Computational Statistics and Data Analysis 53 (2008) 428–435

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

# Modelling by supersaturated designs

Stelios D. Georgiou*

*Department of Statistics and Actuarial-Financial Mathematics, University of the Aegean, Karlovassi 83200, Samos, Greece*

ARTICLE INFO

*Article history:*
Received 18 June 2007
Received in revised form 9 August 2008
Accepted 11 August 2008
Available online 19 August 2008

ABSTRACT

The analysis of supersaturated designs is an interesting problem of great importance since it provides economic estimates. Moreover, this problem is challenging due to the fact that the design matrix has a complicated structure. The identification of the active factors in supersaturated designs is investigated. The singular value decomposition (SVD), principal components analysis and regression analysis are used together in an SVD principal regression method to reveal the hidden true linear model. Special cases are studied by using simulation data under idealized conditions. Simulations are used to investigate the performance of the method and also to compare the proposed method with other known methods from the literature.

## 1. Introduction

Supersaturated designs form an important class of factorial designs. This is because they can be used to investigate a large number of factors using only few experimental runs, and thus achieve a lower cost than traditional factorial designs. We call a factorial design supersaturated if the number of factors (including the mean column) is larger than the number of possible orthogonal columns. A two-level design is said to be supersaturated if the number of factors $m$ is greater than the number of runs $n$. In this case we usually use the symbols 1 and $-1$ to denote the high and low level of each factor respectively. Throughout this paper, we use the linear main effects model

$$y = X\beta + \varepsilon, \qquad \varepsilon \sim N_n(0_n, \sigma^2 I_n), \tag{1}$$

where $y$ is the $n \times 1$ response vector and $X = [x_0, x_1, \ldots, x_m] = [1_n, x_1, \ldots, x_m]$ is the $n \times (m+1)$ model matrix (or design matrix). The first column of the model matrix is $1_n = [1, 1, \ldots, 1]^{\mathrm{T}}$, and this column corresponds to the mean effect. The $j_{\text{th}}$ column of the design matrix is denoted by $x_j = [x_{1j}, x_{2j}, \ldots, x_{nj}]^{\mathrm{T}}$, $x_{ij} \in \{1, -1\}$. This represents the main effect contrast between the high and low level of factor $j$ corresponding to the $j_{\text{th}}$ element of the parameter vector $\beta, j = 0, 1, \ldots, m$.

The experimental error is denoted by $\varepsilon$, and is assumed to be i.i.d. multivariate normal with dimension $n$, zero mean vector and a variance matrix $\Sigma = \sigma^2 I_n$, where $I_n$ is the identity matrix of order $n$. If no confusion is caused, a vector $u$ of length $\ell$ will often be used as an $\ell \times 1$ matrix, and the transpose of $u$ will be denoted by $u^{\mathrm{T}}$. Even though the design matrix and the model matrix usually play different roles, we will use the same symbol for both since, in the case of two-level factors, these matrices will be exactly the same.

A two-level column vector $u$ of length $n$ is said to be *balanced* if each of the levels appears an equal number of times so that $1_n^{\mathrm{T}} u = u^{\mathrm{T}} 1_n = 0$ and $u^{\mathrm{T}} u = n$. A supersaturated design $X = [1_n, x_1, \ldots, x_m]$ is said to be *balanced* if each of its columns $x_j, j = 1, 2, \ldots, m$ is balanced. In the literature, these designs are also called *mean orthogonal* designs. Two columns $x_j$ and $x_k$ are *orthogonal* to each other iff $x_j^{\mathrm{T}} x_k = x_k^{\mathrm{T}} x_j = 0$, are *fully aliased* (or fully confounded) iff $x_j^{\mathrm{T}} x_k = x_k^{\mathrm{T}} x_j = \pm n$, and are *partially aliased* iff $0 < x_j^{\mathrm{T}} x_k = x_k^{\mathrm{T}} x_j < n$. We are not interested in supersaturated designs with fully aliased columns.

* Tel.: +30 2273082329; fax: +30 2273082309.
*E-mail address:* stgeorgiou@aegean.gr.

0167-9473/$ – see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2008.08.015

If the design matrix is non-singular then the matrix $(X^{\mathrm{T}}X)^{-1}$ exists and the coefficients $\beta$ can be estimated by the well-known least squares estimator thus: $\hat{\beta} = (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}y$. In supersaturated designs this approach is unsuitable since the number of factors $m$ is greater than the number of runs $n$. Thus, $X^{\mathrm{T}}X$ is singular and so not invertible. While the construction of supersaturated designs has been widely explored (Booth and Cox, 1962; Bulutoglu and Cheng, 2004; Butler et al., 2001; Cheng, 1997; Eskridge et al., 2004; Lin, 1993, 1995; Liu and Dean, 2004; Liu et al., 2007; Liu and Zhang, 2000; Lu and Meng, 2000), the data analysis aspect of these designs remains primitive. Several approaches have been suggested in the literature for the analysis of supersaturated designs. Some of them are briefly mentioned below.

Srivastava (1975) showed that any set of $p$ active effects may be estimated if all subsets of $2p$ variables, in the model matrix, contain independent columns. Chen and Lin (1998) investigated the identifiability of supersaturated designs. Under normality assumptions, they provided a lower bound for the probability that the factor with the largest estimated effect had, indeed, the largest true effect. Westfall et al. (1998) used the effect sparsity hypothesis and forward-selection multiple test procedures to address the problem of analysing data with supersaturated designs. Abraham et al. (1999) examined supersaturated designs and methods for their analysis. They mentioned that the correlation structure inherent in supersaturated designs can obscure real effects or promote effects. They concluded that this problem could occur whatever method of analysis was used. A two-stage Bayesian model selection strategy for supersaturated designs was proposed by Beattie et al. (2002). Li and Lin (2002) suggested a variable selection method for identifying the active effects in supersaturated designs via non-convex penalized least squares. Holcomb et al. (2003) showed that the contrasts of supersaturated designs follow a permuted multivariate hyper-geometric distribution, which may be approximated by a normal distribution. They compared several methods from the literature and proposed a contrast-based method for analysing data from supersaturated designs. Li and Lin (2003) introduced a variable selection procedure to screen out the active effects in supersaturated designs, and they performed empirical comparison with Bayesian variable selection approaches by using simulation experiments. Lu and Wu (2004) introduced a new strategy of searching active factors in supersaturated screening experiments based on the idea of staged dimensionality reduction. Koukouvinos and Stylianou (2005) suggested a modified contrast variance method for analysing data from supersaturated designs. They used simulation models to compare their method with several others from the literature. Liu et al. (2007) addressed the difficulties of supersaturated designs in detecting the active factors. They investigated the correct identification in several cases, including one- and two-variable linear models.

The correlation of the design matrix and the departure from orthogonality are of huge importance and have a catastrophic influence in the detection of the true active factors. The best-known and most-used criteria for comparing designs are the $E(s^2)$ criterion and the max $|s_{ij}|$ criterion. For more details on these criteria, see, for example, Booth and Cox (1962) or Bulutoglu and Cheng (2004).

In this paper, the identification of the active factors in supersaturated designs is investigated. Singular value decomposition, principal components, and regression analysis are used together to reveal the true linear model. Special cases are studied by using simulation data under idealized conditions. Simulation results are used to enable a comparison between the proposed method and some known methods from the literature. In Section 2 we present a modified principal components regression analysis, suitable for two-level supersaturated designs. The method is developed theoretically and is then explained with some analytical examples. A serious modification of the method might be suitable for analysing mixed-level supersaturated designs. Further research is needed in this direction, since only limited results are available in the literature. In Section 3 we apply the proposed method, using simulation data from the literature.

## 2. The proposed method

In this section we present a method for analysing two-level balanced supersaturated designs. A serious modification of the method might be suitable for analysing mixed-level supersaturated designs. Further research is needed in this direction, since only limited results are available in the literature. Throughout this paper we assume that we have a model of the form (1) and that $m \geq n$. The classical assumptions when analysing supersaturated designs are:

- *Effect sparsity:* Only a few, say $p$, of the $m$ potentially active factors are really active.
- The coefficients of the active effects ($\beta$s) are large enough to be distinguished from the error.
- The columns in the model matrix are not pairwise fully aliased.

The proposed method is briefly described by the following synoptic steps:

(1) Compute the standardized contrast (univariate standard regression coefficient) of each of the variables.
(2) Form a reduced design matrix constituted of the $p$ variables which correspond to the $p$ largest absolute contrasts.
(3) Compute the principal components of the reduced design matrix by calculating its singular value decomposition.
(4) Use the computed principal components and the original response vector to estimate a linear main effect regression model.
(5) Transform the result back to the original variables.
(6) Run significance tests and retain only significant factors.

Since the principal components provide a new uncorrelated reduced matrix, we expect to have better results when we apply regression analysis on the new orthogonal columns. Moreover, to obtain more precise results we choose which columns of the original design matrix we should include in the reduced matrix before applying the principal components analysis. So, we choose to include only those columns with the strongest estimated contrast correlation with the response $y$. We now give a detailed description of the method and present an illustrative example.

Using the design matrix $X = [x_0, x_1, \ldots, x_m]$, we define the normalized model matrix

$$D = [d_0, d_1, \ldots, d_m] = \left[\frac{x_0}{\|x_0\|}, \frac{x_1}{\|x_1\|}, \ldots, \frac{x_m}{\|x_m\|}\right],$$

where $\|x_j\| = \sqrt{x_j^{\mathrm{T}} x_j}$. The elements of $x_j$ are $\pm 1$, and thus $\|x_j\| = \sqrt{n}$, for all $j = 0, 1, \ldots, m$. Let

$$C = D^{\mathrm{T}} y = [c_0, c_1, \ldots, c_m]^{\mathrm{T}}$$

be the vector of standardized contrasts. This means that $c_j$ is the sum of the responses when $d_j$ is at the high level minus the sum of the responses when $d_j$ is at the low level, and that result is divided by $\sqrt{n}$, for $j = 0, 1, \ldots, m$. It is noted that the normalization of each column of the design matrix $X$ is not necessary, but is retained to transform $D^{\mathrm{T}}D$ into the form of correlation matrix. The method can be applied, in the same way, even if normalization is skipped. Normalization will be necessary in one tries to generalize the method to the case of mixed-level supersaturated designs.

To be more precise, we should include a scale estimate $\hat{\sigma}$ in each of the $c_j, j = 1, 2, \ldots, m$, but this is omitted since this estimate is common in all contrasts, and it will thus have no influence in the comparison of the contrast magnitudes.

In the first stage we carefully apply a raw screening procedure with the condition that the type II error should be zero or near zero. At this stage we are not interested in the magnitude of the type I error since any active variables not selected in this stage cannot be identified later, while any inactive variables selected might be eliminated. So, at this stage of the method we choose $r$ variables (columns of $X$) that will be included in a new reduced $n \times r$ model matrix denoted by $X_r$. If the number of true active factors ($p$) was known we would have chosen $r = p$. Since $p$ is unknown, and based on the assumption of effect sparsity, we suggest $r$ to be as large as $\frac{n}{2}$ (Srivastava, 1975). The new design matrix $X_r$ consist of variables (columns of $X$) $x_j$ with higher absolute contrast values $|c_j|$. One selection criterion might be to include all columns for which the corresponding absolute value of their contrasts is greater than $\xi$, i.e., we select all $x_j$ for which $|c_j| \geq \xi$, where $\xi$ could be estimated by cross-validation. One other, simpler and effective way (which is adapted in this paper) is to select $x_j, j = 0, i_1, \ldots, i_r$ for which $|c_{i_1}| \geq |c_{i_2}| \geq \cdots \geq |c_{i_r}|$, i.e., select the $r$ columns that correspond to the largest absolute values of contrast.

We can then write the singular value decomposition of $X_r$ as

$$X_r = U_r D_r V_r^{\mathrm{T}},$$

where $X_r$ is an $n \times r$ matrix of rank $t \leq r$, $U_r$ is an orthogonal $n \times n$ matrix, $D_r$ is an $n \times r$ matrix with the singular values ordered $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_t > 0$, and $V_r$ is an $r \times r$ matrix.

Since the rank of $X_r$ is $t \leq r$, there are exactly $t$ non-zero (positive) singular values. Due to this fact we project the matrices onto a smaller dimension space by removing columns and rows corresponding to zero singular values. What we achieve by our use of the SVD is to eliminate any singularities in the $X_r$ matrix by simply dropping terms. One point that needs to be clear is exactly which terms should be dropped. If we have a set of terms in $X_r$ that are not linearly independent, any one of them might be dropped in an attempt to get back to a linearly independent set (but just which term(s) is(are) dropped might depend on how we have ordered the terms in $X_r$ and just how our SVD routine works). Thus we need to order the terms in $X_r$ from strongest contrast to weakest and proceed sequentially. If the current term generates a singularity, drop it and go on to the next term; if not, pass it on to $X_t$. Checking whether we have a singularity is easy — just regress the new term on the ones already in $X_t$ and see if we get a perfect fit. In this way, the final reduced $n \times t$ matrix $X_t$ will be

$$X_t = U_t D_t V_t^{\mathrm{T}},$$

where $U_t$ is an $n \times t$ orthogonal matrix, consisting of the $t$ left singular vectors that correspond to non-zero singular values, $D_t$ is an $t \times t$ diagonal matrix with the singular values ordered $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_t > 0$, and $V_t$ is a $t \times t$ orthogonal matrix.

The standard tactic when using principal components analysis is to keep the first few principal components and work with them, but in the case of supersaturated designs this approach would be unsatisfactory. This is because the first few principal components will not explain a sufficient volume of the total variance and this reduction would result an important loss of information. So, we continue by using all $t$ of the principal components. We might have no further reduction in the dimensions of the data but we have achieved the construction of uncorrelated regressors (i.e., the left singular vectors).

We then apply linear regression analysis using the left singular vectors as independent factors and with the original response data vector $y$ as the dependent variable. Let $U_t = [u_{t,1}, u_{t,2}, \ldots, u_{t,t}]$ be the matrix which contains the left singular vectors $u_{t,j}, j = 1, 2, \ldots, t$. We have that $\mathrm{Var}(u_{t,1}) \geq \mathrm{Var}(u_{t,2}) \geq \ldots \geq \mathrm{Var}(u_{t,t})$ and that $\mathrm{Var}(u_{t,j}) = d_{t,j}, j = 1, 2, \ldots, t$, where $d_{t,j}$ is the $(j, j)$-element of $D_t$ and $\mathrm{Var}(u_{t,j})$ is the variance of $u_{t,j}$. The fitted linear model is

$$y = \gamma_1 u_{t,1} + \gamma_2 u_{t,2} + \cdots + \gamma_t u_{t,t} + \varepsilon = U_t \gamma + \varepsilon, \tag{2}$$

and

$$\hat{\gamma} = (U_t^{\mathrm{T}} U_t)^{-1} U_t^{\mathrm{T}} y \Rightarrow \hat{\gamma}_j = \frac{u_{t,j}^{\mathrm{T}} y}{\|u_{t,j}\|} = u_{t,j}^{\mathrm{T}} y \in (a_j, b_j) \tag{3}$$

is the least squares estimate of the coefficient $\gamma_j$, where $(a_j, b_j)$ is the $(1-\alpha)100\%$ confidence interval for coefficient $\gamma_j$.

At this stage we have estimated a linear model where the dependent variables are the principal components and the coefficients are the $\gamma_j$. Now we need to switch back to the original predictors and conclude the final linear model with factors $x_j$ (the original variables).

**Lemma 1.** *When the model and the estimate of the coefficients are as (2) and (3) respectively, then $E(\hat{y}) = X_t\beta$.*

**Proof.** We have that

$$X_t = U_t D_t V_t^{\mathrm{T}} \Rightarrow X_t V_t = U_t D_t \Rightarrow U_t = X_t V_t D_t^{-1} = X_t Z_t,$$

where $Z_t = V_t D_t^{-1} = [z_{t,1}, z_{t,2}, \ldots, z_{t,t}]$. Using this transformation, the derived model can be converted to

$$\begin{aligned} E[\hat{y}] &= \gamma_1 u_{t,1} + \gamma_2 u_{t,2} + \cdots + \gamma_t u_{t,t} = \gamma_1 X_t z_{t,1} + \gamma_2 X_t z_{t,2} + \cdots + \gamma_t X_t z_{t,t} \\ &= X_t[\gamma_1 z_{t,1} + \gamma_2 z_{t,2} + \cdots + \gamma_t z_{t,t}] = X_t\beta, \end{aligned}$$

where $\beta = \gamma_1 z_{t,1} + \gamma_2 z_{t,2} + \cdots + \gamma_t z_{t,t} = V_t D_t^{-1}\gamma$. $\square$

Following the proof of Lemma 1, we can fit a linear model using the original predictors. We then need to check which of the included variables (the $x_i$s) are important for a specific significance level $\alpha$. One way to achieve this is to test the null hypothesis $H_0 : \beta_i = 0$ against the general alternatives $H_1 : \beta_i \neq 0, i = 1, 2, \ldots, t$.

The test statistic we used was

$$F = \frac{\hat{\beta}_i^2}{S^2 g_{ii}^2} \sim F_{1,n-t,\alpha},$$

where $\hat{\beta}_i$ is the estimate of the $i_{\text{th}}$ coefficient of the model $S^2 = \frac{(Y-X\hat{\beta})^{\mathrm{T}}(Y-X\hat{\beta})}{n-t}$ and $g_{ii}^2$ is the $i_{\text{th}}$ diagonal element of the $(X_t^{\mathrm{T}} X_t)^{-1}$ matrix. We have that

$$(X_t^{\mathrm{T}} X_t)^{-1} = \left[(V_t D_t^{\mathrm{T}} U_t^{\mathrm{T}})(U_t D_t V_t^{\mathrm{T}})\right]^{-1} = (V_t D_t^{\mathrm{T}} D_t V_t^{\mathrm{T}})^{-1} = \operatorname{diag}\left[\frac{1}{d_{11}^2}, \frac{1}{d_{22}^2}, \ldots, \frac{1}{d_{tt}^2}\right]$$

and thus the test statistic is transformed to

$$F = \frac{\hat{\beta}_i^2 d_{ii}^2}{S^2} \sim F_{1,n-t,\alpha},$$

where $\hat{\beta}_i$ and $S^2$ are as before, and $d_{ii}^2 = \lambda_i^2$ is the element $(i, i)$ of $D$, i.e., the $i_{\text{th}}$ singular value.

If the null hypothesis cannot be rejected then $\beta_i$ is not important at significance level $\alpha$. We now remove the non-active effects and obtain the final linear model that fits the data. Thus, the final estimated model will be

$$\hat{y} = \hat{\beta}_{i_1} x_{i_1} + \hat{\beta}_{i_2} x_{i_2} + \cdots + \hat{\beta}_{i_s} x_{i_s} + \varepsilon = X_s\beta + \varepsilon,$$

where $\hat{\beta} = [\hat{\beta}_{i_1}, \hat{\beta}_{i_2}, \ldots, \hat{\beta}_{i_s}]^{\mathrm{T}}$, $\hat{\beta}_{i_k} \in \{\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_t\}$, $k = 1, 2, \ldots, s$, and $s \leq t$.

To illustrate the above methodology we present a detailed example.

### 2.1. An illustrative example

We shall perform all steps, as they are described in the proposed method, by using a supersaturated design with six runs and ten factors. The design matrix (or model matrix) is

$$X = \begin{bmatrix} 1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & -1 & -1 & -1 \\ 1 & 1 & 1 & 1 & -1 & -1 & -1 & 1 & 1 & 1 \\ -1 & 1 & -1 & -1 & 1 & 1 & -1 & 1 & 1 & -1 \\ -1 & -1 & 1 & -1 & 1 & -1 & 1 & 1 & -1 & 1 \\ -1 & -1 & -1 & 1 & -1 & 1 & 1 & -1 & 1 & 1 \end{bmatrix} = [x_1, x_2, \ldots, x_{10}].$$

To the above matrix, we add a first column $x_0$ with all entries equal to 1. This column corresponds to the overall mean. In order to test our results we use simulated data obtained from the linear model

$$y = \beta_0 x_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_{10} x_{10} + \varepsilon = 5x_0 + 4x_2 + 3x_5 + \varepsilon,$$

where $\varepsilon \sim N_6(0_6, I_6)$ (i.e., $\varepsilon$ is i.i.d. to a multivariate normal distribution with mean vector zero and with a variance matrix the identity matrix). A response $y$, obtained by using the above simulated model, is

$$y^{\mathrm{T}} = [-1.54, 12.02, 6.82, 12.44, 4.62, -1.21].$$

We initialize the method by using significance level $\alpha = 1\%$ and $r = \frac{n}{2} = 3$. With these inputs, the final model will probably include the constant term and up to three other variables selected from the set $\{x_1, x_2, \dots, x_{10}\}$.

Using the design matrix $X = [x_0, x_1, \dots, x_{10}]$ we define

$$D = [d_0, d_1, \dots, d_{10}] = \begin{bmatrix}
\frac{1}{\sqrt{6}} & \frac{1}{\sqrt{6}} & -\frac{1}{\sqrt{6}} & -\frac{1}{\sqrt{6}} & -\frac{1}{\sqrt{6}} & -\frac{1}{\sqrt{6}} & -\frac{1}{\sqrt{6}} & -\frac{1}{\sqrt{6}} & -\frac{1}{\sqrt{6}} & -\frac{1}{\sqrt{6}} & -\frac{1}{\sqrt{6}} \\
\frac{1}{\sqrt{6}} & \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{6}} & -\frac{1}{\sqrt{6}} & -\frac{1}{\sqrt{6}} & -\frac{1}{\sqrt{6}} \\
\frac{1}{\sqrt{6}} & \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{6}} & -\frac{1}{\sqrt{6}} & -\frac{1}{\sqrt{6}} & -\frac{1}{\sqrt{6}} & \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{6}} \\
\frac{1}{\sqrt{6}} & -\frac{1}{\sqrt{6}} & \frac{1}{\sqrt{6}} & -\frac{1}{\sqrt{6}} & -\frac{1}{\sqrt{6}} & \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{6}} & -\frac{1}{\sqrt{6}} & \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{6}} & -\frac{1}{\sqrt{6}} \\
\frac{1}{\sqrt{6}} & -\frac{1}{\sqrt{6}} & -\frac{1}{\sqrt{6}} & \frac{1}{\sqrt{6}} & -\frac{1}{\sqrt{6}} & \frac{1}{\sqrt{6}} & -\frac{1}{\sqrt{6}} & \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{6}} & -\frac{1}{\sqrt{6}} & \frac{1}{\sqrt{6}} \\
\frac{1}{\sqrt{6}} & -\frac{1}{\sqrt{6}} & -\frac{1}{\sqrt{6}} & -\frac{1}{\sqrt{6}} & \frac{1}{\sqrt{6}} & -\frac{1}{\sqrt{6}} & \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{6}} & -\frac{1}{\sqrt{6}} & \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{6}}
\end{bmatrix},$$

and the vector of standardized contrasts is $C = D^{\mathrm{T}}y = [13.53, 0.59, 12.01, 5.62, 0.86, 10.21, 5.45, -0.94, 5.97, 1.21, -5.18]^{\mathrm{T}}$. Note that $\frac{c_0}{\|x_0\|} = \frac{\sum_{i=1}^{6} y_i}{6}$ is the mean of the response.

The $r = 3$ larger absolute standardized contrasts (not including $|c_0|$) are $|c_2|$, $|c_5|$, and $|c_8|$. So, the reduced matrix $X_r$ is

$$X_r = [x_0, x_2, x_5, x_8] = \begin{bmatrix}
1 & -1 & -1 & -1 \\
1 & 1 & 1 & -1 \\
1 & 1 & -1 & 1 \\
1 & 1 & 1 & 1 \\
1 & -1 & 1 & 1 \\
1 & -1 & -1 & -1
\end{bmatrix}.$$

We compute the singular value decomposition of $X_r = U_r D_r V_r^{\mathrm{T}}$, and find that

$$U_r = \begin{bmatrix}
0.548 & -0.408 & 0.000 & 0.000 & -0.344 & -0.644 \\
-0.183 & -0.408 & 0.408 & 0.707 & 0.355 & -0.086 \\
-0.183 & -0.408 & 0.408 & -0.707 & 0.355 & -0.086 \\
-0.548 & -0.408 & 0.000 & 0.000 & -0.710 & 0.172 \\
-0.183 & -0.408 & -0.817 & 0.000 & 0.355 & -0.086 \\
0.548 & -0.408 & 0.000 & 0.000 & -0.011 & 0.730
\end{bmatrix} = [u_1, \dots, u_6],$$

$$D_r = \begin{bmatrix}
3.162 & 0.000 & 0.000 & 0.000 \\
0.000 & 2.450 & 0.000 & 0.000 \\
0.000 & 0.000 & 2.000 & 0.000 \\
0.000 & 0.000 & 0.000 & 2.000 \\
0.000 & 0.000 & 0.000 & 0.000 \\
0.000 & 0.000 & 0.000 & 0.000
\end{bmatrix},$$

$$V_r = \begin{bmatrix}
0.000 & -1.000 & 0.000 & 0.000 \\
-0.577 & 0.000 & 0.817 & 0.000 \\
-0.577 & 0.000 & -0.408 & 0.707 \\
-0.577 & 0.000 & -0.408 & -0.707
\end{bmatrix}.$$

There are exactly $t = 4$ non-zero (positive) singular values. We therefore project the matrices into a smaller space by removing columns and rows corresponding to zero singular values. The final reduced matrix $X_t$ will be

$$X_t = U_t D_t V_t^{\mathrm{T}},$$

where $U_t$ consists of the first four columns of $U_r$ (i.e., $U_t = [u_1, u_2, u_3, u_4]$), $D_t$ consists of the first four rows of $D_r$ (i.e., $D_t = \mathrm{diag}(3.162, 2.450, 2.000, 2.000)$), and $V_t$ is the same as $V_r$ (i.e., $V_t = V_r$).

We then apply linear regression analysis, using the left singular vectors as predictors and the original vector $y$ as the response. The linear model obtained is

$$\hat{y} = \hat{\gamma}_1 u_{t,1} + \hat{\gamma}_2 u_{t,2} + \hat{\gamma}_3 u_{3,t} + \hat{\gamma}_4 u_{4,t} + \varepsilon,$$

**Table 1**
The Williams' data — rubber age data

| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 | 24 | $y$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| + | + | + | − | − | − | + | + | + | + | + | − | − | − | + | + | − | − | + | − | − | − | + | 133 |
| + | − | − | − | − | − | + | + | + | − | − | − | + | + | + | − | + | − | − | + | + | − | − | 62 |
| + | + | − | + | + | − | − | − | − | + | − | + | + | + | + | + | − | − | − | − | + | + | − | 45 |
| + | + | − | + | − | + | − | − | − | + | + | − | − | + | + | − | + | + | + | − | − | − | − | 52 |
| − | − | + | + | + | + | − | + | + | − | − | − | − | + | + | + | − | − | + | − | + | + | + | 56 |
| − | − | + | + | + | + | + | − | + | + | + | − | + | + | − | + | + | + | + | + | + | − | − | 47 |
| − | − | − | − | + | − | − | + | − | + | − | + | + | − | + | + | + | + | + | + | − | − | + | 88 |
| − | + | + | − | − | + | − | + | − | + | − | − | − | − | − | − | − | + | − | + | + | + | − | 193 |
| − | − | − | − | − | + | + | − | − | − | + | + | − | + | − | + | + | − | − | − | − | + | + | 32 |
| + | + | + | + | − | + | + | + | − | − | − | + | + | + | − | + | − | + | − | + | − | − | + | 53 |
| − | + | − | + | + | − | − | + | + | − | + | − | + | − | − | − | + | + | − | − | − | + | + | 276 |
| + | − | − | − | + | + | + | − | + | + | + | + | − | − | + | − | − | + | − | + | + | + | + | 145 |
| + | + | + | + | + | − | + | − | + | − | − | + | − | − | − | − | + | − | + | + | − | + | − | 130 |
| − | − | + | − | − | − | − | − | − | − | + | + | + | − | − | − | − | − | + | − | + | − | − | 127 |

where

$$\hat{\gamma} = [-12.6057, -13.5328, 3.9229, 3.6749]^{\mathrm{T}}.$$

We switch back to the original variables to obtain estimates for the $\beta$s and to decide which of the original factors are significant. We have that

$$\hat{\beta} = V_t D_t^{-1} \hat{\gamma} = [5.5249, 3.9030, 2.8000, 0.2014]^{\mathrm{T}}.$$

To find out which of the $\beta_i$s are significantly different from zero, we test the hypothesis $H_0 : \beta_i = 0$ with alternatives $H_1 : \beta_i \neq 0$, for $i = 1, 2, 3, 4$. The value of the F distribution with $\alpha = 1\%$, with 1 degree of freedom for the numerator and with $n - t = 2$ degrees of freedom for the denominator, is 98.5. The values of the test statistics for $\hat{\beta}_i$, $i = 1, 2, 3, 4$, are 10762.14, 3222.69, 1105.73, and 5.72 respectively. Thus, the constant and the first two estimated coefficients (the $\hat{\beta}$s) are important at significance level $\alpha = 1\%$. So, the final estimated model is

$$\hat{y} = 5.5249x_0 + 3.9030x_2 + 2.8000x_5 + \varepsilon,$$

which provides a very good approximation of the original simulated true model ($R^2 = 0.97$).

In the next section we use simulation experiments to investigate the performance of the proposed SVD principal regression method.

## 3. Simulation study

As with any decision problem, errors of various types must be balanced against cost. In screening designs, there is the cost of declaring an inactive factor to be active (type I error), and the cost of declaring an active variable to be inactive (type II error). In model selection techniques, it would be ideal if a method could identify all the true active factors (zero type II error) and ignore those that are inactive (zero type I error). Usually, supersaturated designs are used to perform a first-stage screening experiment in order to reduce the cost of the experiment. A follow-up procedure is then applied to reveal the true model. Obviously, it is critical to have a tiny or zero type II error in the screening stage. On the other hand, it is important to keep the type I error as low as possible in order to avoid unnecessary cost in the follow-up experiments.

In this section we investigate the performance of the proposed method using some simulation data, which have been used in a number of research papers. This permits a comparison between the proposed method and other methods from the literature.

### 3.1. The Williams' data — rubber age data

The first simulated data set we shall use is that of Williams (1968) (rubber age data). We use the half-fraction of the 28-run Plackett and Burman design as a model matrix. Columns 13 and 16 in the original design matrix were identical and column 13 was therefore removed; however, we retain the same factor labelling. The corrected design matrix and the rubber age data are presented in Table 1. This example has been studied in almost every paper that deals with the analysis of supersaturated designs. The results obtained by our method and by many methods from the literature are summarized below.

Using the proposed SVD principal regression method, with $p \leq 7$ and $a = 0.05$, only factor 15 seems to be influential.

In the next section we use several simulation experiments from the literature to investigate the performance of the proposed method and also to enable a comparison with other methods.

**Table 2**
The simulation results

| Md no. | $\mathbf{a}_{\text{opt}}$ | Error | Cross val. | | Estim model infr | | | Error in Ref | |
|---|---|---|---|---|---|---|---|---|---|
| | | Type I, II | $\mathbf{I}_{\text{cross}}$ | $\mathbf{II}_{\text{cross}}$ | Identif | Mean | Median | Type I | Type II |
| 1 | 0.04 | 0.01 | 0.02 | 0.05 | 814 | 3.28 | 3.00 | 0.20 | 0.12 |
| 2 | 0.37 | 0.04 | 0.02 | 0.06 | 483 | 2.35 | 2.00 | 0.06 | 0.07 |
| 3 | 0.01 | 0.00 | 0.00 | 0.00 | 996 | 1.01 | 1.00 | 0.08 | 0.14 |
| 4 | 0.01 | 0.00 | 0.00 | 0.00 | 999 | 2.00 | 2.00 | 0.01 | 0.01 |
| 5 | 0.01 | 0.00 | 0.00 | 0.00 | 1000 | 3.00 | 3.00 | 0.00 | 0.00 |
| 6 | 0.07 | 0.02 | 0.03 | 0.01 | 887 | 2.13 | 2.00 | 0.09 | 0.18 |
| 7 | 0.09 | 0.01 | 0.03 | 0.12 | 899 | 3.00 | 3.00 | 0.02 | 0.11 |
| 8 | 0.07 | 0.00 | 0.07 | 0.01 | 972 | 3.99 | 4.00 | 0.08 | 0.10 |
| 9 | 0.06 | 0.00 | 0.03 | 0.09 | 998 | 5.00 | 5.00 | 0.11 | 0.17 |
| 10 | 0.35 | 0.26 | 0.06 | 0.31 | 3 | 3.20 | 2.00 | 0.12 | 0.22 |

### 3.2. Further simulation experiments

Abraham et al. (1999), Beattie et al. (2002), Li and Lin (2002, 2003), and Westfall et al. (1998) used one of Lin's design matrices (that is given in Table 1) and generated data from, among others, the following simulation models:

True Model 1: $y \sim N(15x_1 + 8x_5 - 6x_9 + 3x_5x_9, I_{14})$ (Beattie et al., 2002),
True Model 2: $y \sim N(8x_1 + 5x_{12}, I_{14})$ (Li and Lin, 2002),
True Model 3: $y \sim N(20x_1, I_{14})$ (Abraham et al., 1999; Beattie et al., 2002; Li and Lin, 2003),
True Model 4: $y \sim N(20x_2 + 20x_7, I_{14})$ (Abraham et al., 1999),
True Model 5: $y \sim N(14x_2 + 20x_7 + 20x_{16}, I_{14})$ (Abraham et al., 1999),
True Model 6: $y \sim N(5x_1 + 5x_2, I_{14})$ (Westfall et al., 1998),
True Model 7: $y \sim N(5x_1 + 5x_2 + 5x_3, I_{14})$ (Westfall et al., 1998),
True Model 8: $y \sim N(5x_1 + 5x_2 + 5x_3 + 5x_4, I_{14})$ (Westfall et al., 1998),
True Model 9: $y \sim N(5x_1 + 5x_2 + 5x_3 + 5x_4 - 5x_5, I_{14})$ (Westfall et al., 1998),
True Model 10: $y \sim N(10x_1 + 9x_2 + 2x_3, I_{14})$ (Li and Lin, 2002).

In Model 3, Beattie et al. (2002) and Li and Lin (2003) set the coefficient of the only active factor $x_1$ equal to 10 rather than 20, but the results for both values were similar.

We generated 1000 experiments from each of the above models. In these examples we included an error term taken from a normal distribution with zero mean and standard deviation 1. We present the results we obtained in Table 2. The method seems to be very powerful and accurate when all the coefficients of the active variables are larger than four times the standard deviation of the error ($>4\sigma$). Most of the time, any variables with small coefficients (less or equal to $2\sigma$) are not estimable by this method since they cannot be distinguished from the experimental error (see, for example, Model 10). In some cases (see, for example, Model 1) the method is not prevented from screening out the true main effects, by the existence of interactions in the model, but this is an exception to the rule. In comparison with other known methods, we can conclude that the efficiency of the proposed method is superior when all model coefficients are sufficiently large. When there is one small coefficient or more the efficiency of the method is comparable with other known methods.

We need two parameters to apply the method: the initial number of active variables and the desirable significance level. To achieve best performance of the method, we suggest that the initial number of active variables be equal to $n/2$, where $n$ is the number of runs (if $p > n/2$ there will be an identification (estimability) problem, thus there is no point to investigate $p > n/2$ (Srivastava, 1975)). The significance level can be selected either by a cross-validation method or by the researcher. We have run the above simulation experiments using a range of significance levels varying from $a = 0.005$ to $a = 0.400$ (with a step of 0.005) as well as significance level, estimated by a "leave-one-out" cross-validation procedure for each of the 1000 experiments. The results are summarized in Table 2 and Fig. 1.

The first column "Md no." of this table refers to the model number, while the second column "$a_{\text{opt}}$" presents the value of the significance level for which the type I error rates become equal to the type II error rates. In this case the common value of the error rates is indicated in the third column "Error Type I, II" of the table. Columns four, "$\text{I}_{\text{cross}}$", and five, "$\text{II}_{\text{cross}}$", give the type I and type II errors respectively, when the significance level is estimated by a "leave-one-out" cross-validation method. In the next three columns of the table we present some descriptives for the estimated model: in particular, the frequency of true model identification "Identif", the mean "mean" and median "median" of the number of selected active variables. In the last two columns, "Type I" and "Type II", of Table 2, we present the type I and type II error rates respectively for each simulation model, using the methods proposed in the literature. When a model was used in more than one paper this column presents the mean of the results obtained by the methods proposed in the literature.

To save space, we only present the figure showing type I and type II error rates for one of the simulated model. Figures for all other models are similar to this one. Results from the simulations indicate that best value for the significance level is usually between 0.03 and 0.13. As expected, as the significance level grows larger, the type II error rates get smaller and the type I error rates get larger. It is easy to see in the figure that there exists an optimal value for the significance level $a_{\text{opt}}$ which will balance the type I and type II error rates ($a_{\text{opt}}$ is the point where the type I error and type II error lines meet). As

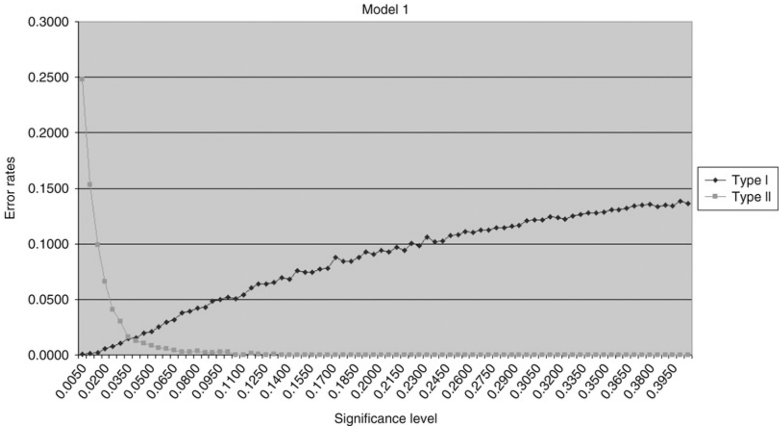

**Fig. 1.** Mean type I and type II error rates from simulation model 1.

we can see, for significance levels between 0.03 and 0.4, both error types are quite small; in our simulated experiments the error rates were usually less than 15%, and in most of the experiments errors of both types are less than 5%.

The efficiency of some methods might be influenced by the design and the simulated model used, by the number of active variables and by the magnitudes of the coefficients. Certainly, all methods give uncertain results and take a serious risk in the identification of the true model. One way in which confidence in the results might be improved would be to analyse the data using several methods from the literature and find the set of active factors for each method. If all sets of active factors are the same then one can trust the results with a large probability of being correct. Otherwise, the results should not be trusted. In this case, only variables identified as active by all methods have much probability of being truly active. In conclusion, we can say that one should be very careful and suspicious when using any method for analysing data from supersaturated designs.

## Acknowledgments

The author would like to thank the two anonymous referees for their useful comments and suggestions that substantially improved the results and the appearance of the paper.

## References

Abraham, B., Chipman, H., Vijayan, K., 1999. Some risks in the construction and analysis of supersaturated designs. Technometrics 41, 135–141.
Beattie, S.D., Fong, D.K.H., Lin, D.K.J., 2002. A two-stage Bayesian model selection strategy for supersaturated designs. Technometrics 44, 55–63.
Booth, K.H.V., Cox, D.R., 1962. Some systematic supersaturated designs. Technometrics 4, 489–495.
Bulutoglu, D.A., Cheng, C.S., 2004. Construction of $E(s^2)$-optimal supersaturated designs. Annals of Statistics 32, 1662–1678.
Butler, N., Mead, R., Eskridge, K.M., Gilmour, S.G., 2001. A general method of constructing $E(s^2)$-optimal supersaturated designs. Journal of Royal Statistical Society 63, 621–632.
Cheng, C.S., 1997. $E(s^2)$-optimal supersaturated designs. Statist. Sinica 7, 929–939.
Chen, J., Lin, D.K.J., 1998. On the identifiability of a supersaturated designs. Journal of Statistics and Planning Inference 72, 99–107.
Eskridge, K.M., Gilmour, S.G., Mead, R., Butler, N.A., Travnicek, D.A., 2004. Large supersaturated designs. Journal of Statistical Computation and Simulation 74, 525–542.
Holcomb, D.R., Montgomery, D.C., Carlyle, W.M., 2003. Analysis of supersaturated designs. Journal of Quality Technology 35, 13–27.
Koukouvinos, C., Stylianou, S., 2005. A method for analyzing supersaturated designs. Communications in Statistics-Simulation and Computation 34, 929–937.
Li, R., Lin, D.K.J., 2002. Data analysis in supersaturated designs. Statistics and Probability Letters 59, 135–144.
Li, R., Lin, D.K.J., 2003. Analysis methods for supersaturated designs: Some comparisons. Journal of Data Science 1, 249–260.
Lin, D.K.J., 1993. A new class of supersaturated designs. Technometrics 35, 28–31.
Lin, D.K.J., 1995. Generating systematic supersaturated designs. Technometrics 37, 213–225.
Liu, Y.F., Dean, A.M., 2004. k-circulant supersaturated designs. Technometrics 46, 32–43.
Liu, Y.F., Ruan, S., Dean, A.M., 2007. Construction and analysis of $Es^2$ efficient supersaturated designs. Journal of Statistics and Planning Inference 137, 1516–1529.
Liu, M., Zhang, R., 2000. Construction of $E(s^2)$ optimal supersaturated designs using cyclic BIBDs. Journal of Statistics and Planning Inference 91, 139–150.
Lu, X., Meng, Y., 2000. A new method in the construction of two-level supersaturated designs. Journal of Statistics and Planning Inference 86, 229–238.
Lu, X., Wu, X., 2004. A strategy of searching active factors in supersaturated screening experiments. Journal of Quality Technology 36, 392–399.
Srivastava, N.J., 1975. Designs for searching for non-negligible effects. In: A Survey of Statistical Designs and Linear Models. North-Holland, Amsterdam, pp. 507–519.
Westfall, P.H., Young, S.S., Lin, D.K.J., 1998. Forward selection error control in the analysis of supersaturated designs. Statist. Sinica 8, 101–117.
Williams, K.R., 1968. Designed Experiments. Rubber Age 100, 65–71.