Statistics and Probability Letters 79 (2009) 1224–1230

Contents lists available at ScienceDirect

Statistics and Probability Letters

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

Probability of correct model identification in supersaturated design

Angshuman Sarkar [a], Dennis K.J. Lin [b,*], Kashinath Chatterjee [a]

[a] *Department of Statistics, Visva-Bharati University, West Bengal 731235, India*
[b] *Department of Supply Chain & Information Systems, The Pennsylvania State University, University Park, PA 16802, USA*

A R T I C L E I N F O

Article history:
Received 12 June 2008
Received in revised form 21 December 2008
Accepted 18 January 2009
Available online 30 January 2009

A B S T R A C T

A criterion for comparing two competitive designs is proposed for symmetric factorial experiment based on the probability of correct identification of active factors. An algorithm for constructing multi-level supersaturated designs which maximize the probability of correct model identification is presented.

1. Introduction

Using a large-scale system in industrial experimentation, a large number of factors are typically involved during design and operation stages. In reality, however, it is a common phenomenon that only a few of these factors have a *substantial effect*, also known as *factor sparsity* (Box and Meyer, 1986). Supersaturated designs are constructed to identify factors having non-negligible effects from a large set of potential factors on the basis of a small number of observations (or runs).

The construction of supersaturated designs dates back to Satterthwaite (1959) and later gained increasing attention by Lin (1992, 1993). Wu (1993), Lin (1995, 1998, 2000), Nguyen (1996), Cheng (1997), Deng et al. (1999), Liu and Dean (2004), Sarkar (2007) and Chatterjee et al. (2008) among others considered the construction and the properties of two-level supersaturated designs. In this context, for more than two-level, the work by Yamada and Lin (1999, 2002), Fang et al. (2000, 2003, 2004), Chatterjee and Gupta (2003), Xu and Wu (2001, 2005) and Xu (2003) are worth mentioning.

While constructing a supersaturated design, one must ensure that the design has a high probability of correct identification of both active and inactive factors. Srivastava (1975) gave a necessary and sufficient condition for a plan to search and estimate the true non-negligible effects under the noiseless case. Following Shirakura et al. (1996), Chatterjee et al. (2008) introduced the concept of searching probabilities for two-level supersaturated designs. The stochastic behavior of supersaturated designs are also studied by authors like Chen and Lin (1998), Allen and Bernshteyn (2003) and Sarkar (2007).

The present work studies the stochastic properties of multi-level supersaturated designs in terms of the probability of correct model identification. A new criterion, based on the lower bound of the probability of correct model identification, is proposed for comparing two competitive balanced supersaturated designs for the symmetric factorials. Furthermore, using genetic algorithm, an algorithm is proposed for constructing supersaturated designs for general symmetric and asymmetric factorial experiments. It aims at maximizing the probability of correct model identification of the design. Genetic algorithms are evolutionary search strategies based on simplified rules, biological population genetics and theories of evolution (see, for example, Gen and Cheng (2000)). For optimization purpose, genetic algorithms are attractive not only because they are relatively easy to implement, but also because they do not require differentiable objective functions (even though they can be applied in optimizing stochastic objective functions).

The paper is organized as follows. Section 2 discuses the notations and preliminaries and introduces the concept of probability of correct model identification for multi-level supersaturated designs. Section 3 proposes a new criterion for

* Corresponding author. Tel.: +1 814 8650377; fax: +1 814 8637067.
E-mail address: DKL5@psu.edu (D.K.J. Lin).

0167-7152/$ – see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.spl.2009.01.017

comparing balanced supersaturated designs, for the symmetric factorial experiments, on the basis of the lower bound of the probability of correct model identification. Using genetic algorithm, Section 4 presents the construction algorithm of supersaturated designs for the general symmetric and asymmetric factorials. Finally, Section 5 presents some concluding remarks.

2. Notations and preliminaries

Consider a factorial experiment with n factors $F_1, F_2, \ldots, F_n$ at the levels $m_1, m_2, \ldots, m_n (\geq 2)$ respectively. For $1 \leq j \leq n$, the levels of F_j are coded as $0, 1, \ldots, m_j - 1$. A typical level combination is denoted as $a_1 a_2 \ldots a_n$, $0 \leq a_j \leq m_j - 1$, $1 \leq j \leq n$. Under *factor sparsity* situation, we assume that only k out of n factors are active.

For $1 \leq j \leq n$, let 1_j be the $m_j \times 1$ vector with all elements unity, $I^*_{m_j - 1}$ be the identity matrix of order $m_j - 1$, and $P_j = [p_j(0), \ldots, p_j(m_j - 1)]$ be an $(m_j - 1) \times m_j$ matrix such that

$$P_j P_j' = I^*_{mj-1}, \quad P_j 1_j = 0,$$

where 0 is a null vector of appropriate order. Furthermore, let Z_j be the $N \times (m_j - 1)$ matrix with rows $\sqrt{\frac{m_j}{v}} p_j(a_j)'$, where $v = \prod m_j$. For $1 \leq k \leq n$, let $H(k)$ be the collection of all sets of k factors and v be the cardinality of $H(k)$. For any $h \in H(k)$, let $M(h)$ be the model consisting of the general mean and the main effects of the k factors belonging to h.

Let $\mathcal{D}(m_1 \times m_2 \times \cdots \times m_n)$ be the class of N-run supersaturated designs corresponding to a $m_1 \times m_2 \times \cdots \times m_n$ factorial experiment and let d be a design in $\mathcal{D}(m_1 \times m_2 \times \cdots \times m_n)$. Then the linear model corresponding to d is given by

$$y = b_N \mu + Z_1 \theta_1 + \cdots + Z_n \theta_n + \epsilon, \tag{1}$$

where y is the $N \times 1$ vector of observations, b_N is the $N \times 1$ vector with each element $N^{-1/2}$, μ is the general mean effect, θ_j, $1 \leq j \leq n$, is the $(m_j - 1) \times 1$ vector of unknown main effect contrasts corresponding to the factor F_j and ϵ is the $N \times 1$ vector of random error components. The components of ϵ are assumed to be independently and identically distributed with mean 0 and variance σ^2. It is to be noted that under the model $M(h)$, Z_i, $1 \leq i \leq n$, will be included in (1) provided the factor F_i is included in h.

For any $h \in H(k)$, let h include the k factors $F_{i_1}, F_{i_2}, \ldots, F_{i_k}$, $1 \leq i_1 < i_2 < \cdots < i_k \leq n$. Then under $M(h)$, (1) reduces to

$$y = b_N \mu + Z_{i_1} \theta_{i_1} + \cdots + Z_{i_k} \theta_{i_k} + \epsilon. \tag{2}$$

The sum of squares due to error corresponding to the model $M(h)$, denoted by $S(h)^2$, can be obtained as

$$S(h)^2 = y'(I_N^* - Q_h) y, \tag{3}$$

where

$$B_h = [\; a_N \quad Z_{i_1} \quad \ldots \quad Z_{i_k} \;] \quad \text{and} \quad Q_h = B_h (B_h' B_h)^{-1} B_h'.$$

Following Srivastava (1975), choose $M(h)$ as the true model if and only if

$$S(h)^2 = \min_{h^* \in H(k)} S(h^*)^2.$$

Definition 1. The probability of correct identification of a model through a design $d \in \mathcal{D}(m_1 \times m_2 \times \cdots \times m_n)$ is the expected probability of identifying a true model where the expectation is taken over all possible models.

Corresponding to $h_1, h_2 \in H(k)$, $h_1 \neq h_2$, define an event $E_{h_1 h_2}$ as

$$E_{h_1 h_2} = (S(h_1)^2 \leq S(h_2)^2).$$

Following the searching procedure proposed by Srivastava (1975), probability of identifying $M(h_1)$ as the true model with the help of the design d will then be given by

$$\mathrm{P}_{h_1}(d) = \mathrm{P}\left(\bigcap_{h_2 (\neq h_1) \in H(k)} E_{h_1 h_2} | M(h_1) \right), \tag{4}$$

Thus, the probability of correct identification of the true model through the design d, assuming that all possible v models are equally likely to be the true model, is given by

$$\mathrm{P}(d) = \frac{1}{v} \sum_{h_1 \in H(k)} \mathrm{P}_{h_1}(d). \tag{5}$$

For any $h \in H(k)$, define $L_h = [Z_{i_1} \ldots Z_{i_k}]$ and $u_h = y' L_h (L_h' L_h)^{-1} L_h' y$. Then following Shirakura et al. (1996), we can simplify $\mathrm{P}(E_{h_1 h_2} | M(h_1))$ as

$$\mathrm{P}(E_{h_1 h_2} | M(h_1)) = \mathrm{P}(u_{h_1} > u_{h_2} | M(h_1)), \quad \text{where } h_1, h_2 \in H(k), h_1 \neq h_2. \tag{6}$$

Assuming L_h to be of full column rank, we can find a non-singular matrix R_h such that $(L_h' L_h)^{-1} = R_h R_h'$.

Define $z_{h_1} = R'_{h_1} L'_{h_1} y$ and $z_{h_2} = R'_{h_2} L'_{h_2} y$. It is to be noted that, for $i = 1, 2$, z_{h_i} is a random vector of order $r_i \times 1$, where r_i is the number of columns of L_{h_i}. The following lemma will be helpful in proving Theorem 1.

Lemma 1. *For any $h_1, h_2 \in H(k)$, $h_1 \neq h_2$*

$$P(E_{h_1h_2}|M(h_1)) = P\left(\sum_{\alpha=1}^{r_1} z^2_{h_1\alpha} > \sum_{\alpha=1}^{r_2} z^2_{h_2\alpha}\right), \tag{7}$$

where, $z_{h_1} = ((z_{h_1\alpha})) \sim N_{r_1}((R_{h_1})^{-1}\theta_{h_1}, \sigma^2 I^*_{r_1})$, $z_{h_2} = ((z_{h_2\alpha})) \sim N_{r_2}(R'_{h_2}L'_{h_2}L_{h_1}\theta_{h_1}, \sigma^2 I^*_{r_2})$. Moreover, here we have $\mathrm{cov}(z_{h_1}, z_{h_2}) = \sigma^2 R'_{h_1} L'_{h_1} L_{h_2} R_{h_2}$.

3. The case $k = 1$ and symmetric factorial set-up

This section considers the case $k = 1$. Suppose $m_1 = \cdots = m_n = m$. For $1 \leq i \leq n$, let us write the matrix L_i as $L_i = [l_1^{(i)} \ldots l_{m-1}^{(i)}]$ and the vector θ_i as $\theta_i = [\theta_{i1} \ldots \theta_{i(m-1)}]$. Then, for any $h \in H(1)$ and for a column balanced supersaturated design d, the matrix R_h is given by $R_h = \mathrm{diag}(\sqrt{\nu/N}, \ldots, \sqrt{\nu/N})$, where $\nu = m^n$. It is to be noted that, for $k = 1$, the cardinality of the set $H(1)$ is n. The following corollary follows immediately from Lemma 1.

Corollary 1. *For $1 \leq \alpha \leq m-1$, the random variables $z_{h_1\alpha}$ and $z_{h_2\alpha}$ have the normal distributions $N(\mu_{h_1\alpha}, \sigma^2)$ and $N(\mu_{h_2\alpha}, \sigma^2)$, respectively, where*

$$\mu_{h_1\alpha} = (\sqrt{N/\nu})\theta_{h_1\alpha}, \qquad \mu_{h_2\alpha} = (\sqrt{\nu/N}) \sum_{\alpha'=1}^{m-1} l_\alpha^{(h_2)'} l_{\alpha'}^{(h_1)} \theta_{h_1\alpha'},$$

and $\mathrm{cov}(z_{h_1\alpha}, z_{h_2\alpha}) = ((\nu/N) l_\alpha^{(h_1)'} l_\alpha^{(h_2)})\sigma^2$.

For $h_1, h_2 \in H(1)$, $h_1 \neq h_2$, $1 \leq \alpha \leq m-1$, if we define the event $B_\alpha^{(h_1h_2)} = \{z^2_{h_1\alpha} > z^2_{h_2\alpha}\}$, then it is easy to note that $\bigcap_{\alpha=1}^{(m-1)} B_\alpha^{(h_1h_2)} \Rightarrow E_{h_1h_2}$. The main result is presented below.

Theorem 1. *For any design $d \in \mathcal{D}(m^n)$,*

$$P(d) \geq \frac{1}{n} \sum_{h_1=1}^{n} \sum_{h_2(\neq h_1)=1}^{n} \sum_{\alpha=1}^{m-1} p_\alpha^{(h_1h_2)} - (n-1)(m-2) - (n-2),$$

where

$$p_\alpha^{(h_1h_2)} = 1 - \Phi\left(\frac{\eta_{x_1\alpha}}{\sigma_{x_1\alpha}}\right) - \Phi\left(\frac{\eta_{x_2\alpha}}{\sigma_{x_2\alpha}}\right) + 2\Phi\left(\frac{\eta_{x_1\alpha}}{\sigma_{x_1\alpha}}\right)\Phi\left(\frac{\eta_{x_2\alpha}}{\sigma_{x_2\alpha}}\right), \quad \textit{with}$$

$$\eta_{x_1\alpha} = \sqrt{\frac{N}{\nu}}\left(\theta_{h_1\alpha} + \frac{\nu}{N}\sum_{\alpha'=1}^{m-1} l_\alpha^{(h_2)'} l_{\alpha'}^{(h_1)} \theta_{h_1\alpha'}\right), \qquad \sigma^2_{x_1\alpha} = 2\sigma^2\left(1 + \frac{\nu}{N} l_\alpha^{(h_2)'} l_\alpha^{(h_1)}\right)$$

$$\eta_{x_2\alpha} = \sqrt{\frac{N}{\nu}}\left(\theta_{h_1\alpha} - \frac{\nu}{N}\sum_{\alpha'=1}^{m-1} l_\alpha^{(h_2)'} l_{\alpha'}^{(h_1)} \theta_{i\alpha'}\right), \quad \textit{and} \quad \sigma^2_{x_2\alpha} = 2\sigma^2\left(1 - \frac{\nu}{N} l_\alpha^{(h_2)'} l_\alpha^{(h_1)}\right).$$

An outline of the proof of the theorem is given in Appendix.

For the purpose of comparing different designs belonging to $\mathcal{D}(m^n)$, we define the measure,

$$LP(d) = \frac{1}{n} \sum_{h_1=1}^{n} \sum_{h_2(\neq h_1)=1}^{n} \sum_{\alpha=1}^{m-1} p_\alpha^{(h_1h_2)}.$$

Definition 2. Among two designs $d_1, d_2 \in \mathcal{D}(m^n)$, d_1 will be better than d_2 in terms of probability of identifying the true model if $LP(d_1) \geq LP(d_2)$.

Consider two designs $d_1, d_2 \in \mathcal{D}(3^7)$, with 9 runs. d_1 is due to Xu and Wu (2005) and d_2 has been constructed by permuting the elements of last two columns of d_1. The comparison of performance, in terms of both $LP(d)$ and $P(d)$, of the two designs is presented in Table 1. For the sake of space complexity we only provide the values corresponding to $\theta_{i\alpha}/\sigma = \rho$, $1 \leq \alpha \leq m-1$ and $1 \leq i \leq n$. It is evident from Table 1 that $LP(d)$ can be used as a good surrogate of $P(d)$ for comparing different competitive designs in the same class for $k = 1$. The computation procedure of $P(d)$ is given in the next section. In the above calculation we have set $\sigma_1 = 5$ (as will be mentioned in the next section).

Table 1
Comparison of d_1 and d_2 in terms of $LP(d)$ and $P(d)$ for $k = 1$.

Design		$\rho = 30$	$\rho = 35$	$\rho = 40$	$\rho = 45$
d_1	$LP(d_1)$	10.9915	11.6322	12.1224	12.4812
	$P(d_1)$	0.6843	0.7167	0.7423	0.7816
d_2	$LP(d_2)$	10.5331	11.0998	11.5278	11.8340
	$P(d_2)$	0.6783	0.6999	0.7288	0.7696

The next section considers the case $k > 1$ and provides computation of probability of correct model identification through simulation. Moreover, construction of supersaturated designs through genetic algorithm is presented in the next section.

4. A construction algorithm for general $k (k > 1)$

4.1. Computation of probability of correct model identification using simulation

The following notations will be helpful in developing the computational procedure of the probability of correct model identification through simulation.

For $1 \le i \le n$, define

$$\begin{aligned} z_i &= 1 \quad \text{if } i\text{th factor is active,} \\ &= 0 \quad \text{otherwise.} \end{aligned} \tag{8}$$

In this present context exactly k of the z_i's are unity. Following Allen and Bernshteyn (2003), let us make the following distributional assumptions

$$\theta_{ij}|z_i \sim \begin{cases} \mathrm{N}(0, \sigma_1^2) & \text{if } z_i = 0, \\ \xi\,\mathrm{sign}(\zeta) + \zeta, \text{ where } \zeta \sim N(0, \sigma_1^2) & \text{if } z_i = 1. \end{cases} \tag{9}$$

The threshold value ξ is considered to make the probability of false discovery a minimum. For simplicity, we assume that $\epsilon \sim N(0, 1)$.

For a particular model $M(h)$, we generate the values of θ_{ij}'s from the distribution mentioned above for a particular choice of the parameters. Subsequently, N values of y are generated and the error sum of squares of all possible models is calculated. We repeat the above procedure 20,000 times and the proportion of times the error sum of squares due to model $M(h)$ is less than the error sum of squares due to the other models is calculated. This proportion gives an estimate of the probability of correct identification of the model $M(h)$. The above procedure is repeated for all the models to obtain the probability of correct identification of the given design. It is to be noted that when the matrix $L_h'L_h$ is singular, the problem of singularity is tackled by setting few effects of few factors at zero level and replacing $(L_h'L_h)^{-1}$ by a g-inverse of $L_h'L_h$.

4.2. The construction algorithm

Bernshteyn (2001) considered the construction of supersaturated designs using genetic algorithm. Following Bernshteyn (2001), we use the Confidence Interval Elitist Genetic Algorithm (CIEGA) for constructing general symmetric and asymmetric supersaturated designs with the objective of maximizing the probability of correct model identification given by $P(d)$. The main activities of the elitist version of genetic algorithm is to clone (or copy) the chromosomes having higher fittest values in a particular generation to the next generation. The goal of this procedure is to ensure that the value of the objective function will improve in every subsequent generation. Moreover, the identification of correct elitist subset minimizes the computational effort of the algorithm. Allen and Bernshteyn (2003) constructed two-level supersaturated designs maximizing the coverage probability using CIEGA with realistic assumptions and obtained some good results.

Let T be the set of all possible level combinations lexicographically ordered and $I = \{1, \ldots, \nu\}$, where, $\nu = m_1 \times \cdots \times m_n$. It is easy to note that there is a one-to-one correspondence between T and the set I. Suppose, in the ith generation we have r different chromosomes $C_1^i, C_2^i, \ldots, C_r^i$, consisting of N genes. A typical chromosome is actually a collection of N integers from the set $I = \{1, \ldots, \nu\}$. The integers in a chromosome denote the treatment combinations of the set T to be included in the plan.

For example if we consider a 3^3 factorial experiment, then the following provides a one-to-one correspondence between the set T and the set I.

000	$\leftrightarrow$	1	001	$\leftrightarrow$	2	002	$\leftrightarrow$	3	010	$\leftrightarrow$	4	011	$\leftrightarrow$	5	012	$\leftrightarrow$	6
020	$\leftrightarrow$	7	021	$\leftrightarrow$	8	022	$\leftrightarrow$	9	100	$\leftrightarrow$	10	101	$\leftrightarrow$	11	102	$\leftrightarrow$	12
110	$\leftrightarrow$	13	111	$\leftrightarrow$	14	112	$\leftrightarrow$	15	120	$\leftrightarrow$	16	121	$\leftrightarrow$	17	122	$\leftrightarrow$	18
200	$\leftrightarrow$	19	201	$\leftrightarrow$	20	202	$\leftrightarrow$	21	210	$\leftrightarrow$	22	211	$\leftrightarrow$	23	212	$\leftrightarrow$	24
220	$\leftrightarrow$	25	221	$\leftrightarrow$	26	222	$\leftrightarrow$	27									

Table 2
A 3^7 supersaturated design with 9 runs.

0	0	1	1	1	1	2	2	2
0	1	0	0	1	2	0	0	0
2	1	0	2	0	2	0	0	1
0	0	1	2	0	0	0	2	1
2	1	1	2	2	0	2	0	0
1	0	2	1	1	1	0	0	1
0	2	1	0	0	0	1	2	0

A typical chromosome and the corresponding design of a 3^3 factorial experiment for $N = 6$ is shown below.

Typical chromosome						Plan (Columns as runs)					
						0	0	1	1	2	2
1	5	12	16	23	27	0	1	0	2	1	2
						0	1	2	0	1	2

From the above mentioned r chromosomes $C_1^i, C_2^i, \ldots, C_r^i$, we have to select the elitist subset that is those chromosomes which will be cloned in the next generation. This screening procedure is to be performed in different stages. It is to be noted that the inferior chromosomes should be screened with a few samples in the earlier stages, while the comparisons of competitive chromosomes require more stages. Thus screening in stages reduces the simulation effort. Let n_{ij}^{γ} be the number of samples to be drawn in the jth stage of the ith generation corresponding to the γth gene. The values of n_{ij}^{γ} are sufficiently large and grow with stages. Let $\overline{P_{\gamma}^{ij}}$ and $\overline{S_{\gamma}^{ij}}$ be respectively the mean and standard deviation of the probability of correct model identification of C_{γ}^i at the jth stage in the ith generation, $1 \le \gamma \le r$. On the basis of $\overline{P_{\gamma}^{ij}}$ and $\overline{S_{\gamma}^{ij}}$ we construct a confidence interval for γth chromosome as $(L_{\gamma}^{ij}, U_{\gamma}^{ij})$, where

$$L_{\gamma}^{ij} = \overline{P_{\gamma}^{ij}} - t_{[1-(1-\alpha)^{1/r}]/2,deg} \frac{S^{ij}}{\sqrt{n_{ij}^{\gamma}}},$$

$$U_{\gamma}^{ij} = \overline{P_{\gamma}^{ij}} + t_{[1-(1-\alpha)^{1/r}]/2,deg} \frac{S^{ij}}{\sqrt{n_{ij}^{\gamma}}},$$

S^{ij} is the pooled estimate of standard deviation, assuming the population standard deviations of chromosomes are all equal and deg stands for the degrees of freedom of S^{ij}. Define $M^{ij} = \max_{\gamma} L_{\gamma}^{ij}$ and construct the set $F^{ij} = \{C_{\gamma}^i | U_{\gamma}^{ij} > M^{ij}\}$. Then F^{ij} is the desired set of elitist chromosomes. In our implementation, the cardinality of F^{ij} decreases with an increase in n_{ij}^{γ}. After having a desired number of elitist chromosomes we apply the genetic operation, crossover and mutation, among them to generate a new generation. We continue the whole procedure till there is no significant achievement in terms of probability.

The schema of our genetic algorithm is as follows.

1. Create an initial generation.
2. Apply the screening operations in different stages to select the desired number of elitist chromosomes.
3. Reproduce to construct a new generation by applying the genetic operations of crossover and mutation to the elitist subset.
   (a) Crossover, that is, a pair of chromosomes is split at a random position and the head of one is combined with the tail of other and vice-versa.
   (b) Mutation, that is, the state of a randomly chosen gene (number within a chromosome) is changed. This helps the search avoid being trapped into local optima.
4. Repeat 2 and 3 until some convergence criterion is met or some pre-assigned number of generations have passed.

We used the above algorithm to construct the designs in Tables 2 and 4.

Table 2 gives a 3^7 design with 9 runs, where the columns are considered as runs. The design is constructed under the choice $\sigma_1 = 1$, $\xi = 1$ and $k = 4$. It is evident from the table that the newly constructed design outperforms in all cases. A comparison of the efficiency, in terms of the probability of correct model identification, of this design with that of a 3^7 design with 9 runs constructed by Theorem 4 of Xu and Wu (2005) is presented in Table 3. These probabilities are calculated for different choices of σ_1 and ξ. Next we consider the construction of a mixed level design. Table 4 presents a $2 \times 3 \times 4 \times 3 \times 2$ design with 10 runs generated by the above algorithm. This design is constructed by maximizing the probability of correct model identification for at most two active factors, that is, $k = 2$.

Table 3
Comparison in terms of probability of correct model identification of design in Table 2 (the first row) with the 9 runs design given by Xu and Wu (2005) corresponding to 3^7 factorial experiments (the second row).

σ_1	ξ					
	1	1.5	2	2.5	3	3.5
1	0.4615	0.6034	0.7337	0.8395	0.9241	0.9802
	0.3830	0.5389	0.6801	0.7943	0.8704	0.9225
1.5	0.3763	0.4720	0.5664	0.6615	0.7483	0.8190
	0.2871	0.3863	0.4925	0.6022	0.6887	0.7686
2	0.3438	0.4041	0.4742	0.5440	0.6146	0.6852
	0.2449	0.3143	0.3870	0.4712	0.5509	0.6228

Table 4
A $2 \times 3 \times 4 \times 3 \times 2$ supersaturated design with 10 runs.

0	0	0	0	0	1	1	1	1	1
0	0	0	1	2	0	1	1	1	2
2	3	3	1	3	0	1	1	3	0
1	1	2	0	1	2	0	0	1	2
1	0	0	1	0	1	0	1	0	1

Remark 1. It is to be remarked that the probability of correct model identification increases with ξ and decreases with σ_1. The increase in the values of the probability with that of ξ can be attributed as logical, as ξ increases, the supposedly important effects has larger magnitude than the supposedly unimportant effects. So they can be easily identified. The diminishing gain in probability with the increase in σ_1 also can be explained by the fact, as the values of σ_1 increases, the absolute values of the coefficients of negligible effects increases causing the difficulties in identifying the true non-negligible effects.

5. Conclusion

In this present work, the probability of correct model identification of supersaturated designs has been thoroughly studied. We have constructed multi-level supersaturated designs using the elitist version of the Genetic Algorithm by maximizing the probability of correct model identification. These designs are considered as superior to other existing designs in the light of having higher probability of correct model identification. Moreover, in the literature it is hard to see any guideline for construction of a general mixed level supersaturated design. Although Fang et al. (2003) gave a construction procedure of mixed level supersaturated designs, their plan is also of a very special type. This paper provides a meaningful guideline for constructing general mixed level supersaturated designs which is another important achievement of the work.

Appendix. Proof of Theorem 1

From (4), (5) and (7), we get

$$\mathrm{P}(d) \geq \frac{1}{n} \sum_{h_1=1}^{n} \sum_{h_2(\neq h_1)=1}^{n} \sum_{\alpha=1}^{m-1} \mathrm{P}(B_\alpha^{(h_1 h_2)} | M(h_1)) - (n-1)(m-2) - (n-2). \tag{10}$$

Now, let

$$\begin{aligned} p_\alpha^{(h_1 h_2)} &= \mathrm{P}(B_\alpha^{(h_1 h_2)} | M(h_1)) = \mathrm{P}(z_{h_1\alpha}^2 > z_{h_2\alpha}^2 | M(h_1)) = \mathrm{P}(x_{1\alpha} x_{2\alpha} > 0) \\ &= 1 - \Phi\left(\frac{\eta_{x_{1\alpha}}}{\sigma_{x_{1\alpha}}}\right) - \Phi\left(\frac{\eta_{x_{2\alpha}}}{\sigma_{x_{2\alpha}}}\right) + 2\Phi\left(\frac{\eta_{x_{1\alpha}}}{\sigma_{x_{1\alpha}}}\right)\Phi\left(\frac{\eta_{x_{2\alpha}}}{\sigma_{x_{2\alpha}}}\right); \end{aligned} \tag{11}$$

where $x_{1\alpha} = z_{h_1\alpha} + z_{h_2\alpha} \sim N(\eta_{x_{1\alpha}}, \sigma^2_{x_{1\alpha}})$ and $x_{2\alpha} = z_{h_1\alpha} - z_{h_2\alpha} \sim N(\eta_{x_{2\alpha}}, \sigma^2_{x_{2\alpha}})$. The proof of Theorem 1 thus follows. □

References

Allen, T.T., Bernshteyn, M., 2003. Supersaturated designs that maximize the probability of identifying active factors. Technometrics 45, 90–97.
Bernshteyn, M., 2001. Simulation optimization methods that combine multiple comparisons and genetic algorithms with applications in design for computer and supersaturated experiments, Doctoral Dissertation, Ohio State University.
Box, G.E.P., Meyer, R.D., 1986. An analysis for unreplicated fractional factorials. Technometrics 28, 1118.
Chatterjee, K., Gupta, S., 2003. Construction of supersaturated designs involving s-level factors. J. Statist. Plann. Inference 113, 589–595.
Chatterjee, K., Sarkar, A., Lin, D.K.J., 2008. Supersaturated design with high searching probability. J. Statist. Plann. Inference 138, 272–277.
Chen, J., Lin, D.K.J., 1998. On the identifiability of a supersaturated design. J. Statist. Plann. Inference 72, 99–107.
Cheng, C.S., 1997. E(s^2)-optimal supersaturated designs. Statist. Sinica 7, 929–939.
Deng, L.Y., Lin, D.K.J., Wang, J., 1999. A resolution rank criterion for supersaturated designs. Statist. Sinica 9, 605–610.

Fang, K.T., Lin, D.K.J., Ma, C.X., 2000. On the construction of multi-level supersaturated designs. J. Statist. Plann. Inference 86, 239–252.
Fang, K.T., Lin, D.K.J., Liu, M.Q., 2003. Optimal mixed-level supersaturated designs and computer experiment. Metrika 58, 279–291.
Fang, K.T., Gennian, G.E., Liu, M.Q., 2004. Construction of optimal supersaturated design by the packing method. Sci. China Ser. A 47 (1), 128–143.
Gen, M., Cheng, R., 2000. Genetic Algorithm & Engineering Optimization. John Wiley & Sons Inc., New York, NY.
Lin, D.K.J., 1992. Systematic supersaturated design. Technical Report, University of Tennessee.
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.
Lin, D.K.J., 1998. Spotlight interaction effects in 'Main-Effect' plans: A supersaturated design approach. Qual. Eng. 11 (1), 133–139.
Lin, D.K.J., 2000. Supersaturated design: Theory and application. In: Park, Sung H., Vining, G.G. (Eds.), Statistical Process Monitoring and Optimization. Marcel Dekker (Chapter 18).
Liu, Y., Dean, A.M., 2004. K circulant supersaturated designs. Technometrics 46, 32–43.
Nguyen, N.K., 1996. An algorithmic approach to constructing supersaturated designs. Technometrics 38, 69–73.
Sarkar, A., 2007. Probability optimal supersaturated design. Calcutta Statist. Assoc. Bull. 59, 237–255.
Satterthwaite, F., 1959. Random balance experimentation. Technometrics 1, 111–137 (with discussion).
Shirakura, T., Tkahashi, T., Srivastatava, J.N., 1996. Searching probabilities for non zero effects in search designs for the noisy case. Ann. Statist. 24 (6), 2560–2568.
Srivastava, J.N., 1975. Designs for searching non negligible effects. In: Srivastava, J.N. (Ed.), A Survey of Statistical Design and Linear Models. North-Holland, Amsterdam, pp. 507–519.
Wu, C.F.J., 1993. Construction of supersaturated designs through partially aliased interactions. Biometrika 80, 661–669.
Xu, H., 2003. Minimum moment aberration for non regular designs and supersaturated designs. Statist. Sinica 13, 691–708.
Xu, H., Wu, C.F.J., 2001. Generalized minimum aberration for asymmetrical fractional factorial designs. Ann. Statist. 29, 1066–1077.
Xu, H., Wu, C.F.J., 2005. Construction of optimal multi-level supersaturated design. Ann. Statist. 33 (6), 2811–2836.
Yamada, S., Lin, D.K.J., 1999. Three-level supersaturated design. Statist. Probab. Lett. 45, 31–39.
Yamada, S., Lin, D.K.J., 2002. Construction of mixed-Level supersaturated designs. Metrika 56, 205–214.