

Supplementary Material 1: Strategies for Supersaturated Screening: Group Orthogonal and Constrained $Var(s)$ Designs

1 Results of Questionnaire

Of the 63 survey respondents, 55% work in the field of Engineering, Manufacturing and Technology and 26% in Academia or Education; 80% have a Master’s degree or a PhD in statistics (see Figure 1). Figure 2 shows that most respondents use Response Surface Methodology, Fractional factorials/Plackett Burman designs and/or Split plot designs. Regression/ANOVA is unsurprisingly the most common analysis technique with reported use by 95% of the respondents. Interestingly, the next most common analysis is LASSO and/or other penalized regression techniques, followed closely with Bayesian methods and Gaussian Process Models. Only six respondents indicated SSDs as a design technique they use on a regular basis. Most respondents reported choosing a design technique with which they are comfortable and/or which is associated with a straightforward analysis method. Several self-reported reasons for design choice include the particular experimental situation and the efficiency of the design.

With regard to SSDs, 68% said they were familiar with them, and only 13 respondents reported using a SSD in practice. Table 1 provides examples of descriptions of the use of SSDs in practice with some successes and a failed experiment. Nineteen respondents voluntarily provided detailed explanations of concerns they have regarding SSDs; these are categorized into groups and shown in Figure 3. The most common concern with using a SSD is the sparsity assumptions and/or power (Figure 3). For example, one respondent’s concern regarding SSDs was stated as “Not identifying significant factors that are appearing insignificant”; this response was categorized both as “Sparsity/Power” and “Model Misspecification”. Several respondents commented on the complicated

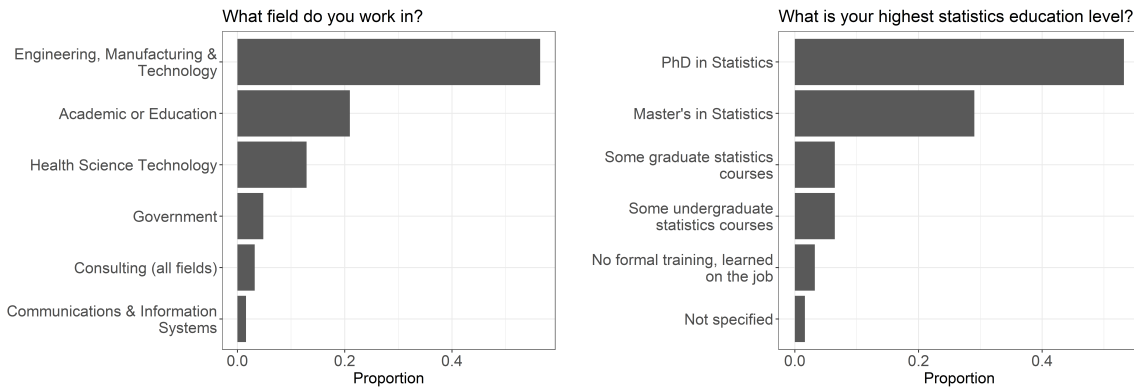


Figure 1: Summary of demographics for the survey respondents.

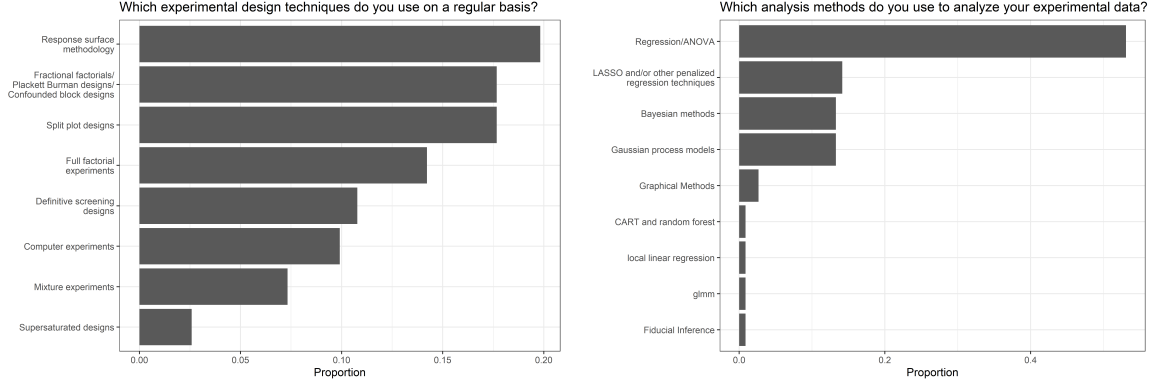


Figure 2: Summary of the design and analysis techniques used in practice for the survey respondents.

SSD User Experience

“100+ factors 64 runs; failed experiment”

“Bayesian D-optimal design with many terms that were unable to be estimated by the design, but were able to be estimated after unimportant factors were removed.”

“Analytical Method Robustness testing. Successful.”

“Testing to characterize drill bit effectiveness as a function of many input parameters. Experiment was successful due to engineering expertise for interpretation.”

“Confounded effects that were not believed to be important or could be estimated in aggregate. Without knowing the truth, I think the design seemed successful.”

Table 1: Summary of SSD use in practice from survey responses.

analysis; “Ambiguity of interpretation if any factors are significant”, “Lack of [Degrees of Freedom]”. One respondent stated that they have never had a situation where a SSD seems to be the appropriate design choice and a few others commented that it was not possible to manipulate many factors simultaneously without “breaking the process.” Although SSDs are not widely used among the survey respondents, 78% stated that they would be interested in learning more about these designs.

Comments such as “Lack of [Degrees of Freedom]” and “You gamble. There is no guarantee that you find the important few [factors]; you can find a few but not necessarily the important ones” gives the impression that some practitioners consider SSDs as a one-shot experiment. Practitioners also believe that sparsity is a necessary requirement for SSDs to be successful and that the analysis is uninformative.

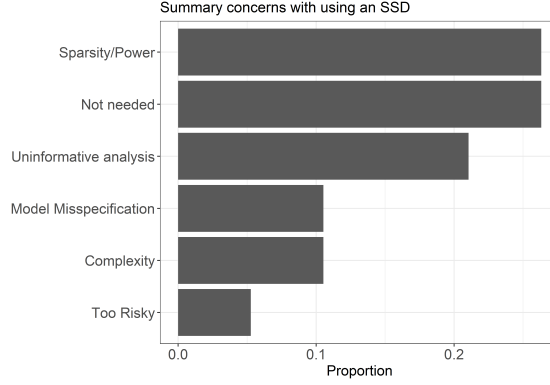


Figure 3: Summary of the concerns that respondents have with using a SSD.

2 Design Characteristics

n	k	Design	$UE(s)$	$UE(s^2)$	$Var(s)$	Mean $ r $	Max $ r $
8	12	$UE(s^2)$	0.0513	3.6923	3.7135	0.2023	0.7746
		$Var(s+)$	1.2308	4.4103	2.9141	0.2334	0.6000
12	12	$UE(s^2)$	0.0513	1.8462	1.8554	0.0321	0.7071
		$Var(s+)$	0.5128	1.8462	1.5934	0.0846	0.1690
		GO-SSD	0.3077	3.6923	3.6208	0.0909	0.3333
12	24	$UE(s^2)$	0.0533	6.7733	6.7818	0.1794	0.7143
		$Var(s+)$	1.6800	8.4533	5.6403	0.1952	0.7071
16	28	$UE(s^2)$	-0.0591	7.9803	7.9866	0.1466	0.5238
		$Var(s+)$	1.7143	9.9310	7.0009	0.1624	0.5222
20	24	$UE(s^2)$	0.1200	5.8133	5.8086	0.0973	0.4141
		$Var(s+)$	0.6133	5.9733	5.6065	0.0965	0.4530
		GO-SSD	0.4800	9.6000	9.3852	0.0783	0.6000
24	28	$UE(s^2)$	0.1281	6.2463	6.2376	0.0832	0.3714
		$Var(s+)$	0.7488	6.7389	6.1859	0.0860	0.3055
		GO-SSD	0.2759	6.6207	6.5527	0.0370	0.3333
40	56	$UE(s^2)$	0.1704	14.5965	14.5720	0.0768	0.3504
		$Var(s+)$	1.3759	15.9900	14.1012	0.0808	0.3572
		GO-SSD	0.4211	16.8421	16.6700	0.0327	0.6000

Table 2: Comparison of design characteristics for all simulation designs

3 Design Size and Factor Sparsity

Recall the insights of Marley and Woods (2010) regarding SSD sample size requirements as a function of number of factors and number of active factors. In general, their simulations suggested that the run size should be at least three times the number of active factors for successful screening with SSDs. They also assert that the level of saturation of a SSD—the number of factors relative to the number of runs, k/n —should be less than 2.

To further investigate the recommendations of Marley and Woods (2010), we briefly present simulation results that include a larger variety of design sizes than previously studied. For a complete description of the simulation, please refer to Section 3.2 of the main article. We consider 22 different (n, k) combinations (Table 3) ranging from (6, 10) to (40, 100), constructed with both the $Var(s+)$ and $UE(s^2)$ criteria (giving a total of 42 SSDs). Three levels of sparsity are considered ($0.25n$, $0.5n$, $0.75n$) and two levels of complexity ($SN=1$ and $SN=3$). We analyze the designs using the Dantzig procedure outlined in Section 3.1 of the main article with $\gamma = 1$.

Figure 4 shows power as a function of n for each of the 42 SSDs. Each design is indicated by the ratio of n to the number of active factors, a , in the simulation. It is clear that the scenarios where the number of runs was at least three times the number of active factors ($n/3 \geq a$) produce higher power than situations where $n/3 < a$. This pattern is present regardless of the complexity of the simulation scenario (compare $SN=1$ to $SN=3$ in Figure 4). We note that for designs with fewer than 25 or so runs, violation of this rule of thumb seems to be particularly problematic for analysis.

Figure 5 examines how the level of saturation, k/n , is associated with power. Average power over both model complexities ($SN = 1$ and $SN = 3$) and design types versus the k/n ratio is shown. The marker size represents the number of runs in the SSD. There is a clear degradation in performance as the level of supersaturation increases. As the number of factors increases, the column correlations will necessarily rise; it is well known that large column correlation can impact ability to detect active factors. The designs with larger n generally have higher power for a similar level of k/n .

Thus, we confirm the recommendation from Marley and Woods (2010) that n/a be larger than 3, but find no clear evidence that $k/n = 2$ is a changepoint. Rather, there is a steady degradation of performance as the level of supersaturation grows larger. We also note that if the run size is small, there is an additional possibility of reduced effectiveness.

n	k	n	k
6	10	15	45
8	12	16	28
9	18	18	22
10	11	20	34
10	15	20	24
10	22	24	28
10	28	26	31
12	12	31	33
12	24	40	56
14	24	40	85
15	35	40	100

Table 3: SSD sizes used in simulations.

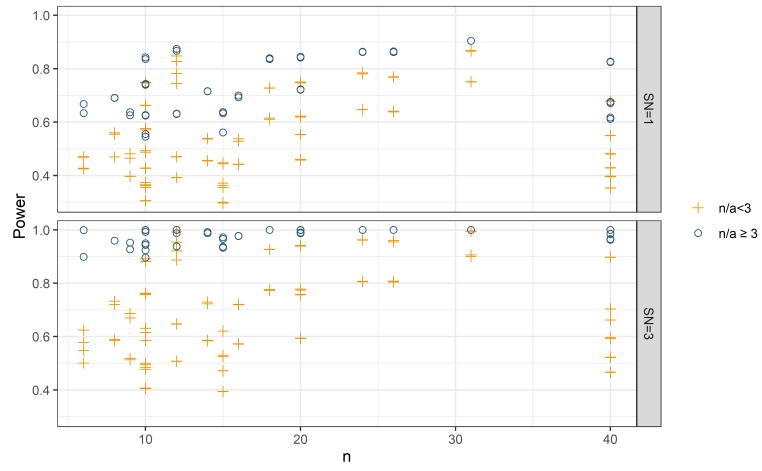


Figure 4: Power vs. run size (40 SSDs shown).

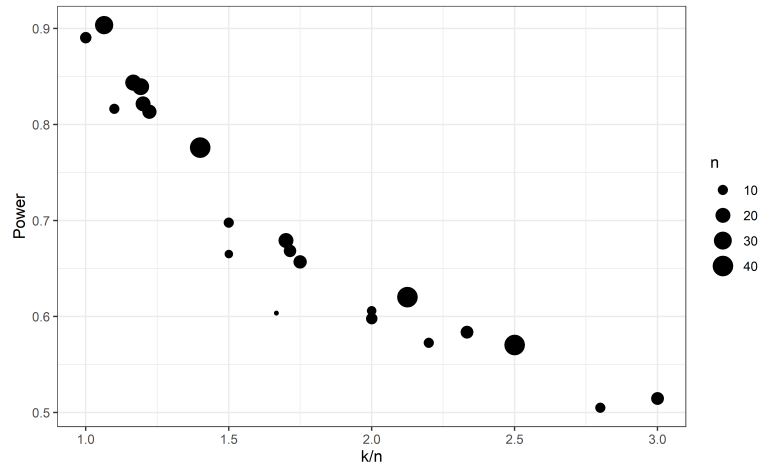


Figure 5: Average Power vs. k/n

4 $Var(s+)$ and Dantzig Selector Simulation Results

This section contains the results for the simulations referred to in Section 4.1 of the main article. We compare three versions of the Dantzig selector ($\gamma = \sigma^2$, $\gamma = 0.1 \times \max|\hat{\beta}_j|$, and $\delta = 0$) for $Var(s+)$ -optimal SSDs.

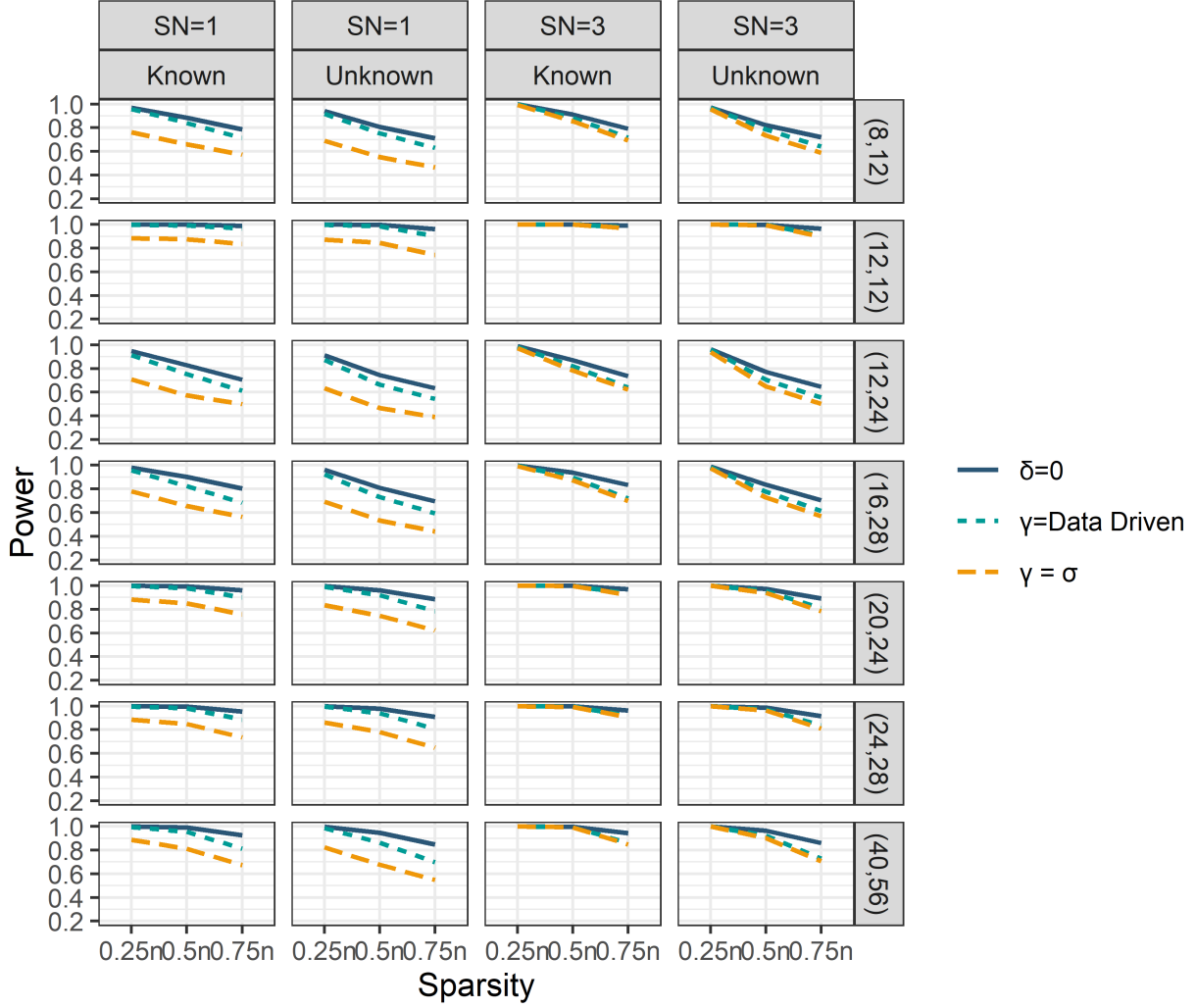


Figure 6: Power vs. Sparsity level for the $Var(s+)$ designs by sign specification (known, unknown) and model complexity (SN) for the different values of γ and the $\delta = 0$ Dantzig solution.

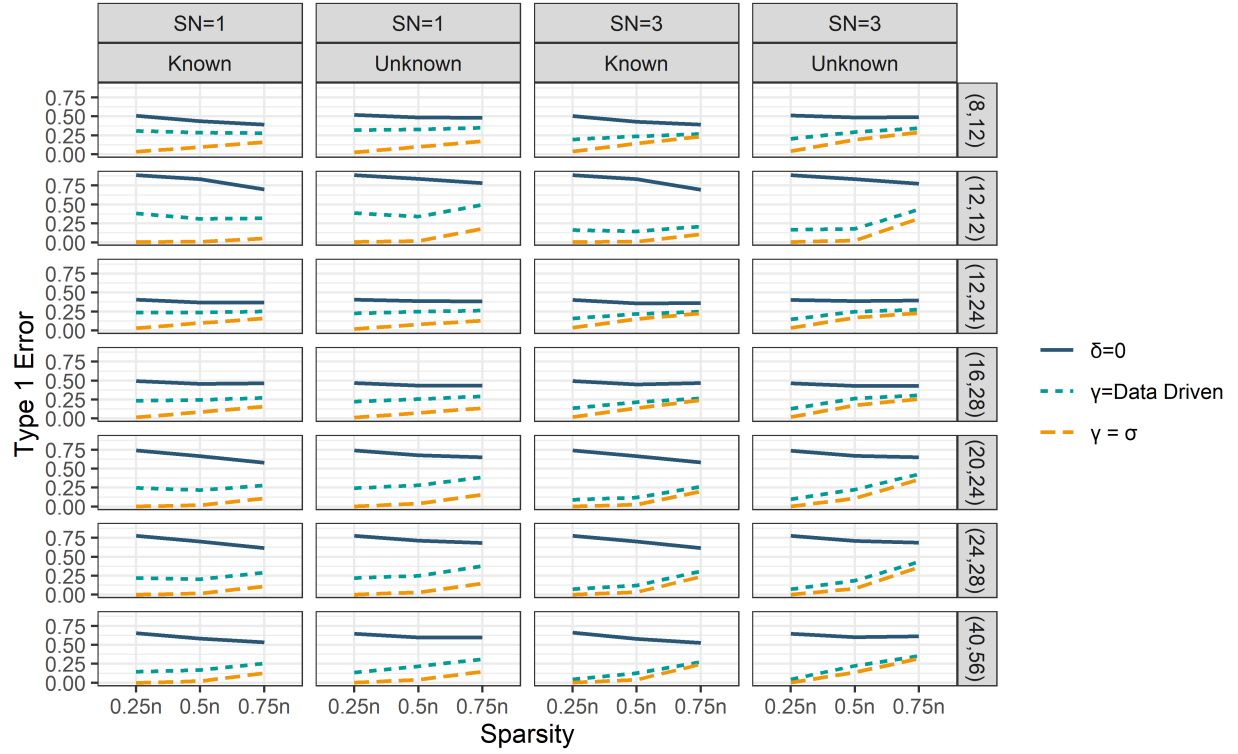


Figure 7: Type I Error vs. Sparsity for $Var(s+)$ designs by sign specification (known, unknown) and model complexity (SN) for the different values of γ and the $\delta = 0$ Dantzig solution.

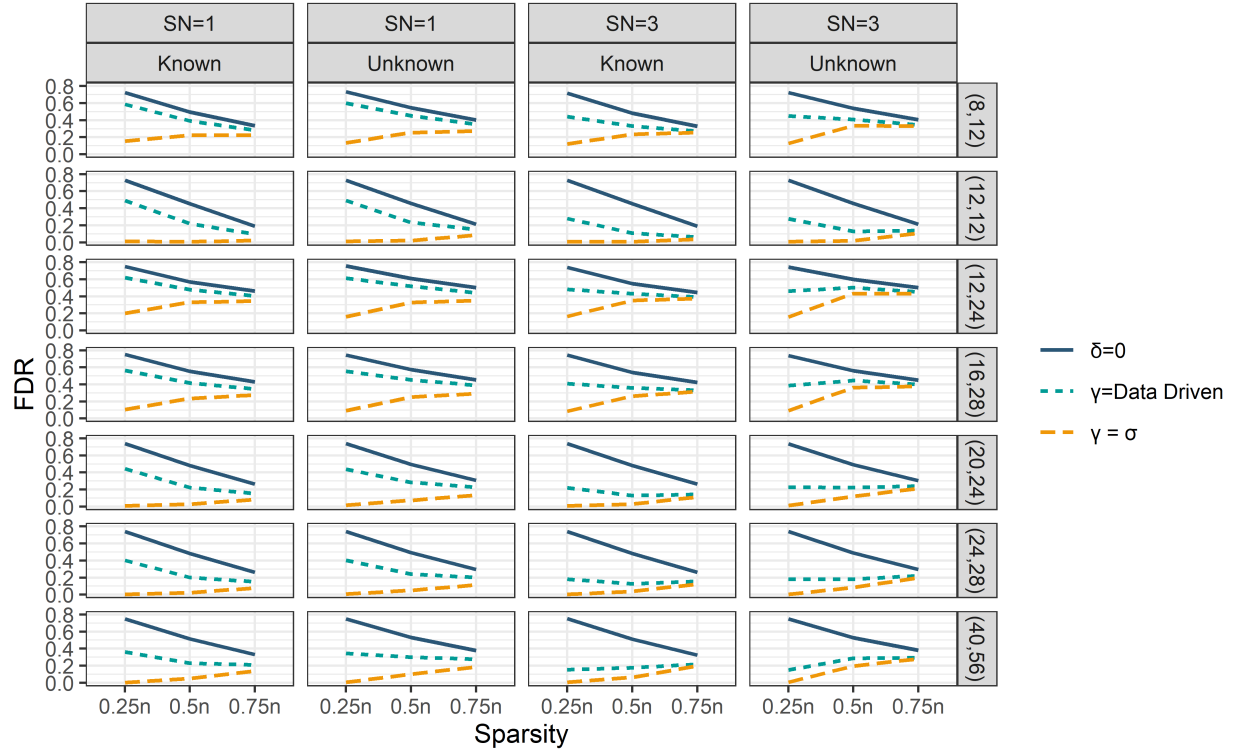


Figure 8: False Discovery Rate vs. Sparsity for $\text{Var}(s+)$ designs by sign specification (known, unknown) and model complexity (SN) for the different values of γ and the $\delta = 0$ Dantzig solution.

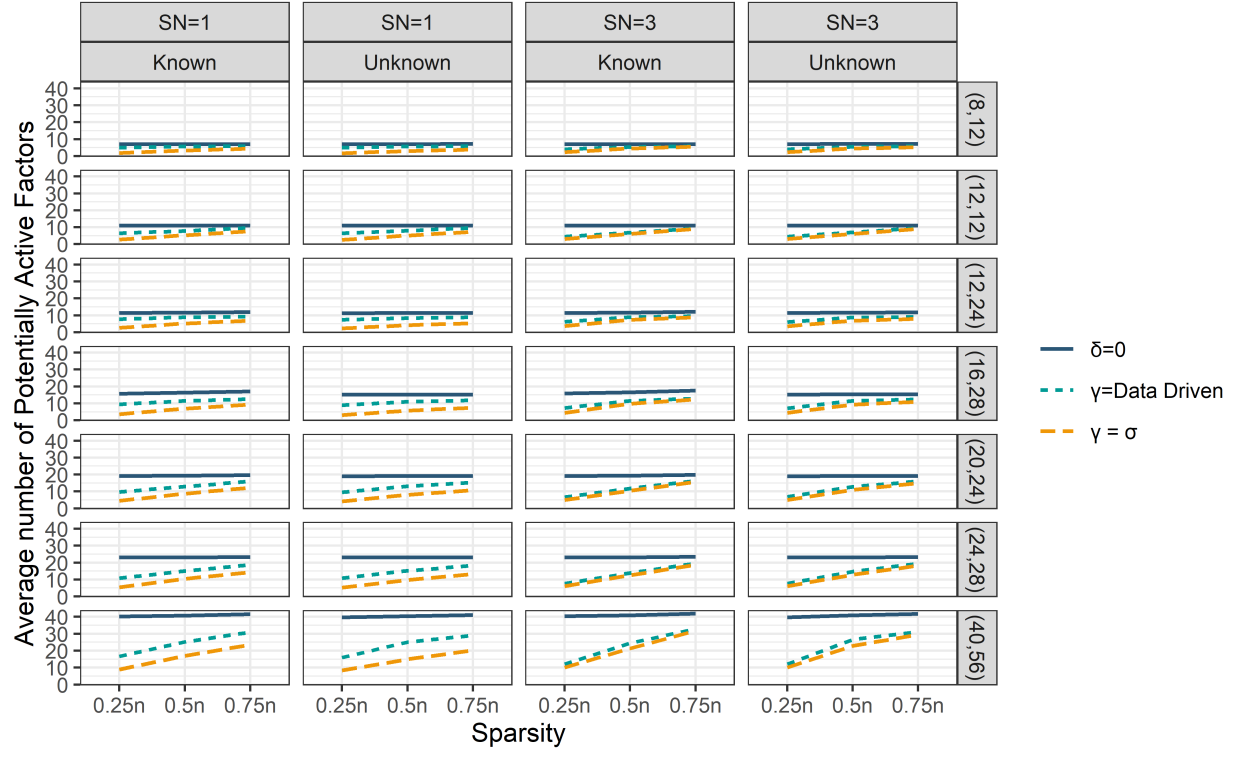


Figure 9: Average number of potentially active factors vs. Sparsity for $Var(s+)$ designs by sign specification (known, unknown) and model complexity (SN) for the different values of γ and the $\delta = 0$ Dantzig solution.

5 Additional GO-SSD Simulation Results

This section contains results for simulations referred to in Section 5.2 of the main article. Figures 10 and 11 show the false discovery rates (FDRs) and average model sizes for the Jones et al. (2019) and MaxPower analysis methods. The FDR's are generally lower for Jones et al. (2019) except for the low sparsity setting, $0.75n$. This is explained by the decrease in the number of inactive factors.

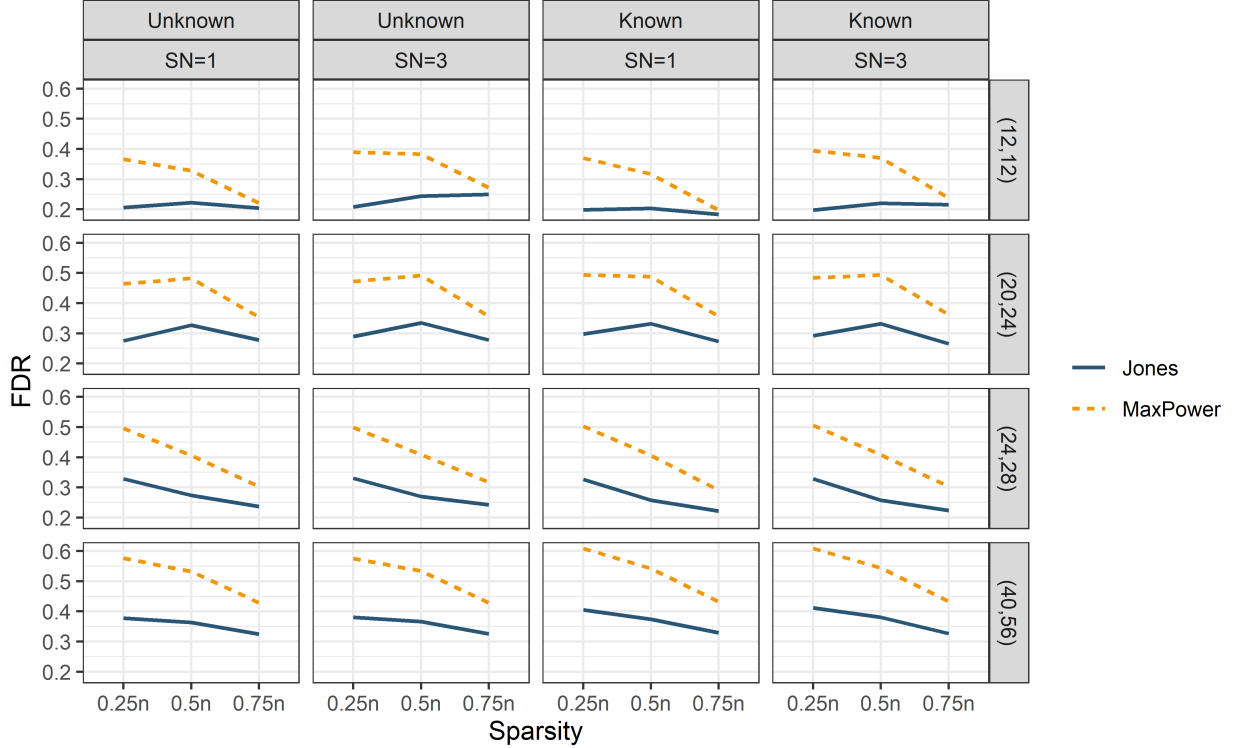


Figure 10: False Discovery Rate vs. Sparsity for GO-SSDs by sign specification (known, unknown) and model complexity (SN) for each analysis method.

Jones et al. (2019) recommend that if lack of fit is detected for the model of size $r - 1$, where r is the rank of a factor group, the first $r - 1$ factors chosen in their sequential approach should be deemed active and the remaining factors in the group should be deemed potentially active. While we take a different classification approach in this paper, one reviewer requested we impose this type of classification system on the MaxPower analysis. To be consistent with Jones et al. (2019), for a given rank $r_1 < r$, we classify a factor as active when it is included in the r_1 -rank model having the minimum MSE_{g1} . If this model's $LOF_{g1} < F(1 - \alpha, r - r_1, df_d)$ then we do not consider a larger rank model. The remaining factors in the group are classified as either potentially active or inactive, depending on whether they are included in other models that exhibit lack of fit.

As a short demonstration, suppose a group has 8 factors with rank 5 and we consider all models of rank $r_1 = 2$. The best model includes factors 1 and 2; denote its LOF by LOF_{g1}^* . Here are a few scenarios leading to different classifications:

1. $LOF_{g1}^* < F(1 - \alpha, r - 2, df_d)$ and no other rank- r_1 model has a $LOF_{g1} < F(1 - \alpha, r - 2, df_d)$.

Factors 1 and 2 are deemed active, while factors 3 to 8 are deemed inactive. The classification would be the same for both analysis approaches.

2. $LOF_{g1}^* < F(1 - \alpha, r - 2, df_d)$ but the model with factors 1 and 3 also has $LOF_{g1} < F(1 - \alpha, r - 2, df_d)$. Factors 1 and 2 are deemed active, factor 3 is potentially active, factors 4 to 8 are deemed inactive. The classifications are different between the two analysis approaches.
3. $LOF_{g1}^* > F(1 - \alpha, r - 2, df_d)$ so all models exhibit lack of fit. Since $r_1 = 2 = \lfloor \frac{5}{2} \rfloor$, MaxPower declares all factors in the group as potentially active. Under this new classification system, factors 1 and 2 are deemed active, while factors 3 to 8 are deemed potentially active. Jones et al. (2019) would instead consider all models of size 3.

Unlike MaxPower, Jones et al. (2019) will rarely designate factors as potentially active. This can only happen when all rank- $r - 1$ models exhibit lack of fit. We argued through simulation in Section 5.1 of the main article that this will rarely occur.

Figures 12 and 13 show the power and type 1 error rates, respectively, for the active and potentially active sets for the simulation scenarios under the GO-SSD with $n = 20$ and $k = 24$. Figures 3 and 4 of the main article pool these two groups together. The first takeaway is that the Jones et al. (2019) approach focuses its efforts on constructing the active factor group set, having higher power than the MaxPower active group but also higher Type 1 error. This can be explained by its consideration of models with rank greater than $\lfloor \frac{r}{2} \rfloor$. We can also see that MaxPower's large overall power is explained by its larger potentially active factor set. While this greatly improves power, it also significantly inflates the overall Type 1 error.

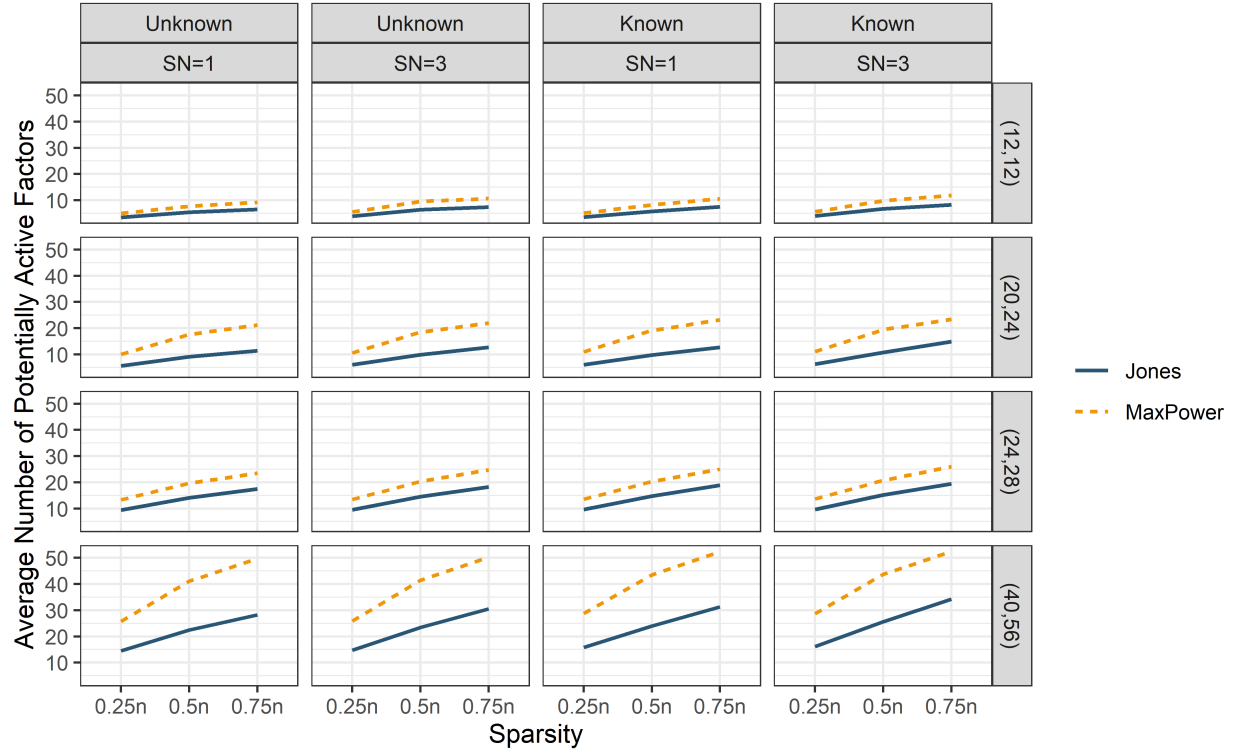


Figure 11: Average number of potentially active factors vs. Sparsity for GO-SSDs by sign specification (known, unknown) and model complexity (SN) for each analysis method.

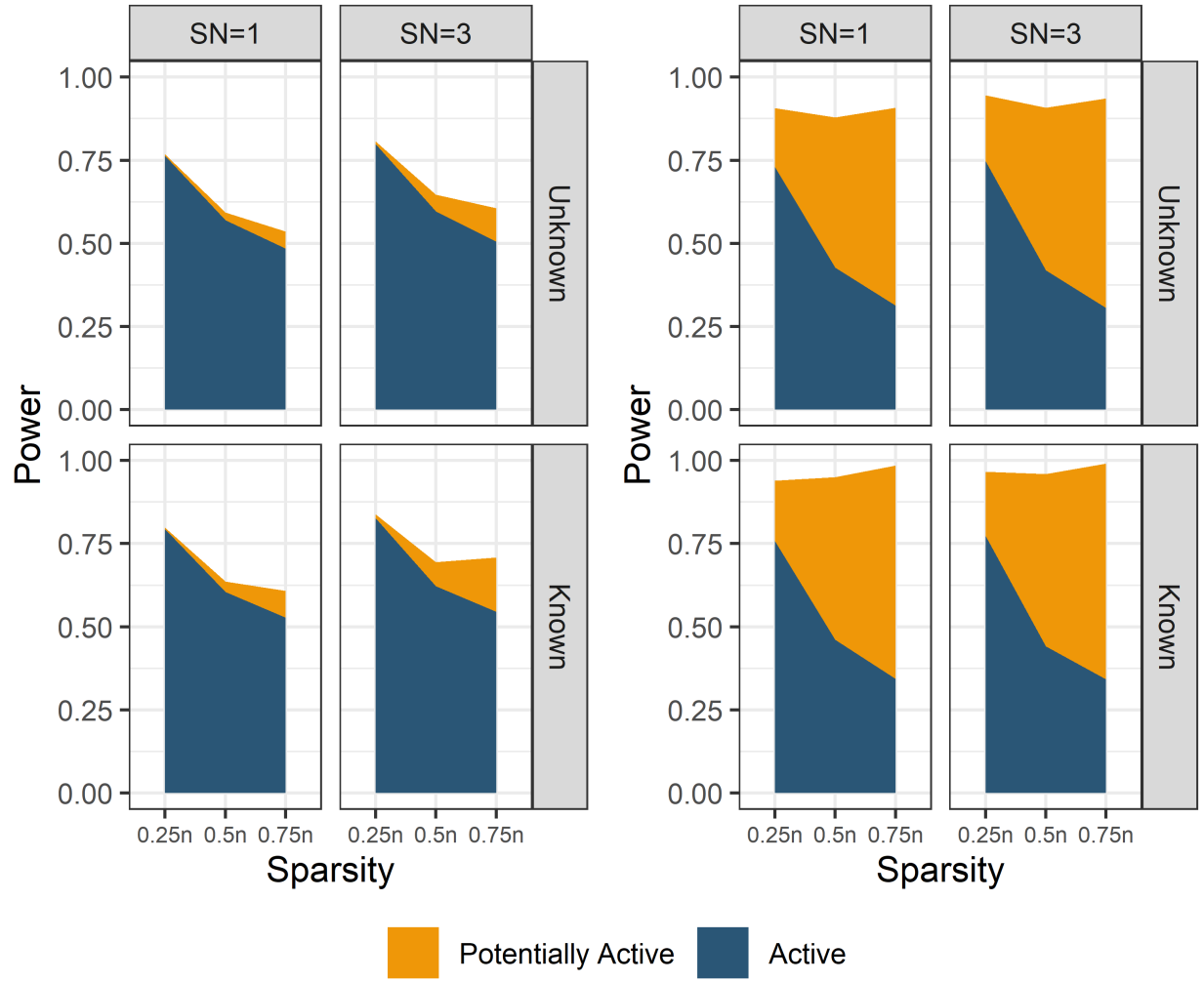


Figure 12: Average Power classification for the Jones et al. (2019) method (left) and for the MaxPower method (right) for the (20,24) GOSSD by sparsity, effect direction and S/N .

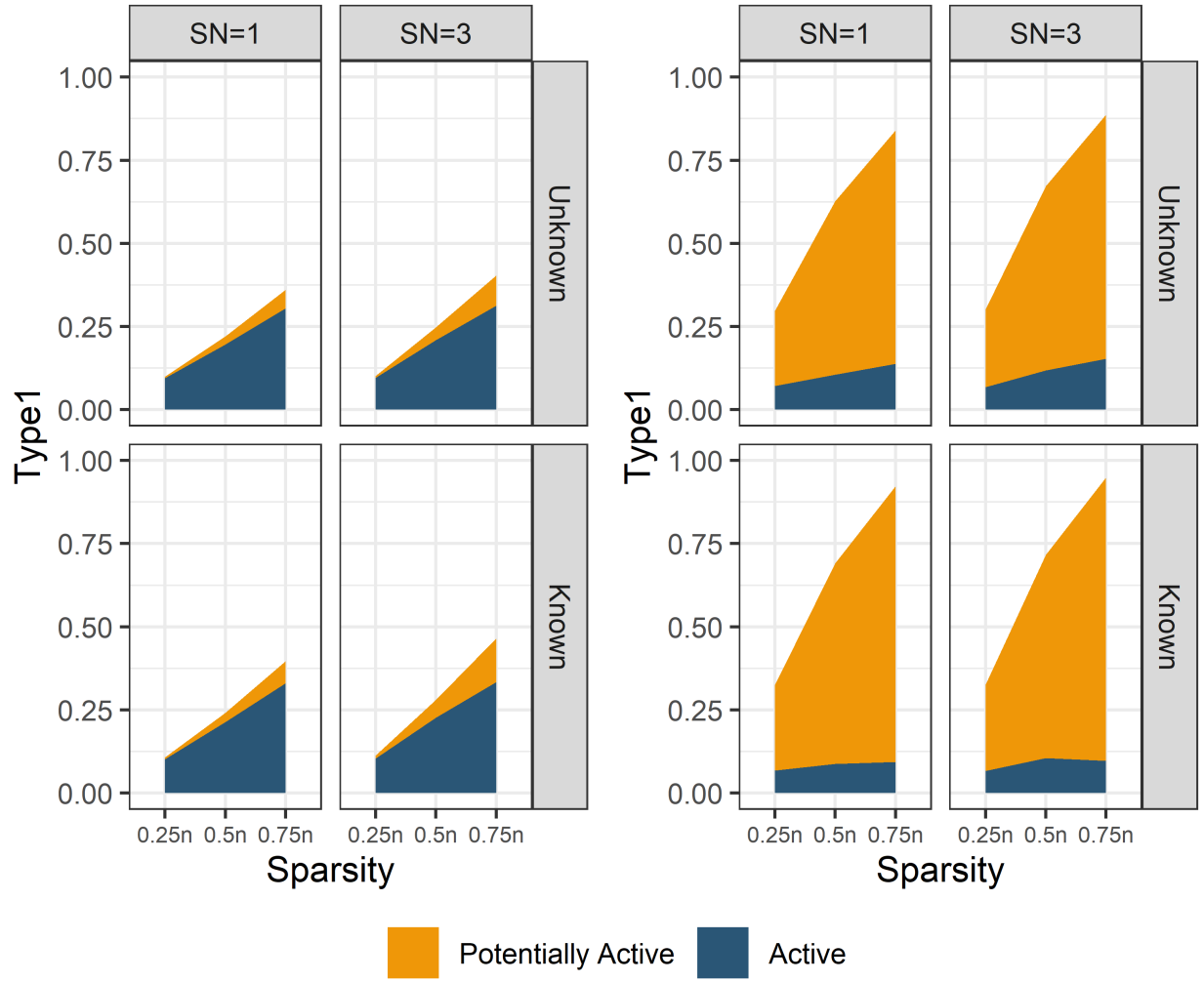


Figure 13: Average type 1 error classification for the Jones et al. (2019) method (left) and for the MaxPower method (right) for the (20,24) GOSSD by sparsity, effect direction and S/N .

6 Additional Simulation results comparing $Var(s+)$ -optimal, $UE(s^2)$ -optimal and GO-SSDs

This section contains results referred to in Section 6 of the main article.

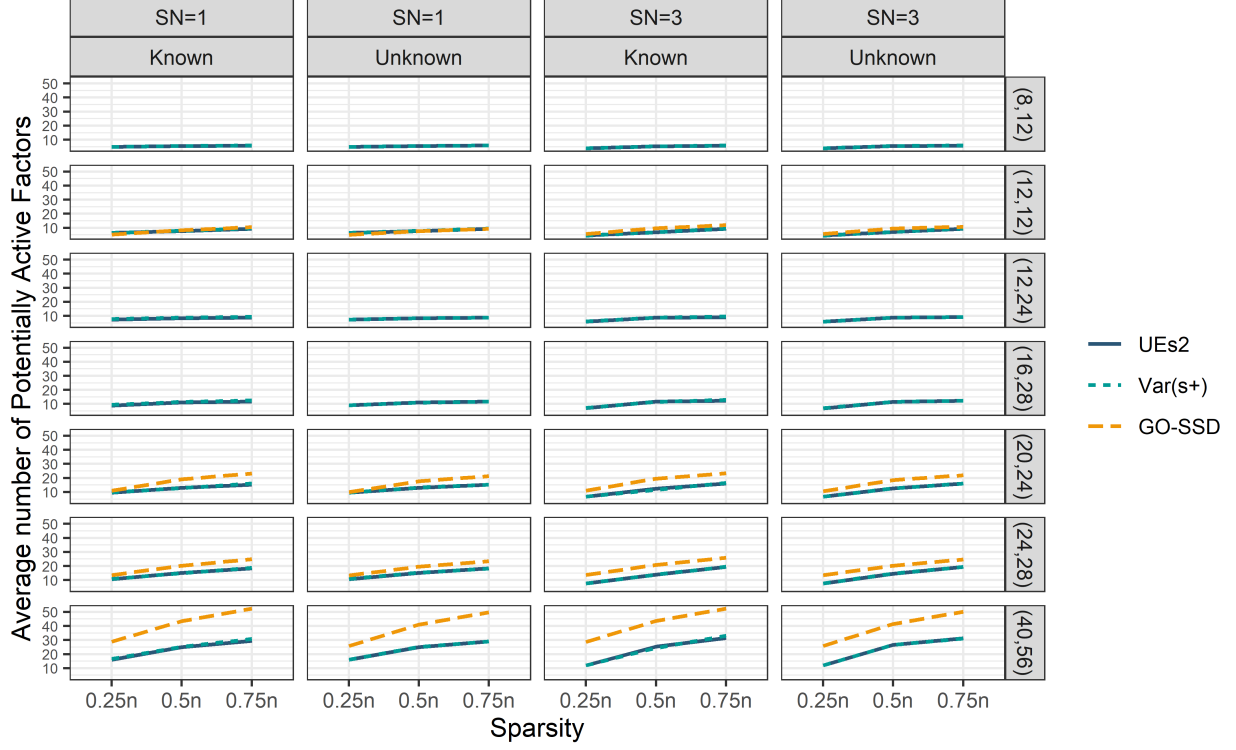


Figure 14: Average number of potentially active factors vs. design size $UE(s^2)$ and $Var(s+)$ -optimal SSDs using the data-driven Dantzig selector, and the MaxPower GoSSD.

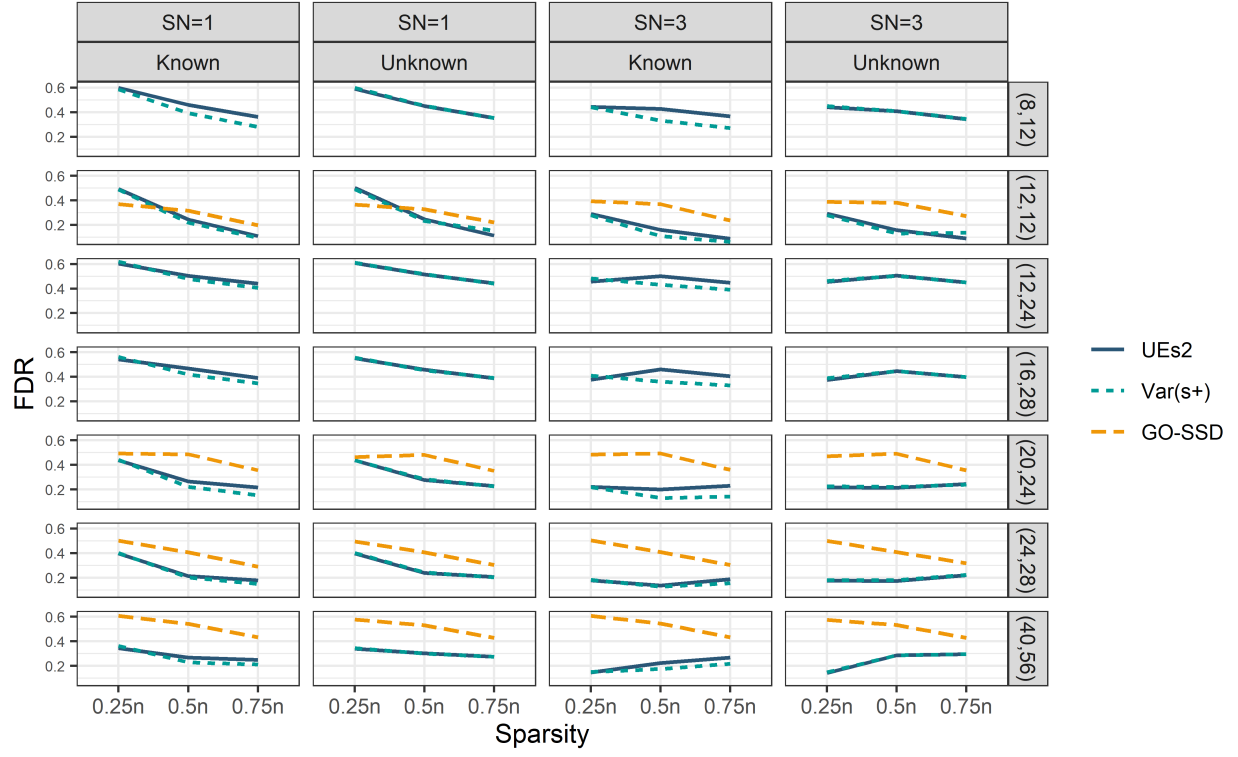


Figure 15: False Discovery Rate vs. design size $UE(s^2)$ and $Var(s+)$ -optimal SSDs using the data-driven Dantzig selector, and the MaxPower GoSSD.

References

- Jones, B., Lekivetz, R., Majumdar, D., Nachtsheim, C. J., and Stallrich, J. W. (2019). Construction, properties, and analysis of group-orthogonal supersaturated designs. *Technometrics*, 0(ja):1–31.
- Marley, C. J. and Woods, D. C. (2010). A Comparison of design and model selection methods for supersaturated experiments. *Computational Statistics and Data Analysis*, 54:3158–3167.