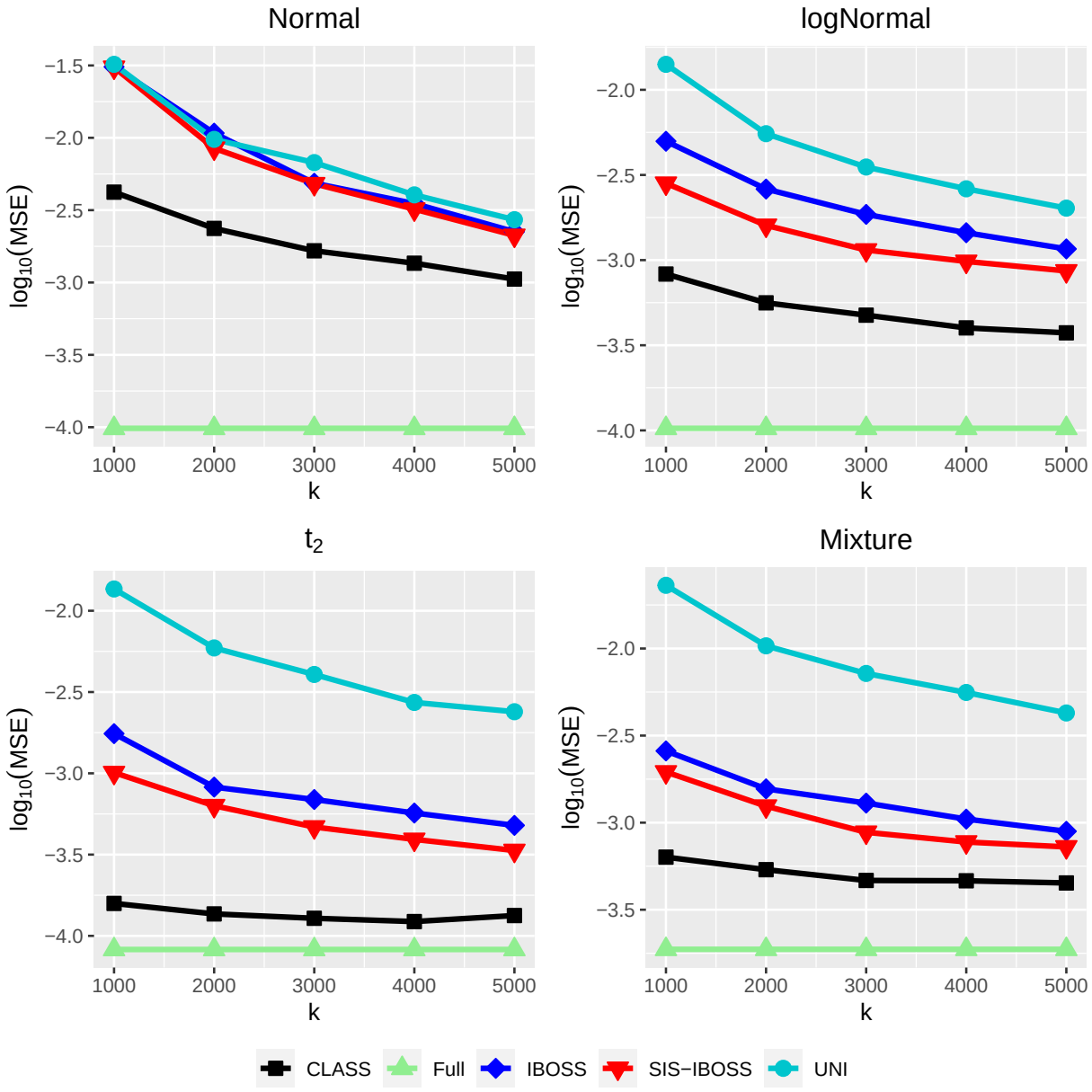


Figure 38: Variable selection performance for $n = 10^5$, $p = 500$, $p_1 = 10$, and $\mathbf{\Sigma} = (0^{I(i \neq j)})$. The solid lines represent the power, whereas the dashed lines represent the error.

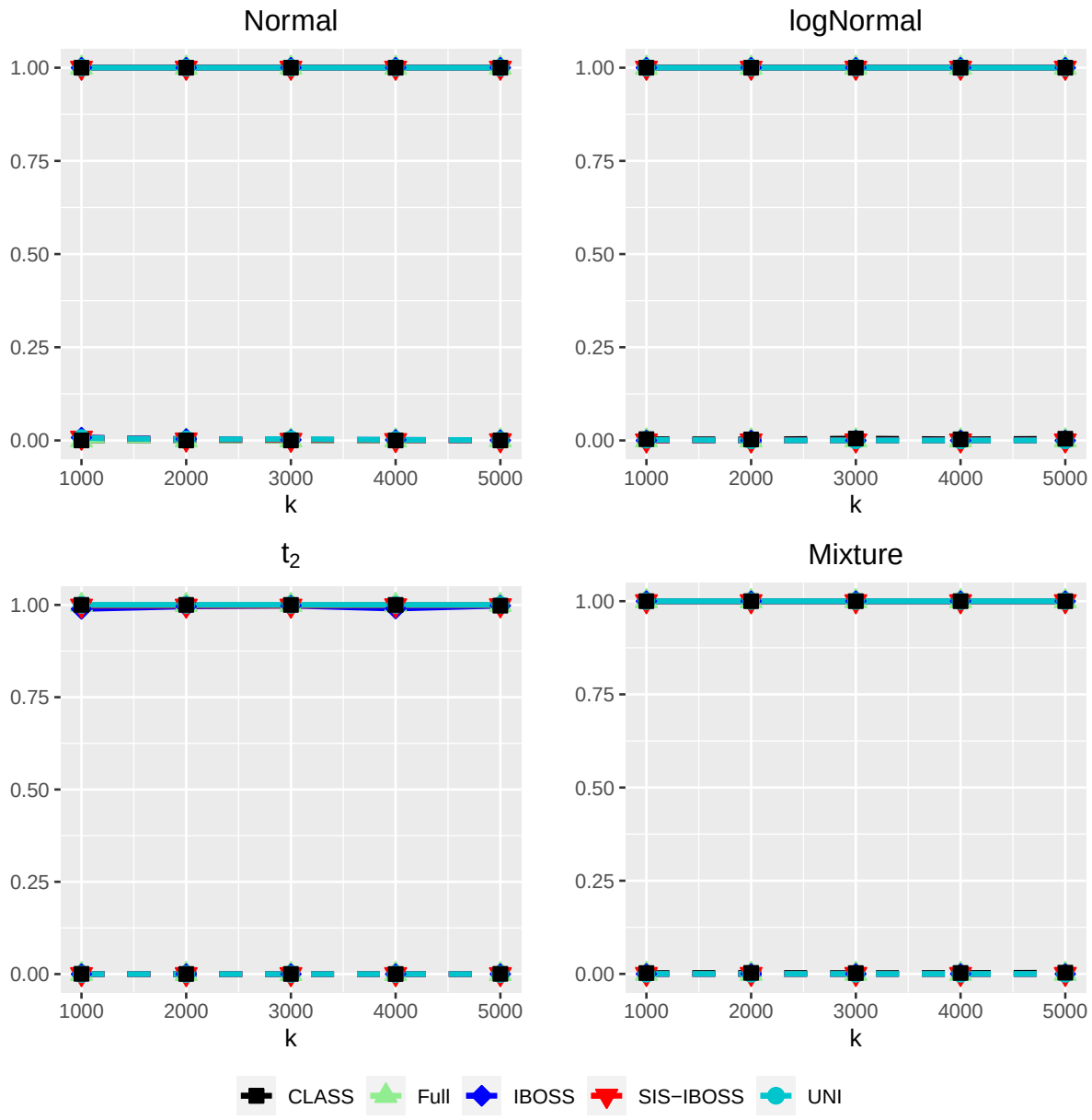


Figure 39: MSE for $n = 10^5$, $p = 500$, $p_1 = 25$, and $\mathbf{\Sigma} = (0^{I(i \neq j)})$.

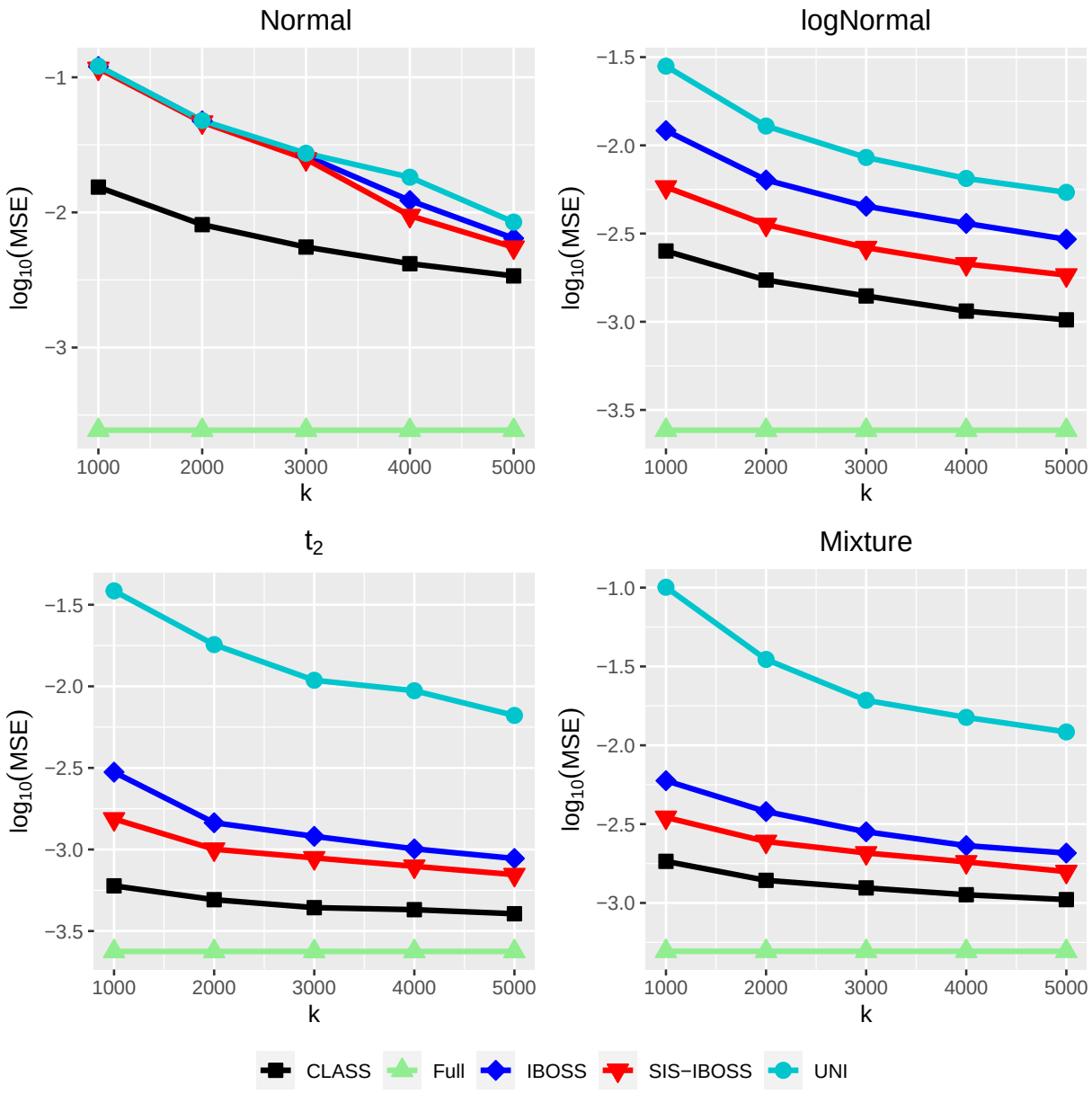


Figure 40: Variable selection performance for $n = 10^5$, $p = 500$, $p_1 = 25$, and $\mathbf{\Sigma} = (0^{I(i \neq j)})$. The solid lines represent the power, whereas the dashed lines represent the error.

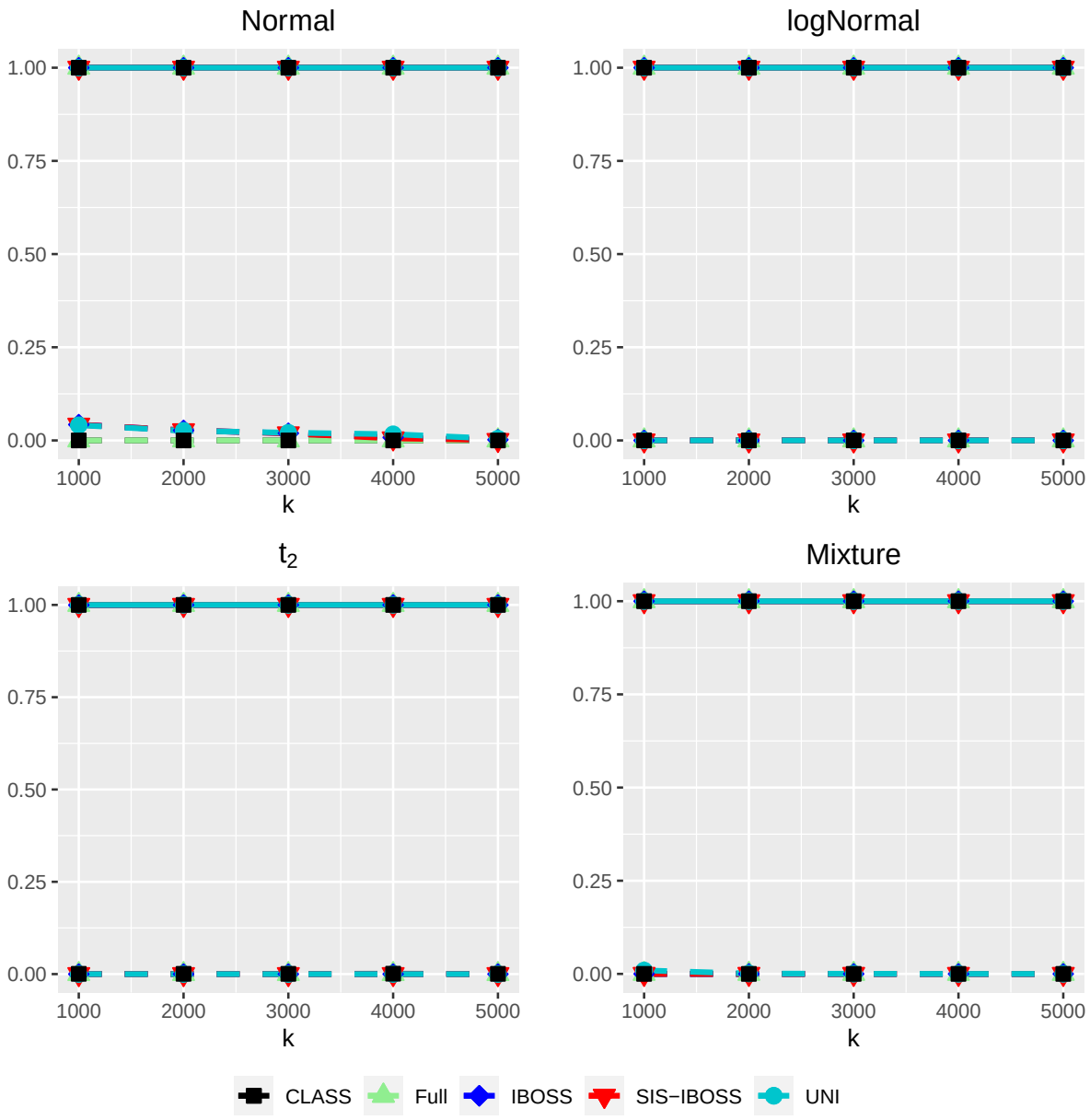


Figure 41: MSE for $n = 10^5$, $p = 500$, $p_1 = 50$, and $\mathbf{\Sigma} = (0^{I(i \neq j)})$.

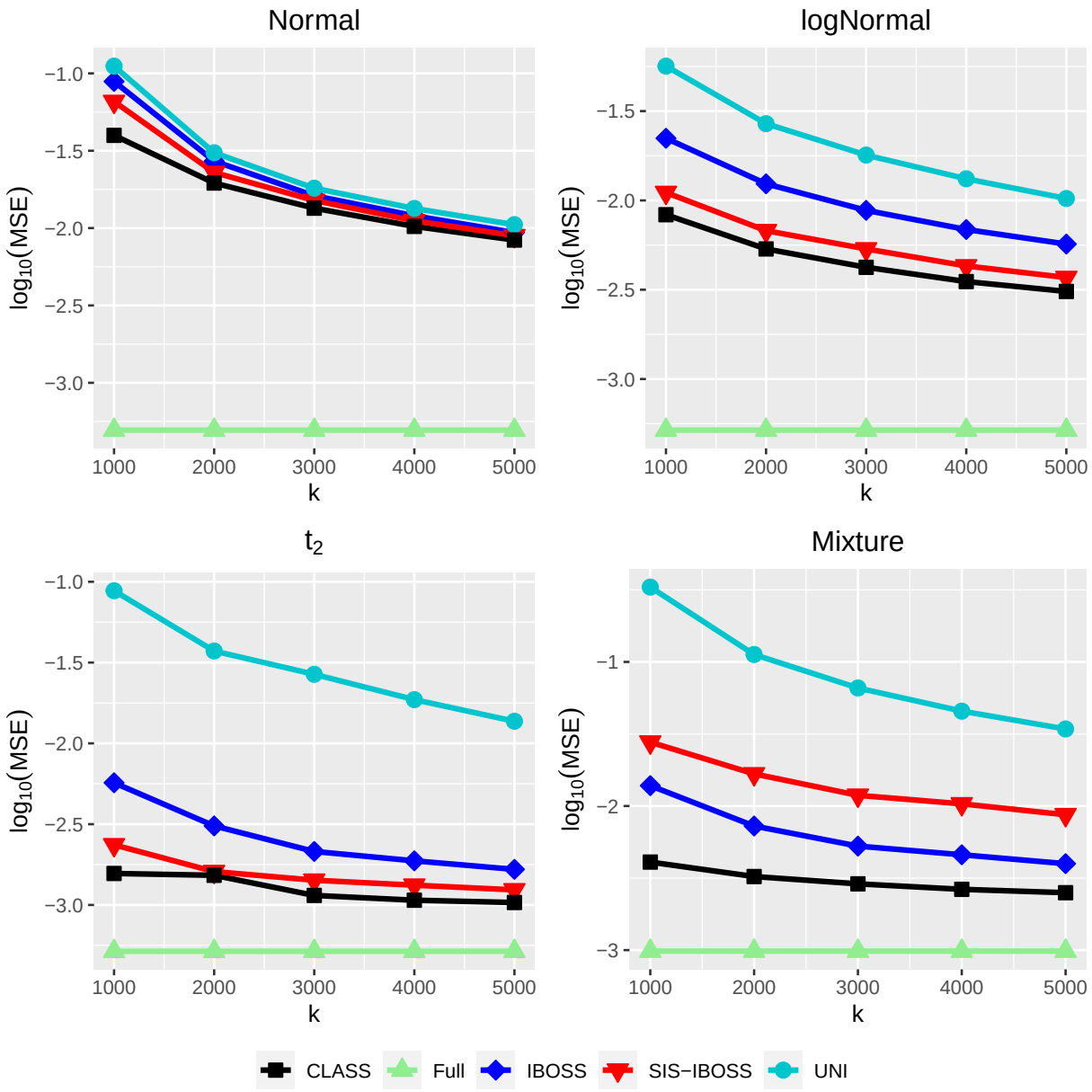


Figure 42: Variable selection performance for $n = 10^5$, $p = 500$, $p_1 = 50$, and $\mathbf{\Sigma} = (0^{I(i \neq j)})$. The solid lines represent the power, whereas the dashed lines represent the error.

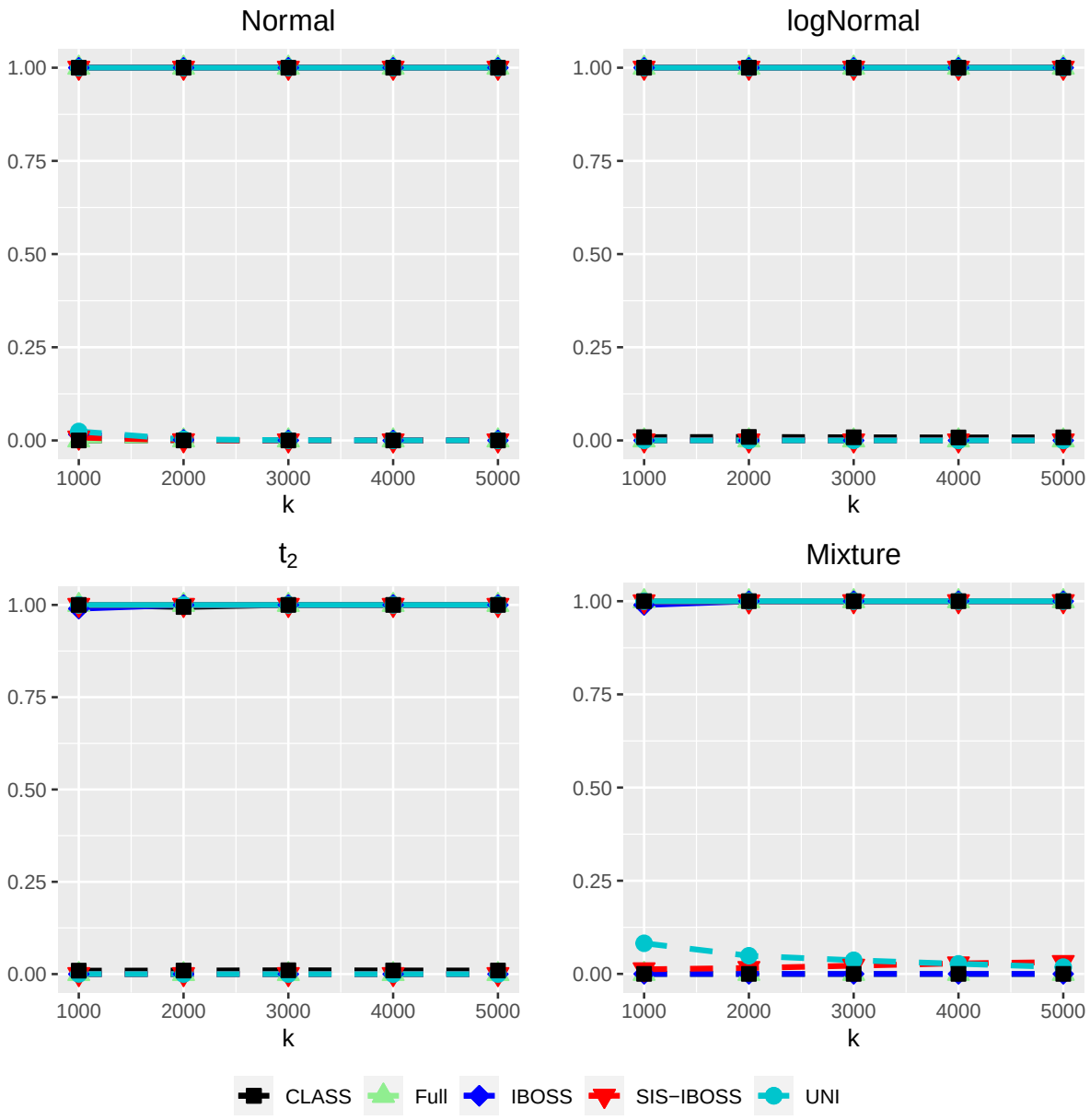


Figure 43: MSE for $n = 10^5$, $p = 500$, $p_1 = 10$, and $\mathbf{\Sigma} = (0.5^{I(i \neq j)})$.

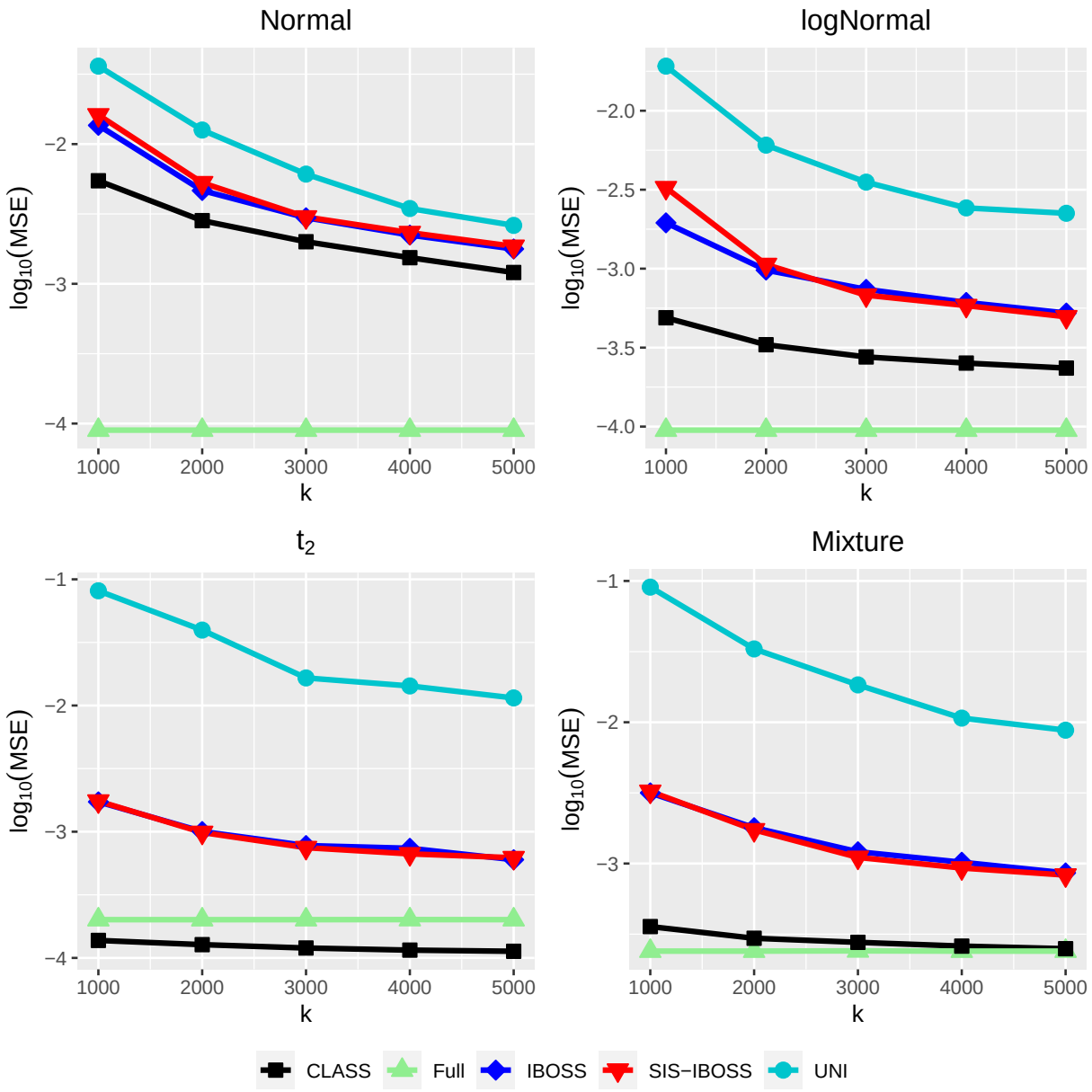


Figure 44: Variable selection performance for $n = 10^5$, $p = 500$, $p_1 = 10$, and $\mathbf{\Sigma} = (0.5^{I(i \neq j)})$. The solid lines represent the power, whereas the dashed lines represent the error.

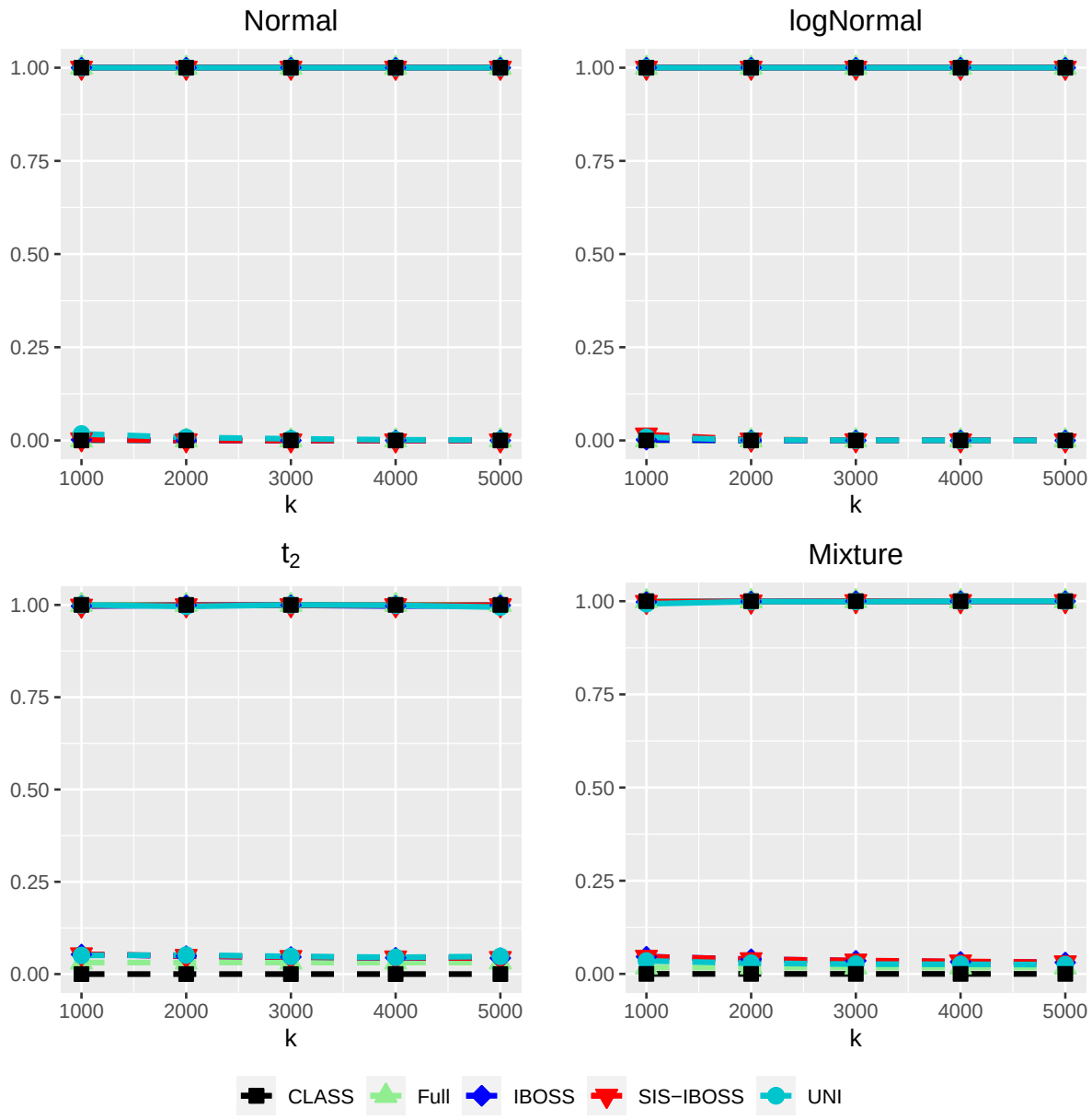


Figure 45: MSE for $n = 10^5$, $p = 500$, $p_1 = 25$, and $\mathbf{\Sigma} = (0.5^{I(i \neq j)})$.

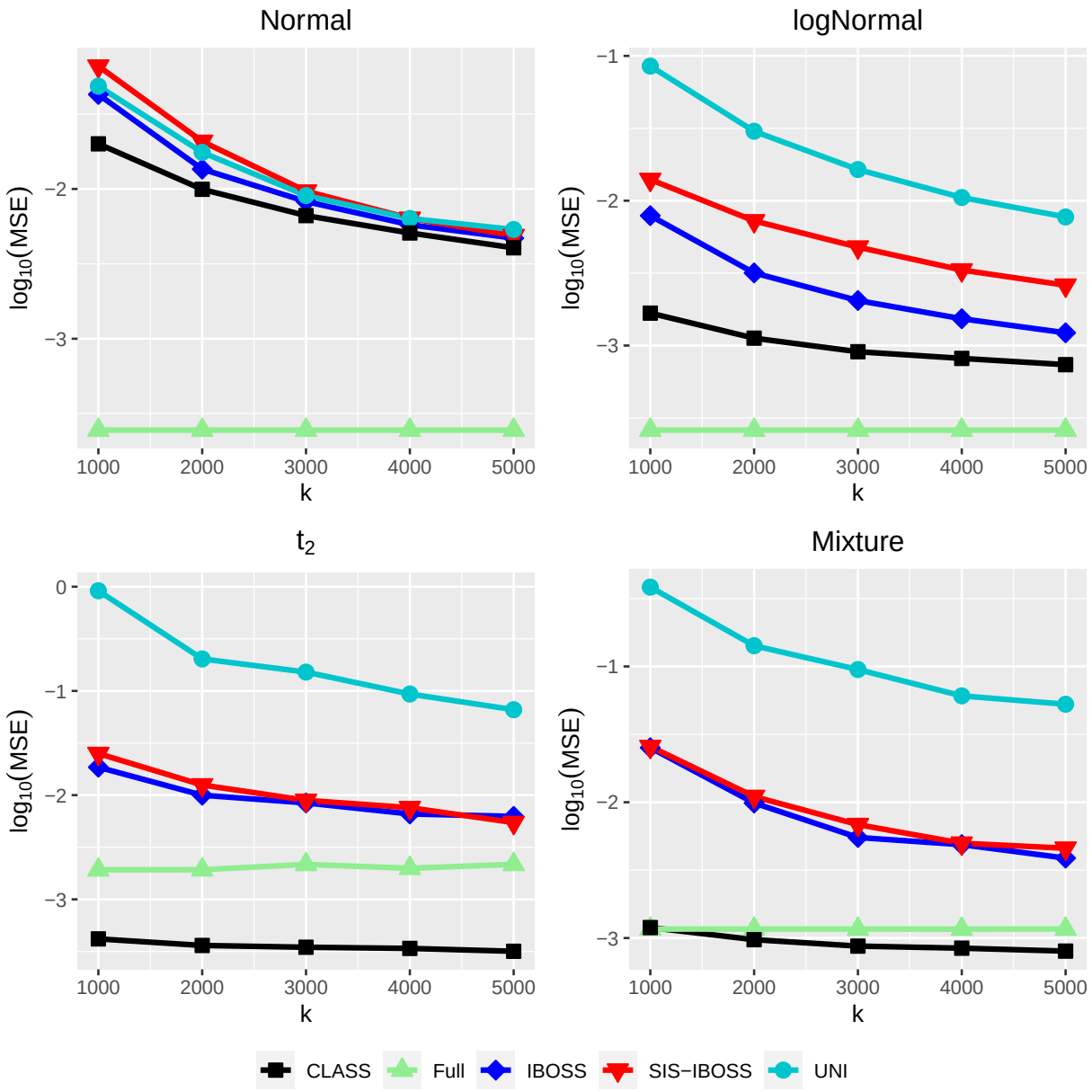


Figure 46: Variable selection performance for for $n = 10^5$, $p = 500$, $p_1 = 25$, and $\mathbf{\Sigma} = (0.5^{I(i \neq j)})$. The solid lines represent the power, whereas the dashed lines represent the error.

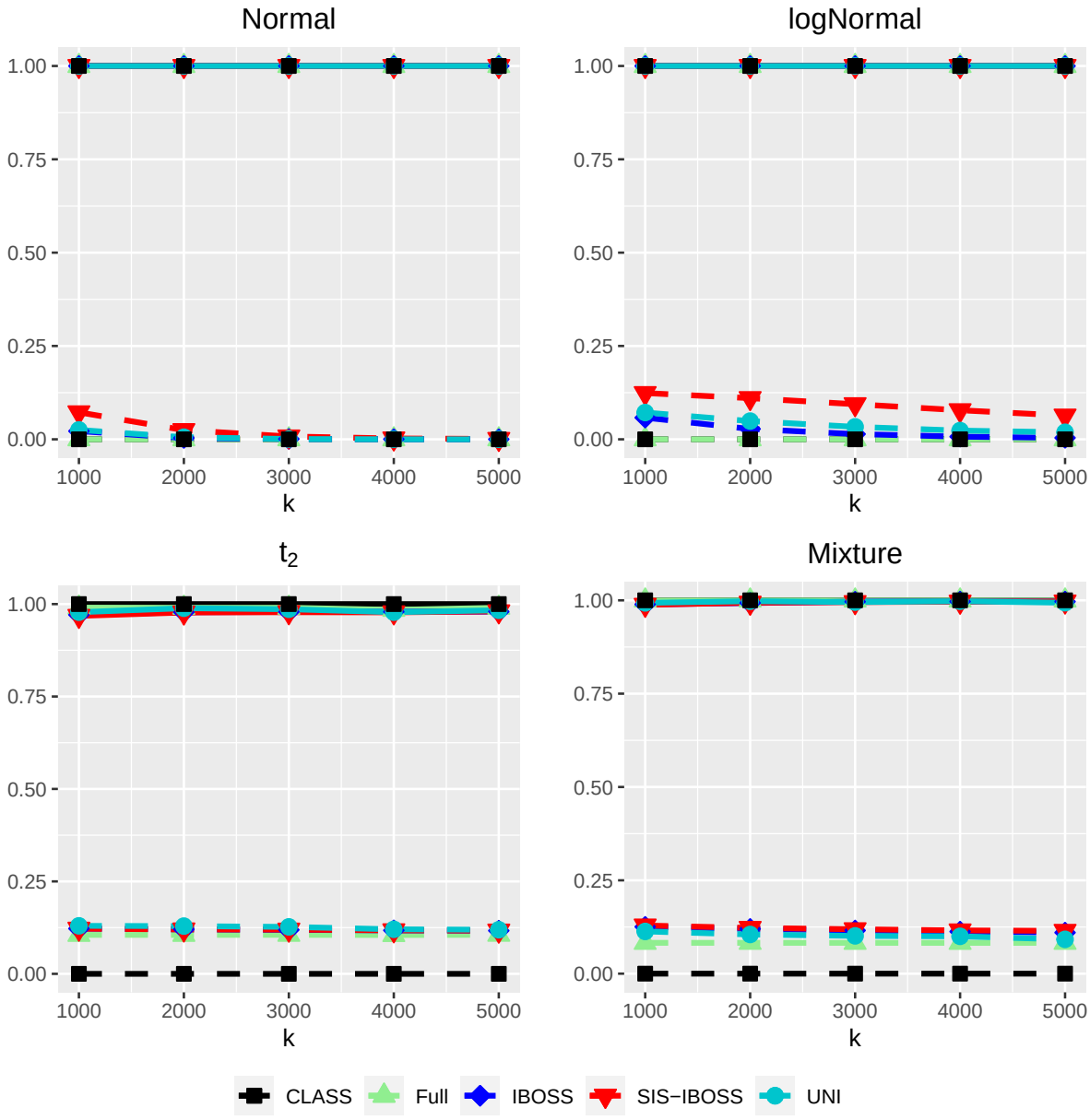


Figure 47: MSE for $n = 10^5$, $p = 500$, $p_1 = 50$, and $\mathbf{\Sigma} = (0.5^{I(i \neq j)})$.

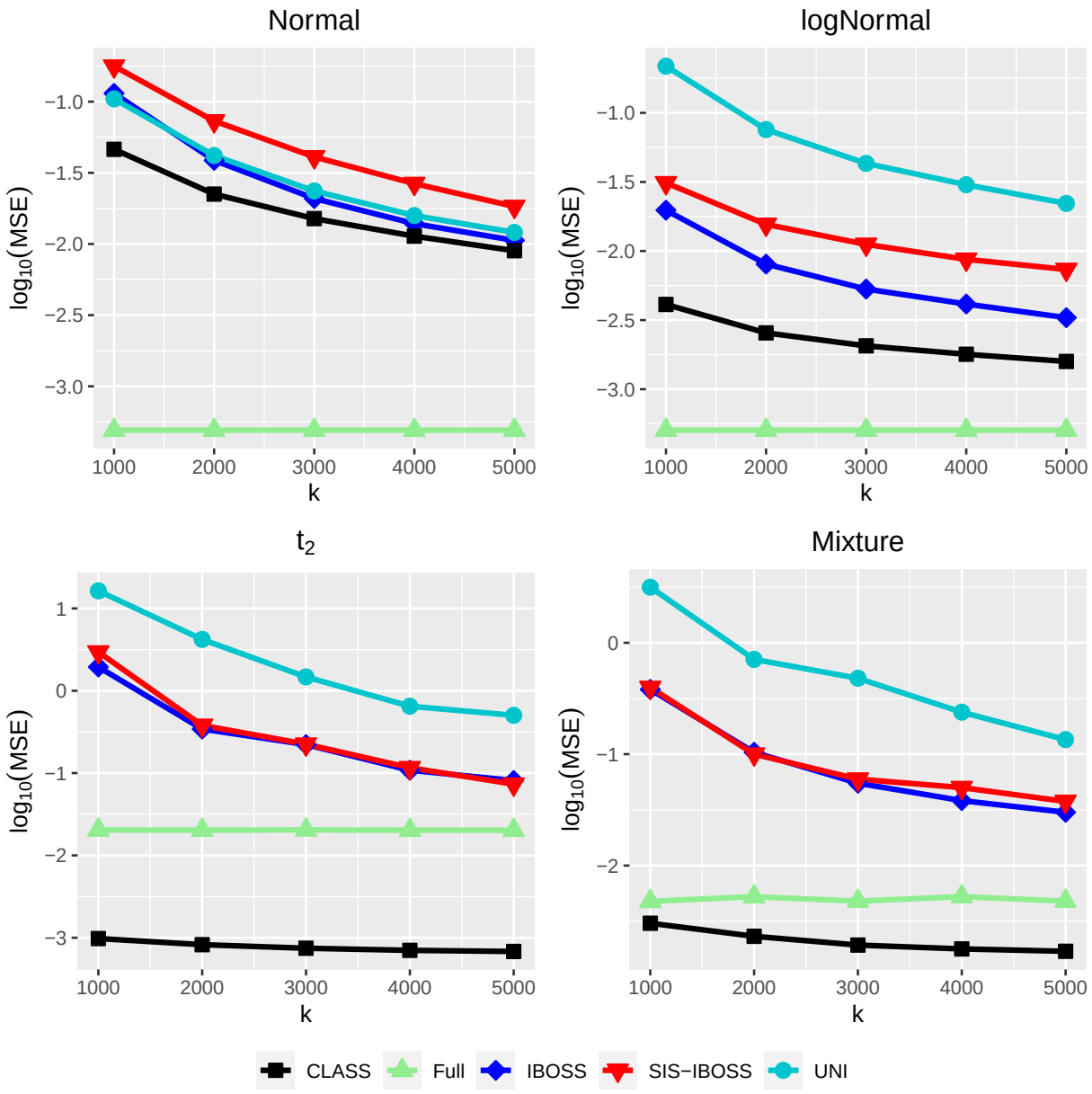


Figure 48: Variable selection performance for $n = 10^5$, $p = 500$, $p_1 = 50$, and $\mathbf{\Sigma} = (0.5^{I(i \neq j)})$. The solid lines represent the power, whereas the dashed lines represent the error.

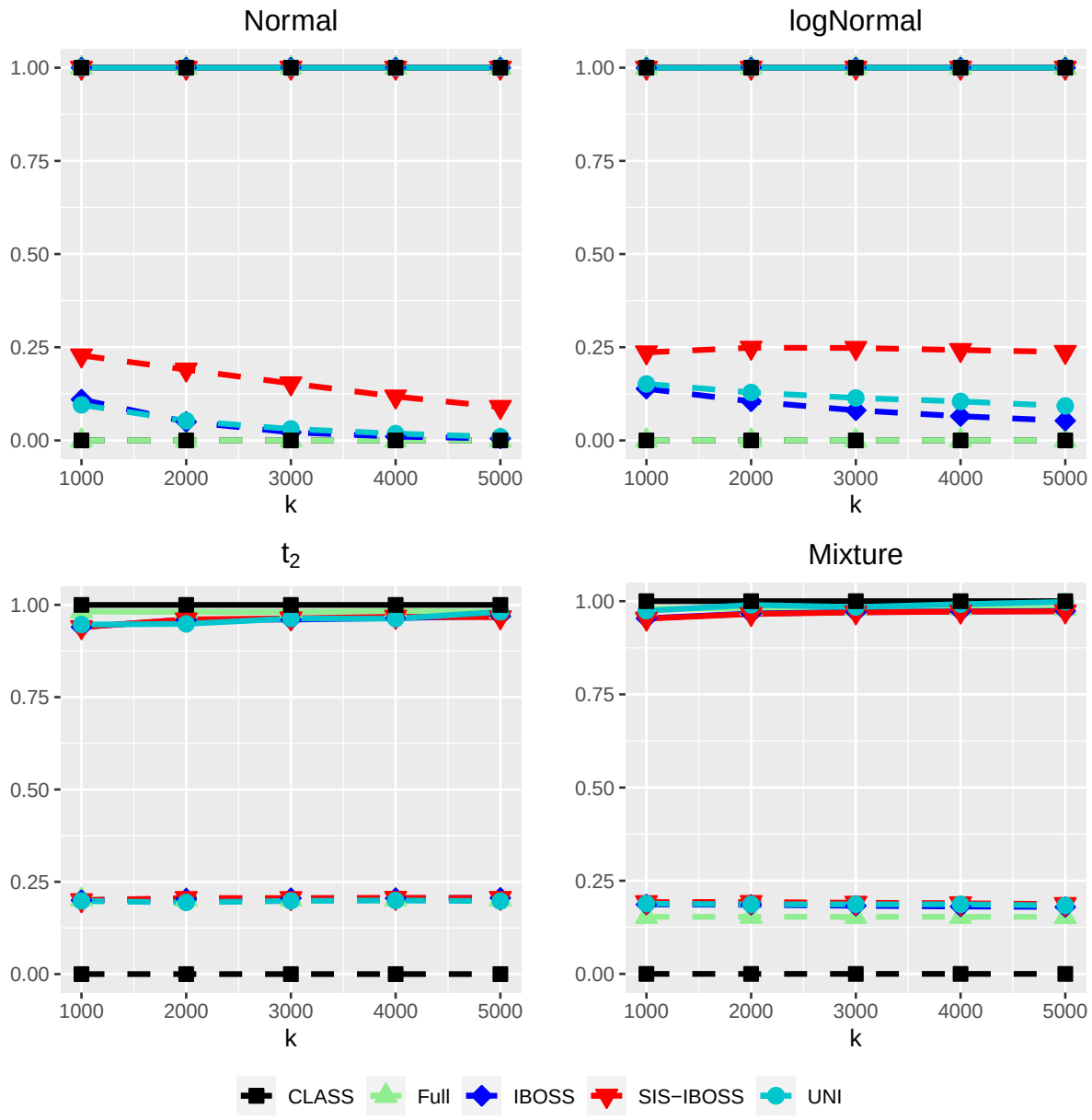


Figure 49: MSE for $n = 10^5$, $p = 500$, $p_1 = 10$, and $\mathbf{\Sigma}$ is Random.

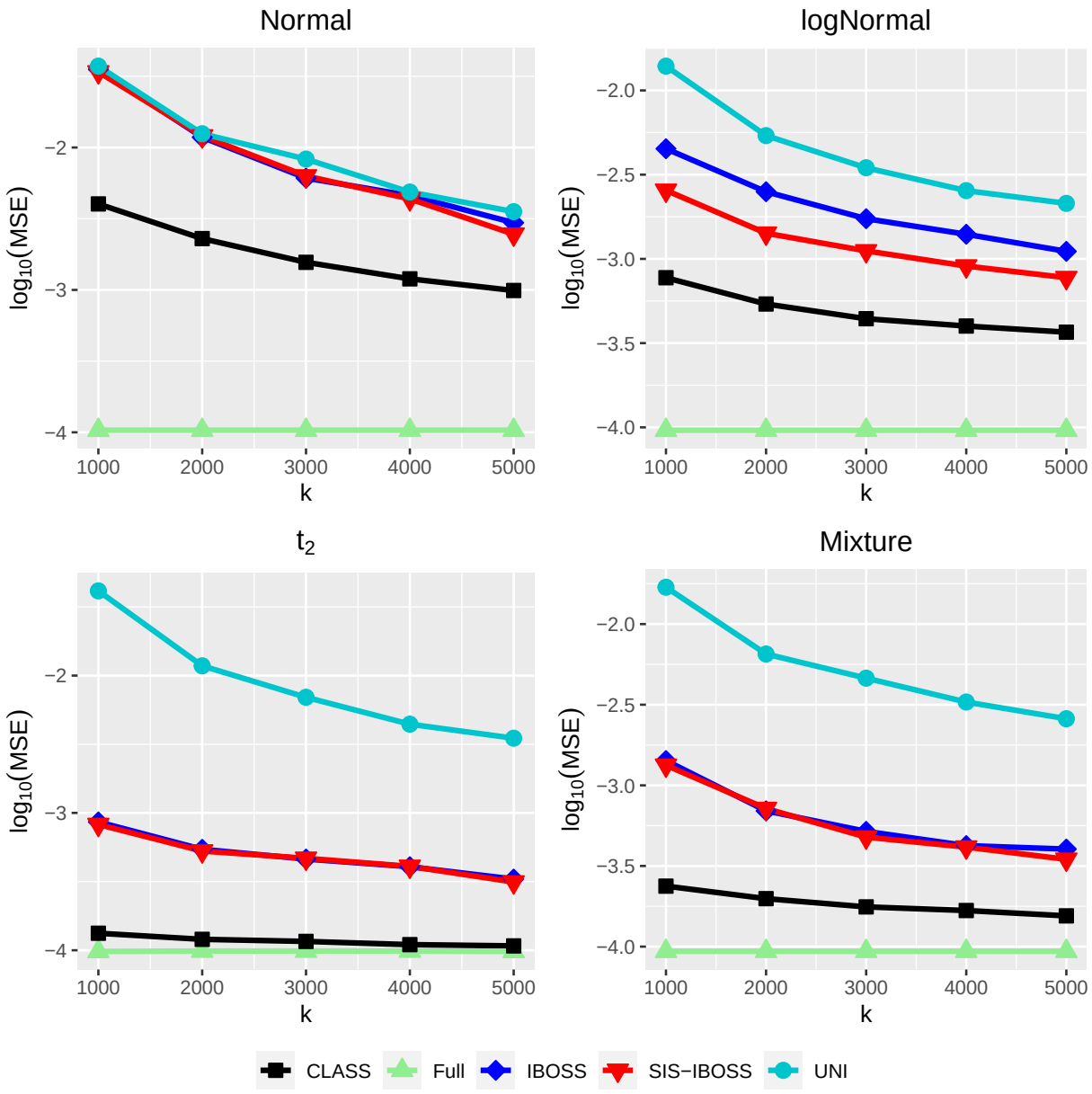


Figure 50: Variable selection performance for $n = 10^5$, $p = 500$, $p_1 = 10$, and $\mathbf{\Sigma}$ is Random. The solid lines represent the power, whereas the dashed lines represent the error.

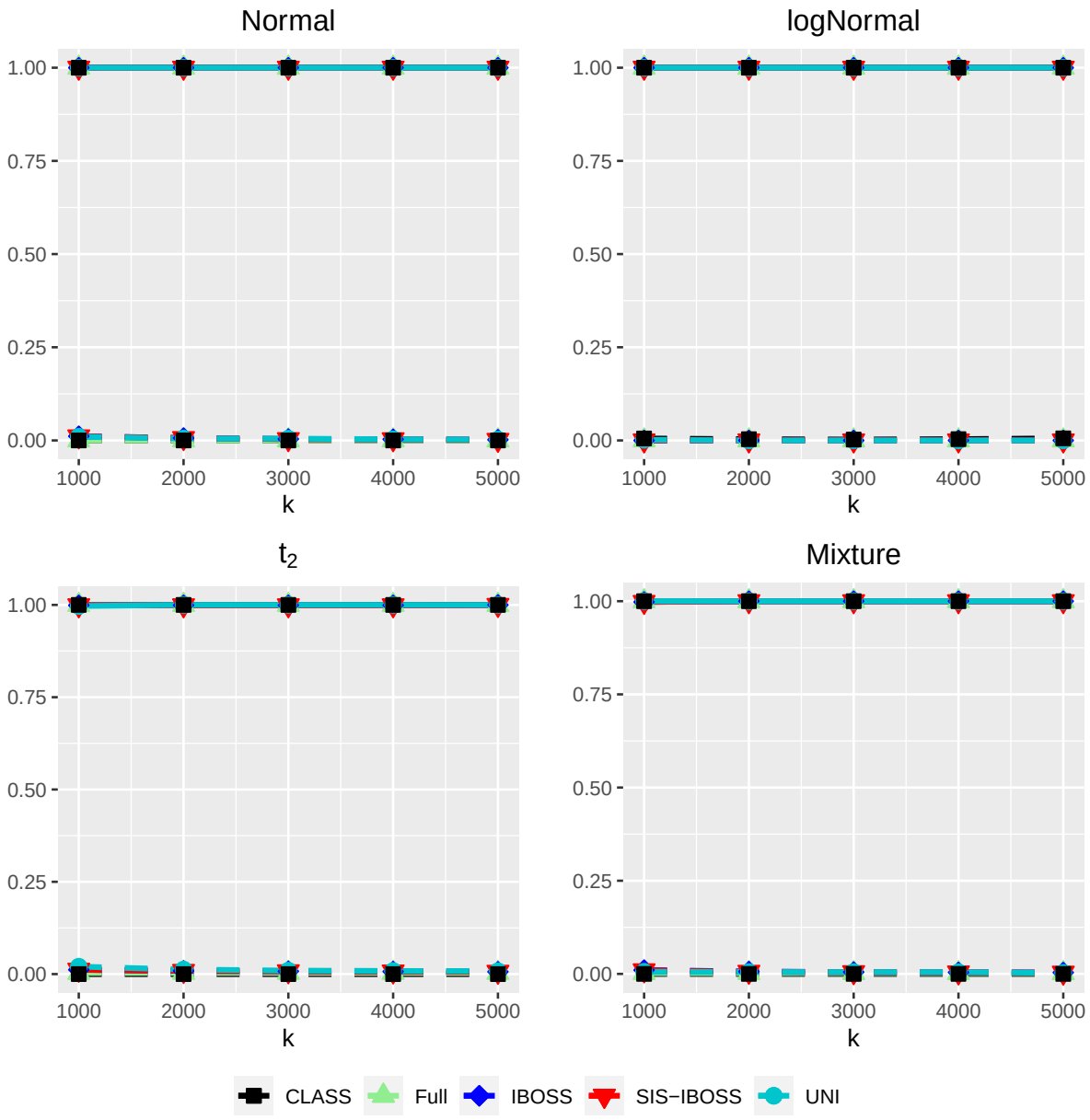


Figure 51: MSE for $n = 10^5$, $p = 500$, $p_1 = 25$, and $\mathbf{\Sigma}$ is Random.

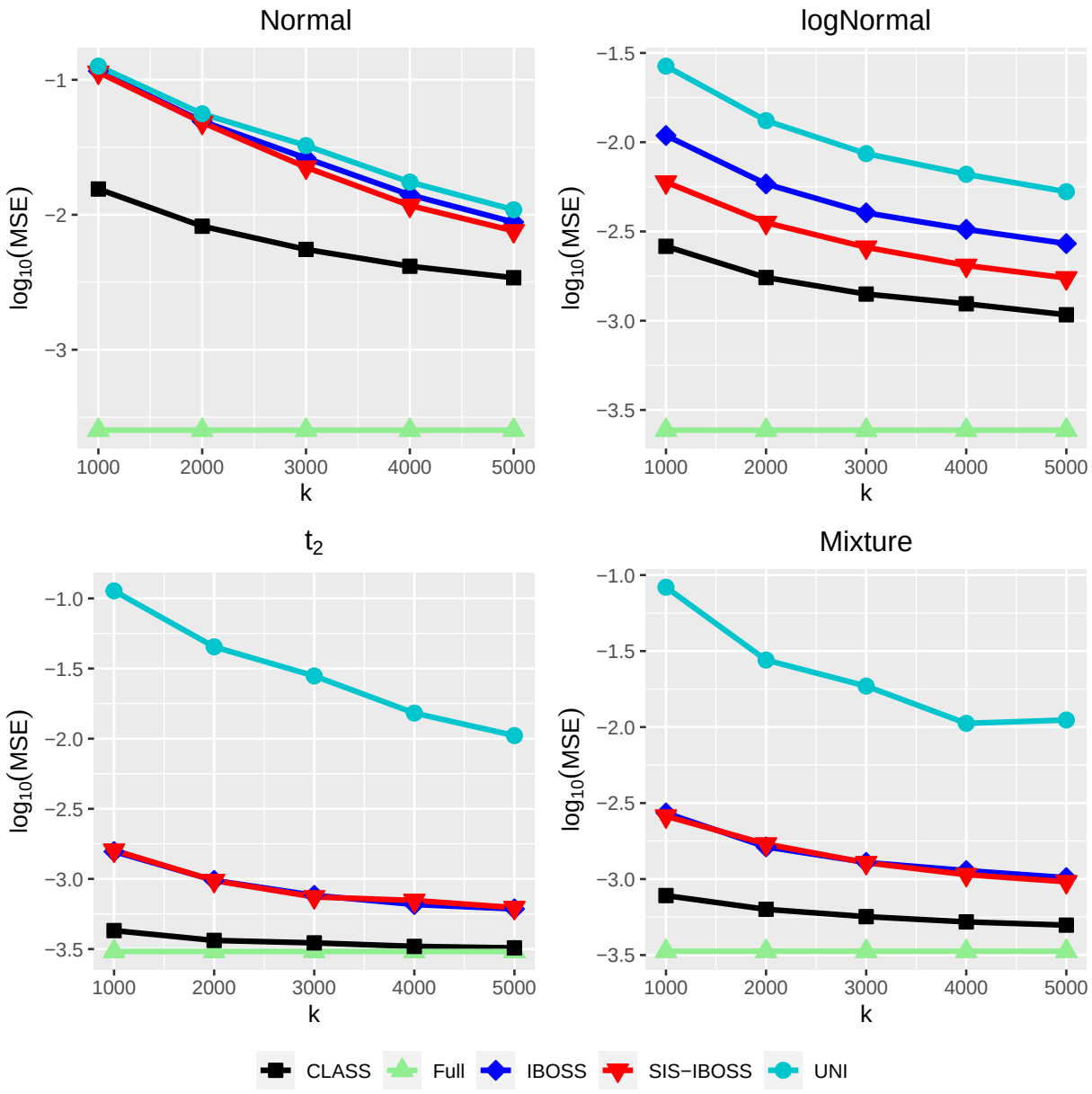


Figure 52: Variable selection performance for $n = 10^5$, $p = 500$, $p_1 = 25$, and $\mathbf{\Sigma}$ is Random. The solid lines represent the power, whereas the dashed lines represent the error.

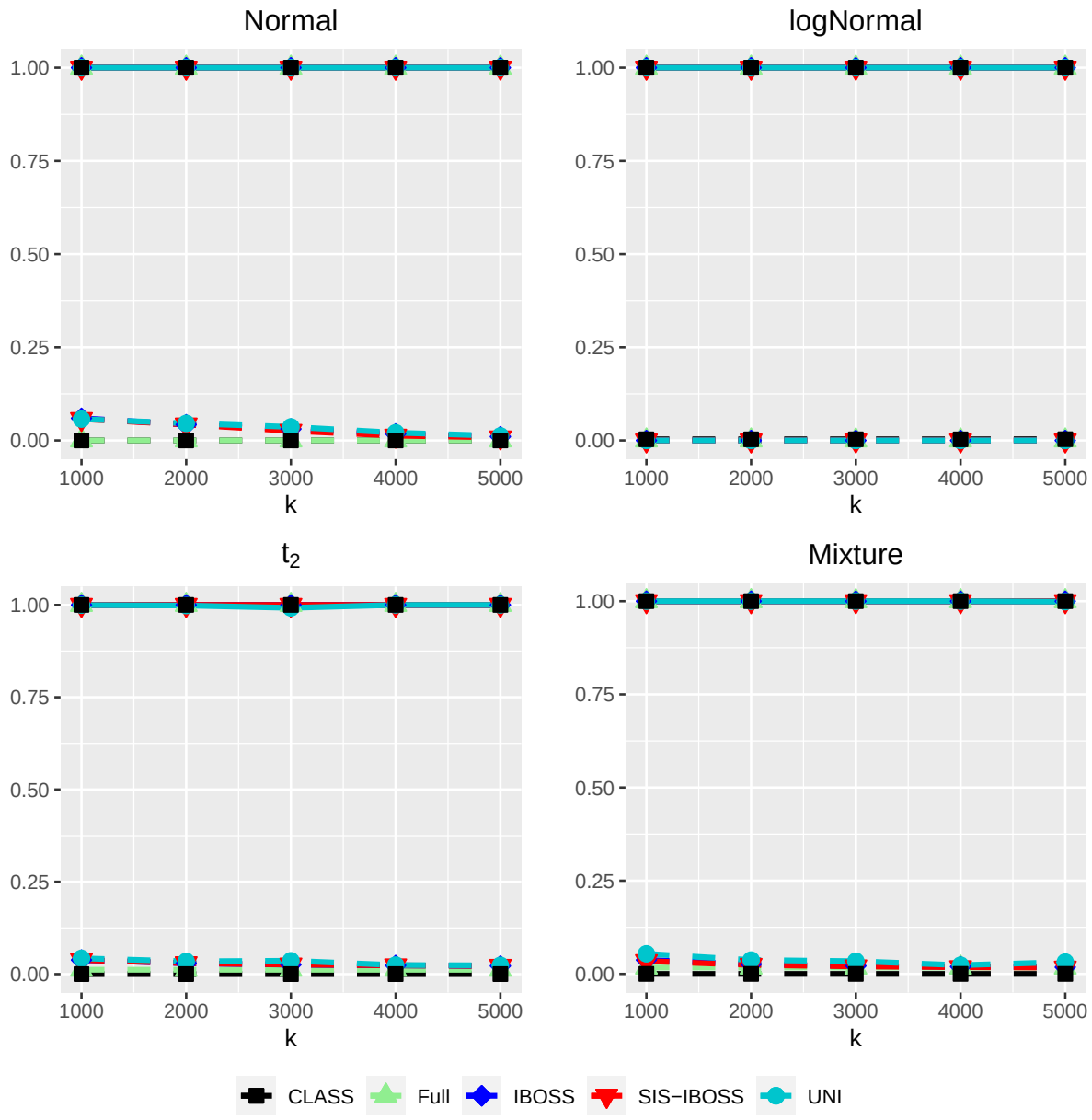


Figure 53: MSE for $n = 10^5$, $p = 500$, $p_1 = 50$, and $\mathbf{\Sigma}$ is Random.

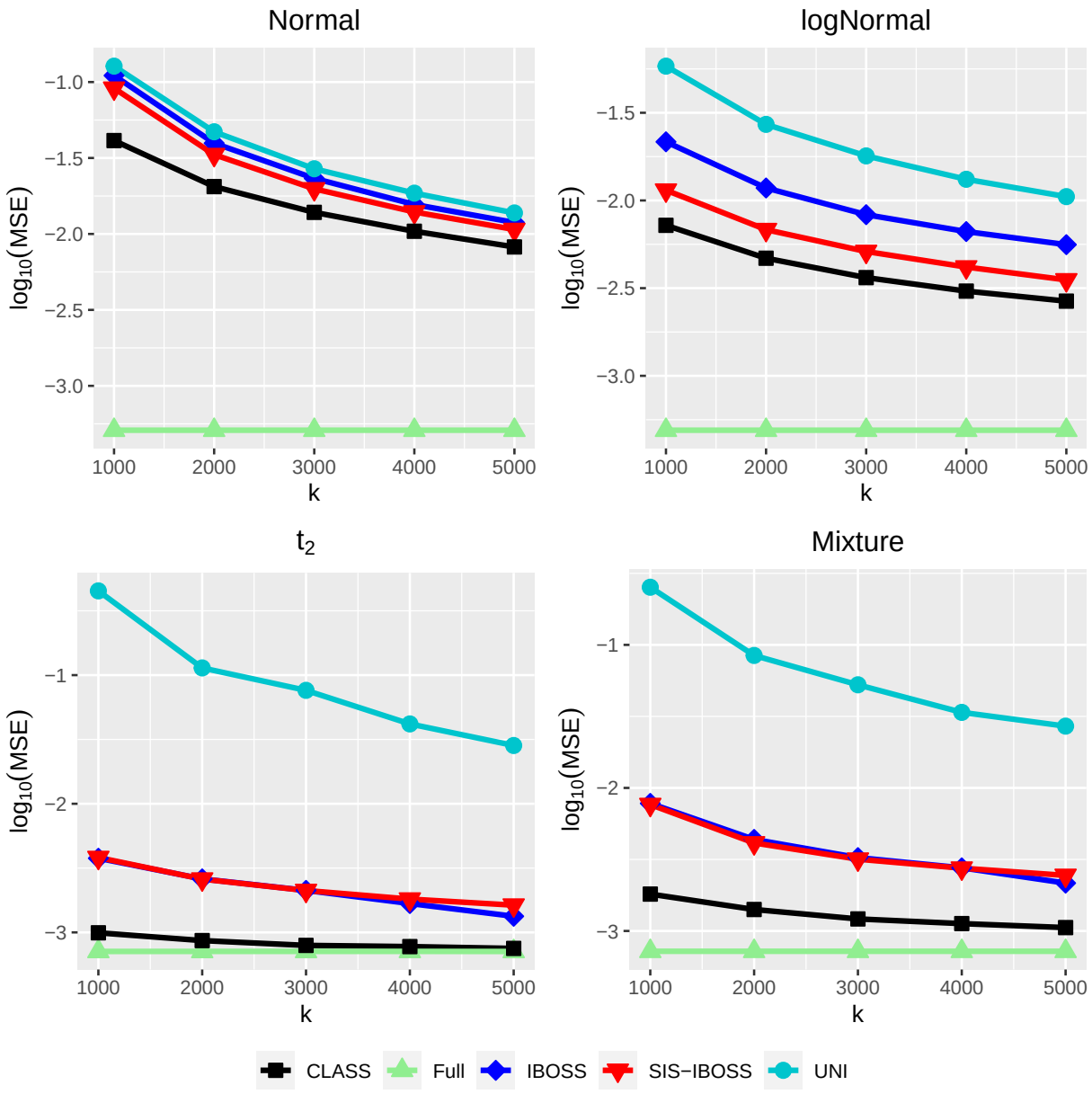
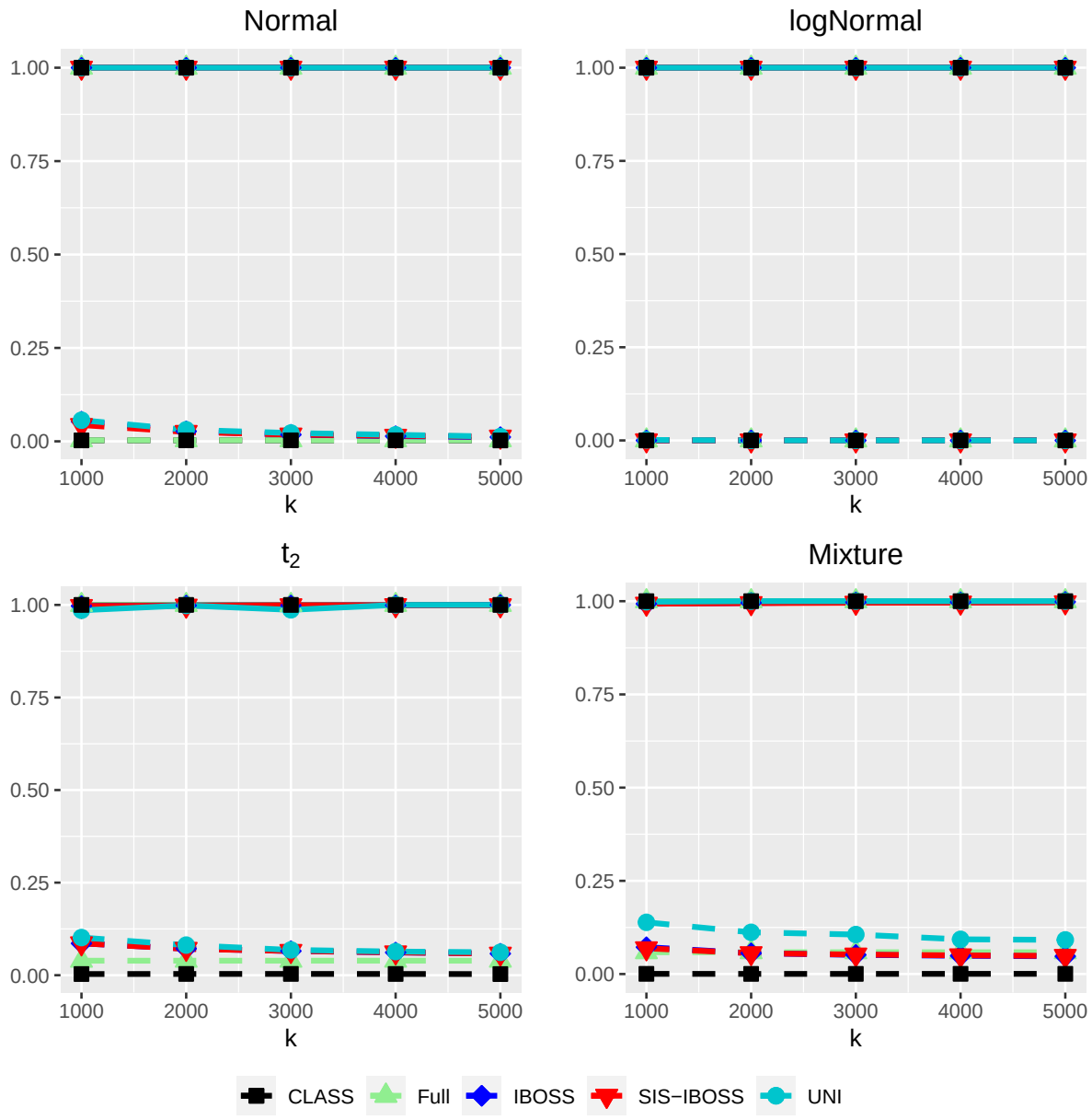


Figure 54: Variable selection performance for $n = 10^5$, $p = 500$, $p_1 = 50$, and $\mathbf{\Sigma}$ is Random. The solid lines represent the power, whereas the dashed lines represent the error.



5 For $p = 5000$ with error standard deviation equals 1

Figure 55: MSE for $k = 1000$, $p = 5000$, $p_1 = 25$, and $\mathbf{\Sigma} = (0^{I(i \neq j)})$.

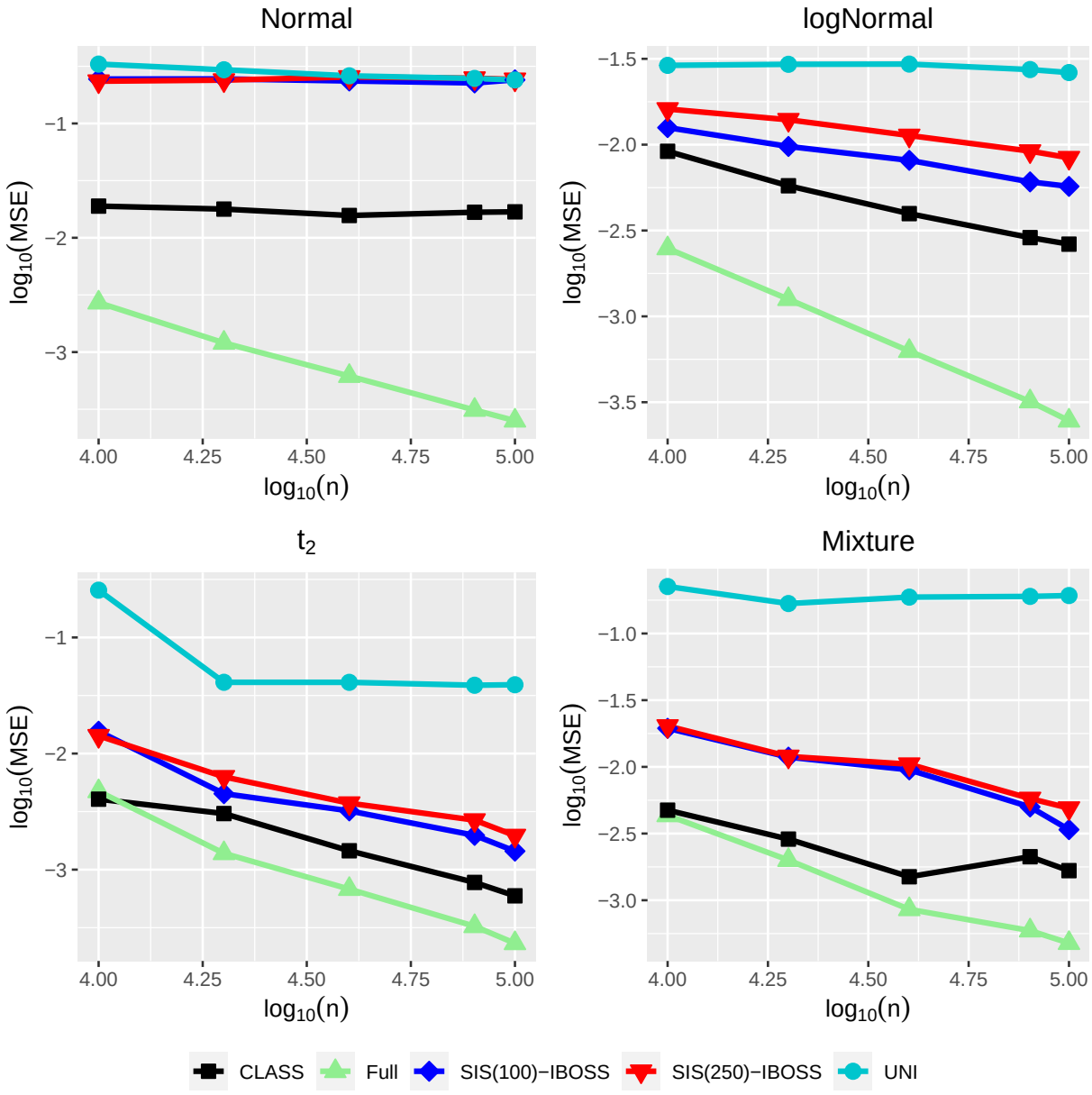


Figure 56: Variable selection performance for $k = 1000$, $p = 5000$, $p_1 = 25$, and $\mathbf{\Sigma} = (0^{I(i \neq j)})$. The solid lines represent the power, whereas the dashed lines represent the error.

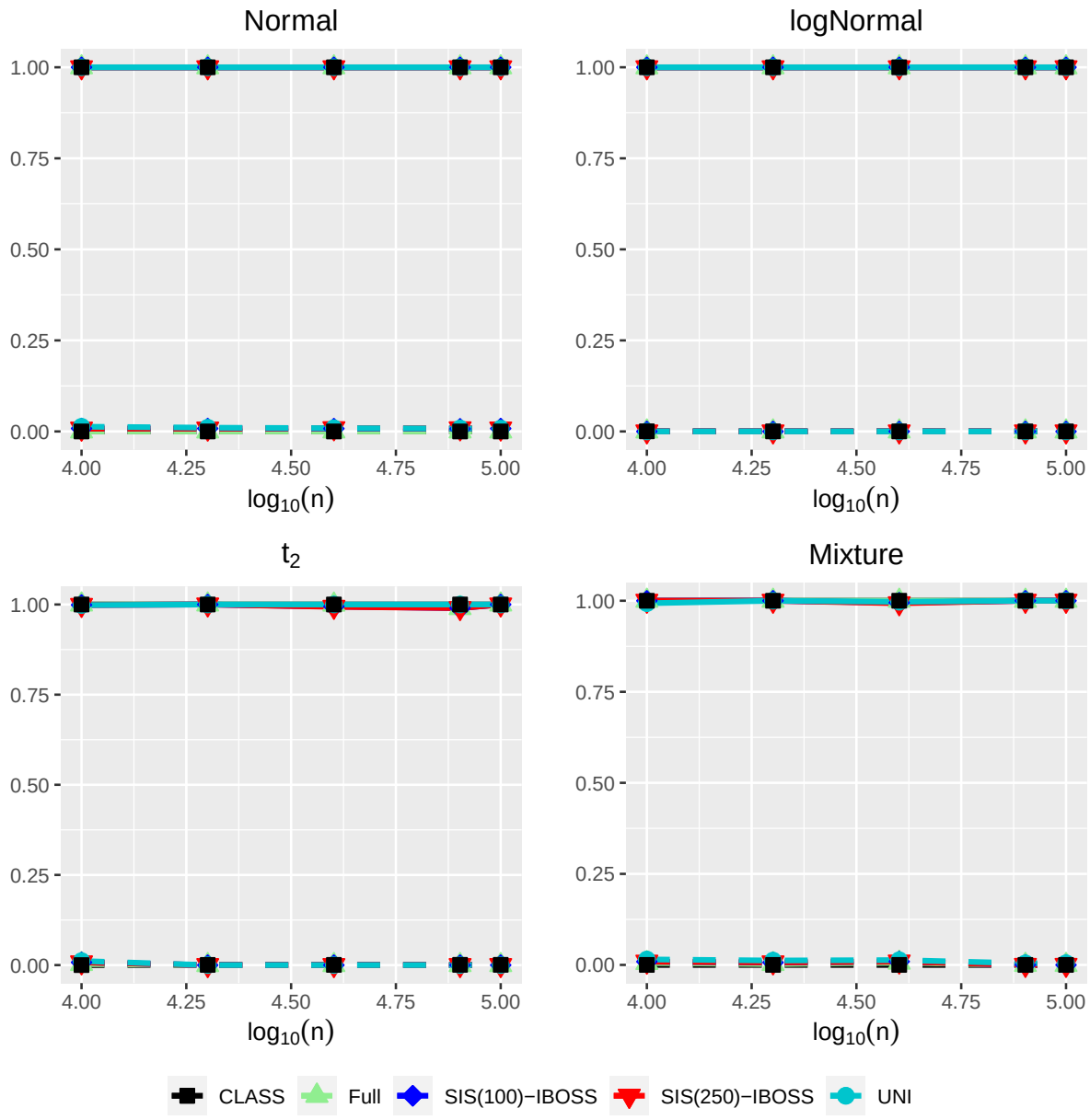


Figure 57: MSE for $k = 1000$, $p = 5000$, $p_1 = 50$, and $\mathbf{\Sigma} = (0^{I(i \neq j)})$.

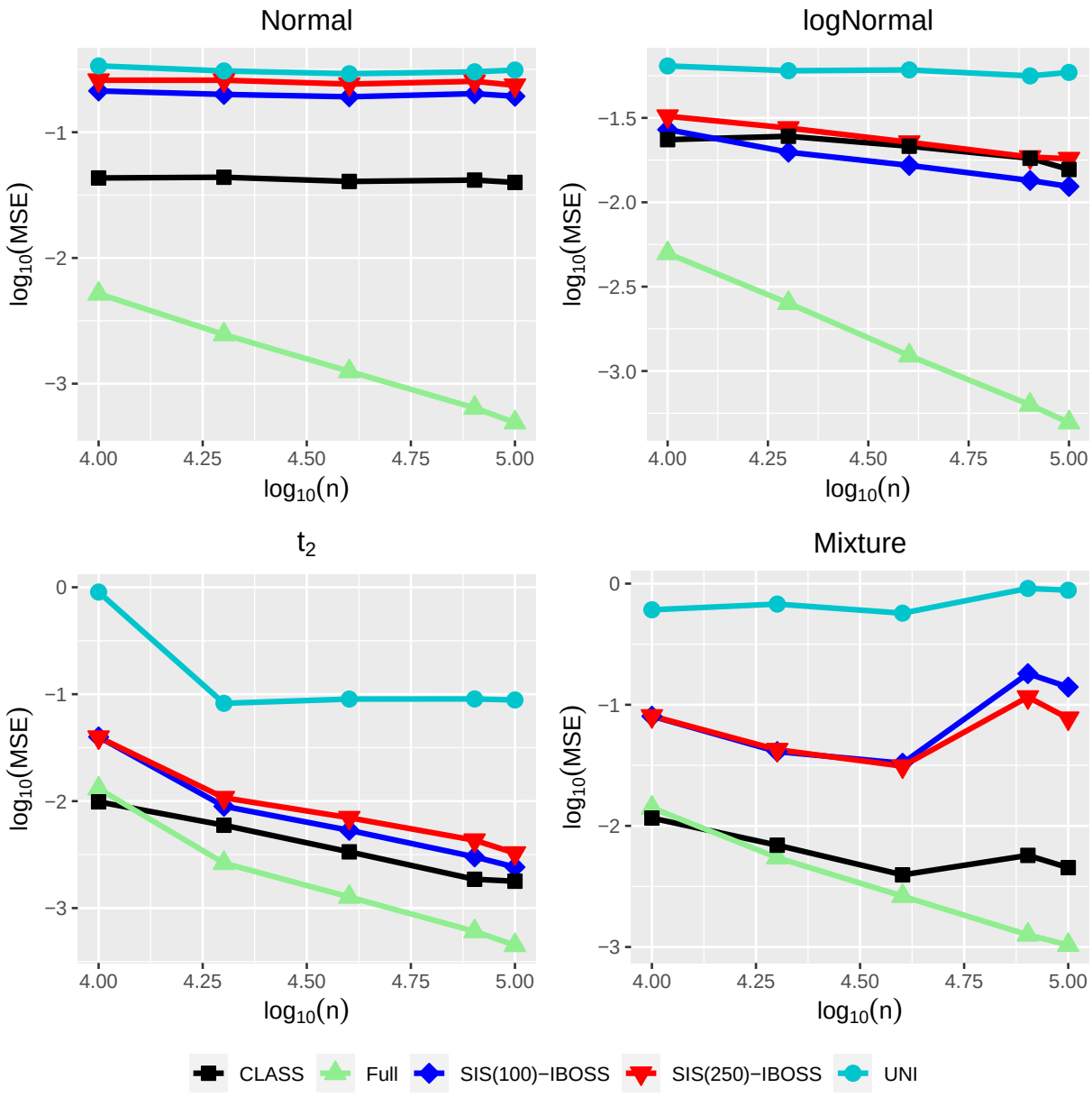


Figure 58: Variable selection performance for $k = 1000$, $p = 5000$, $p_1 = 50$, and $\mathbf{\Sigma} = (0^{I(i \neq j)})$. The solid lines represent the power, whereas the dashed lines represent the error.

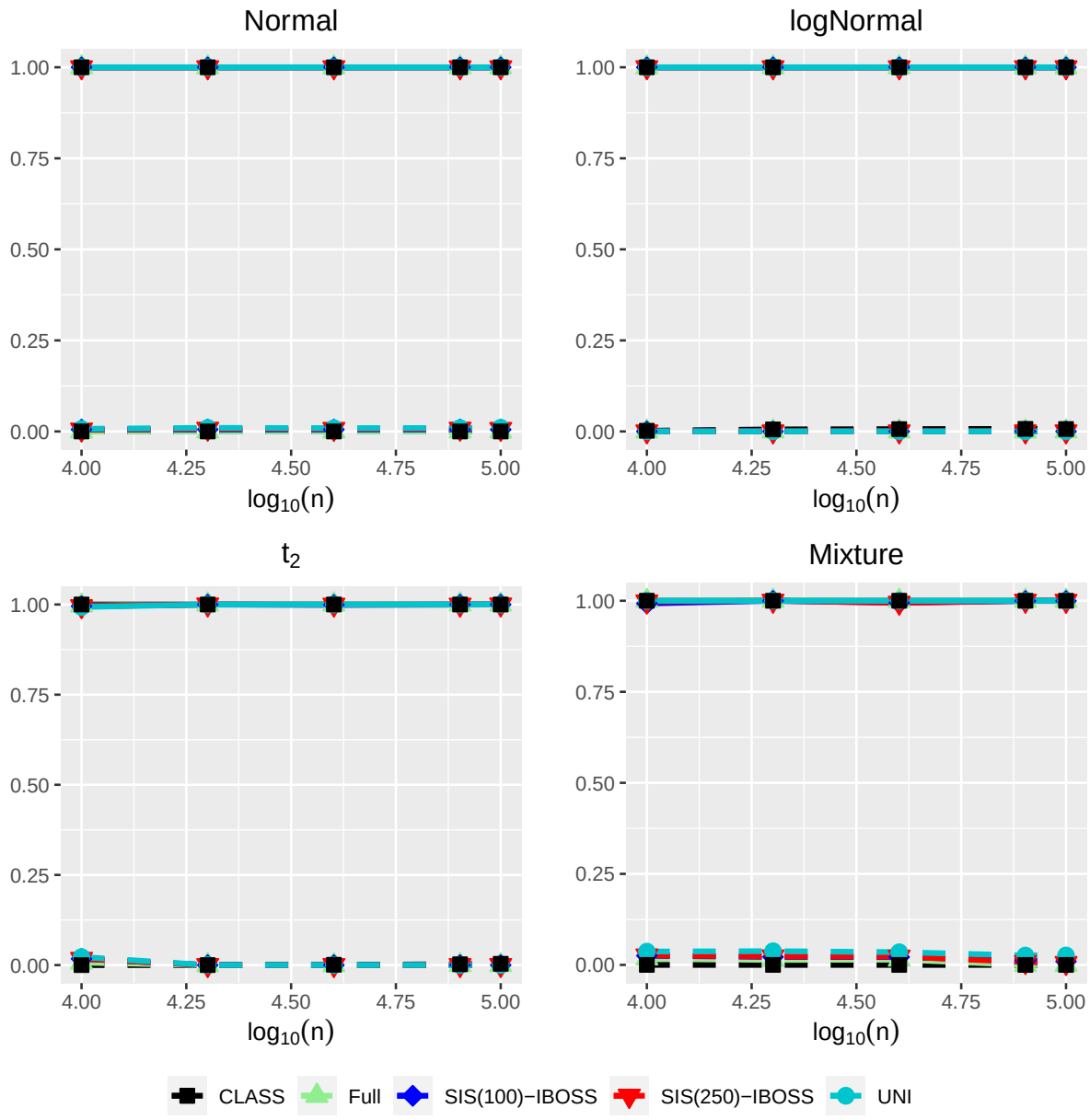


Figure 59: MSE for $k = 1000$, $p = 5000$, $p_1 = 75$, and $\mathbf{\Sigma} = (0^{I(i \neq j)})$.

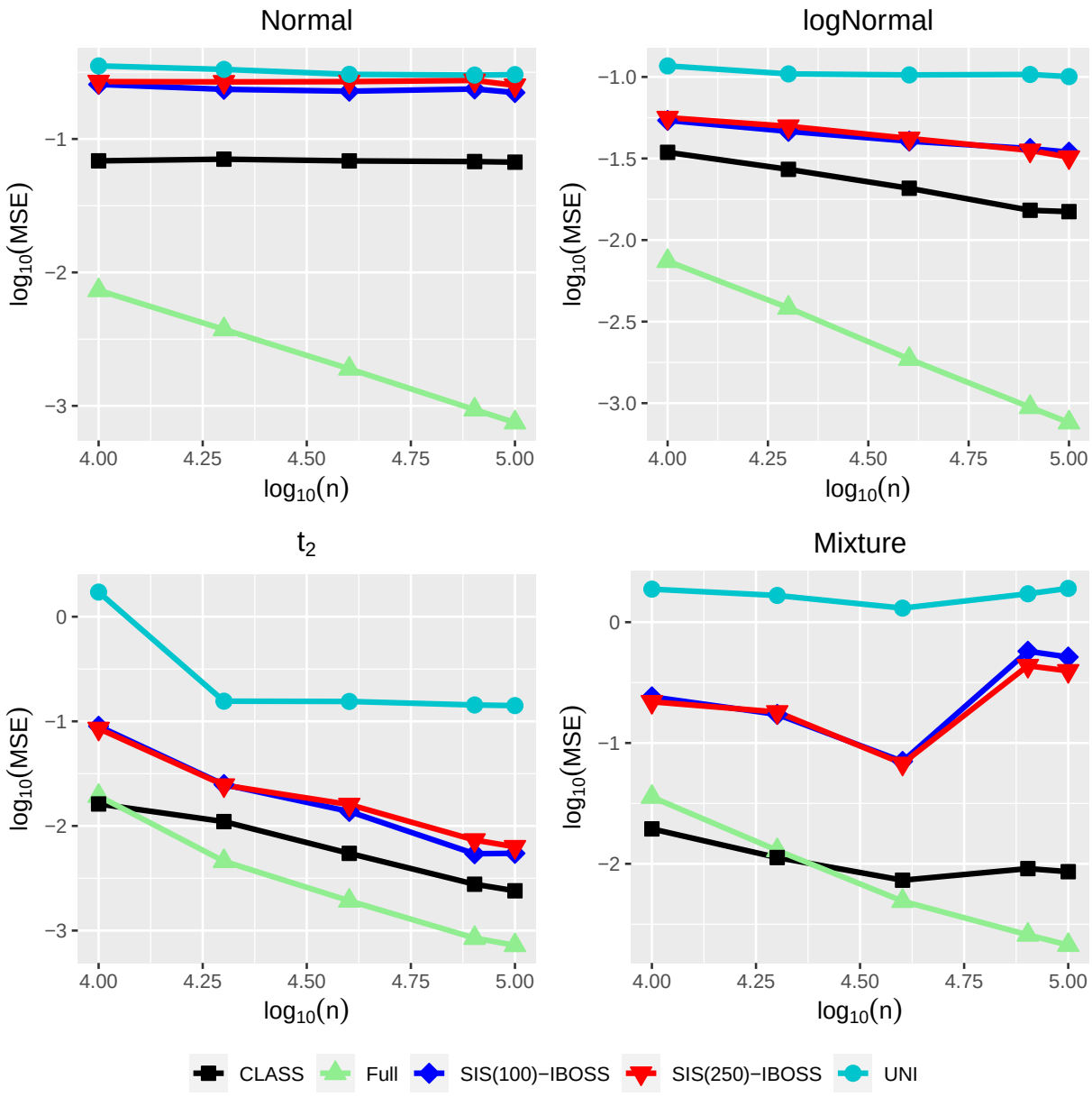


Figure 60: Variable selection performance for $k = 1000$, $p = 5000$, $p_1 = 75$, and $\mathbf{\Sigma} = (0^{I(i \neq j)})$. The solid lines represent the power, whereas the dashed lines represent the error.

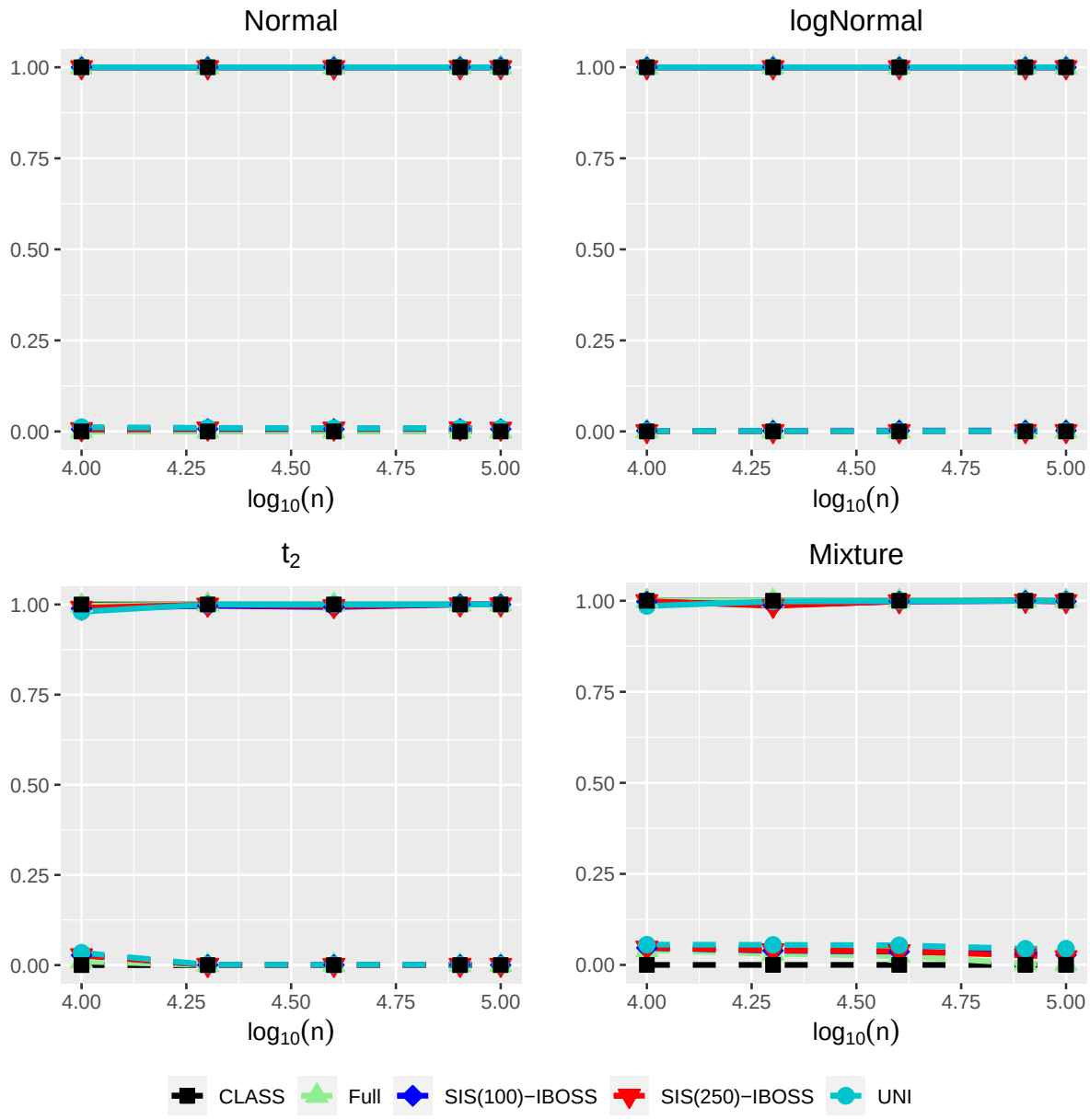


Figure 61: MSE for $k = 1000$, $p = 5000$, $p_1 = 25$, and $\mathbf{\Sigma} = (0.5^{I(i \neq j)})$.

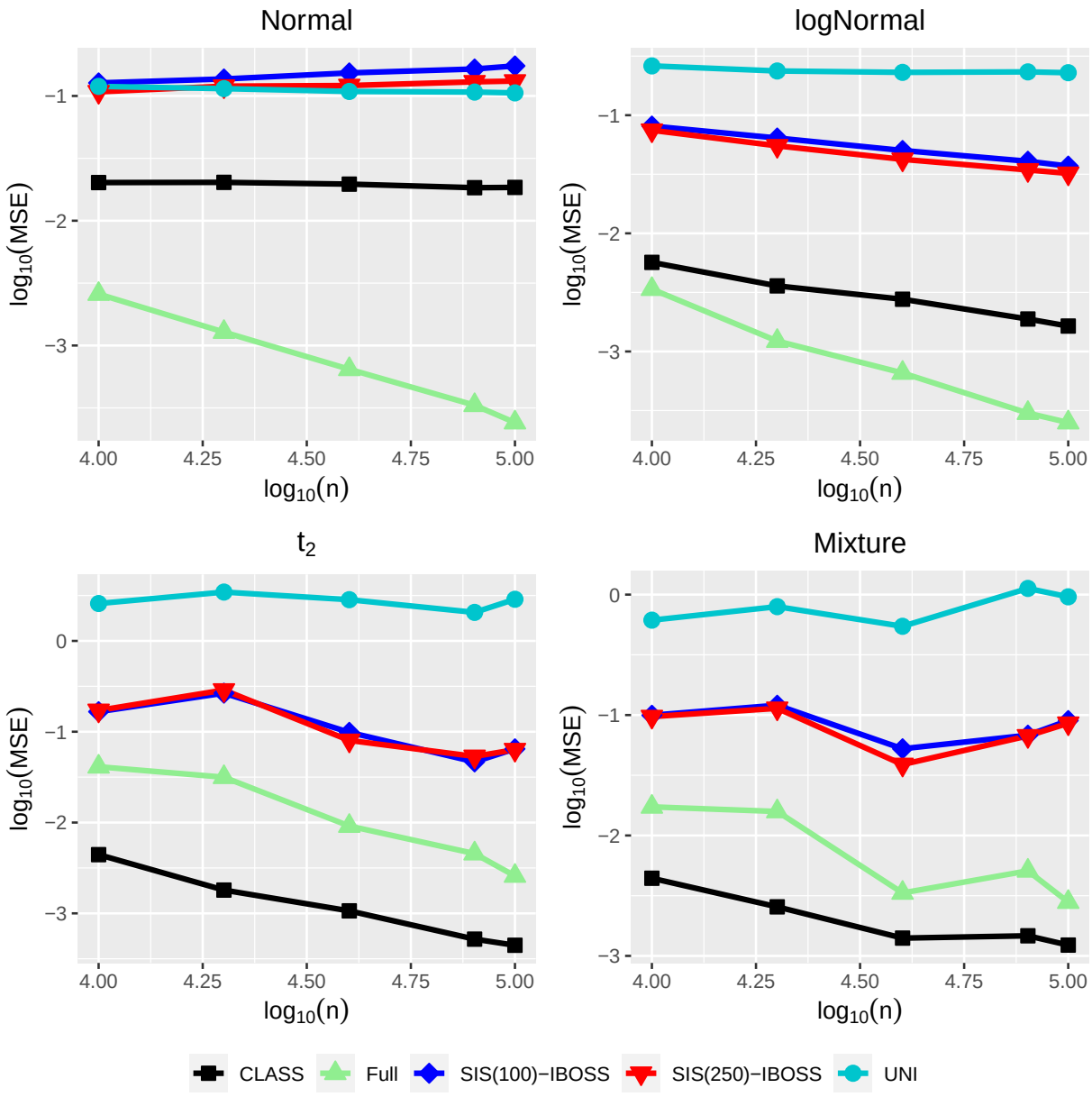


Figure 62: Variable selection performance for $k = 1000$, $p = 5000$, $p_1 = 25$, and $\mathbf{\Sigma} = (0.5^{I(i \neq j)})$. The solid lines represent the power, whereas the dashed lines represent the error.

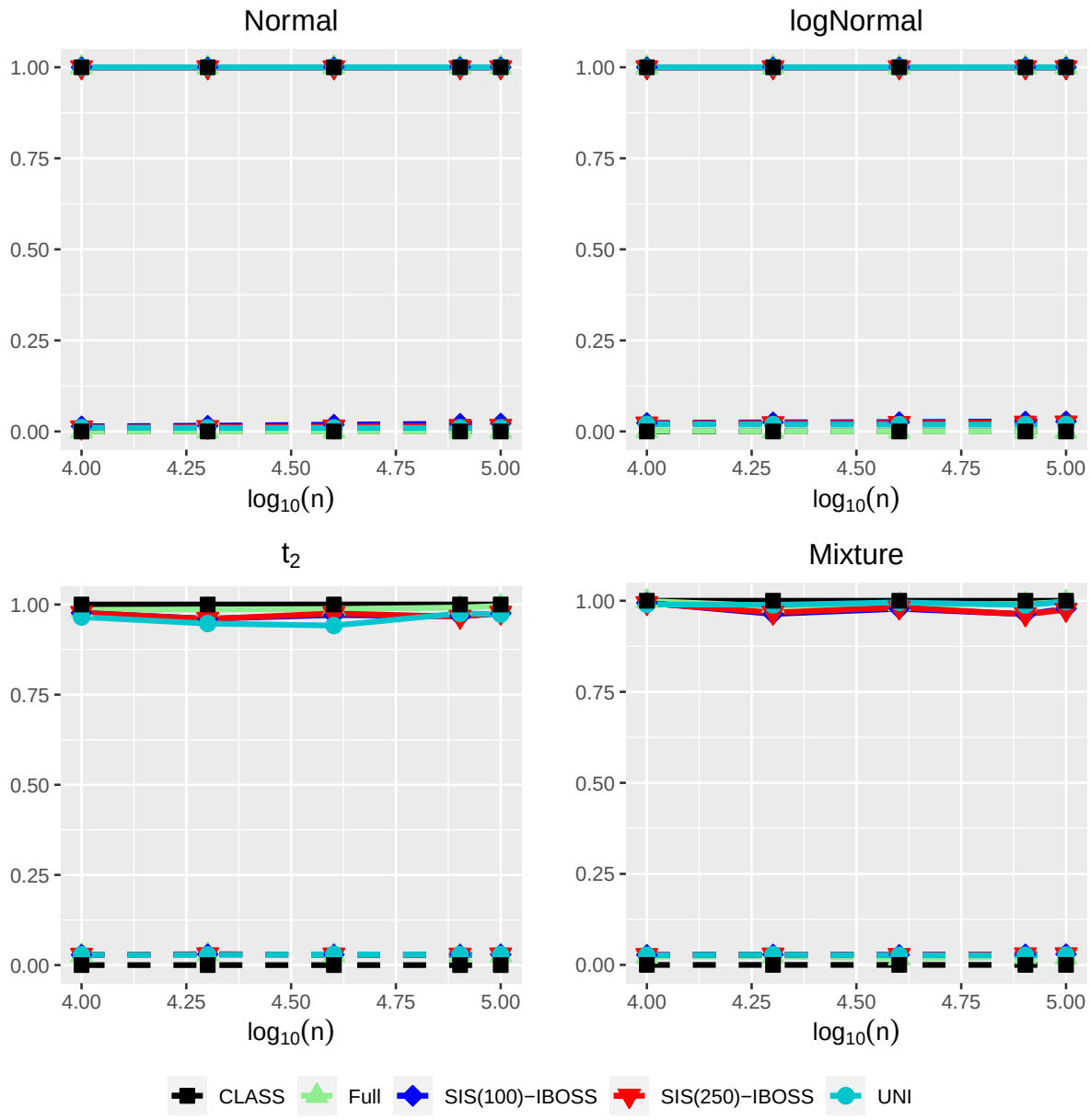


Figure 63: MSE for $k = 1000$, $p = 5000$, $p_1 = 50$, and $\mathbf{\Sigma} = (0.5^{I(i \neq j)})$.

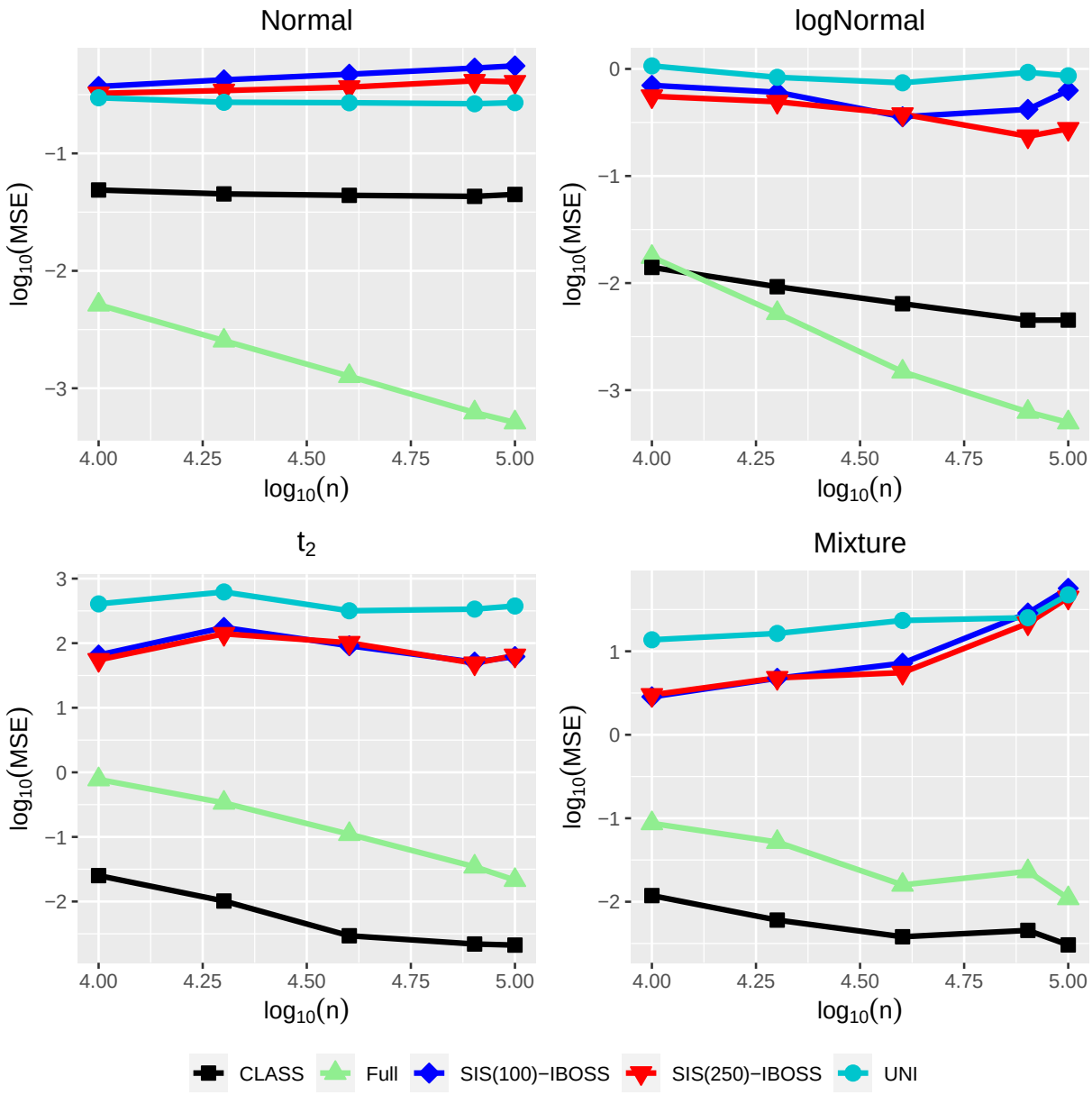


Figure 64: Variable selection performance for $k = 1000$, $p = 5000$, $p_1 = 50$, and $\mathbf{\Sigma} = (0.5^{I(i \neq j)})$. The solid lines represent the power, whereas the dashed lines represent the error.

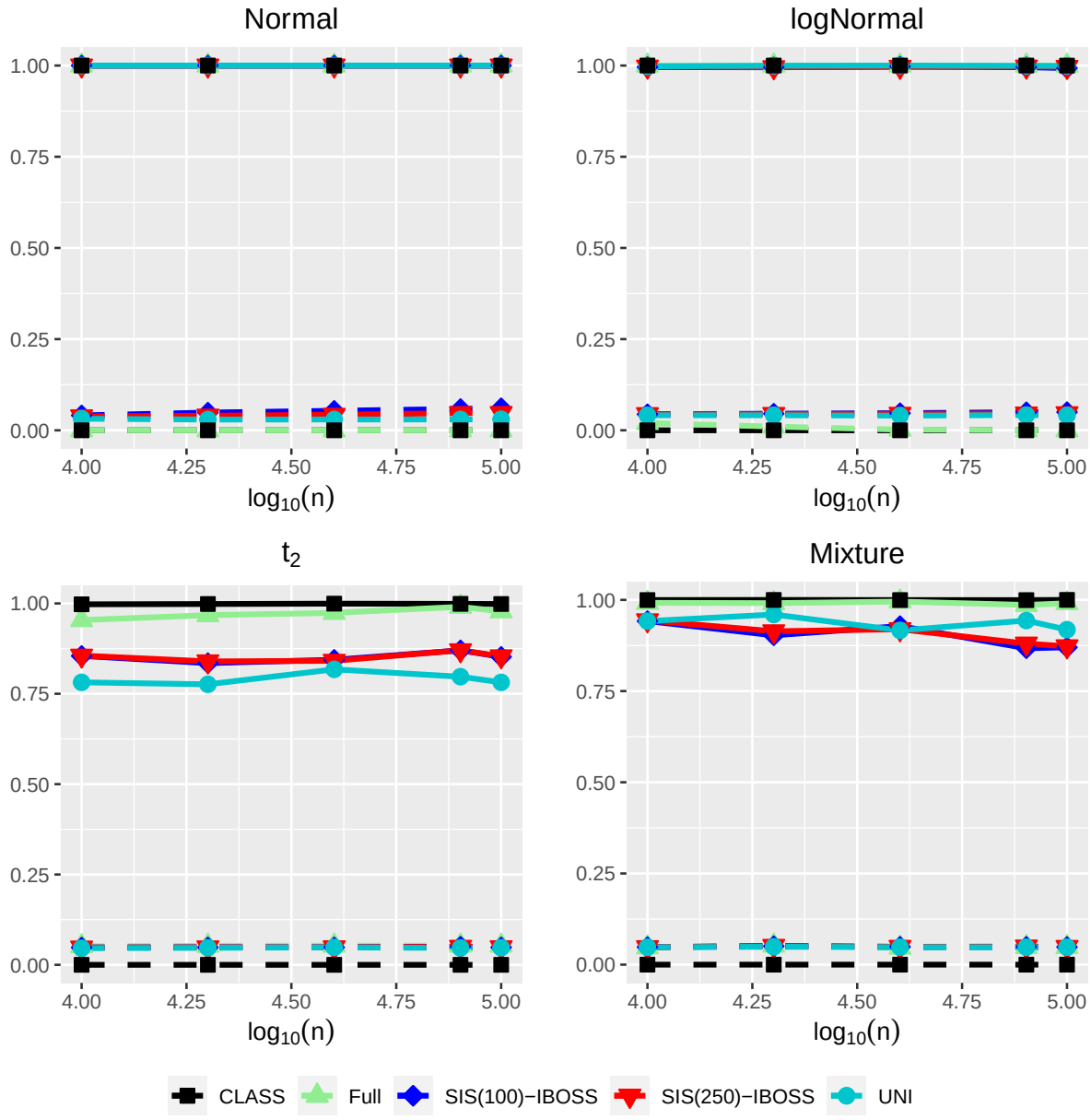


Figure 65: MSE for $k = 1000$, $p = 5000$, $p_1 = 75$, and $\mathbf{\Sigma} = (0.5^{I(i \neq j)})$.

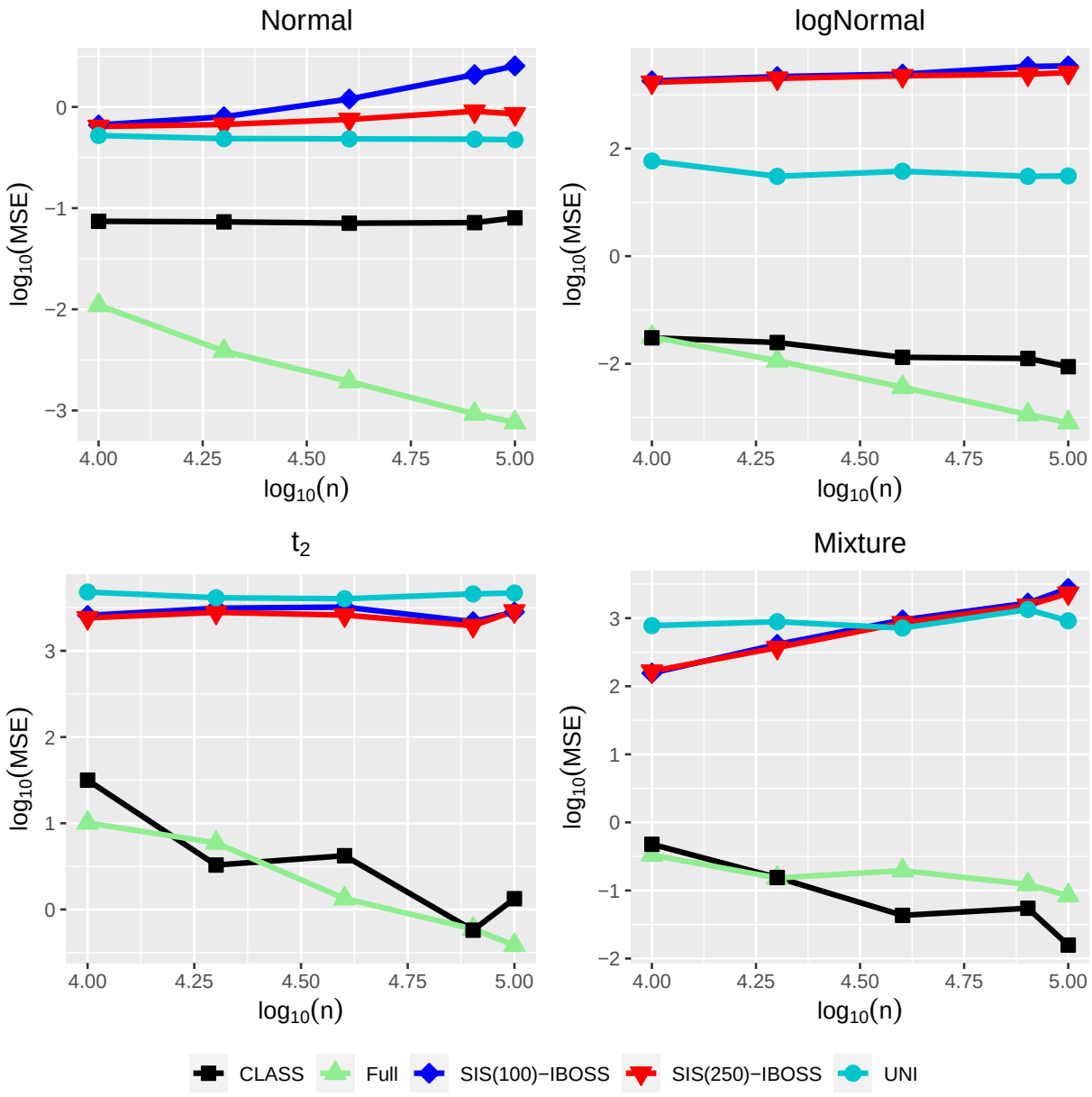
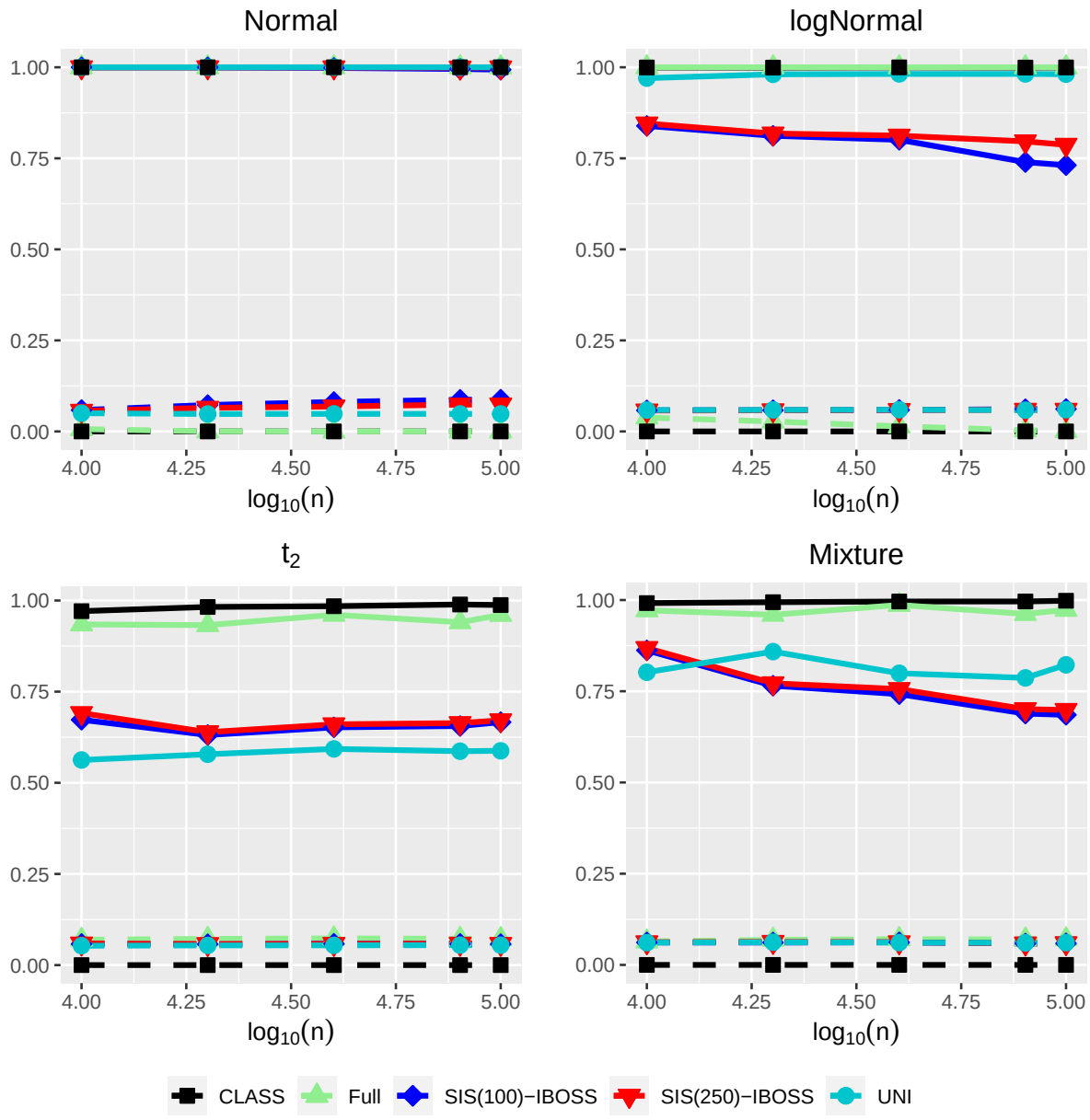


Figure 66: Variable selection performance for $k = 1000$, $p = 5000$, $p_1 = 75$, and $\mathbf{\Sigma} = (0.5^{I(i \neq j)})$. The solid lines represent the power, whereas the dashed lines represent the error.



6 Tuning parameters

Active effects coefficients come from $N(5,1)$, and error is $N(0,1)$. The legend has entries in the form of $nsample - ntimes$. We see that the black, green, and blue lines perform far worse as compared to the red lines. For four choices in red (that is, for $nsample = 1000$), $ntimes = 10$ is far worse as compared to the rest of the choices of $ntimes$. Our choices for $nsample = 1000$ and $ntimes = 100$ seems like a reasonable choice.

Figure 67: MSE for $k = 1000$, $p = 500$, $p_1 = 50$, and $\mathbf{\Sigma} = (0.5^{I(i \neq j)})$.

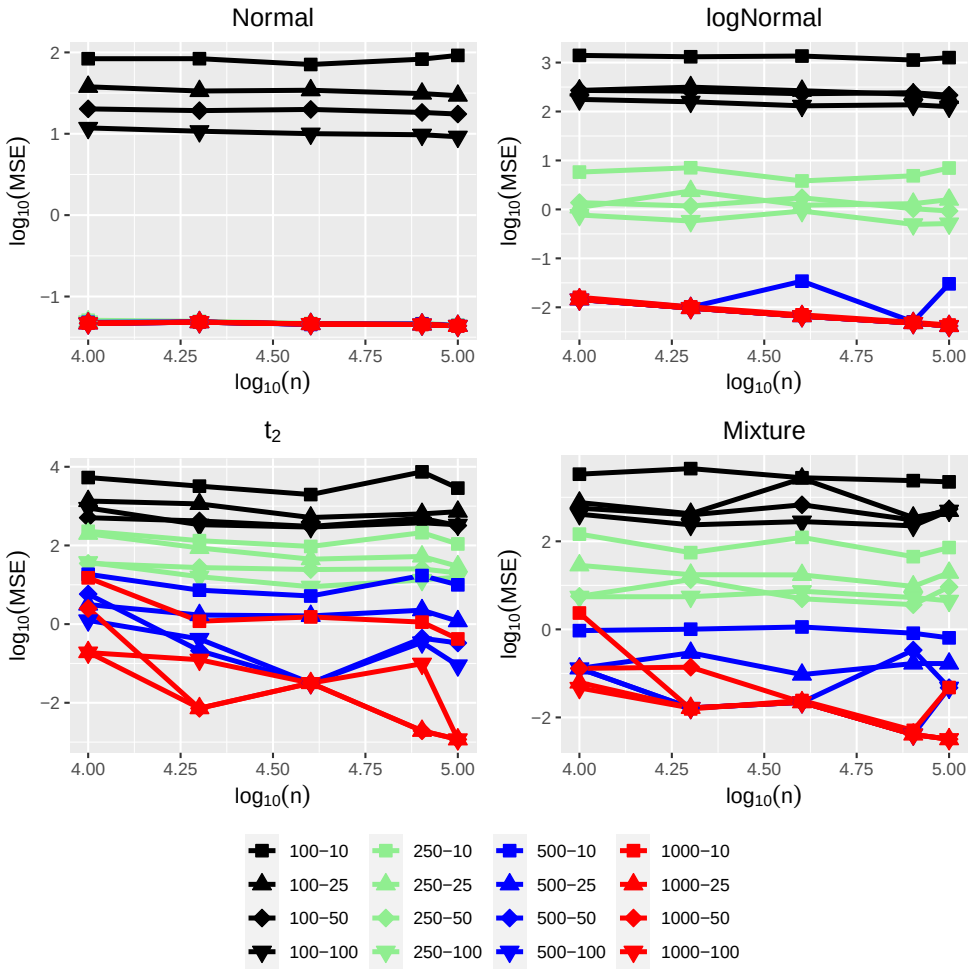
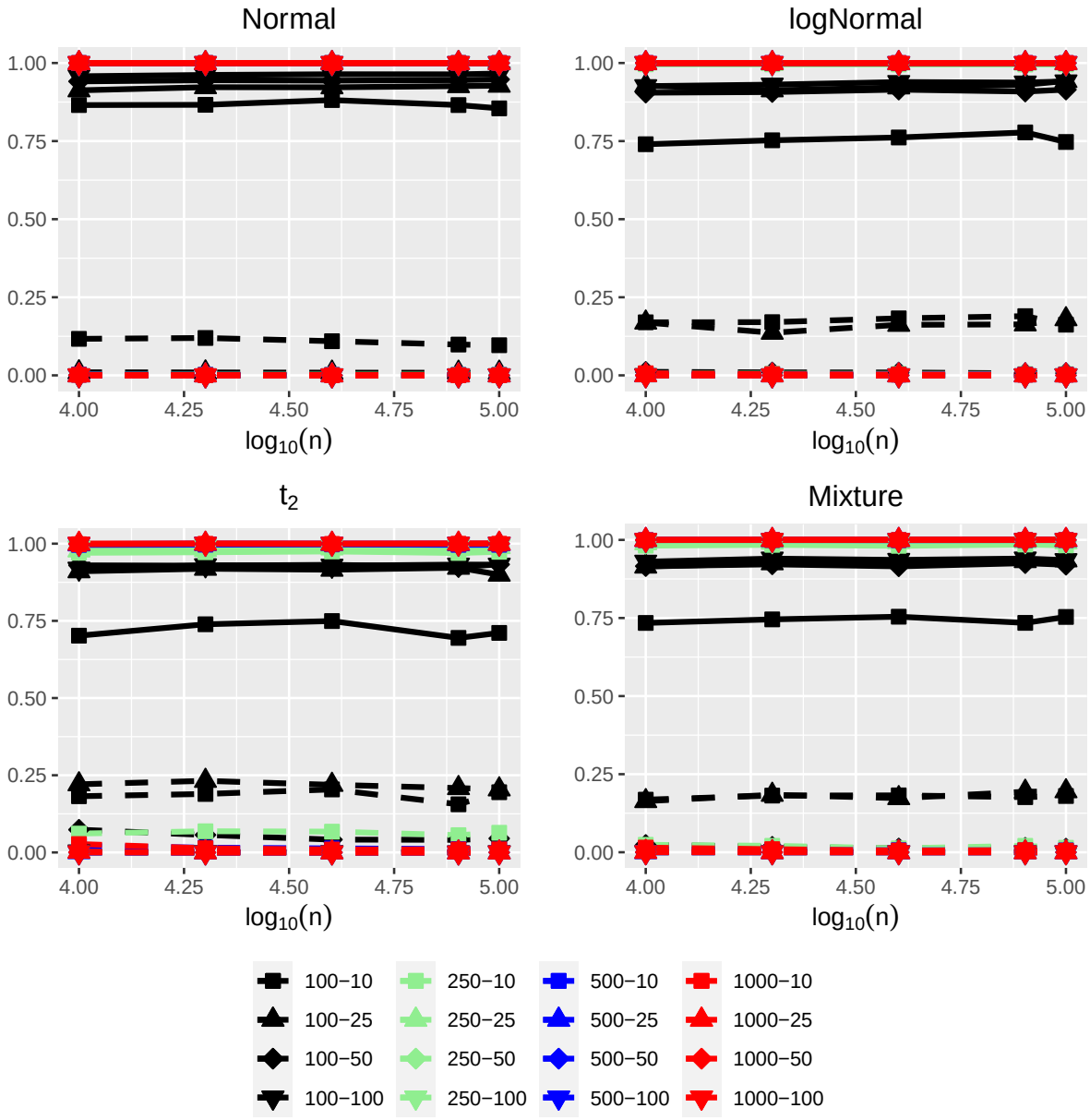


Figure 68: Variable selection performance for $k = 1000$, $p = 500$, $p_1 = 50$, and $\mathbf{\Sigma} = (0.5^{I(i \neq j)})$. The solid lines represent the power, whereas the dashed lines represent the error. The legend has entries as $nsample - ntimes$



7 Performance of Methods when CPU Time is Comparable

Table 1: MSE and variable selection performance for different subdata methods with approximately equal CPU times, different n and k, $p = 500$, $p_1 = 50$, $\mathbf{\Sigma} = (0.5^{I(i \neq j)})$ and joint variable distribution logNormal

Method	k	time (s)	MSE	Power	Error
	For $n = 10^5$				
UNI	90000	48.36	0.001064	1	0.00113
IBOSS	60000	45.32	0.000632	1	0.00004
SIS(100)-IBOSS	75000	39.68	0.000564	1	0.00004
CLASS	1000	46.20	0.004264	1	0.00000
	For $n = 10^6$				
UNI	80000	42.19	0.000699	1	0.0000
IBOSS	50000	54.07	0.000204	1	0.0000
SIS(100)-IBOSS	70000	77.68	0.000596	1	0.2840
CLASS	1000	47.83	0.001579	1	0.0000

Table 2: MSE and variable selection performance for different subdata methods with approximately equal CPU times, different n and k, $p = 500$, $p_1 = 50$, $\mathbf{\Sigma} = (0.5^{I(i \neq j)})$ and joint variable distribution Mixture

Method	k	time (s)	MSE	Power	Error
	For $n = 10^5$				
UNI	90000	71.46	8.63465	0.998	0.17678
IBOSS	60000	61.63	11.16609	0.9978	0.17431
SIS(100)-IBOSS	75000	59.63	6.979962	0.9982	0.17142
CLASS	1000	51.71	0.003897	1	0.00000
	For $n = 10^6$				
UNI	80000	61.31	35.86625	0.995	0.1677
IBOSS	50000	65.53	548.1969	0.9884	0.1813
SIS(100)-IBOSS	70000	62.75	463.3618	0.989	0.1841
CLASS	1000	52.79	0.000629	1	0.0000