Available online at www.sciencedirect.com
ScienceDirect
Journal of Statistical Planning and Inference 137 (2006) 2068–2079
journal of statistical planning and inference
www.elsevier.com/locate/jspi

A method for screening active effects in supersaturated designs

Qiao-Zhen Zhang, Run-Chu Zhang, Min-Qian Liu*

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

Received 2 November 2005; accepted 27 June 2006
Available online 22 August 2006

Abstract

A supersaturated design (SSD) is a design whose run size is not enough for estimating all the main effects. The goal in conducting such a design is to identify, presumably only a few, relatively dominant active effects with a cost as low as possible. However, data analysis of such designs remains primitive: traditional approaches are not appropriate in such a situation and several methods which were proposed in the literature in recent years are effective when used to analyze two-level SSDs. In this paper, we introduce a variable selection procedure, called the PLSVS method, to screen active effects in mixed-level SSDs based on the variable importance in projection which is an important concept in the partial least-squares regression. Simulation studies show that this procedure is effective.

MSC: Primary, 62K15;; secondary 62J05

Keywords: Mixed-level; Partial least-squares; Supersaturated design; Variable selection; Variable importance in projection

1. Introduction

Many preliminary industrial screening experiments typically contain a large number of potentially relevant factors. Among them, only a few are believed to be active. The supersaturated design (SSD), first constructed systematically by Booth and Cox (1962), has received a great deal of attention since the appearance of Lin (1993). Most studies have focused on two-level and multi-level SSDs. Extensions to mixed-level SSDs include Yamada and Lin (2002), Yamada and Matsui (2002), Fang et al. (2003), Fang et al. (2004), Li et al. (2004), Yamada et al. (2006) and Liu et al. (2006). With the construction of SSDs having been widely explored, the inferential aspect of such designs needs more investigation. Of course, the analysis is very challenging since even a main-effect model is not identifiable.

To find the sparse active effects, variable selection becomes fundamental in the analysis stage of such screening experiments. Some new analysis methods were developed in recent years, of course, all these studies restricted their discussion at two-level SSDs. 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); Li and Lin (2002, 2003) employed penalized least squares with the smoothly clipped absolute deviation (SCAD) penalty to identify the sparse active effects; Holcomb et al. (2003) proposed contrast-based methods; Lu and Wu (2004) proposed a modified stepwise selection based on the idea of staged dimensionality reduction. Yamada (2004) examined type II error (declaring an

* Corresponding author. Tel.: +86 22 23504709; fax: +86 22 23506423.
E-mail address: mqliu@nankai.edu.cn (M.-Q. Liu).

0378-3758/$ - see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.jspi.2006.06.038

active effect to be inactive) in stepwise selection and discussed some guidelines for data analysis. Simulation studies demonstrated that the SCAD method of Li and Lin (2002, 2003) outperforms the other approaches. However, the SCAD method requires a good initial value which is close to the true value, if the initial value given by the stepwise selection is not very close to the true one (that we do not know in fact), perhaps the procedure will not give a satisfactory result. With the construction of multi-level and mixed-level SSDs being discussed so often, the aspect of data analysis needs investigation. However, until now, this problem has not been studied in adequate detail. In this paper, we introduce an approach via partial least-squares (PLS) regression which can be used to screen active effects in mixed-level SSDs.

PLS regression is a technique that generalizes and combines features from principal component analysis, canonical correlation analysis and multiple regression analysis. It is particularly useful when we need to predict a set of response variables from a (very) large set of explanatory variables. By taking advantage from the statistical tests associated with linear regression, it is feasible to select the significant explanatory variables to include in PLS regression and to choose the number of PLS components to retain. Recent work focusing on this topic includes, e.g. Helland (1990), Garthwaite (1994), Butler and Denham (2000), Höskuldsson (2001) and Bastien et al. (2005). This paper proposes a method, called the PLS variable selection (PLSVS) method, for searching active effects in SSDs based on the variable importance in projection (VIP). Simulation studies demonstrate that the method is effective when used to analyze data collected from SSDs with mixed-level, multi-level or two-level factors; in addition, the simulation shows that the PLSVS procedure outperforms the SSVS/IBF approach and can be comparable with the SCAD method when used to screen active effects in two-level SSDs.

The paper is organized as follows. In Section 2, the PLS regression technique is introduced. Section 3 presents the PLSVS procedure for screening active effects. The simulation studies are reported in Section 4. The last section contains some concluding remarks.

2. Background of PLS regression

Let $\mathbf{y0}, \mathbf{x0}_1, \ldots, \mathbf{x0}_k$ be the raw variables and their column centered and normalized patterns are denoted by $\mathbf{y}, \mathbf{x}_1, \ldots, \mathbf{x}_k$.

PLS univariate regression is a model linking a response variable $\mathbf{y}$ to a set $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_k)$ of (numerical or categorical) explanatory variables. It can be obtained as a series of simple and multiple regressions.

The PLS regression model with m components is written as

$$\mathbf{y} = \sum_{h=1}^{m} c_h \left(\sum_{j=1}^{k} w_{hj} \mathbf{x}_j \right) + \text{residual}, \tag{1}$$

with the constraint that the m PLS components $\mathbf{t}_h$'s are orthogonal where

$$\mathbf{t}_h = \sum_{j=1}^{k} w_{hj} \mathbf{x}_j = \mathbf{X}\mathbf{w}_h, \quad \text{for } h = 1, \ldots, m,$$

and $\mathbf{w}_h = (w_{h1}, \ldots, w_{hk})'$. PLS regression is an algorithm for estimating the parameters of model (1). For the detailed introduction, please refer to Bastien et al. (2005) and the references therein. Now let us introduce this algorithm briefly.

Computation of the first PLS component $\mathbf{t}_1$. The component $\mathbf{t}_1$ should bear the information of explanatory variables $\mathbf{X}$ as much as possible and the correlation coefficient $\text{corr}(\mathbf{y}, \mathbf{t}_1)$ is maximal. This means the first goal is to maximize

$$\text{cov}(\mathbf{y}, \mathbf{t}_1) = \text{s}(\mathbf{t}_1) * \text{corr}(\mathbf{y}, \mathbf{t}_1),$$

with the constraints of $\mathbf{t}_1 = \mathbf{X}\mathbf{w}_1$ and $\mathbf{w}_1'\mathbf{w}_1 = 1$.

The optimal solution $\mathbf{w}_1$ is the standard eigenvector of $\mathbf{X}'\mathbf{y}\mathbf{y}'\mathbf{X}$ corresponding to the largest eigenvalue, and then:

$$\mathbf{t}_1 = \frac{1}{\sqrt{\sum_{j=1}^{k} \text{cov}(\mathbf{y}, \mathbf{x}_j)^2}} \sum_{j=1}^{k} \text{cov}(\mathbf{y}, \mathbf{x}_j)\mathbf{x}_j.$$

The weight for $\mathbf{x}_j$ can be written as $\text{corr}(\mathbf{y}, \mathbf{x}_j)$ since $\mathbf{y}$ and $\mathbf{x}_j$ are respectively standardized. So in order for a variable $\mathbf{x}_j$ to be important in building up $\mathbf{t}_1$, it needs to be strongly correlated with $\mathbf{y}$.

Computation of the second PLS component $\mathbf{t}_2$. Firstly, the $k+1$ simple regressions of, respectively, $\mathbf{y}$ and each $\mathbf{x}_j$ on $\mathbf{t}_1$ are run:

$$\mathbf{y} = c_1\mathbf{t}_1 + \mathbf{y}_1,$$
$$\mathbf{x}_j = p_{1j}\mathbf{t}_1 + \mathbf{x}_{1j}, \quad j = 1, \ldots, k.$$

Then the second PLS component $\mathbf{t}_2$ is defined as

$$\mathbf{t}_2 = \frac{1}{\sqrt{\sum_{j=1}^{k} \operatorname{cov}(\mathbf{y}_1, \mathbf{x}_{1j})^2}} \sum_{j=1}^{k} \operatorname{cov}(\mathbf{y}_1, \mathbf{x}_{1j})\mathbf{x}_{1j}.$$

For interpretation purpose, the component $\mathbf{t}_2$ is better expressed as a function of variables $\mathbf{x}_j$'s. This is possible because the residuals $\mathbf{x}_{1j} = \mathbf{x}_j - p_{1j}\mathbf{t}_1$ for $j = 1, \ldots, k$ are functions of $\mathbf{x}_j$'s. When expressed in terms of $\mathbf{x}_j$'s, the component $\mathbf{t}_2$ is written as $\mathbf{t}_2 = \mathbf{X}\mathbf{w}_2$.

Computation of the next PLS components and stopping rule. We follow the same procedure for computing the next components $\mathbf{t}_h = \mathbf{X}\mathbf{w}_h$ for $h \geqslant 3$. The search for new components is stopped either in accordance with a cross-validation procedure or when all partial covariances are not significant. The PLS algorithm converges very quickly, in practice, it will give a satisfactory result when $m = 1$, 2 or 3.

PLS regression formula. In model (1), the coefficients c_h's are estimated by multiple regression of $\mathbf{y}$ on the PLS components $\mathbf{t}_h$'s. The estimated regression equation may be then expressed in terms of the variables $\mathbf{x}_j$'s:

$$\hat{\mathbf{y}} = \sum_{h=1}^{m} \hat{c}_h \left(\sum_{j=1}^{k} w_{hj}\mathbf{x}_j \right) = \sum_{j=1}^{k} \left(\sum_{h=1}^{m} \hat{c}_h w_{hj} \right) \mathbf{x}_j = \sum_{j=1}^{k} \hat{b}_j \mathbf{x}_j.$$

Now if an inverse procedure of standardization is implemented, we will get the regression equation expressed in terms of the raw variables $\mathbf{y0}$ and $\mathbf{x0}_j$'s:

$$\hat{\mathbf{y0}} = \hat{b}^* + \sum_{j=1}^{k} \hat{b}_j^* \mathbf{x0}_j.$$

We can see that if $\mathbf{t}_h$ is strongly correlated with $\mathbf{y}$, and $\mathbf{x}_j$ is important when building up $\mathbf{t}_h$, then $\mathbf{x}_j$ will be important to $\mathbf{y}$. The idea is reflected in the concept of *variable importance in projection* (VIP), for the jth variable $\mathbf{x}_j$, it is defined as

$$\mathrm{VIP}_j = \sqrt{\frac{k}{\mathrm{Rd}(\mathbf{y}; \mathbf{t}_1, \ldots, \mathbf{t}_m)} \sum_{h=1}^{m} \mathrm{Rd}(\mathbf{y}; \mathbf{t}_h) w_{hj}^2}, \tag{2}$$

where $\mathrm{Rd}(\mathbf{y}; \mathbf{t}_1, \ldots, \mathbf{t}_m) = \sum_{h=1}^{m} \mathrm{Rd}(\mathbf{y}; \mathbf{t}_h)$ and $\mathrm{Rd}(\mathbf{y}; \mathbf{t}_h) = [\mathrm{corr}(\mathbf{y}, \mathbf{t}_h)]^2$. Since w_{hj}^2 will take a large value if $\mathbf{x}_j$ is important in building up $\mathbf{t}_h$, and $\mathrm{Rd}(\mathbf{y}; \mathbf{t}_h)$ will be large if $\mathbf{t}_h$ is strongly correlated with $\mathbf{y}$, then VIP_j will be large in the situation. In addition, for given $\mathbf{y}$ and $\mathbf{X}$, $\mathbf{w}_h'\mathbf{w}_h (h = 1, \ldots, m)$ are fixed values, thus

$$\sum_{j=1}^{k} \mathrm{VIP}_j^2 = \frac{k \sum_{h=1}^{m} \mathrm{Rd}(\mathbf{y}; \mathbf{t}_h) \mathbf{w}_h'\mathbf{w}_h}{\mathrm{Rd}(\mathbf{y}; \mathbf{t}_1, \ldots, \mathbf{t}_m)},$$

is a constant. So for the response variable $\mathbf{y}$, if the k explanatory variables have the same explanatory ability, then all the VIP_j's are equal; otherwise, the explanatory variable with larger VIP value will tend to be more important than others. For more detailed discussion of the VIP concept, see Wang (1999).

3. Variable selection procedure

Some related notations are as follows: A mixed-level design of n runs, p factors and levels $s_1, \ldots, s_p$ is denoted by $D(n, s_1 \cdots s_p)$, when some s_j's are equal, it is denoted by $D(n, s_1^{r_1} \cdots s_q^{r_q})$ with $\sum_{j=1}^{q} r_j = p$. A $D(n, s_1 \cdots s_p)$

design is called an orthogonal array of strength t, denoted by $\mathrm{OA}(n, s_1 \cdots s_p, t)$ if all possible level-combinations for any t factors appear equally often. When $k = \sum_{i=1}^{p}(s_i - 1) > n - 1$, orthogonality is not obtainable and the design is supersaturated, denoted by $\mathrm{SSD}(n, s_1 \cdots s_p)$. Next we will introduce the new variable selection procedure and show how the procedure can be used for screening active effects in supersaturated designs.

3.1. The proposed variable selection strategy

For model (1), the regression coefficients can be estimated if we implement the procedure given in Section 2. However, actual active effects are believed to be sparse, and most of the coefficients should be close to zero. In order to build a reasonable model, we wish to select the best variable subset from the raw variables. The index VIP_j defined in (2) can be used to describe how important the variable $\mathbf{x}_j$ is to the response variable $\mathbf{y}$. So the best variable subset selection is based on the VIP values.

Let I be an empty set and $J = \{\mathbf{x}_1, \ldots, \mathbf{x}_k\}$, the PLSVS procedure can be carried out as follows:

Selection of the first important variable. Firstly, sort the k VIP values for $\mathbf{x}_1, \ldots, \mathbf{x}_k$ in increasing order:

$$\mathrm{VIP}_{(1)} \leqslant \mathrm{VIP}_{(2)} \leqslant \cdots \leqslant \mathrm{VIP}_{(k)}.$$

Select the two variables with their VIP values equaling $\mathrm{VIP}_{(k-1)}$ and $\mathrm{VIP}_{(k)}$ respectively, and denote them by $\bar{\mathbf{x}}_1, \bar{\mathbf{x}}_2$. Then their corresponding raw variables $\bar{\mathbf{x}}\mathbf{0}_1, \bar{\mathbf{x}}\mathbf{0}_2$ and $\mathbf{y0}$ are used to compute the Mpress (to be defined in (4) below) values respectively. The minimum Mpress value will be written as Mpress_1. The variable with Mpress_1 will be the first important variable $\mathbf{z0}_1$. The best variable subset now is $I = \{\mathbf{z0}_1\}$. Let $\mathbf{z}_1$ be $\bar{\mathbf{x}}_1$ or $\bar{\mathbf{x}}_2$ depending on whether $\mathbf{z0}_1$ is equal to $\bar{\mathbf{x}}\mathbf{0}_1$ or $\bar{\mathbf{x}}\mathbf{0}_2$.

Selection of the second important variable. At first, a simple regression is run:

$$\mathbf{y} = u_1\mathbf{z}_1 + \mathbf{y}_{re}, \tag{3}$$

where the coefficient is $u_1 = \mathbf{y}'\mathbf{z}_1/\|\mathbf{z}_1\|^2$ and $\mathbf{y}_{re}$ is the regression residual. Now with $\mathbf{y}_{re}$ and $J\backslash\{\mathbf{z}_1\}$, following the procedure given in Section 2, we can compute the m PLS components and $(k-1)$ VIP values of the rest $(k-1)$ variables. Then just as what we have done for selecting the first important variable, select the two variables $\bar{\mathbf{x}}_3, \bar{\mathbf{x}}_4$ with their VIP values being the largest two, their corresponding raw variables are denoted by $\bar{\mathbf{x}}\mathbf{0}_3, \bar{\mathbf{x}}\mathbf{0}_4$ respectively. Let $I_1 = \{\mathbf{z0}_1, \bar{\mathbf{x}}\mathbf{0}_3\}$ and $I_2 = \{\mathbf{z0}_1, \bar{\mathbf{x}}\mathbf{0}_4\}$, with the raw response variable $\mathbf{y0}$, compute their Mpress values. Let Mpress_2 be the minimum of the two Mpress values. Then the best variable subset I will equal I_1 or I_2 depending on whose Mpress is Mpress_2.

Selection of the next important variables and stopping rule. We follow the same procedure for selecting the next important variables. For selecting the rth important variable, we let Mpress_r be the minimum of the two Mpress values. Now we propose the variable selection stopping criterion.

Assume there are l $(0 \leqslant l \leqslant k)$ explanatory variables and their ith observation is $(x0_{i1}, \ldots, x0_{il})'$, $y0_i$ is the corresponding observation of the response, where $i = 1, \ldots, n$. Let $\tilde{\mathbf{x}}\mathbf{0}_i = (1, x0_{i1}, \ldots, x0_{il})'$, then the regression design matrix is $\tilde{\mathbf{X}}\mathbf{0}_{n\times(l+1)} = (\tilde{\mathbf{x}}\mathbf{0}_1, \ldots, \tilde{\mathbf{x}}\mathbf{0}_n)'$. If the ith observation is deleted, we can build up a least-squares (OLS) regression model with the rest $n-1$ observations, let $\hat{y0}_{l(-i)}$ be the predicted value of the ith response under the OLS model. Denote the predicted error of the ith response by,

$$\hat{e}_{l(-i)} = y0_i - \hat{y0}_{l(-i)}, \quad i = 1, \ldots, n,$$

then we can use

$$\mathrm{Press}(l) = \sum_{i=1}^{n} (\hat{e}_{l(-i)})^2,$$

to describe the predictable ability of a model. However, $\mathrm{Press}(l)$ will decrease with the value of l increasing, so it cannot be used as a variable selection criterion. We propose a modified version of $\mathrm{Press}(l)$ by adding a penalty function of l, i.e. the number of explanatory variables in the present model, that is,

$$\mathrm{Mpress}(l) = \frac{\mathrm{Press}(l)}{2(n-l)} + \frac{2l}{n}, \tag{4}$$

to determine when the selection will be stopped. Our simulation results in Section 4 reveal that this modified version works effectively for screening active effects in SSDs. In addition, we have tried to use other modified versions of Press(l), however, simulation results show that they are not so good as this one. Note that from (4), for $l = 0$ we have

$$\text{Mpress}(0) = \sum_{i=1}^{n} (y0_i - \bar{y0}_{(-i)})^2/(2n), \tag{5}$$

where $\bar{y0}_{(-i)}$ is the mean of $n-1$ responses with $y0_i$ being deleted.

For our variable selection method, with the number of variables selected into the best variable subset increasing, Mpress will decrease firstly; then it will increase with the number of variables increasing. The selection will be stopped if $\text{Mpress}_{r+1} > \text{Mpress}_r$ for the first time, where $\text{Mpress}_0 = \text{Mpress}(0)$, which is given in (5). The best variable subset is then obtained, which has r important variables.

3.2. A step-by-step procedure for the PLSVS method

A step-by-step guideline for the proposed procedure can be summarized as the following steps:

1. Column center and normalize the raw variables $\mathbf{y0}$ and $\mathbf{X0} = [\mathbf{x0}_1, \ldots, \mathbf{x0}_k]$, denote their standardized matrices by $\mathbf{y}$ and $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_k]$ respectively.
2. Initialize $r \leftarrow 0$, $\text{Mpress}_0 \leftarrow \text{Mpress}(0)$, $I \leftarrow \Phi$ (empty set), $J \leftarrow \{\mathbf{x}_1, \ldots, \mathbf{x}_k\}$ and $J^* \leftarrow \{\mathbf{x0}_1, \ldots, \mathbf{x0}_k\}$, where I, J and J^* are three variable sets.
3. Set $r \leftarrow r + 1$, for the variables in set J, compute the VIP values based on $\mathbf{y}$ by the PLS procedure.
4. From set J, select the variables with the largest two VIP values: $\bar{\mathbf{x}}_1$, $\bar{\mathbf{x}}_2$, suppose their corresponding raw variables in J^* are $\bar{\mathbf{x}}\mathbf{0}_1$, $\bar{\mathbf{x}}\mathbf{0}_2$ respectively. Set $I_1 \leftarrow I \cup \{\bar{\mathbf{x}}\mathbf{0}_1\}$, $I_2 \leftarrow I \cup \{\bar{\mathbf{x}}\mathbf{0}_2\}$.
5. For $\mathbf{y0}$ and the respective variables in I_1 and I_2, compute the Mpress values, denote them by M_1 and M_2.
6. Set $\text{Mpress}_1 \leftarrow \min(M_1, M_2)$. If $\text{Mpress}_1 < \text{Mpress}_0$, set $\text{Mpress}_0 \leftarrow \text{Mpress}_1$, go to step 7; otherwise, go to step 9.
7. If $M_1 \leqslant M_2$, set $\mathbf{z}_1 \leftarrow \bar{\mathbf{x}}_1$ and $\mathbf{z0}_1 \leftarrow \bar{\mathbf{x}}\mathbf{0}_1$; otherwise, $\mathbf{z}_1 \leftarrow \bar{\mathbf{x}}_2$ and $\mathbf{z0}_1 \leftarrow \bar{\mathbf{x}}\mathbf{0}_2$. Run the simple regression (3).
8. Set $I \leftarrow I \cup \{\mathbf{z0}_1\}$, $J \leftarrow J \backslash \{\mathbf{z}_1\}$, $J^* \leftarrow J^* \backslash \{\mathbf{z0}_1\}$, $\mathbf{y} \leftarrow \mathbf{y}_{re}$. Go to step 3.
9. Output the best variable set I.

3.3. Mixed-level SSDs and ANOVA model

Let $G_i = \{0, \ldots, s_i - 1\}$ and $H = G_1 \times \cdots \times G_p$. For an SSD($n, s_1 \ldots s_p$), consider the following main-effect ANOVA model,

$$\mathbf{Y} = \mathbf{1}_n \beta_0 + \mathbf{X}_c \boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{6}$$

where $\mathbf{Y}$ is the vector of n observations of the response, β_0 is the general mean and $\boldsymbol{\beta}$ is a vector of k treatment contrasts (or factorial main effects), $\mathbf{X}_c = [\chi_u(x)]_{x \in D, wt(u)=1}$ is the matrix of contrast coefficients for $\boldsymbol{\beta}$ and $wt(u) = 1$ means for all $u \in H$ with one nonzero element, $\boldsymbol{\epsilon}$ is the vector of errors with distribution $\text{N}(\mathbf{0}_n, \sigma^2 \mathbf{I}_n)$.

In this paper, we only consider contrasts defined by tensor products:

$$\chi_u(x) = \prod_{i=1}^{p} \chi_{u_i}^{(s_i)}(x_i) \quad \text{for } u = (u_1, \ldots, u_p) \in H \text{ and } x = (x_1, \ldots, x_p) \in H,$$

where $\{\chi_{u_i}^{(s_i)}(x_i), u_i \in G_i\}$ are orthogonal polynomial contrasts for the ith factor which has s_i levels, and $\chi_0^{(s_i)}(x_i) = 1$, for any $x_i \in G_i$. In addition, the condition of orthonormal is required, that is for all i,

$$\sum_{x_i \in G_i} \chi_{u_i}^{(s_i)}(x_i) \chi_{v_i}^{(s_i)}(x_i) = |G_i| \delta_{u_i, v_i}.$$

The computation of $\chi_u(x)$ is illustrated by the following example.

Example 1. Consider the following SSD(6, $2^1 3^3$) design D:

$$D = \begin{pmatrix} 0 & 0 & 1 & 1 \\ 0 & 1 & 2 & 0 \\ 0 & 2 & 0 & 2 \\ 1 & 0 & 2 & 2 \\ 1 & 1 & 0 & 1 \\ 1 & 2 & 1 & 0 \end{pmatrix},$$

which can be constructed from Fang et al.'s (2003) fractions of saturated orthogonal arrays (FSOA) method. In design D, each column represents a factor and each row represents a run.

It is known that for a two-level factor, the orthonormal contrast coefficient vector is $(-1, 1)'$, while for a three-level factor, the orthogonal polynomial coefficient vectors are $(-1, 0, 1)'$ and $(1, -2, 1)'$. Consequently the orthonormal contrast coefficients are $(\chi_1^{(2)}(0), \chi_1^{(2)}(1)) = (-1, 1)$ for the first factor, $(\chi_1^{(3)}(0), \chi_1^{(3)}(1), \chi_1^{(3)}(2)) = (-\sqrt{6}/2, 0, \sqrt{6}/2)$ and $(\chi_2^{(3)}(0), \chi_2^{(3)}(1), \chi_2^{(3)}(2)) = (\sqrt{2}/2, -\sqrt{2}, \sqrt{2}/2)$ for the last three factors, respectively. Then for any given $u \in H$ and $x \in H$, where $H = \{0, 1\} \times \{0, 1, 2\} \times \{0, 1, 2\} \times \{0, 1, 2\}$ for design D, the orthonormal contrast $\chi_u(x)$ can be calculated out. For example, for $x = (0, 1, 2, 0)$, $u = (0, 2, 0, 0)$,

$$\chi_u(x) = \chi_0^{(2)}(0)\chi_2^{(3)}(1)\chi_0^{(3)}(2)\chi_0^{(3)}(0) = \chi_2^{(3)}(1) = -\sqrt{2},$$

which is the (2, 3)th element of the corresponding contrast coefficient matrix $\mathbf{X_c}$ of design D, where

$$\mathbf{X}_c = \begin{pmatrix} -1 & -\sqrt{6}/2 & \sqrt{2}/2 & 0 & -\sqrt{2} & 0 & -\sqrt{2} \\ -1 & 0 & -\sqrt{2} & \sqrt{6}/2 & \sqrt{2}/2 & -\sqrt{6}/2 & \sqrt{2}/2 \\ -1 & \sqrt{6}/2 & \sqrt{2}/2 & -\sqrt{6}/2 & \sqrt{2}/2 & \sqrt{6}/2 & \sqrt{2}/2 \\ 1 & -\sqrt{6}/2 & \sqrt{2}/2 & \sqrt{6}/2 & \sqrt{2}/2 & \sqrt{6}/2 & \sqrt{2}/2 \\ 1 & 0 & -\sqrt{2} & -\sqrt{6}/2 & \sqrt{2}/2 & 0 & -\sqrt{2} \\ 1 & \sqrt{6}/2 & \sqrt{2}/2 & 0 & -\sqrt{2} & -\sqrt{6}/2 & \sqrt{2}/2 \end{pmatrix}.$$

Corresponding to $\mathbf{X}_c$, we have $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_7)'$, where β_1 is the main effect of the first factor, i.e. the two-level factor, β_{2i-2} and β_{2i-1} are the linear and quadratic main effects respectively for the ith factor, $i = 2, 3, 4$.

For model (6), let us take the k columns of $\mathbf{X}_c$ as the k raw explanatory variables: $\mathbf{x0}_1, \ldots, \mathbf{x0}_k$ and take $\mathbf{Y}$ as the raw response variable $\mathbf{y0}$, then the variable selection procedure proposed in last subsection can be used to screen active effects in mixed SSDs.

4. Simulation study and example

In this section, some simulations and comparisons are carried out. Firstly, the PLSVS procedure will be adopted to screen active effects in a mixed-level SSD. Then a real data set will be analyzed. Finally, we will compare the performance of the PLSVS method with the SCAD method and the SSVS/IBF method by simulations. As Li and Lin (2002, 2003) have done, we assess the performance of these variable selection procedures in terms of their abilities of identifying the true model and all the active effects, and the size of selected model.

Example 2. Construction of optimal mixed-level SSDs has been discussed in the literature in recent years, however, there is still not a paper studying the data analysis of such designs. Now we conduct some simulations to show the performance of the PLSVS method when it is used to analyze mixed-level SSDs, which include multi-level and two-level SSDs as special cases.

Fang et al.'s (2003) FSOA method for constructing mixed-level SSDs is an extension of Lin's (1993) half-fractions of Hadamard matrices method. As an illustration, given a saturated OA(27, 3^{13}, 2), taking any factor as the branching factor, one can obtain three one-third fractions according to the levels of the branching factor, any two-third fraction is an SSD(18, $2^1 3^{12}$) and all these designs are optimal according to Fang et al. (2003) $E(f_{\mathrm{NOD}})$ criterion, one of these designs is shown in Table 1, where the rows represent factors and the columns represent runs. From this design, the

Table 1
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	0	2	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

steps of setting active effects and generating data are as follows:

1. For this SSD(18, $2^1 3^{12}$), compute the orthonormal main-effect contrast matrix $\mathbf{X}_c$, which has 18 rows and 25 columns.
2. Randomly assign f active effects to $\mathbf{X}_c$. Here the number of active effects, f, is determined from the set $\{1, \ldots, 5\}$, according to the effect sparsity principle. For example when $f = 3$, three columns are randomly chosen from $\mathbf{X}_c$.
3. For the f active effects, set their model coefficients, i.e. the corresponding f components of $\boldsymbol{\beta}$, to be $\beta^{(1)}, \ldots, \beta^{(f)}$ respectively, and set the other components of $\boldsymbol{\beta}$ to be zero. Here we consider three cases of relative magnitude of coefficients $\beta^{(1)}, \ldots, \beta^{(f)}$. In each case, the different levels of f active effects ascend in constant intervals from low to high. For example, the three cases for the coefficients of four active effects are respectively: 1, 2, 3, 4; 2, 4, 6, 8 and 3, 6, 9, 12. Generally, in Case $i\,(i = 1, 2, 3)$,$(\beta^{(1)}, \ldots, \beta^{(f)}) = (i, 2i, \ldots, fi)$.
4. Generate $\mathbf{Y}$, the vector of 18 observations of the response, from the linear model

$$\mathbf{Y} = \mathbf{X}_c\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

where $\boldsymbol{\epsilon}$ has the distribution $\mathrm{N}(\mathbf{0}_{18}, \mathbf{I}_{18})$.

Simulation results for PLSVS based on 1000 replicates are summarized in Tables 2 and 3. Table 2 lists the percent of 1000 simulations succeeding in identification when there are 1, 3 and 5 active effects with m, the number of components in PLS procedure, ranging from 1 to 4; while Table 3 displays simulation results with $f = 1, \ldots, 5$ and $m = 3$. In both tables, "f" stands for the number of active effects, "Case" refers to the relative magnitude of coefficients, "m" denotes the number of components in the PLS regression, and the entries followed by "$[f, f + 2]$" are the rates of identifying the model size between f and $f + 2$.

From Table 2, we can see that if only a single effect is active, the performance is not sensitive to the choice of m providing $m = 1, 2, 3, 4$. However, if 3 or 5 effects are active, the rate of identifying the true model increases with m increasing for Case 2 or 3 of the magnitude of coefficients. Especially when there are 3 active effects with coefficients being 3, 6 and 9, the rate of identifying the true model is 45% and the median of model size is 4 when $m = 1$; however, they are respectively 50% and 3 when $m = 2$ or 3, and 53% and 3 when $m = 4$.

Furthermore, for Case 1 of the magnitude of coefficients, i.e. the coefficients are $1, \ldots, f$, if we choose $m = 4$, the performance is not better than that of $m = 1, 2, 3$. Particularly, the performance of $m = 4$ is worse than that of $m = 3$. So selecting relatively large number of components in PLS procedure will prevent us from identifying small active effects. Generally $m = 3$ is a better choice.

When there is a single active effect, the performances of different cases of the magnitude of coefficients vary little. However, if there are two or more active effects, the magnitude of coefficients have a strong effect on the performance

Table 2
Summary of simulation results in Example 2 with $f = 1, 3, 5$ and $m = 1, 2, 3, 4$

f	Case	m	True model identified rate (%)	Active effects identified rate (%)	Model size identified	
					Median	$[f, f+2]$ (%)
	1	1	57	96	1	97
		2	56	96	1	96
		3	60	97	1	98
		4	57	96	1	98
1	2	1	59	100	1	97
		2	58	100	1	96
		3	59	100	1	98
		4	63	100	1	97
	3	1	59	100	1	96
		2	57	100	1	97
		3	60	100	1	98
		4	61	100	1	96
	1	1	41	89	4	90
		2	41	91	4	90
		3	40	91	4	90
		4	38	90	4	88
3	2	1	44	97	4	92
		2	44	96	4	91
		3	48	97	4	93
		4	49	96	4	92
	3	1	45	96	4	91
		2	50	96	3	94
		3	50	97	3	92
		4	53	97	3	94
	1	1	29	75	6	81
		2	30	74	6	83
		3	32	75	6	85
		4	31	75	6	84
5	2	1	46	82	5	91
		2	50	85	5	91
		3	53	84	5	91
		4	52	82	5	91
	3	1	56	83	5	93
		2	57	84	5	91
		3	58	84	5	92
		4	59	84	5	91

and this effect increases with the number of active effects increasing. This point can be seen easily from the rates of identifying the true model in Table 3, and can also be observed from those data in Table 2.

As expected, the PLSVS procedure performs better when there are less active effects providing the same magnitude of coefficients, and it also performs better with a larger magnitude of coefficients when the numbers of active effects are the same. Particularly from these two tables, we notice that the procedure performs worse in Case 1 than with other cases in terms of the rate of identifying all the active effects; however, this rate varies little when the magnitude of coefficients ranges from Cases 2 to 3. For two-level SSDs, Lin (1995) pointed out that "to detect effects with magnitudes in the range of 2–3σ in the presence of many factors, however, is a very difficult task", now we can see that the same problem exists when we want to screen the active effects in the mixed-level SSDs.

In almost all the cases, the PLSVS method is effective in identifying active effects and determining the correct model size. Hence we conclude that our strategy is efficient and effective.

Table 3
Summary of simulation results in Example 2 when $m = 3$

f	Case	True model identified rate (%)	Active effects identified rate (%)	Model size identified	
				Median	$[f, f+2]$ (%)
1	1	60	97	1	98
	2	59	100	1	98
	3	60	100	1	98
2	1	48	93	2	93
	2	50	100	2	94
	3	54	100	2	95
3	1	40	91	4	90
	2	48	97	4	93
	3	50	97	3	92
4	1	33	85	5	87
	2	47	92	5	92
	3	54	92	4	92
5	1	32	75	6	81
	2	49	83	5	91
	3	58	84	5	93

Remark 1. We have tried to select variables with the largest three VIP values in the simulation study. The results demonstrate that the procedure needs more computations while the improvement is not so evident; When a single variable with the largest VIP value is selected, simulation results show that the rate of identifying all the active effects will be lower, and hence the type II error will become larger.

Example 3. The rubber data has been analyzed in many studies, e.g. Lin (1993, 1995), Chipman et al. (1997), Westfall et al. (1998), Abraham et al. (1999), Beattie et al. (2002), Li and Lin (2002, 2003), Lu and Wu (2004). The original experiment investigated 24 factors in a 28-run Plackett-Burman design, see Williams (1968). Lin (1993) took a half-fraction of the original data set as if the experiment had only 14 runs. Factor 16 is deleted since it is fully aliased with factor 13, but the other factors' original labels are kept unchanged. The data set is listed in Table 4. The last row is the responses and other rows represent factors and the columns represent runs.

We apply the PLSVS method to this data set. The final model identifies {15, 12, 20, 4} as the active effects when $m = 1, 2$ or 3. This is consistent with the conclusion of Li and Lin (2003), that {15, 20, 12, 4} were selected as active factors, a little difference is their order of importance.

Example 4. To compare the performance of the PLSVS method with that of the SCAD method and the SSVS/IBF method by simulations, we consider the same models with Li and Lin (2002, 2003).

Consider the design matrix $\mathbf{X}$ displayed in Table 4. Generate data from the linear model

$$\mathbf{Y} = \mathbf{X}'\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

where the vector of random errors $\boldsymbol{\epsilon}$ has the distribution $\mathrm{N}(\mathbf{0}_{14}, \mathbf{I}_{14})$. Consider the following three cases for $\boldsymbol{\beta}$:

Case I: One active effect, $\beta_1 = 10$ and all other components of $\boldsymbol{\beta}$ equal zero;

Case II: Three active effects, $\beta_1 = -15$, $\beta_5 = 8$, $\beta_9 = -2$, and all other components of $\boldsymbol{\beta}$ equal zero;

Case III: Five active effects, $\beta_1 = -15$, $\beta_5 = 12$, $\beta_9 = -8$, $\beta_{13} = 6$, $\beta_{17} = -2$, and all other components of $\boldsymbol{\beta}$ equal zero.

For these three cases, as the coefficients being relatively large, we just set $m = 1$ for the PLS regression. Simulation results for PLSVS based on 1000 replicates are summarized in Table 5 and are compared with the other two methods. In this table, "SCAD" stands for the SCAD method, and "SSVS(0.10, 500)/IBF" refers to the SSVS/IBF method with given parameters 0.10 and 500 (see, Beattie et al., 2002).

Table 4
SSD(14, 2^{23}): half-fraction of Williams (1968) 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
	133	62	45	52	56	47	88	193	32	53	276	145	130	127

Table 5
Summary of simulation results in Example 4

Method	True model identified rate (%)	Smallest effect identified rate (%)	Average size	
			Median	Mean
Case I: One active effect				
SSVS(0.10, 500)/IBF	61	98	1	2.5
SCAD	75.6	100	1	1.7
PLSVS($m = 1$)	61	100	1	1.5
Case II: Three active effects				
SSVS(0.10, 500)/IBF	8.0	28	3	4.2
SCAD	74.7	98.5	3	3.3
PLSVS($m = 1$)	76.4	97.7	3	3.3
Case III: Five active effects				
SSVS(0.10, 500)/IBF	40.7	75	5	5.6
SCAD	69.7	99.4	5	5.4
PLSVS($m = 1$)	73.6	95	5	5.2

The PLSVS method includes the smallest active effect with a high probability ($\geqslant 95\%$) in any of these three cases, that is to say, type II errors are all controlled within the level 0.05. In terms of the model size, the PLSVS method performs quite well. It is clear that both the SCAD and PLSVS perform better than the SSVS/IBF method. In Cases II and III, the PLSVS method identifies the true model with the highest probabilities. In Case I, there is only a single strong effect, the probability of getting the exact model with the PLSVS method is a little lower than that of the SCAD. However, with the only one active factor being 100% identified, the average model size is smaller than those resulted from the other two methods, in this sense the method is more efficient than the other two.

5. Concluding remarks

This paper proposes the PLSVS method for selecting active effects in SSDs. Simulation performance and a real data set analysis demonstrate that the PLSVS method is efficient. Note that all the existing data analysis methods in the literature are for two-level SSDs only. The PLSVS approach can be used for screening active effects collected by SSDs with two-level, multi-level and even mixed-level factors. In addition, it is easy to understand and implement.

In this paper, the best variable subset selection is based on the VIP values where index VIP_j reflects how important the variable $\mathbf{x}_j$ is to the response variable $\mathbf{y}$. In general, a variable with the largest VIP value will tend to give the greatest contribution to $\mathbf{y}$, so in principle, it should be selected as the most important variable in the present stage. However as we can see, for an SSD, correlation exists among the k columns of $\mathbf{X_c}$ in model (6), which may cause the inconsistency between the order of the VIP values and the explanatory ability of the variables. Thus in our proposed procedure, the variables with the largest two VIP values are selected first, and then their Mpress values determine which one will be kept in the best variable subset. As discussed in Remark 1, simulation results also show that selecting a single variable with the largest VIP value or the variables with the largest three VIP values in the procedure performs not so well as selecting the variables with the largest two VIP values.

As we have mentioned, PLS regression is particularly useful when we need to predict a set of response variables from a (very) large set of explanatory variables, so the PLSVS method can be used in the situation when there are several response variables.

The screening of active effects and data analysis in multi-level and mixed-level SSDs still need further investigations.

Acknowledgements

The authors are grateful to the Executive-Editor, two anonymous referees and Professor Rahul Mukerjee for their valuable comments and constructive suggestions. This work was partially supported by the NNSF of China Grants 10301015 and 10571093, and the SRFDP of China Grant 20050055038. Liu's research was also supported by the Science and Technology Innovation Fund of Nankai University and the Visiting Scholar Program at Chern Institute of Mathematics.

References

Abraham, B., Chipman, H., Vijiayan, H., 1999. Some risks in the construction and analysis of supersaturated designs. Technometrics 41 (2), 135–141.

Bastien, P., Vinzi, V.E., Tenenhaus, M., 2005. PLS generalized linear regression. Comput. Statist. Data Anal. 48 (1), 17–46.

Beattie, S.D., Fong, D.K.H., Lin, D.K.J., 2002. A two-stage Bayesian model selection strategy for supersaturated designs. Technometrics 44 (1), 55–63.

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

Butler, N.A., Denham, M.C., 2000. The peculiar shrinkage properties of partial least squares regression. J. Roy. Statist. Soc. Ser. B 62 (3), 585–593.

Chipman, H., Hamada, H., Wu, C.F.J., 1997. A Bayesian variable-selection approach for analyzing designed experiments with complex aliasing. Technometrics 39 (4), 372–381.

Fang, K.T., Lin, D.K.J., Liu, M.Q., 2003. Optimal mixed-level supersaturated design. Metrika 58 (3), 279–291.

Fang, K.T., Ge, G., Liu, M.Q., Qin, H., 2004. Combinatorial constructions for optimal supersaturated designs. Discrete Math. 279, 191–202.

Garthwaite, P.H., 1994. An interpretation of partial least squares. J. Amer. Statist. Assoc. 89 (425), 122–127.

Helland, I.S., 1990. Partial least squares regression and statistical models. Scand. J. Statist. 17 (2), 97–114.

Holcomb, D.R., Montgomery, D.C., Carlyle, W.M., 2003. Analysis of supersaturated designs. J. Quality Tech. 35 (1), 13–27.

Höskuldsson, A., 2001. Variable and subset selection in PLS regression. Chemometrics & Intelligent Laboratory Systems 55, 23–38.

Li, R., Lin, D.K.J., 2002. Data analysis of supersaturated designs. Statist. Probab. Lett. 59 (2), 135–144.

Li, R., Lin, D.K.J., 2003. Analysis methods for supersaturated design: some comparisons. J. Data Sci. 1 (3), 249–260.

Li, P.F., Liu, M.Q., Zhang, R.C., 2004. Some theory and the construction of mixed-level supersaturated designs. Statist. Probab. Lett. 69 (1), 105–116.

Lin, D.K.J., 1993. A new class of supersaturated designs. Technometrics 35 (1), 28–31.

Lin, D.K.J., 1995. Generating systematic supersaturated designs. Technometrics 37 (2), 213–225.

Liu, M.Q., Fang, K.T., Hickernell, F.J., 2006. Connections among different criteria for asymmetrical fractional factorial designs. Statist. Sinica 16 (4),

Lu, X., Wu, X., 2004. A strategy of searching active factors in supersaturated screening experiments. J. Quality Technol. 36 (4), 392–399.

Wang, H.W., 1999. Partial least-squares regression method and applications. National Defence Industry, Beijing.

Westfall, P.H., Young, S.S., Lin, D.K.J., 1998. Forward selection error control in the analysis of supersaturated design. Statist. Sinica 8 (1), 101–117.

Williams, K.R., 1968. Designed experiments. Rubber Age 100, 65–71.
Yamada, S., 2004. Selection of active factors by stepwise regression in the data analysis of supersaturated design. Qual. Eng. 16 (4), 501–513.
Yamada, S., Lin, D.K.J., 2002. Construction of mixed-level supersaturated design. Metrika 56 (3), 205–214.
Yamada, S., Matsui, T., 2002. Optimality of mixed-level supersaturated designs. J. Statist. Plann. Inference 104 (2), 459–468.
Yamada, S., Matsui, M., Matsui, T., Lin, D.K.J., Takahashi, T., 2006. A general construction method for mixed-level supersaturated design. Comput. Statist. Data Anal. 50 (1), 254–265.