A method for screening active effects in supersaturated designs

Min-Qian Liu
(刘民千)

Nankai University, China

mqliu@nankai.edu.cn

(Joint work with Qiao-Zhen Zhang and Run-Chu Zhang)

Thanks to . . .

- **your kind support to the conference!**
- Grants from NNSF of China, SRFDP of China and Nankai University;
- the Visiting Scholar Program at Chern Institute of Mathematics.

Thanks to . . .

- **your kind support to the conference!**
- Grants from NNSF of China, SRFDP of China and Nankai University;
- the Visiting Scholar Program at Chern Institute of Mathematics.

Outline

1. Introduction
2. Background of PLS regression
3. Variable selection procedure
4. Simulation study and example
5. Concluding remarks

Introduction

What's supersaturated design?

- **Supersaturated design (SSD):**
  factorial design in which $\#\{\text{main effects}\} \geq \#\{\text{runs}\}$.
  Screening active effects under the assumption of **effect sparsity**.
- **Construction:**
  - 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, 2004), Li et al. (2004), Yamada et al. (2006) and Liu et al. (2006).
- **Data analysis:**
  To find the sparse active effects, variable selection becomes fundamental in the analysis stage of such screening experiments.

What's supersaturated design?

- **Supersaturated design (SSD):**
  factorial design in which $\#\{\text{main effects}\} \geq \#\{\text{runs}\}$.
  Screening active effects under the assumption of **effect sparsity**.
- **Construction:**
  - 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, 2004), Li et al. (2004), Yamada et al. (2006) and Liu et al. (2006).
- **Data analysis:**
  To find the sparse active effects, variable selection becomes fundamental in the analysis stage of such screening experiments.

Some recent analysis methods

All restricted at two-level SSDs.

- Bayesian variable selection approach: Chipman et al. (1997)
- error control skill in forward selection: Westfall et al. (1998)
- two-stage Bayesian model selection strategy (SSVS/IBF): Beattie et al. (2002)
- **smoothly clipped absolute deviation** (SCAD) method: Li and Lin (2002, 2003)
- contrast-based methods: Holcomb et al. (2003)
- modified stepwise selection based on the idea of staged dimensionality reduction: Lu and Wu (2004)

Simulation studies demonstrated that the SCAD method outperforms the other approaches.

Motivation

- The aspect of data analysis of multi-level and mixed-level SSDs has not been studied in adequate detail.
- This talk will introduce an approach via **Partial least-squares** (PLS) regression, called the PLS variable selection (PLSVS) method, for searching active effects in SSDs based on the **variable importance in projection** (VIP).
- PLSVS can be used to analyze data collected from SSDs with mixed-level, multi-level or two-level factors.

Motivation

- The aspect of data analysis of multi-level and mixed-level SSDs has not been studied in adequate detail.
- This talk will introduce an approach via **Partial least-squares** (PLS) regression, called the PLS variable selection (PLSVS) method, for searching active effects in SSDs based on the **variable importance in projection** (VIP).
- PLSVS can be used to analyze data collected from SSDs with mixed-level, multi-level or two-level factors.

Background of PLS regression

- $\mathbf{y0}, \mathbf{x0}_1, \ldots, \mathbf{x0}_k$: raw variables
  $\mathbf{y}, \mathbf{x}_1, \ldots, \mathbf{x}_k$: column centered and normalized patterns
  $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_k)$
  $\mathbf{w}_h = (w_{h1}, \ldots, w_{hk})'$
  $\mathbf{t}_h = \sum_{j=1}^{k} w_{hj}\mathbf{x}_j = \mathbf{X}\mathbf{w}_h,\ h = 1, \ldots, m,$
- PLS regression model with m components:

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

  s.t. the m PLS components $\mathbf{t}_h$'s are orthogonal.
- PLS regression is an algorithm for estimating the parameters of model (1) (Bastien et al., 2005).

Computation of the first PLS component $\mathbf{t}_1$

- maximize $\mathrm{cov}(\mathbf{y}, \mathbf{t_1}) = \mathrm{s}(\mathbf{t_1}) * \mathrm{corr}(\mathbf{y}, \mathbf{t}_1)$, *s.t.* $\mathbf{t}_1 = \mathbf{X}\mathbf{w}_1$ and $\mathbf{w}_1'\mathbf{w}_1 = 1$.
- $\mathbf{w}_1$ is the standard eigenvector of $\mathbf{X}'\mathbf{y}\mathbf{y}'\mathbf{X}$ corresponding to the largest eigenvalue, and then

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

- $\mathrm{cov}(\mathbf{y}, \mathbf{x}_j) = \mathrm{corr}(\mathbf{y}, \mathbf{x}_j)$, 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 $\mathbf{t}_2$

- Run the $k+1$ simple regressions:

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

- Then $\mathbf{t}_2$ is defined as

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

- It can be expressed as a function of variables $\mathbf{x}_j$'s: $\mathbf{t}_2 = \mathbf{X}\mathbf{w}_2$.

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 $h \geq 3$.
- The search of 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

- Estimate c_h's in model (1) by multiple regression of $\mathbf{y}$ on the PLS components $\mathbf{t}_h$'s.
- Then

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

- 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:

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

PLS regression formula

- Estimate c_h's in model (1) by multiple regression of $\mathbf{y}$ on the PLS components $\mathbf{t}_h$'s.
- Then

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

- 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:

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

Variable importance in projection (VIP)

For $\mathbf{x}_j$, its VIP is defined as:

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

$\mathrm{Rd}(\mathbf{y};\mathbf{t}_h) = [\mathrm{corr}(\mathbf{y},\mathbf{t}_h)]^2$, $\mathrm{Rd}(\mathbf{y};\mathbf{t}_1,\ldots,\mathbf{t}_m) = \sum_{h=1}^{m}\mathrm{Rd}(\mathbf{y};\mathbf{t}_h)$.

- For given $\mathbf{y}$ and $\mathbf{X}$, $\sum_{j=1}^{k}\mathrm{VIP}_j^2$ is a constant.
- For the response variable $\mathbf{y}$, the explanatory variable with larger VIP value will tend to be more important than others (Wang, 1999).

Variable selection procedure

Notations

- mixed-level design: $D(n, s_1 \cdots s_p)$, $D(n, s_1^{r_1} \cdots s_q^{r_q})$
- orthogonal array of strength t: $\mathrm{OA}(n, s_1 \cdots s_p, t)$
- supersaturated design: $\mathrm{SSD}(n, s_1 \cdots s_p)$, where

  $k = \sum_{i=1}^{p}(s_i - 1) > n - 1$

For an SSD, actual active effects are believed to be sparse, and most of the coefficients in model (1) should be close to zero.

The index VIP_j defined in (2) can be used to describe how important $\mathbf{x}_j$ is to $\mathbf{y}$. So the best variable subset selection is based on the VIP values.

Mpress – a new variable selection criterion

Assume there are l $(0 \le l \le k)$ explanatory variables,

- $y0_i$, $x0_{i1}, \ldots, x0_{il}$: i-th observation, $i = 1, \ldots, n$
- $\tilde{\mathbf{x0}}_i = (1, x0_{i1}, \ldots, x0_{il})'$, $\tilde{\mathbf{X0}}_{n\times(l+1)} = (\tilde{\mathbf{x0}}_1, \ldots, \tilde{\mathbf{x0}}_n)'$
- $\hat{y0}_{l(-i)}$: predicted value of $y0_i$ under the OLS model after deleting the i-th observation
- $\hat{e}_{l(-i)} = y0_i - \hat{y0}_{l(-i)}, \quad i = 1, \ldots, n,$
- $\mathrm{Press}(l) = \sum_{i=1}^{n} (\hat{e}_{l(-i)})^2$ will decrease with the value of l increasing, so it can not be used as a variable selection criterion.

Mpress

- Mpress: a modified version of $\mathrm{Press}(l)$, i.e.

$$\mathrm{Mpress}(l) = \frac{\mathrm{Press}(l)}{2(n-l)} + \frac{2l}{n}. \tag{3}$$

- Simulation results reveal that this modified version works effectively for screening active effects in SSDs.
- Other modified versions of $\mathrm{Press}(l)$ have been tried, however, simulation results show that they are not so good as Mpress.
- 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 proposed variable selection strategy

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

1. **Selection of the first important variable**
   - For the variables in set J, compute the VIP values based on $\mathbf{y}$ by the PLS procedure.
   - Select the variables with the largest two VIP values: $\bar{\mathbf{x}}_1$, $\bar{\mathbf{x}}_2$, suppose their corresponding raw variables are $\bar{\mathbf{x}}\mathbf{0}_1$, $\bar{\mathbf{x}}\mathbf{0}_2$ resp.
   - For $\bar{\mathbf{x}}\mathbf{0}_1$, $\bar{\mathbf{x}}\mathbf{0}_2$ and $\mathbf{y0}$, compute the Mpress values respectively.
   - The variable with the minimum Mpress, say 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 $\bar{\mathbf{x}}\mathbf{0}_2$.

2 Selection of the second important variable

- Run a simple regression $\mathbf{y} = u_1\mathbf{z}_1 + \mathbf{y}_{re}$, where $u_1 = \mathbf{y}'\mathbf{z}_1/\|\mathbf{z}_1\|^2$ and $\mathbf{y}_{re}$ is the regression residual.
- With $\mathbf{y}_{re}$ and $J\backslash\{\mathbf{z}_1\}$, compute the m PLS components and the VIP values of the rest $(k-1)$ variables.
- Select the two variables $\bar{\mathbf{x}}_3$ and $\bar{\mathbf{x}}_4$ with the largest two VIP's, suppose their corresponding raw variables are $\bar{\mathbf{x}}\mathbf{0}_3$, $\bar{\mathbf{x}}\mathbf{0}_4$ resp.
- 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.
- The best variable subset I will equal I_1 or I_2 depending on whose Mpress is Mpress_2.

3. **Selection of the next important variables and stopping rule**

- Follow the same procedure for selecting the next important variables.
- For selecting the r-th important variable, let Mpress_r be the minimum of the two Mpress values.
- The selection will be stopped if $\mathrm{Mpress}_{r+1} > \mathrm{Mpress}_r$ for the first time. The best variable subset is then obtained, which has r important variables.

Mixed-level SSDs and ANOVA model

For an SSD($n, s_1 \cdots s_p$), consider the following main-effect ANOVA model

$$\mathbf{Y} = \mathbf{1}_n \beta_0 + \mathbf{X_c} \boldsymbol{\beta} + \varepsilon, \tag{4}$$

- $\mathbf{Y}$ is the vector of n observations of the response,
- β_0 is the general mean,
- $\boldsymbol{\beta}$ is a vector of k treatment contrasts (or factorial main effects),
- $\mathbf{X_c}$ is the matrix of contrast coefficients for $\boldsymbol{\beta}$,
- $\varepsilon \sim N(\mathbf{0}_n, \sigma^2 \mathbf{I}_n)$.

Example 1

SSD$(6, 2^1 3^3)$ design D constructed from Fang et al.'s (2003) fractions of saturated orthogonal arrays (FSOA) method:

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

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

Example 1

SSD$(6, 2^1 3^3)$ design D constructed from Fang et al.'s (2003) fractions of saturated orthogonal arrays (FSOA) method:

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

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

Example 1

SSD$(6, 2^1 3^3)$ design D constructed from Fang et al.'s (2003) fractions of saturated orthogonal arrays (FSOA) method:

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

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

Example 1

SSD$(6, 2^1 3^3)$ design D constructed from Fang et al.'s (2003) fractions of saturated orthogonal arrays (FSOA) method:

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

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

Example 2

An $\mathrm{SSD}(18, 2^1 3^{12})$ constructed from Fang et al.'s (2003) FSOA method:

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

- $\mathbf{X}_c$ has 18 rows and 25 columns.
- Given β, $\mathbf{Y}$ can be generated from the linear model $\mathbf{Y} = \mathbf{X}_c\beta + \varepsilon$, where $\varepsilon \sim N(\mathbf{0}_{18}, \mathbf{I}_{18})$.
- f: the number of randomly chosen active effects, $f = 1, 2, 3, 4, 5$;
- *Case*: the relative magnitude of coefficients, in *Case* $i (i = 1, 2, 3)$, the coefficients of f active effects are $(i, 2i, \ldots, fi)$;
- m: the number of components in the PLS regression, $m = 1, 2, 3, 4$;
- Simulation results for PLSVS based on 1000 replicates show that $m = 3$ is a better choice.

- $\mathbf{X}_c$ has 18 rows and 25 columns.
- Given β, $\mathbf{Y}$ can be generated from the linear model $\mathbf{Y} = \mathbf{X}_c\beta + \varepsilon$, where $\varepsilon \sim N(\mathbf{0}_{18}, \mathbf{I}_{18})$.

- f: the number of randomly chosen active effects, $f = 1, 2, 3, 4, 5$;
- *Case*: the relative magnitude of coefficients, in *Case* $i (i = 1, 2, 3)$, the coefficients of f active effects are $(i, 2i, \dots, fi)$;
- m: the number of components in the PLS regression, $m = 1, 2, 3, 4$;
- Simulation results for PLSVS based on 1000 replicates show that $m = 3$ is a better choice.

Simulation results in Example 2 when $m = 3$

f	*Case*	True Model Identified Rate	Active Effects Identified Rate	Model Size Identified: Median	Model Size Identified: $[f, f+2]$
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%

where "$[f, f+2]$" denotes the rates of identifying the model size between f and $f+2$.

Summary of simulation results in Example 2

- PLSVS performs better when there are less active effects providing the same magnitude of coefficients;
- PLSVS performs better with larger magnitude of coefficients when the numbers of active effects are the same;
- In almost all the cases, PLSVS is effective in identifying active effects and determining the correct model size.

  Hence we conclude that our strategy is efficient and effective.

- Simulation results 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.

Example 3: (Williams Rubber Experiment)

The rubber data has been analyzed in many studies, e.g. Lin (1993, 1995), ...

- PLSVS identifies $\{15, 12, 20, 4\}$ as the active effects when $m = 1, 2$ or 3. This is consistent with the conclusion of
- Lin (1993): $\{15, 12, 20, 4\}$,
- Li and Lin (2002): $\{4, 12, 15, 20\}$,
- Li and Lin (2003): $\{15, 20, 12, 4\}$,

a little difference is their order of importance.

Example 4: Comparisons with SCAD and SSVS/IBF

Consider the same models with Li and Lin (2002, 2003):

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \varepsilon, \text{ where } \varepsilon \sim N(\mathbf{0}_{14}, \mathbf{I}_{14}),$$

$\mathbf{X}$ is an $\mathrm{SSD}(14, 2^{23})$, i.e. half-fraction of Williams' (1968) data.

1. Case I: $\beta_1 = 10$ and all other components of $\boldsymbol{\beta}$ equal zero;
2. Case II: $\beta_1 = -15, \beta_5 = 8, \beta_9 = -2$, and all other components of $\boldsymbol{\beta}$ equal zero;
3. Case III: $\beta_1 = -15, \beta_5 = 12, \beta_9 = -8, \beta_{13} = 6, \beta_{17} = -2$, and all other components of $\boldsymbol{\beta}$ equal zero.

Summary of simulation results in Example 4

Method	True Model Identified Rate	Smallest Effect Identified Rate	Average Size Median	Average Size 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

- PLSVS includes the smallest active effect with a high probability ($\geq 95\%$);
- PLSVS performs quite well in terms of the model size.
- Both the SCAD and PLSVS perform better than the SSVS/IBF method.

Concluding remarks

- The existence of correlation among the k columns of $\mathbf{X_c}$ in model (4) may cause the inconsistent between the order of the VIP values and the explanatory ability of the variables, so we proposed the PLSVS method;
- Simulation performance and a real data set analysis demonstrate that the PLSVS method is efficient;
- PLSVS can be used for screening active effects collected by SSDs with two-level, multi-level and even mixed-level factors;
- 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.

Any question or comment?

Enjoy the Chinese banquet tonight!