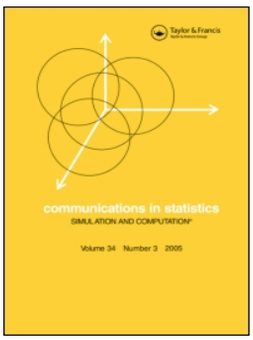

Taylor & Francis
Taylor & Francis Group

Communications in Statistics - Simulation and Computation

ISSN: 0361-0918 (Print) 1532-4141 (Online) Journal homepage: http://www.tandfonline.com/loi/lssp20

A three-stage variable selection method for supersaturated designs

Ai-Jun Qi, Zong-Feng Qi, Jian-Feng Yang & Qiao-Zhen Zhang

To cite this article: Ai-Jun Qi, Zong-Feng Qi, Jian-Feng Yang & Qiao-Zhen Zhang (2017) A three-stage variable selection method for supersaturated designs, Communications in Statistics - Simulation and Computation, 46:4, 2601-2610, DOI: 10.1080/03610918.2015.1053927

To link to this article: http://dx.doi.org/10.1080/03610918.2015.1053927

Accepted author version posted online: 06 Jul 2015.
Published online: 06 Jul 2015.

COMMUNICATIONS IN STATISTICS—SIMULATION AND COMPUTATION®
2017, VOL. 46, NO. 4, 2601–2610
http://dx.doi.org/10.1080/03610918.2015.1053927

Taylor & Francis
Taylor & Francis Group

A three-stage variable selection method for supersaturated designs

Ai-Jun Qi[a,b], Zong-Feng Qi[a], Jian-Feng Yang[b], and Qiao-Zhen Zhang[b]

[a]The State Key Laboratory of Complex Electromagnetic Environment Effects on Electronics and Information System (CEMEE), Luoyang, China; [b]LPMC and Institute of Statistics, Nankai University, Tianjin, China

ABSTRACT

A supersaturated design (SSD) is a design whose run size is not enough for estimating all main effects. Such a design is commonly used in screening experiments to screen active effects based on the effect sparsity principle. Traditional approaches, such as the ordinary stepwise regression and the best subset variable selection, may not be appropriate in this situation. In this article, a new variable selection method is proposed based on the idea of staged dimensionality reduction. Simulations and several real data studies indicate that the newly proposed method is more effective than the existing data analysis methods.

ARTICLE HISTORY
Received 11 February 2015
Accepted 7 May 2015

KEYWORDS
Best subset; Stepwise regression; Supersaturated design; Variable selection

MATHEMATICS SUBJECT CLASSIFICATION
Primary 62K15; Secondary 62J05

1. Introduction

Many preliminary industrial screening experiments typically contain a large number of potentially relevant factors. Based on the effects sparsity assumption, only a few of them are believed to be active. A supersaturated design (SSD) is a design whose run size is not enough for estimating all the main effects. The construction of SSDs can date back to Satterthwaite (1959), who suggested the use of random balanced designs, and Booth and Cox (1962), who proposed an algorithm to construct systematic SSD. Since the appearance of Lin (1993) and Wu (1993), lots of criteria and methods have been proposed for the construction of SSDs. A comprehensive list of early works can be found in Liu et al. (2006), Liu and Liu (2011), and Sun, Lin, and Liu (2011). Compared with the construction of SSDs, the inferential aspect of such designs needs more investigation. Unfortunately, the data analysis is challenging because even a main effect model is not identifiable for SSDs.

Finding the sparse active effects is a common and fundamental application of SSDs. Some new analysis methods were developed in recent years. At first, Lin (1993) used the stepwise selection method to screen active factors; Chipman et al. (1997) proposed a Bayesian variable selection approach for analyzing experiments with complex aliasing; Westfall et al. (1998) developed an error control skill in forward selection; Beattie et al. (2002) gave a two-stage Bayesian model selection strategy (SSVS/IBF) by combining the SSVS with the intrinsic Bayes factor (IBF); Li and Lin (2003) employed the penalized least squares with the smoothly clipped absolute deviation (SCAD) penalty; Holcomb et al. (2003) gave contrast-based methods; Lu and Wu (2004) proposed a strategy based on the idea of staged dimensionality; Zhang

CONTACT Jian-Feng Yang jfyang@nankai.edu.cn LPMC and Institute of Statistics, Nankai University, Tianjin 300071, China.

et al. (2007) employed a partial least-square regression; Georgiou (2008) gave a singular value decomposition (SVD) regression method for SSDs; Phoa et al. (2009) studied the Dantzig selector approach to screen active effects; Li et al. (2010) gave the contrast-orthogonality cluster analysis (COCA) method; Yin et al. (2013) proposed a multivariate partial least-square-stepwise regression method to select active effects in SSDs with multiple responses; Phoa (2014) proposed the stepwise response refinement screener (SRRS) and Chen et al. (2013) proposed a componentwise Gibbs sampler (CGS) method. Simulation studies show that the SRRS and CGS outperform the other approaches. In this article, we introduce a three-stage variable selection (TSVS) method, which not only is much more effective than SRRS and CGS, but also is much easier to be understood.

We denote the total set of factors by $\mathcal{A}$. The TSVS method chooses the active effects by three stages. The first stage contains narrowing the active factors to $\mathcal{B}$ by a stepwise selection procedure. The second stage screens out the factors with small absolute coefficient one by one by regressing the response on $\mathcal{B}$, and then we get the candidate active factors set $\mathcal{C}$. The third stage performs the best subset search from candidate set $\mathcal{C}$. Simulation studies show that the method is effective for analyzing data collected from SSDs, and in most cases the method outperforms the methods SRRS and CGS.

The article is organized as follows. Section 2 introduces the screening procedure of the TSVS method and discusses the main idea of the method. Section 3 compares the performance of TSVS with other existing methods. The last section contains some concluding remarks.

2. The screening procedure

Consider a linear regression model $y = X\beta + \epsilon$, where y is an $n \times 1$ vector of responses, $X = (x_1, \ldots, x_k)$ is an $n \times k$ model matrix, $\beta = (\beta_1, \ldots, \beta_k)'$ is a $k \times 1$ vector of unknown coefficients, and ϵ is an $n \times 1$ vector of noise that follows a multivariate normal distribution with mean zero and covariance matrix $\sigma^2 I_n$.

The TSVS method is divided into three procedures: stepwise selection, screening out factors with small absolute coefficient, and all subset search. We denote the total set of factors by $\mathcal{A}$, the set of factors after first procedure by $\mathcal{B}$, the set of factors after the second procedure by $\mathcal{C}$, and the set of final active factors by $\mathcal{D}$.

(i) Stepwise screening

Step 1 Standardize the data y and X so that y has mean 0 and each column of X has length one.

Step 2 Perform the stepwise selection procedure on the factors in $\mathcal{A}$ by choosing the $\alpha_{\text{in}} = 0.05$, $\alpha_{\text{out}} = 0.1$, and output the factors set $\mathcal{B}$.

(ii) Screening out factors with small absolute coefficients

Step 3 For all the factors in $\mathcal{B}$,

(a) Regress y on $\mathcal{B}$ to obtain the estimate $\beta_{\mathcal{B}}$;

(b) Set a threshold of noise level $\gamma > 0$, do $\beta^1_{\mathcal{B}} = |\beta_{\mathcal{B}}| - \gamma$;

(c) Rank the elements in $\beta^1_{\mathcal{B}}$, and drop the smallest one to obtain new $\mathcal{B}$;

(d) Repeat (a) to (c) until the absolute of any element in $\beta_{\mathcal{B}}$ is greater than γ, then output $\mathcal{B}$ as $\mathcal{C}$.

(iii) All subset search

Step 4 Perform all subset search for the factors in $\mathcal{C}$. For all possible models, compute the modified AIC (mAIC) value defined by

$$\text{mAIC} = n/q \log(\text{RSS}/n) + q^2/\sqrt{n},$$

where n is the number of observations, q is the number of factors in the model, and RSS $= \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ is the residual sum of squares.

Step 5 The final model is chosen as the one with the smallest mAIC value among all models. Then output the final factors set as $\mathcal{D}$.

Step 1 is a standard normalization on the response y and the columns of X, so that the columns of X have equal length. Step 2 is a usual stepwise selection procedure. Through it, we can get the factors set $\mathcal{B}$, which are significant under statistic F at $\alpha_{in} = 0.05$ and $\alpha_{out} = 0.1$. This step can ensure that the whole model could reach some significant level, however, it tends to select too many inactive factors in most cases, so comes Step 3. Step 3 can select out the factors with large absolute coefficients. Unless specified, we select γ to be 10% of the largest $|\hat{\beta}_i|$ in the model, that is, $\gamma = 0.1 \max |\hat{\beta}_i|$. The factors with the absolute coefficient lower than the 10% of the largest absolute coefficient are considered inactive. Here, we choose 10% just as Phoa (2014) recommended. If γ is too large, the factors set $\mathcal{C}$ may lose some active factors; if γ is too small, the factors set $\mathcal{C}$ may contain too many inactive factors. Therefore, it is important to choose an appropriate value of γ. Simulations show that for most cases $\gamma = 0.1 \max |\hat{\beta}_i|$ can produce satisfactory results, and the value can be adjusted according to our previous experience. Steps 4 and 5 are the adjustment of the above steps. The final model with the factors in $\mathcal{C}$ is chosen to minimize the mAIC value, where the mAIC criterion is chosen to increase the penalty of model size than that of the traditional AIC criterion.

3. Performance of the TSVS method

In order to show the performance of the TSVS method, we revisit six examples that have been analyzed in existing literature.

Example 3.1. We apply our method to the SSD first considered by Lin (1993). There are 24 factors and 14 responses but the factors 13 and 16 are identical. Therefore, we delete the factor 13 and rename factors 14 to 24 as 13 to 23. The design matrix and the response are shown in Table 1 .

By applying the procedure of TSVS, the output of Step 2 is $\mathcal{B} = \{14\}$. Then regress y on $\mathcal{B}$ to obtain the estimate $|\hat{\beta}_{14}| = 53.21$. So TSVS selects the factor 14 as the active factor.

This example was analyzed by several other researchers. Westfall et al. (1998), Beattie et al. (2002), Phoa (2014), and Chen et al. (2013) suggested that only factor 14 is active. Li and Lin (2003) and Zhang et al. (2007) chose 14, 12, 19, 4 as active factors. Table 2 lists different methods and the corresponding active factors selected for the data in Table 1. This table shows that the method TSVS is comparable with the existing methods.

Example 3.2. The second example is the cast fatigue experiment with 7 factors and 12 observations (Wu and Hamada, 2009). The design and observations are listed in Table 3 . To use TSVS, we set $\gamma = 0.045$, which is 0.1 times the largest $|\hat{\beta}_F|$. The final output shows that the factor F is active, which is the same as the results of Wu and Hamada (2009), Phoa et al. (2009), and Chen et al. (2013).

If all two-factor interactions are also considered in the model, we set $\gamma = 0.0458$, which is 0.1 times the largest $|\hat{\beta}_{FG}|$. The final output shows that the main effect F and two-factor interaction FG are active, which is the same as the results of Wu and Hamada (2009), Westfall et al. (1998), Hamada and Hamada (2010), and Chen et al. (2013).

Table 1. The two-level SSD in Lin (1993).

	Factors																						Response
Run	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	Y
1	1	1	1	−1	−1	−1	1	1	1	1	1	−1	−1	−1	1	1	−1	−1	1	−1	−1	−1	1	133
2	1	−1	−1	−1	−1	−1	1	1	1	−1	−1	−1	1	1	1	−1	1	−1	−1	1	1	−1	−1	62
3	1	1	−1	1	1	−1	−1	−1	−1	1	−1	1	1	1	1	1	−1	−1	−1	−1	1	1	−1	45
4	1	1	−1	1	−1	1	−1	−1	−1	1	1	−1	−1	1	1	−1	1	1	1	−1	−1	−1	−1	52
5	−1	−1	1	1	1	1	−1	1	1	−1	−1	−1	−1	1	1	1	−1	−1	1	−1	1	1	1	56
6	−1	−1	1	1	1	1	1	−1	1	1	1	−1	1	1	−1	1	1	1	1	1	1	−1	−1	47
7	−1	−1	−1	−1	1	−1	−1	1	−1	1	−1	1	1	−1	1	1	1	1	1	1	−1	−1	1	88
8	−1	1	1	−1	−1	1	−1	1	−1	1	−1	−1	−1	−1	−1	−1	−1	1	−1	1	1	1	−1	193
9	−1	−1	−1	−1	−1	1	1	−1	−1	−1	1	1	−1	1	−1	1	1	−1	−1	−1	−1	1	1	32
10	1	1	1	1	−1	1	1	1	−1	−1	−1	1	1	1	−1	1	−1	1	−1	1	−1	−1	1	53
11	−1	1	−1	1	1	−1	−1	1	1	−1	1	−1	1	−1	−1	−1	1	1	−1	−1	−1	1	1	276
12	1	−1	−1	−1	1	1	1	−1	1	1	1	1	−1	−1	1	−1	−1	1	−1	1	1	1	1	145
13	1	1	1	1	1	−1	1	−1	1	−1	−1	1	−1	−1	−1	−1	1	−1	1	1	−1	1	−1	130
14	−1	−1	1	−1	−1	−1	−1	−1	−1	−1	1	1	1	−1	−1	−1	−1	−1	1	−1	1	−1	−1	127

Table 2. Summary of the selection results for Example 3.1.

Selection method	Active factors
Beattie et al. (2002)	14
Phoa et al. (2009)	14
Li and Lin (2003)	14, 12, 19, 4
Zhang et al. (2007)	14, 12, 19, 4
Westfall et al. (1998)	14
Phoa (2014)	14
Chen et al. (2013)	14
TSVS	14

Example 3.3. This example considers the SSD with 31 factors and 18 observations from Rais et al. (2009). The design matrix and observations are shown in Table 4 In the screening procedure, we set $\gamma = 0.781$, which is 0.1 times the largest $|\hat{\beta}_{24}|$. The final output shows that the factors 4, 24, 27, 28 are active. Rais et al. (2009) suggested that 13, 18, 19, 20, 24, 27, 28, 29, 30 are nine significant factors. By using the SRRS method, Phoa (2014) suggested that 24 and 27 are two significant factors. By using the CGS method, Chen et al. (2013) suggested that 19, 24, 27, and 28 are four significant factors. However, all the methods select 24 and 27 to be significant. This shows that the method TSVS can select correct important factors with a moderate model size. .

In the following three examples, we make some comparisons with other methods.

Example 3.4. In order to judge the performance of the TSVS method, we compare the TSVS with the following eight methods in the literature:

1. SSVS, the Bayesian variable selection procedure in George and McCulloch (1993) and Chipman et al. (1997);
2. SSVS/IBF, the two-stage Bayesian procedure in Beattie et al. (2002);
3. SCAD, the penalized least-square approach in Li and Lin (2003);
4. PLSVS, the partial least-square regression method in Zhang et al. (2007);
5. DS, the Dantzig selector in Phoa et al. (2009);
6. COCA, the contrast-orthogonality cluster analysis method in Li et al. (2010);
7. SRRS, the stepwise response refinement screener in Phoa (2014);
8. CGS, the componentwise Gibbs sampler in Chen et al. (2013).

The design used here is from Lin (1993) as listed in Table 1. Consider the following three models:

Model 1: $y = 10x_1 + \epsilon$,

Table 3. Design matrix and observations for the cast fatigue experiment.

Run	*A*	*B*	*C*	*D*	*E*	*F*	*G*	Response
1	1	1	−1	1	1	1	−1	6.058
2	1	−1	1	1	1	−1	−1	4.733
3	−1	1	1	1	−1	−1	−1	4.625
4	1	1	1	−1	−1	−1	1	5.899
5	1	1	−1	−1	−1	1	−1	7.000
6	1	−1	−1	−1	1	−1	1	5.752
7	−1	−1	−1	1	−1	1	1	5.682
8	−1	−1	1	−1	1	1	−1	6.607
9	−1	1	−1	1	1	−1	1	5.818
10	1	−1	1	1	−1	1	1	5.917
11	−1	1	1	−1	1	1	1	5.863
12	−1	−1	−1	−1	−1	−1	−1	4.809

Table 4. A two-level SSD in Rais et al. (2009).

	Factors															
Run	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1	−1	1	−1	1	1	1	−1	−1	−1	1	1	1	1	1	−1	1
2	1	−1	−1	1	−1	1	1	1	−1	−1	−1	1	1	1	1	1
3	−1	1	−1	−1	1	−1	1	1	1	−1	−1	−1	1	1	1	1
4	1	1	−1	−1	1	−1	−1	1	−1	1	1	1	−1	−1	−1	1
5	1	−1	1	1	−1	−1	1	−1	−1	1	−1	1	1	1	−1	−1
6	1	−1	1	−1	1	1	−1	−1	1	−1	−1	1	−1	1	1	1
7	−1	1	−1	1	−1	1	1	−1	−1	1	−1	−1	1	−1	1	1
8	−1	−1	1	−1	1	−1	1	1	−1	−1	1	−1	−1	1	−1	1
9	−1	−1	−1	1	−1	1	−1	1	1	−1	−1	1	−1	−1	1	−1
10	1	−1	−1	−1	−1	1	−1	1	−1	1	1	−1	−1	1	−1	−1
11	−1	1	−1	−1	−1	−1	1	−1	1	−1	1	1	−1	−1	1	−1
12	1	1	1	−1	−1	1	−1	−1	−1	−1	1	−1	1	−1	1	1
13	1	1	1	1	1	−1	1	1	1	−1	−1	1	−1	−1	−1	−1
14	−1	1	1	1	1	1	−1	1	1	1	−1	−1	1	−1	−1	−1
15	−1	−1	1	1	1	1	1	−1	1	1	1	−1	−1	1	−1	−1
16	1	1	1	−1	−1	−1	1	1	1	1	1	−1	1	1	1	−1
17	1	−1	1	1	1	−1	−1	−1	1	1	1	1	1	−1	1	1
18	−1	−1	−1	−1	−1	−1	−1	−1	−1	−1	−1	−1	−1	−1	−1	−1

	Factors															
Run	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	y
1	1	1	−1	−1	1	−1	−1	−1	−1	1	−1	1	−1	1	1	76.2
2	−1	1	1	1	−1	−1	1	−1	−1	−1	−1	1	−1	1	−1	82.6
3	1	−1	1	1	1	−1	−1	1	−1	−1	−1	−1	1	−1	1	99.4
4	1	1	1	1	−1	1	1	1	−1	−1	1	−1	−1	−1	−1	80.5
5	−1	1	1	1	1	1	−1	1	1	1	−1	−1	1	−1	−1	103.5
6	−1	−1	−1	1	1	1	1	1	−1	1	1	1	−1	−1	1	52.1
7	1	−1	−1	−1	1	1	1	1	1	−1	1	1	1	−1	−1	73.8
8	1	1	−1	−1	−1	1	1	1	1	1	−1	1	1	1	−1	89.8
9	1	1	1	−1	−1	−1	1	1	1	1	1	−1	1	1	1	100.7
10	1	−1	1	1	1	−1	−1	−1	1	1	1	1	1	−1	1	59.8
11	−1	1	−1	1	1	1	−1	−1	−1	1	1	1	1	1	−1	62.8
12	−1	−1	1	−1	−1	1	−1	1	1	1	−1	−1	−1	1	1	95.0
13	1	−1	1	−1	1	1	−1	−1	1	−1	−1	1	−1	1	1	74.9
14	−1	1	−1	1	−1	1	1	−1	−1	1	−1	−1	1	−1	1	84.4
15	−1	−1	1	−1	1	−1	1	1	−1	−1	1	−1	−1	1	−1	86.7
16	−1	1	−1	−1	−1	−1	1	−1	1	−1	1	1	−1	−1	1	58.3
17	1	−1	−1	1	−1	−1	−1	−1	1	−1	1	−1	1	1	−1	71.1
18	−1	−1	−1	−1	−1	−1	−1	−1	−1	−1	−1	−1	−1	−1	−1	64.8

Model 2: $y = -15x_1 + 8x_5 - 2x_9 + \epsilon$,
Model 3: $y = -15x_1 + 12x_5 - 8x_9 + 6x_{13} - 2x_{16} + \epsilon$,
where ϵ is the random error from $N(0, 1)$. Four measurements, TMIR, SEIR, MEDIAN, and MEAN, are used to compare the performances of different methods, where TMIR represents the rate that the selected model is the same as the true model, SEIR represents the rate that the selected model includes the true model, MEDIAN represents the median of the number of selected factors, and MEAN represents the average of the number of selected factors.

For each model, according to our experience, we set $\gamma = 1$ and repeat 1,000 times. The results are shown in Table 5, where the results of SSVS, SCAD, PLSVS, DS, SRRS are from Phoa (2014); the results of CGS are from Chen et al. (2013); the results of COCA are from Li et al. (2010).

For Model 1, all methods SCAD, PLSVS, DS, COCA, SRRS, CGS, and TSVS have a 100% SEIR. From the TMIR and MEAN, the methods SRRS and DS have better results than CGS. However, our method TSVS has a much better result than all other methods in terms of TMIR,

Table 5. Comparison results in Example 3.4.

Model	Method	TMIR	SEIR	MEDIAN	MEAN
	SSVS(1/10,500)	0.405	0.990	2	3.1
	SSVS(1/10,500)/IBF	0.610	0.980	1	2.5
	SCAD	0.756	1	1	1.7
	PLSVS($m=1$)	0.610	1	1	1.5
1	DS($\gamma=1$)	0.994	1	1	1.0
	COCA	0.892	1	1	1.20
	SRRS($\gamma=0.75$)	0.907	1	1	1.0
	SRRS($\gamma=1$)	0.998	1	1	1.0
	CGS	0.958	1	NA	1.045
	TSVS($\gamma=1$)	0.996	1	1	1.004
	SSVS(1/10,500)	0.086	0.300	3	4.7
	SSVS(1/10,500)/IBF	0.08	0.280	3	4.2
	SCAD	0.756	0.985	3	3.3
	PLSVS($m=1$)	0.764	1	3	3.3
2	DS($\gamma=1$)	0.844	0.853	3	2.9
	COCA	0.804	0.991	3	3.21
	SRRS($\gamma=0.75$)	0.898	0.925	3	3.0
	SRRS($\gamma=1$)	0.842	0.852	3	2.9
	CGS	0.947	0.985	NA	3.033
	TSVS($\gamma=1$)	0.987	0.997	3	3.008
	SSVS(1/10,500)	0.364	0.840	6	8.0
	SSVS(1/10,500)/IBF	0.407	0.750	5	5.6
	SCAD	0.697	0.994	5	5.4
	PLSVS($m=1$)	0.736	0.950	5	5.2
3	DS($\gamma=1$)	0.791	0.912	5	5.1
	COCA	0.852	0.992	5	5.14
	SRRS($\gamma=0.75$)	0.966	0.966	5	5.0
	SRRS($\gamma=1$)	0.953	0.953	5	5.0
	CGS	0.96	0.994	NA	5.040
	TSVS($\gamma=1$)	0.990	0.996	5	5.004

SEIR, and the model size. For Model 2, the method of CGS is better than all the other selection methods except TSVS. Our method has a 98.7% of TMIR and the model size is about 3. For Model 3, the CGS and SRRS have better results than other methods, with TMIRs being 96% and 96.6%, MEANs being 5.04 and 5, respectively. Our method, however, has a more satisfactory result with TMIR being 99% and MEAN being 5.004.

Example 3.5. This example is modified from Example 3.4. We randomly assign the active factors to the columns of X. Consider the following three cases:

Case 1: One active factor with the coefficient $\beta = (10)$;

Case 2: Three active factors with the coefficient vector $\beta = (-15, 8, -2)$;

Case 3: Five active factors with the coefficient vector $\beta = (-15, 12, -8, 6, -2)$.

For each case, we compare the TSVS with methods CGS, DS in Chen et al. (2013), and SRRS in terms of the measurements TMIR, SEIR, AEIR, IEIR, and MEAN, where AEIR is the average rate of active factors identified correctly and IEIR is the average rate of inactive factors that are included in the selected model. For each case, we repeat 1,000 times, generate 1,000 models and, by setting $\gamma = 1$ according to our experience, get the means of the five measurements. The results are shown in Table 6, where the results of CGS and DS are from Chen et al. (2013).

From Table 6, we can see that for Case 1, the Dantzig method has the best performance in all measurements. The CGS selects more inactive factors and has 96.3% in TMIR. The TSVS and SRRS methods have about 99.5% and 99.4% in TMIR, respectively. Considering the other four measurements, it is easy to see that TSVS is slightly better than SRRS. For

Table 6. Comparison results in Example 3.5.

Case	Method	TMIR(%)	SEIR(%)	AEIR(%)	IEIR(%)	MEAN
	CGS	96.3	100	100	3.84	1.040
1	DS($\gamma = 1$)	100	100	100	0.00	1.000
	SRRS	99.4	100	100	0.03	1.010
	TSVS	99.5	100	100	0.02	1.000
	CGS	95.9	99.2	99.73	1.48	3.037
2	DS($\gamma = 1$)	93.2	94.3	98.10	0.71	2.964
	SRRS	95.1	95.4	98.5	0.06	2.970
	TSVS	98.5	99.6	99.9	0.07	3.010
	CGS	92.3	96.6	99.22	1.25	5.024
3	DS($\gamma = 1$)	64.8	66.7	87.16	2.09	4.451
	SRRS	64.1	64.2	85.7	0.58	4.390
	TSVS	87.4	87.9	91.8	1.21	4.810

Case 2, the values of TMIR, AEIR, SEIR under TSVS are higher than those under other methods, and the model size of TSVS is more accurate, which shows that TSVS is the best. For Case 3, considered from all five measurements, the TSVS method is better than the methods SRRS and DS, and the CGS works a little better than the TSVS.

Example 3.6. The final example is more complicated than Example 3.5, where the coefficients of the inactive factors are set to be zero. Following Marley and Woods (2010), the coefficients of the inactive factors are drawn from $N(0, 0.2)$ and the coefficients of the active factors are drawn from $N(\beta, 0.2I_c)$ in this example. We repeat the simulations 1,000 times for each of the 1,000 models, and the performance of TSVS is evaluated by means of TMIR, SEIR, AEIR, IEIR, MEAN for the 1,000 models. Because the coefficients in this example are not fixed, we set $\gamma = 0.1\max(|\hat{\beta}_i|)$ and $\gamma = 0.2\max(|\hat{\beta}_i|)$ for Case 1 and $\gamma = 0.1\max(|\hat{\beta}_i|)$ for Cases 2 and 3. The results are shown in Table 7, where the results of CGS are from Chen et al. (2013).

In Case 1, we can see that when $\gamma = 0.1\max(|\hat{\beta}_i|)$, the results are not satisfactory, however, when $\gamma = 0.2\max(|\hat{\beta}_i|)$, the results are improved greatly, giving a TMIR of 93.3%, sharing 100% for AEIR and SEIR, and 0.32% for IEIR, and a more accurate model size 1.071.

In Case 2, the method TSVS has the best results among the three methods in terms of the measurements TMIR, SEIR, AEIR, and MEAN. For IEIR, the method TSVS is still better than CGS, while a little worse than the method SRRS. The IEIR of SRRS is 0.44%, while the IEIR of TSVS is 2.198%. This shows that TSVS might overselect inactive factors.

In Case 3, we can see that the TMIR of TSVS is 43.61%, higher than those of CGS and SRRS. The TSVS is also capable of identifying the smallest factors, whose SEIR is 48.62%. The AEIR is a little lower than CGS, and higher than SRRS, which means that the power of TSVS is

Table 7. Comparison results in Example 3.6.

Case	Method	TMIR(%)	SEIR(%)	AEIR(%)	IEIR(%)	MEAN
	CGS	89.2	100.0	100.0	10.55	1.118
	SRRS($\gamma = 0.1\max(\lvert\hat{\beta}_i\rvert)$)	37.0	100.0	100.0	3.30	1.725
1	TSVS($\gamma = 0.1\max(\lvert\hat{\beta}_i\rvert)$)	24.6	100.0	100.0	5.73	2.261
	SRRS($\gamma = 0.2\max(\lvert\hat{\beta}_i\rvert)$)	92.0	100.0	100.0	0.37	1.081
	TSVS($\gamma = 0.2\max(\lvert\hat{\beta}_i\rvert)$)	93.3	100.0	100.0	0.32	1.071
	CGS	32.6	41.3	80.43	7.48	2.608
2	SRRS	37.5	39.6	79.9	0.44	2.484
	TSVS	43.2	60.5	86.8	2.20	3.045
	CGS	26	30.7	85.5	4.68	4.485
3	SRRS	9.61	9.78	73.65	0.63	3.795
	TSVS	43.61	48.62	83.00	2.66	4.629

better than SRRS. The value of IEIR of TSVS is between that of CGS and SRRS, which means for the type one error, the SRRS is better than TSVS, and both are better than CGS. From the point of model size, the TSVS gives 4.629, which is the most accurate one.

4. Concluding remarks

Supersaturated designs (SSDs) are very useful for screening experiments and various methods are available for analyzing data from such designs. In this article, we propose a three-stage variable selection (TSVS) method for screening active factors in SSDs. Real examples and simulation studies show that the TSVS method has a satisfactory performance compared with other existing methods although it cannot be guaranteed to be the best in any case.

For multi-level SSDs, the proposed TSVS method still works. The only difference is that each factor column should be first replaced by its contrasts. For example, for a three-level factor, the orthogonal polynomial contrast coefficient vectors are $(-1, 0, 1)'$ and $(1, -2, 1)'$. The remaining procedures for screening active effects are the same with that of two-level case.

In the procedure of TSVS, the choice of γ is an important issue. In this article, we just approximately set $\gamma = 1$ or $\gamma = 0.1 \max(|\hat{\beta}_i|)$. Other methods, such as cross-validation, may be helpful for the choice of γ, however, the time of computation may be greatly increased. This is an open problem for further study.

Acknowledgments

The authors are grateful to the associate editor and two referees for their insightful comments and constructive suggestions.

Funding

This work was supported by State Key Laboratory of CEMEE (CEMEE2015K0301A) and Project 613319. The authorship are listed in alphabetic order.

References

Beattie, S. D., Fong, D. K. H., Lin, D. K. J. (2002). A two-stage Bayesian model selection strategy for supersaturated designs. *Technometrics* 44:55–63.

Booth, K. H. V., Cox, D. R. (1962). Some systematic supersaturated designs. *Technometrics* 4:489–495.

Chen, R. B., Weng, J. Z., Chu, C. H. (2013). Screening procedure for supersaturated designs using a Bayesian variable selection method. *Quality and Reliability Engineering International* 29:89–101.

Chipman, H., Hamada, H., Wu, C. F. J. (1997). A Bayesian variable-selection approach for analyzing designed experiments with complex aliasing. *Technometrics* 39:372–381.

George, E. I., McMulloch, R. E. (1993). A statistical view of some chemometrics regression tools. *Technometrics* 35:109–148.

Georgiou, S. D. (2008). Modelling by supersaturated designs. *Computational Statistics & Data Analysis* 53:428–435.

Hamada, C. A., Hamada, M. S. (2010). All-subsets regression under effect heredity restrictions for experimental designs with complex aliasing. *Quality and Reliability Engineering International* 26:75–81.

Holcomb, D. R., Montgomery, D. C., Carlyle, W. M. (2003). Analysis of supersaturated designs. *Journal of Quality Technology* 35:13–27.

Li, P., Zhao, S. L., Zhang, R. C. (2010). A cluster analysis selection strategy for supersaturated designs. *Computational Statistics & Data Analysis* 54:1605–1612.

Li, R., Lin, D. K. J. (2003). Analysis methods for supersaturated design: Some comparisons. *Journal of Data Science* 1:249–260.
Lin, D. K. J. (1993). A new class of supersaturated designs. *Technometrics* 35:28–31.
Liu, M. Q., Fang, K. T., Hickernell, F. J. (2006). Connections among different criteria for asymmetrical fractional factorial designs. *Statistica Sinica* 16:1285–1297.
Liu, Y., Liu, M. Q. (2011). Construction of optimal supersaturated design with large number of levels. *Journal of Statistical Planning and Inference* 141:2035–2043.
Lu, X., Wu, X. (2004). A strategy of searching active factors in supersaturated screening experiments. *Journal of Quality Technology* 36:392–399.
Marley, C. J., Woods, D. C. (2010). A comparison of design and model selection methods for supersaturated experiments. *Computational Statistics & Data Analysis* 54:3158–3167.
Phoa, F. K. H. (2014). The stepwise response refinement screener (SRRS). *Statistica Sinica* 24:197–210.
Phoa, F. K. H., Pan, Y. H., Xu, H. (2009). Analysis of supersaturated designs via the Dantzig selector. *Journal of Statistical Planning and Inference* 139:2362–2372.
Rais, F., Kamoun, A., Chaabouni, M., Claeys-Bruno, M., Phan-Tan-Luu, R., Sergent, M. (2009). Supersaturated design for screening factors in sequencing the preparation of sulfated amides of olive pomace oil fatty acids. *Chemometrics and Intelligent Laboratory Systems* 99:71–78.
Satterthwaite, F. E. (1959). Random balance experimentation. *Technometrics* 1:111–137.
Sun, F. S., Lin, D. K. J., Liu, M. Q. (2011). On construction of optimal mixed-level supersaturated designs. *Annals of Statistics* 39:1310–1333.
Westfall, P. H., Young, S. S., Lin, D. K. J. (1998). Forward selection error control in the analysis of supersaturated designs. *Statistica Sinica* 8:101–117.
Wu, C. F. J. (1993). Construction of supersaturated designs through partially aliased interactions. *Biometrika* 80:661–669.
Wu, C. F. J., Hamada, M. (2009). *Experiments: Planning, Analysis, and Optimization*. 2nd ed. New York: Wiley.
Yin, Y. H., Zhang, Q. Z., Liu, M. Q. (2013). A two-stage variable selection strategy for supersaturated designs with multiple responses. *Frontiers of Mathematics in China* 8:717–730.
Zhang, Q. Z., Zhang, R. C., Liu, M. Q. (2007). A method for screening active effects in supersaturated designs. *Journal of Statistical Planning and Inference* 137:2068–2079.