Front. Math. China 2013, 8(3): 717–730
DOI 10.1007/s11464-012-0255-9

RESEARCH ARTICLE

A two-stage variable selection strategy for supersaturated designs with multiple responses

Yuhui YIN, Qiaozhen ZHANG, Min-Qian LIU

Department of Statistics, School of Mathematical Sciences and LPMC, Nankai University, Tianjin 300071, China

© Higher Education Press and Springer-Verlag Berlin Heidelberg 2012

Abstract A supersaturated design (SSD), whose run size is not enough for estimating all the main effects, is commonly used in screening experiments. It offers a potential useful tool to investigate a large number of factors with only a few experimental runs. The associated analysis methods have been proposed by many authors to identify active effects in situations where only one response is considered. However, there are often situations where two or more responses are observed simultaneously in one screening experiment, and the analysis of SSDs with multiple responses is thus needed. In this paper, we propose a two-stage variable selection strategy, called the multivariate partial least squares-stepwise regression (MPLS-SR) method, which uses the multivariate partial least squares regression in conjunction with the stepwise regression procedure to select true active effects in SSDs with multiple responses. Simulation studies show that the MPLS-SR method performs pretty good and is easy to understand and implement.

Keywords Multivariate partial least squares (MPLS), supersaturated design (SSD), stepwise regression, variable selection, variable importance in projection
MSC 62K15, 62J05

1 Introduction

In preliminary industrial screening experiments, it is typical to study a large number of potentially relevant factors simultaneously, but only a few are believed to be active among these interested factors (a phenomenon commonly recognized as effect sparsity). Generally, cost consideration and time constraint

Received September 13, 2012; accepted October 25, 2012
Corresponding author: Min-Qian LIU, E-mail: mqliu@nankai.edu.cn

make it impractical to do a traditional fractional factorial design. Faced with this reality, the supersaturated design (SSD), which investigates d factors of levels $s_1, \ldots, s_d$, respectively, by n experimental runs ($\sum_{i=1}^{d}(s_i - 1) > n - 1$), has attracted many authors' attention for its strong and powerful competition in run-size economy. The construction of SSDs dates back to Booth and Cox [3]. But in the following thirty years, SSDs have received little attention until Lin [16]. Since then the construction of SSDs has been widely explored. A comprehensive list of early works can be found in [17,18,22], where Sun et al. [22] also provided an extensive review of the existing methods for mixed-level SSDs. However, relative to the rapid development of the construction, the analysis of SSDs needs more investigation.

The analysis of SSDs aims at screening active effects from a large number of potential important ones correctly and economically. The complex correlation structure inherent in SSDs results in many traditional methods inapplicable. In recent years, some new analysis methods have been developed, Chipman et al. [6] provided a Bayesian variable selection approach for analyzing designed experiments with complex aliasing; Westfall et al. [24] proposed an error control skill in forward selection; Beattie et al. [2] put forward a two-stage Bayesian model selection strategy (SSVS/IBF); Li and Lin [14,15] employed penalized least squares with the smoothly clipped absolute deviation (SCAD) penalty to identify the sparse active effects; Holcomb et al. [11] introduced contrast-based methods; Lu and Wu [19] proposed a modified stepwise selection based on the idea of staged dimensionality reduction; Zhang et al. [26] proposed a method based on partial least squares (PLSVS); Phoa et al. [21] studied a variable selection method via the Dantzig selector (DS); Li et al. [13] introduced a contrast-orthogonality cluster analysis (COCA) method. All these methods concentrate on the situations where only one response is considered. However in practice, SSDs with multiple responses are often encountered. For example, we need to consider more than one response for a type of laundry detergent to study the decontamination ability for different types of stains. Meanwhile, because of the limit of time and money, an SSD becomes a suitable choice.

For SSDs with multiple responses, the forgoing methods for SSDs with a single response can be used directly to select active effects for one response at one time. However, the main disadvantage is that the information lying in the matrix of observations for responses will be neglected. In this paper, we propose an approach of the multivariate partial least squares (MPLS) in conjunction with the stepwise regression to reveal the importance of effects with regard to individual responses when the responses have varied correlations. Simulation studies and comparisons show that this procedure is effective.

MPLS regression, extensively used in applied sciences, is a method for building predictive models when there are many factors that are highly collinear. It bears some relations to the principal component analysis, canonical correlation analysis and multiple regression analysis, and is a particularly useful tool when the number of factors is large compared to the number of observations. Recent work focusing on partial least squares regression includes,

e.g., [1,5,8,10,12]. In this paper, we consider screening active effects in SSDs with multiple responses via the idea of two-stage dimensionality reduction. In the first stage, the MPLS method is used to find a set S_i including all active effects with a high probability for the ith response, and in the second stage, the stepwise regression method is used to find out the active effects from the selected set S_i with low Type I and Type II errors for the ith response. We call this method the MPLS-stepwise regression (MPLS-SR) method.

This paper is organized as follows. In Section 2, some preliminaries of the selection technique are introduced. Section 3 presents the use of the MPLS selection method in conjunction with the stepwise regression procedure for screening active effects. Section 4 shows the simulation and comparison results. Concluding remarks are provided in Section 5.

2 Preliminaries

Let $\mathrm{SSD}(n, s_1^{d_1} \cdots s_l^{d_l})$ denote an SSD of n runs with $d = d_1 + \cdots + d_l$ factors, among which there are d_i factors with s_i levels, say $\{0, 1, \ldots, s_i - 1\}$. Let $X = (x_1, \ldots, x_m)$ be the matrix of orthonormal main-effect contrasts of an $\mathrm{SSD}(n, s_1^{d_1} \cdots s_l^{d_l})$, where

$$m = \sum_{i=1}^{l} d_i(s_i - 1),$$

and $x_i = (x_{1i}, \ldots, x_{ni})^{\mathrm{T}}$ for $i = 1, \ldots, m$ are called the main-effect contrasts, or main effects for simplicity.

Suppose that for an $\mathrm{SSD}(n, s_1^{d_1} \cdots s_l^{d_l})$, there are q responses each with n observations, denoted by

$$y_i = (y_{1i}, \ldots, y_{ni})^{\mathrm{T}}, \quad i = 1, \ldots, q.$$

Let $Y = (y_1, \ldots, y_q)$ be the matrix of observations. For an $\mathrm{SSD}(n, s_1^{d_1} \cdots s_l^{d_l})$, the underlying model between the responses and effects is supposed to be

$$Y = 1_n \beta_0^{\mathrm{T}} + X(\beta_1, \ldots, \beta_q) + \varepsilon, \tag{1}$$

where 1_n denotes the $n \times 1$ vector with all elements unity, $\beta_0 = (\beta_{10}, \ldots, \beta_{q0})^{\mathrm{T}}$ is the vector of grand means, $\beta_i = (\beta_{i1}, \ldots, \beta_{im})^{\mathrm{T}}$ is the vector of unknown coefficients for y_i, $\varepsilon = (\varepsilon_{ij}) = (\varepsilon_1, \ldots, \varepsilon_q)$ is an $n \times q$ matrix of random errors consisting of n independent samples $(\varepsilon_{i1}, \ldots, \varepsilon_{iq}) \sim MN(\mathbf{0}, \Sigma)$ for $i = 1, \ldots, n$, and Σ is a positive semi-definite matrix with all the diagonal elements being one, i.e., the variances of all the responses are equal to one, here X is also called the model matrix.

Given matrices Y and X, we want to screen active effects for each response. In SSDs, because X is non-full-rank, the matrix $(X^{\mathrm{T}}X)^{-1}$ does not exist. Therefore, the analysis of SSDs is a challenging task. Generally, the following assumptions are needed for the analysis of SSDs [9].

(i) Effect sparsity: only a few factors are really active among the potential ones.

(ii) The coefficients of the active effects are large enough to be distinguished from the error.

(iii) The columns in X are not pairwise fully aliased.

Under these assumptions, we propose the MPLS variable selection combined with the stepwise regression procedure to appropriately screen the active effects with regard to individual responses.

Given an SSD$(n, s_1^{d_1} \cdots s_l^{d_l})$, let $E_0 = X$ be the model matrix and $F_0 = Y$ be the corresponding $n \times q$ matrix of observations. For simplicity, we suppose that $E_0 = X$ and $F_0 = Y$ are all column centered and normalized. The procedure of the MPLS is descried as follows [23].

Computing first MPLS component t_1 First, we find two linear functions

$$t_1 = E_0\omega_1, \quad u_1 = F_0c_1,$$

where the Euclidean norms of ω_1 and c_1 are both equal to one, i.e.,

$$\|\omega_1\| = \|c_1\| = 1.$$

Then we solve ω_1 and c_1 which maximize $\mathrm{cov}(t_1, u_1)$, i.e., the covariance between t_1 and u_1. After obtaining ω_1 and c_1, consider the linear regressions of E_0 on t_1 and F_0 on t_1, respectively, and we get

$$E_0 = t_1p_1^{\mathrm{T}} + E_1, \quad F_0 = t_1q_1^{\mathrm{T}} + F_1,$$

where E_1 and F_1 are the vectors of residuals, respectively. The variable t_1 is called the first MPLS component of the regression.

Computing second MPLS component t_2 Find a second pair of linear functions

$$t_2 = E_1\omega_2, \quad u_2 = F_1c_2$$

such that $\mathrm{cov}(t_2, u_2)$ is maximized, where

$$\|\omega_2\| = \|c_2\| = 1.$$

We then consider the regressions of E_1 on t_2 and F_1 on t_2 to get the new vectors of residuals E_2 and F_2, respectively,

$$E_1 = t_2p_2^{\mathrm{T}} + E_2, \quad F_1 = t_2q_2^{\mathrm{T}} + F_2.$$

t_2 is called the second MPLS component.

Computation of next MPLS components and stopping rule Continue the same procedure for computing the next components

$$t_h = E_{h-1}\omega_h, \quad h \geqslant 3.$$

We can figure out all the MPLS components. However, here the point is how to find the right number of MPLS components which are powerful in improving the prediction of the MPLS regression model. In this paper, a proper h is decided following a cross-validation procedure.

Variable importance in projection (VIP) Given the h MPLS components $T = (t_1, t_2, \ldots, t_h)$, we can easily obtain that

$$t_k = E_0 \omega_k^*, \tag{2}$$

where

$$\omega_k^* = (\omega_{k1}^*, \ldots, \omega_{km}^*)^{\mathrm{T}} = \prod_{i=1}^{k-1} (I - \omega_i p_i^{\mathrm{T}}) \omega_k.$$

Then the regression of F_0 on E_0 can be shown to be

$$\hat{F}_0 = T\hat{B},$$

where

$$\hat{B} = (T^{\mathrm{T}}T)^{-1} T^{\mathrm{T}} F_0$$

is the ordinary least squares estimation.

From (2), the kth MPLS component t_k is a function of X. Therefore, if t_k contributes to Y very much, and x_j is important when building up t_k, then it is reasonable to believe that x_j will be important to Y. The score of VIP can describe this idea. The VIP score of the jth effect is defined to be

$$\mathrm{VIP}_j = \left(\frac{m}{\mathrm{Rd}\{Y; t_1, t_2, \ldots, t_h\}} \sum_{k=1}^{h} \mathrm{Rd}(Y; t_k) (\omega_{kj}^*)^2 \right)^{1/2},$$

where

$$\mathrm{Rd}(Y; t_1, t_2, \ldots, t_h) = \sum_{i=1}^{h} \mathrm{Rd}(Y; t_i), \quad \mathrm{Rd}(Y; t_i) = \sum_{j=1}^{q} \frac{\mathrm{Rd}(y_j; t_k)}{q},$$
$$\mathrm{Rd}(y_j; t_k) = r^2(y_j; t_k),$$

and r represents the correlation coefficient between y_j and t_k. A larger VIP score generally means the corresponding effect is more important to the responses. In this paper, we will use VIP scores to select a set of active effects.

Stepwise regression With the availability of statistical packages, the stepwise regression is now a most commonly used method for building models. However, during the stepwise regression process, how to choose the p-value for entering or removing a variable is arbitrary. In fact, the choice of p-values is important to reveal the true relationship between the responses and the variables. Phoa et al. [21] suggested a model selection criterion via a modified version of the Akaike information criterion (AIC), called mAIC,

$$\mathrm{mAIC} = n \log \frac{\mathrm{RSS}}{n} + 2t^2,$$

where RSS is the residual sum of squares, n is the number of runs, and t is the number of parameters in the model. The mAIC imposes heavier penalty on the model complexity than the AIC. Therefore, the mAIC is more rational in the analysis of SSDs because of their factor sparsity assumption. In this paper, we use a procedure which combines the mAIC and the stepwise regression for model selection in the analysis of SSDs.

3 MPLS-SR variable selection procedure

Given an SSD($n, s_1^{d_1} \cdots s_l^{d_l}$), consider the linear model (1). Without loss of generality, Y and X are assumed to be column centered and normalized. Then model (1) becomes

$$Y = X(\beta_1, \ldots, \beta_q) + \varepsilon.$$

We now propose the following steps to simultaneously screen the active effects for each response from the m potential important effects $x_1, \ldots, x_m$.

Proposed variable selection strategy The two-stage analysis strategy, MPLS-SR, is carried out as follows. Denote the set of all effects by

$$S = \{x_1, \ldots, x_m\}.$$

The first stage aims at eliminating the inactive effects as many as possible, and getting the sets of candidate active effects, denoted by S_i for each y_i, $i = 1, \ldots, q$. Therefore, S_i should include all the elements of the true active effects set for y_i (denoted by TA_i) with high probability and be substantially smaller than S. In the second stage, S_i is further screened, the stepwise regression procedure is used to identify the active effects for y_i from S_i, and the set consisting of the identified active effects is denoted by IA_i.

Selection of S_i using MPLS First, obtain the h MPLS components $T = \{t_1, \ldots, t_h\}$, build a regression equation $\hat{Y} = T\hat{B}$, and then, according to (2), obtain a regression equation between Y and X, denoted by $\hat{Y} = X\hat{\beta}$. For each y_i, choose the effects with the largest $\lfloor n/2 \rfloor - 1$ estimated coefficients in $\hat{\beta}_i = (\hat{\beta}_{i1}, \ldots, \hat{\beta}_{im})^{\mathrm{T}}$ to form a set of the potential active effects, denoted by S_i^{M}. Second, calculate the VIP values of all the m effects and sort them in descending order:

$$\mathrm{VIP}_{(1)} \geqslant \mathrm{VIP}_{(2)} \geqslant \cdots \geqslant \mathrm{VIP}_{(m)},$$

and the effects with the largest $\lfloor n/2 \rfloor - 1$ VIP values are chosen to form a set S_{VIP}. Third, define

$$S_i = S_i^{\mathrm{M}} \cup S_{\mathrm{VIP}}.$$

Note that the number of the elements in S_i is usually far smaller than $n - 2$, though the largest possible one is $n - 2$. Then we efficiently narrow the scope of candidate active effects.

Selection of IA$_i$ using stepwise regression A stepwise regression usually tends to not only select too many inactive factors but also miss some active ones, as discussed in [19]. However, after we have efficiently narrowed the set of potentially active effects in the first stage, the stepwise procedure will then give a much better result. The true active effects identified rate can reach a high level. In the stepwise procedure, the choice of p-to-remove and p-to-enter affects the Type I and Type II errors. Here, the stepwise regression method is combined with the mAIC to choose proper p-values automatically. We choose p-values (p-to-remove = p-to-enter) from interval $[0.01, 0.1]$ with a step 0.01, and run a stepwise regression between y_i and the effects in S_i, then the mAIC is used to pick out a best suitable model. The effects included in the chosen model are identified as the elements of IA$_i$.

A step-by-step procedure for MPLS-SR A step-by-step guideline for the proposed procedure can be summarized as follows.

Step 1 Let

$$F_0 = Y = (y_1, \ldots, y_q), \quad E_0 = X = (x_1, \ldots, x_m),$$

which are supposed to be column centered and normalized.

Step 2 Let r be the rank of X, obtain the first r MPLS components $t_1, \ldots, t_r$, and build regressions of Y on $t_1, \ldots, t_j$ for $1 \leqslant j \leqslant r$. From these r models, find the model with the best prediction performance using the cross-validation method. Suppose that there are h MPLS components $t_1, \ldots, t_h$ in this selected model. Then transform the equation between $\hat{Y}$ and $t_1, \ldots, t_h$ into a equation between $\hat{Y}$ and X, and choose the effects with the largest $\lfloor n/2 \rfloor - 1$ estimated coefficients in $\hat{\beta}_i$ to form a set S_i^{M} for y_i, $i = 1, \ldots, q$.

Step 3 Use the h MPLS components $t_1, \ldots, t_h$ chosen in Step 2 to compute the VIP scores for the m effects, denoted by VIP$_j$, $j = 1, \ldots, m$. Sort the VIP scores in descending order:

$$\mathrm{VIP}_{(1)} \geqslant \mathrm{VIP}_{(2)} \geqslant \cdots \geqslant \mathrm{VIP}_{(m)},$$

and choose the x_j's with

$$\mathrm{VIP}_j \geqslant \mathrm{VIP}_{(\lfloor n/2 \rfloor - 1)}$$

to form a set S_{VIP}.

Step 4 For $i = 1, \ldots, q$, let

$$S_i = S_i^{\mathrm{M}} \cup S_{\mathrm{VIP}}.$$

Then run stepwise regressions between y_i and the effects in S_i using ten p-values (p-to-enter = p-to-remove). Choose the one which has the smallest mAIC among the above ten regression models as the final suitable model. The effects included in the chosen model are identified as active effects, which form a set IA$_i$.

Step 5 Output the set IA$_i$ for $i = 1, \ldots, q$.

4 Simulations and comparisons

In this section, some simulations will be firstly carried out to display the performance of the MPLS-SR method, and then we will focus on the comparisons between the proposed MPLS-SR method and the other methods for SSDs with a single response.

4.1 Simulation studies

For simplicity, we consider SSDs with three responses, y_1, y_2, and y_3. In each simulation, the covariance, Σ, of $(\varepsilon_{i1}, \varepsilon_{i2}, \varepsilon_{i3})$ is a randomly generated semi-definite matrix with all the diagonal elements being one. We simulate 5 types of models, Type j denotes models in each of which any of the three responses has the number of the true active effects no more than j. Note that Box and Meyer [4] defined effect sparsity as 20% or fewer of the effects being active, and Marley and Woods [20] defined effect sparsity as the number of active effects being at most a third of the sample size. Here, for an SSD with an $n \times m$ orthonormal main-effect contrast matrix X, the maximum of j is taken to be the integer part of $\min\{n/3, m/5\}$. The number of true active effects and the true active effects for each response are randomly decided under the constraint of each type, i.e., under Type j, the number of true active effects is randomly determined from the set $\{1, \ldots, j\}$ for each response and then the true active effects are randomly arranged on the potential interested effects.

We demonstrate the performance of the proposed method by the two-level SSD(14, 2^{23}) shown in Table 1 (cf. [16]) and the mixed-level SSD($18, 2^1 3^{12}$) in Table 2 (cf. [7]), where the rows represent factors and the columns represent runs.

In each simulation, suppose that there are N_i active effects for y_i, set their model coefficients, i.e., the corresponding N_i components of β_i, to be $\beta_i^{(1)}, \ldots, \beta_i^{(N_i)}$, respectively, and set the other components of β_i to be zero. Here, we consider the following four cases of relative magnitude of coefficients $\beta_i^{(1)}, \ldots, \beta_i^{(N_i)}$ for $i = 1, 2, 3$.

Case 1 The coefficients of active effects are 4, and the signs are randomly assigned.

Case 2 The minimal coefficient of active effect is 3, and the coefficients of other active effects ascend in constant intervals of 1 from low to high. For example, the coefficients of three active effects are $3, 4$, and 5, respectively.

Case 3 The minimal coefficient of active effect is 3, and the coefficients of other active effects ascend in constant intervals of 3 from low to high. For example, the coefficients of four active effects are $3, 6, 9$, and 12, respectively.

Case 4 The coefficients of active effects are randomly drawn from the uniform distribution $U(2, 10)$.

Simulation results based on 1000 replicates are summarized in Table 3 for the SSD($14, 2^{23}$) and in Table 4 for the SSD($18, 2^1 3^{12}$). In these tables, 'Case' refers to the relative magnitude of coefficients, 'Type' denotes the maximum

Table 1 SSD(14, 2^{23}): half-fraction of Williams [25] data*

factor	run													
	1	2	3	4	5	6	7	8	9	10	11	12	13	14
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
3	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
5	−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
7	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
9	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
11	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
13	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
15	−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
18	−1	1	−1	1	−1	1	1	−1	1	−1	1	−1	1	−1
19	−1	−1	−1	1	−1	1	1	1	−1	1	1	1	−1	−1
20	1	−1	−1	1	1	1	1	−1	−1	−1	−1	−1	1	1
21	−1	1	−1	−1	−1	1	1	1	−1	1	−1	1	1	−1
22	−1	1	1	−1	1	1	−1	1	−1	−1	−1	1	−1	1
23	−1	−1	1	−1	1	−1	−1	1	1	−1	1	1	1	−1
24	1	−1	−1	−1	1	−1	1	−1	1	1	1	1	−1	−1

* Note that 16th factor is deleted since it is fully aliased with 13th factor

Table 2 SSD(18, $2^1 3^{12}$): two-third of an $OA(27, 3^{13}, 2)$

factor	run																	
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0
2	0	0	0	1	1	1	2	2	2	0	0	0	1	1	1	2	2	2
3	0	1	2	0	1	2	0	1	2	0	1	2	0	1	2	0	1	2
4	1	0	2	2	1	0	0	2	1	2	1	0	0	2	1	1	0	2
5	0	1	2	1	2	0	2	0	1	0	1	2	1	2	0	2	0	1
6	0	2	1	1	0	2	2	1	0	0	2	1	1	0	2	2	1	0
7	1	0	2	2	0	1	2	1	0	2	1	0	1	0	2	0	2	1
8	2	0	1	2	0	1	2	0	1	1	2	0	1	2	0	1	2	0
9	1	1	1	2	2	2	0	0	0	2	2	2	0	0	0	1	1	1
10	2	1	0	1	0	2	0	2	1	1	0	2	0	2	1	2	1	0
11	1	1	1	0	0	0	2	2	2	2	2	2	1	1	1	0	0	0
12	2	1	0	2	1	0	2	1	0	1	0	2	1	0	2	1	0	2
13	1	2	0	0	1	2	2	0	1	2	0	1	1	2	0	0	1	2

number of true active effects for the responses, 'AEIR1$_i$' stands for the all active effects identified rate for y_i in the first stage, 'TMIR$_i$' stands for the true model identified rate for y_i, and 'AEIR2$_i$' shows the all active effects identified rate for y_i in the second stage.

From Table 3, we have the following observations.

Table 3 Simulation results of SSD(14,2^{23})

Case	Type	AEIR1$_1$	AEIR1$_2$	AEIR1$_3$	TMIR$_1$	TMIR$_2$	TMIR$_3$	AEIR2$_1$	AEIR2$_2$	AEIR2$_3$
1	1	1	1	1	0.789	0.781	0.779	1	1	1
	2	1	1	1	0.797	0.801	0.79	0.991	0.994	0.99
	3	0.996	0.997	0.998	0.732	0.727	0.721	0.95	0.955	0.955
	4	0.979	0.981	0.99	0.682	0.673	0.659	0.862	0.867	0.856
2	1	1	1	1	0.745	0.771	0.788	1	1	1
	2	0.998	1	0.999	0.813	0.825	0.812	0.997	1	0.996
	3	0.976	0.995	0.991	0.796	0.818	0.825	0.965	0.986	0.979
	4	0.937	0.95	0.94	0.735	0.728	0.741	0.895	0.909	0.91
3	1	1	1	1	0.776	0.79	0.788	1	1	1
	2	0.989	0.985	0.986	0.821	0.801	0.817	0.989	0.983	0.984
	3	0.942	0.965	0.959	0.779	0.774	0.793	0.934	0.957	0.947
	4	0.917	0.908	0.891	0.718	0.734	0.696	0.885	0.895	0.865
4	1	1	1	1	0.799	0.763	0.781	1	1	1
	2	0.982	0.989	0.982	0.811	0.81	0.783	0.982	0.989	0.982
	3	0.925	0.937	0.927	0.794	0.784	0.768	0.925	0.937	0.927
	4	0.864	0.87	0.87	0.719	0.727	0.743	0.862	0.869	0.867

Table 4 Simulation results of SSD(18,$2^1 3^{12}$)

Case	Type	AEIR1$_1$	AEIR1$_2$	AEIR1$_3$	TMIR$_1$	TMIR$_2$	TMIR$_3$	AEIR2$_1$	AEIR2$_2$	AEIR2$_3$
1	1	1	1	1	0.529	0.511	0.536	0.92	0.916	0.918
	2	1	1	1	0.618	0.603	0.623	0.862	0.852	0.868
	3	0.999	0.998	0.999	0.624	0.628	0.614	0.798	0.802	0.792
	4	0.992	0.986	0.989	0.604	0.565	0.583	0.746	0.739	0.729
	5	0.965	0.969	0.975	0.521	0.514	0.51	0.658	0.669	0.669
2	1	1	1	1	0.537	0.548	0.551	0.912	0.929	0.932
	2	1	1	0.999	0.634	0.635	0.634	0.863	0.856	0.862
	3	0.997	0.99	0.99	0.646	0.635	0.573	0.829	0.819	0.787
	4	0.973	0.969	0.96	0.611	0.589	0.618	0.76	0.736	0.758
	5	0.909	0.91	0.918	0.551	0.548	0.543	0.665	0.657	0.684
3	1	1	1	1	0.553	0.557	0.538	0.907	0.908	0.923
	2	0.989	0.992	0.989	0.626	0.614	0.626	0.874	0.875	0.887
	3	0.969	0.978	0.973	0.644	0.657	0.638	0.805	0.836	0.809
	4	0.896	0.914	0.904	0.58	0.586	0.581	0.728	0.735	0.717
	5	0.79	0.804	0.808	0.521	0.509	0.495	0.64	0.622	0.611
4	1	1	1	1	0.54	0.545	0.571	0.918	0.908	0.93
	2	0.996	0.996	0.995	0.631	0.642	0.662	0.875	0.864	0.879
	3	0.985	0.979	0.984	0.635	0.66	0.628	0.808	0.815	0.793
	4	0.95	0.938	0.943	0.604	0.584	0.6	0.748	0.73	0.746
	5	0.882	0.875	0.882	0.544	0.543	0.563	0.66	0.662	0.671

(i) If only a single effect is active for the responses, i.e., under Type 1, the performance of the MPLS-SR method is perfect and not sensitive to the choices of the magnitude of the coefficients, and the all active effects identified rates (AEIR1s and AEIR2s) are 100%, meanwhile, the true model identified rates (TMIRs) are all higher than 74.5%.

(ii) When the maximum number of the true active effects is no more than 3, i.e., under Types 1, 2, and 3, the performance of the MPLS-SR is still perfect,

the AEIR1s and AEIR2s are higher than 92%, and the TMIRs are higher than 72%. When the maximum number of the true active effects is 4, i.e., under Type 4, the performance becomes a little worse, some AEIR1s and most of the AEIR2s are lower than 90%, and some TMIRs are lower than 70%.

(iii) In each case, the TMIRs, AEIR1s, and AEIR2s decrease with the number of the true active effects increasing. For different cases, the MPLS-SR performs similarly in AEIR2s and TMIRs, but a little different in AEIR1s, for example, in Case 1, the AEIR1s are all higher than 97%, while in Case 4, some AEIR1s are lower than 88%.

Therefore, we conclude that in the analysis of two-level SSDs with multiple responses, the number of active effects and the magnitude of their coefficients have impacts on the performance of the MPLS-SR method.

From Table 4, we can have the similar observations on the performance of the MPLS-SR method when applied to the mixed-level SSD as those obtained from Table 3. Compared with Table 3, the MPLS-SR method performs stably in AEIR1s, but a little worse in TMIRs and AEIRs. The reason for this may be that the inherent structures of the orthonormal main-effect contrast matrix for mixed-level SSDs are more complicated.

Remark 1 In the simulations, the covariance, Σ, of $(\varepsilon_1, \varepsilon_2, \varepsilon_3)$ is a randomly generated semi-definite matrix with all the diagonal elements being one. How do the correlations among the responses y_1, y_2, and y_3 affect the performance of the MPLS-SR method? In the first stage, i.e., the MPLS stage, we get the set of candidate active effects S_i for y_i by combining the sets S_i^{M} and S_{VIP}. For S_i^{M}, according to the multivariate regression analysis, we know that the correlations among the responses have no effect on estimating the coefficients, and thus, have no effect on the generation of S_i^{M}, as for S_{VIP}, we also have the finding that the correlations among the responses have no effect on the selection of S_{VIP} by some simulations. Therefore, the MPLS-SR method is stable under the random generation of Σ here.

4.2 Comparions

In order to compare the MPLS-SR method with the analysis methods for SSDs with only one response, we now consider the SSD(14, 2^{23}) in Table 1. Suppose that the true model is

$$\begin{cases} y_{i1} = 10x_{i1} + \varepsilon_{i1}, \\ y_{i2} = -15x_{i1} + 8x_{i5} - 2x_{i9} + \varepsilon_{i2}, \\ y_{i3} = -15x_{i1} + 12x_{i5} - 8x_{i9} + 6x_{i,13} - 2x_{i,17} + \varepsilon_{i3}, \end{cases} \quad i = 1, \ldots, 14, \tag{3}$$

where

$$(\varepsilon_{i1}, \varepsilon_{i2}, \varepsilon_{i3}) \sim MN(\mathbf{0}, I)$$

is the vector of random errors.

We generate

$$Y = (y_{ij}) = (y_1, \ldots, y_3)$$

from the above linear model, and run the simulations 1000 times. Note that the three single models in (3) are commonly used in displaying the performance of the analysis methods for the SSD(14, 2^{23}) with only one response y_1, y_2, or y_3. Here, the proposed method, MPLS-SR, can simultaneously screen active effects for all the three models in (3). Table 5 compares the MPLS-SR with other five methods which are used for SSDs with one response. Here, 'TMIR' stands for the true model identified rate, 'SEIR' stands for the smallest effect identified rate, and 'median' and 'mean' are the median and mean sizes of the models.

Table 5 Comparison results

Case	method	TMIR (%)	SEIR (%)	median	mean
I: y_1	SSVS$(1/10, 500)$	40.5	99.0	2	3.1
	SSVS$(1/10, 500)$/IBF	61.0	98.0	1	2.5
	SCAD	75.6	100	1	1.7
	PLSVS$(m = 1)$	61.0	100	1	1.5
	DS$(\gamma = 1)$	99.4	100	1	1.0
	MPLS-SR	78.5	100	1	1.2
II: y_2	SSVS$(1/10, 500)$	8.6	30.0	3	4.7
	SSVS$(1/10, 500)$/IBF	8.0	28.0	3	4.2
	SCAD	75.6	98.5	3	3.3
	PLSVS$(m = 1)$	76.4	100	3	3.3
	DS$(\gamma = 1)$	84.4	85.3	3	3.0
	MPLS-SR	88.6	99.2	3	3.1
III: y_3	SSVS$(1/10, 500)$	36.4	84.0	6	8.0
	SSVS$(1/10, 500)$/IBF	40.7	75.0	5	5.6
	SCAD	69.7	99.4	5	5.4
	PLSVS$(m = 1)$	73.6	95.0	5	5.2
	DS$(\gamma = 1)$	79.1	91.2	5	5.1
	MPLS-SR	82.5	97.6	5	5.1

From Table 5, we can see that the MPLS-SR identifies the true models with the highest probabilities in Cases II and III. In Case I, the MPLS-SR method shares 100% perfect identification rate with the SCAD, PLSVS, and DS methods in identifying the smallest effect. In Cases II and III, the performance of the MPLS-SR method shows its stability and powerfulness in identifying the smallest effect. In fact, the MPLS-SR method includes the smallest active effect with a high probability ($> 97.5\%$) in any of these three cases, that is to say, type II errors are all controlled within the level 0.025.

Remark 2 Compared with the five methods, we can see that the MPLS-SR method not only has the ability of dealing with SSDs with multiple responses, but also performs as good as the methods which consider only one response at a time. For the MPLS-SR method, because the information lying in the matrix of observations is considered in the MPLS procedure, the inactive effects are eliminated as many as possible and the set of candidate active effects is effectively narrowed for the stepwise regression procedure, and thus, the proposed method performs powerful as illustrated in Table 5.

5 Concluding remarks

SSDs are very useful in screening experiments because of their economy in run sizes. There are many methods for the analysis of SSDs with only one response. However, there are often screening experiments in which two or more responses are observed simultaneously. In this paper, we propose a two-stage strategy, called the MPLS-SR method, to solve the variable selection problem in SSDs which have multiple responses with varied correlations.

The new method uses the MPLS and stepwise regression procedures to screen active effects for each response. In the first stage, the MPLS regression method is used to eliminate the inactive effects as many as possible, and only less than $n - 2$ effects are left to form a set of candidate active effects for each response for the stepwise regression in the second stage. Obviously, the performance of the MPLS-SR method in the first stage has a direct impact on its performance in the second stage. Our simulations show that the MPLS-SR method performs stably in the first stage in the analysis of the two-level and mixed-level SSDs. In the second stage, the stepwise regression is combined with the mAIC to improve the performance of the MPLS-SR method.

For SSDs with multiple responses, the proposed MPLS-SR method can screen active effects for each response simultaneously. Compared with the analysis methods for SSDs with one response, the information lying in the matrix of observations for the responses is considered in the MPLS stage. The comparison results in Table 5 provide a rationality for the MPLS-SR method.

Acknowledgements This work was supported by the National Natural Science Foundation of China (Grant Nos. 10971107, 11271205), the "131" Talents Program of Tianjin, and the Fundamental Research Funds for the Central Universities (Grant Nos. 65030011, 65011481).

References

1. Bastien P, Vinzi V E, Tenenhaus M. PLS generalized linear regression. Comput Statist Data Anal, 2005, 48(1): 17–46
2. Beattie S D, Fong D K H, Lin D K J. A two-stage Bayesian model selection strategy for supersaturated designs. Technometrics, 2002, 44(1): 55–63
3. Booth K H V, Cox D R. Some systematic supersaturated designs. Technometrics, 1962, 4(4): 489–495
4. Box G, Meyer R. An analysis for unreplicated fractional factorials. Technometrics, 1986, 28(1): 11–18
5. Butler N A, Denham M C. The peculiar shrinkage properties of partial least squares regression. J Roy Statist Soc Ser B, 2000, 62(3): 585–593
6. Chipman H, Hamada H, Wu C F J. A Bayesian variable-selection approach for analyzing designed experiments with complex aliasing. Technometrics, 1997, 39(4): 372–381
7. Fang K T, Lin D K J, Liu M Q. Optimal mixed-level supersaturated design. Metrika, 2003, 58(3): 279–291

8. Garthwaite P H. An interpretation of partial least squares. J Amer Statist Assoc, 1994, 89(425): 122–127
9. Georgiou S D. Modelling by supersaturated designs. Comput Statist Data Anal, 2008, 53(2): 428–435
10. Helland I S. Partial least squares regression and statistical models. Scand J Statist, 1990, 17(2): 97–114
11. Holcomb D R, Montgomery D C, Carlyle W M. An alysis of supersaturated designs. J Quality Technol, 2003, 35(1): 13–27
12. Höskuldsson A. Variable and subset selection in PLS regression. Chemometrics & Intelligent Laboratory Systems, 2001, 55: 23–38
13. Li P, Zhao S L, Zhang R C. A cluster analysis selection strategy for supersaturated designs. Comput Statist Data Anal, 2010, 54: 1605–1612
14. Li R, Lin D K J. Data analysis of supersaturated designs. Statist Probab Lett, 2002, 59(2): 135–144
15. Li R, Lin D K J. Analysis methods for supersaturated design: some comparisons. J Data Sci, 2003, 1(3): 249–260
16. Lin D K J. A new class of supersaturated designs. Technometrics, 1993, 35(1): 28–31
17. Liu Y, Liu M Q. Construction of optimal supersaturated design with large number of levels. J Statist Plann Inference, 2011, 141: 2035–2043
18. Liu Y, Liu M Q. Construction of equidistant and weak equidistant supersaturated designs. Metrika, 2012, 75(1): 33–53
19. Lu X, Wu X. A strategy of searching active factors in supersaturated screening experiments. J Quality Technol, 2004, 36(4): 392–399
20. Marley C J, Woods D C. A comparison of design and model selection methods for supersaturated experiments. Comput Statist Data Anal, 2010, 54(12): 3158–3167
21. Phoa F K H, Pan Y H, Xu H. Analysis of supersaturated designs via the Dantzig selector. J Statist Plann Inference, 2009, 139: 2362–2372
22. Sun F S, Lin D K J, Liu M Q. On construction of optimal mixed-level supersaturated designs. Ann Statist, 2011, 39(2): 1310–1333
23. Wang H W. Partial Least-squares Regression Method and Applications. Beijing: National Defence Industry Press, 1999 (in Chinese)
24. Westfall P H, Young S S, Lin D K J. Forward selection error control in the analysis of supersaturated design. Statist Sinica, 1998, 8(1): 101–117
25. Williams K R. Designed experiments. Rubber Age, 1968, 100: 65–71
26. Zhang Q Z, Zhang R C, Liu M Q. A method for screening active effects in supersaturated designs. J Statist Plann Inference, 2007, 137(6): 2068–2079