



ELSEVIER

Available online at www.sciencedirect.com

Journal of Statistical Planning and Inference 138 (2008) 272–277

 journal of
 statistical planning
 and inference

www.elsevier.com/locate/jspi

Supersaturated designs with high searching probability

Kashinath Chatterjee^a, Angshuman Sarkar^a, Dennis K.J. Lin^{b,*}

^aDepartment of Statistics, Visva-Bharati University, Santiniketan, Bolpur, West Bengal, India

^bDepartment of Supply Chain and Information Systems, The Pennsylvania State University, University Park, PA 16802, USA

Available online 16 May 2007

Abstract

A supersaturated design is essentially a fractional factorial design whose number of experimental variables is greater than or equal to its number of experimental runs. Under the effect sparsity assumption, a supersaturated design can be very cost-effective. In this paper, our prime objective is to compare the existing two-level supersaturated designs for the noisy case through the probability of correct searching—a powerful criterion proposed by Shirakura et al. [1996. Searching probabilities for nonzero effects in search designs for the noisy case. *Ann. Statist.* 24, 2560–2568]. An algorithm is proposed to construct supersaturated designs with high probability of correct searching. Examples are given for illustration.

© 2007 Elsevier B.V. All rights reserved.

Keywords: Design matrix; Effect sparsity; Screening experiment

1. Introduction

Many preliminary studies in industrial experimentation involve a large-scale system with a large number of factors during design and operation stages. The cost of conducting such a large-scale system can be prohibitively expensive. Moreover, it is quite common that only a few of these factors have a *substantial effect*—a situation known as *effect sparsity* (see, for example, Box and Meyer, 1986). To address such a challenge posed by this technological trend, extensive research has focused on the construction of supersaturated designs from the viewpoint of run size economy and mathematical novelty. Specifically, a supersaturated design is essentially a fractional factorial design whose number of experimental variables is greater than or equal to its number of experimental runs. Under the effect sparsity assumption, a supersaturated design can be very cost-effective.

The construction of supersaturated designs dates back to Satterthwaite (1959) and Booth and Cox (1962). This area remained dormant, until the appearance of Lin (1991, 1993), and has gained increasing attention evidenced by the growth of the recent literature. Since then, many researchers have considered the construction and the properties of two-level supersaturated designs. See, for example, Wu (1993), Lin (1995, 1998), Nguyen (1996) and Cheng (1997), among others. More than two-level supersaturated designs can be found in Yamada and Lin (1999, 2002), Fang et al. (2000, 2003), and Chatterjee and Gupta (2003). Algorithms for constructing supersaturated designs have also been thoroughly studied (see, for example, Lin, 1995; Nguyen, 1996; Li and Wu, 1997; Tang and Wu, 1997; Yamada and Lin, 1997; Deng et al., 1999; Fang et al., 2000; Liu and Dean, 2004) and many more. While constructing a supersaturated

* Corresponding author. Tel.: +1 814 865 0377; fax: +1 814 863 7067.

E-mail address: dk15@psu.edu (D.K.J. Lin).

design, one must be sure that the design should have an ability of identifying both active and inactive factors with high probabilities. In this direction, mention may be made of [Chen and Lin \(1998\)](#).

In this paper we compare available supersaturated designs from the point of view of *Searching Probabilities* along the line of [Shirakura et al. \(1996\)](#). The paper is organized as follows. Section 2 gives notations and preliminaries. In Section 3 searching probabilities for one nonzero effect are discussed. Section 4 deals with the probability of correct searching for the available supersaturated designs that are capable of searching at most two nonzero effects. An algorithm for construction of supersaturated designs with high searching probability is proposed in Section 5. A brief conclusion is given in Section 6.

2. Notations and preliminaries

Throughout this paper we consider a first-order model, i.e., a model which includes only the main effects ([Lin, 2000](#)). Consider a design d consisting of m factors each at two (high and low) levels and n treatment combinations or runs ($m > n$). Then d can be represented by a *design matrix* or *treatment matrix*, T , whose rows represent the n treatment combinations to be observed and whose columns designate the m factors to be examined. The (i, j) th element of T determines the level at which the factor j is to be observed in the i th treatment combination. Here we code the two levels (high and low) of the factors as $+1$ and -1 in T . Suppose prior knowledge is available regarding the non-negligibility (i.e., active) of at most k main effects, where k is small compared to m . Of course, it is not known which of the k main effects are nonzero.

Under the assumptions stated above and following [Srivastava \(1975\)](#), we have the search linear model

$$\mathbf{y} = \mathbf{1}\mu + T\xi_2 + \mathbf{e}, \quad V(\mathbf{e}) = \sigma^2 I, \tag{1}$$

where \mathbf{y} is the $n \times 1$ observational vector, μ is the general mean, ξ_2 is the $m \times 1$ vectors of main effects, \mathbf{e} is the $n \times 1$ error vector and I is the identity matrix of order n . [Srivastava \(1975\)](#) gave the following fundamental formulation of the search problem.

Theorem 1. *Under the model (1) with $\sigma^2 = 0$, a necessary and sufficient condition that the problem stated above can be completely solved is that for every submatrix T_1^* of order $n \times 2k$ of T ,*

$$\text{rank}[\mathbf{1} \ T_1^*] = 1 + 2k \tag{2}$$

holds.

Obviously, $n \geq 1 + 2k$. It is to be noted that condition (2) remains necessary for the case $\sigma^2 > 0$. Now consider a supersaturated design d with n runs. Suppose ζ is a $k \times 1$ vector of nonzero effects of ξ_2 . Then model (1) reduces to

$$\mathbf{y} = \mathbf{1}\mu + T_1(\zeta)\zeta + \mathbf{e}, \quad V(\mathbf{e}) = \sigma^2 I, \tag{3}$$

where $T_1(\zeta)$ is the $n \times k$ submatrix of T corresponding to ζ ($\subset \xi_2$). Following [Shirakura et al. \(1996\)](#), the sum squares error (SSE) for the model (3), denoted by $s(\zeta)^2$, can be obtained as

$$s(\zeta)^2 = \mathbf{y}'(I - Q(\zeta))\mathbf{y}, \tag{4}$$

where $Q(\zeta) = A(\zeta)[A(\zeta)'A(\zeta)]^{-1}A(\zeta)'$, and $A(\zeta) = [\mathbf{1} \ T(\zeta)]$.

[Srivastava \(1975\)](#) proposed the following procedure for searching out the nonzero effects of ξ_2 :

Step 1: For an observation vector \mathbf{y} , calculate $s(\zeta)^2$ for all possible $k \times 1$ vectors ζ of ξ_2 .

Step 2: Find a vector, say ζ^* , for which $s(\zeta)^2$ is minimized. Take ζ^* as the possibly nonzero vector of ξ_2 .

It is to be noted that $s(\zeta)^2$ can be expressed as

$$s(\zeta)^2 = \mathbf{y}'(I - Q)\mathbf{y} - h(\zeta, \mathbf{y}), \tag{5}$$

where $Q = (1/n)J_n$, J_n is an $n \times n$ matrix with all elements unity and $h(\zeta, \mathbf{y}) = \mathbf{y}'(Q(\zeta) - Q)\mathbf{y}$.

Now suppose the true model (1) to be

$$\mathbf{y} = \mathbf{1}\mu + T_1(\zeta_0)\zeta_0 + \mathbf{e}, \quad V(\mathbf{e}) = \sigma^2 I, \tag{6}$$

where ζ_0 is the nonzero $k \times 1$ vector of ξ_2 . In view of the above discussions, it is desirable that if $\zeta^* (\subset \xi_2)$ maximizes $h(\zeta, \mathbf{y})$, then $\zeta^* = \zeta_0$ exactly. However, this is not ensured with non-zero σ^2 . Hence we calculate $P(d)$, the probability of correct searching, as

$$P(d) = \min_{\zeta_0 \subset \xi_2} \min_{\zeta \in \mathcal{A}(\xi_2, \zeta_0)} P(h(\zeta_0, \mathbf{y}) > h(\zeta, \mathbf{y})), \tag{7}$$

where $\mathcal{A}(\xi_2, \zeta_0)$ denotes the set of all possible ζ of ξ_2 of which at least one element is not in ζ_0 . Shirakura et al. (1996) suggested the following for comparing two competing designs:

Let d and d^* be two competing supersaturated designs. Calculate the searching probabilities $P(d)$ and $P(d^*)$ for these designs. Then the design d will be said to perform better than the design d^* if $P(d) > P(d^*)$. The next two sections develop the searching probability of a design d with $k = 1$ and 2, respectively.

3. Searching probabilities for one nonzero effect

Consider the case $k = 1$, i.e., there is at most one nonzero main effect in ξ_2 . Then the model (3) becomes

$$\mathbf{y} = \mathbf{1}\mu + t_1(\zeta)\zeta + \mathbf{e}, \tag{8}$$

where $t_1(\zeta)$ is the $n \times 1$ column of T corresponding to the nonzero effect ζ in ξ_2 . We assume that the components of random error vector \mathbf{e} are distributed independently with a normal distribution $N(0, \sigma^2)$. The expression of $h(\zeta, \mathbf{y})$ can then be simplified to

$$h(\zeta, \mathbf{y}) = [t_1(\zeta)'(I - Q)\mathbf{y}]^2 / r(\zeta),$$

where $r(\zeta) = t_1(\zeta)'(I - Q)t_1(\zeta)$ is positive for any ζ . It is to be noted that the available supersaturated designs have the equal occurrence property that every column of the design matrix has equal number of plus and minus ones. Also the true model (6) now becomes

$$\mathbf{y} = \mathbf{1}\mu + t_1(\zeta_0)\zeta_0 + \mathbf{e}. \tag{9}$$

We have the following theorem for any design d .

Theorem 2. For any $\zeta_0 \subset \xi_2$ and $\zeta (\neq \zeta_0)$, we have

$$\begin{aligned} P(h(\zeta_0, \mathbf{y}) > h(\zeta, \mathbf{y})) &= G_d(x, \rho) \\ &= 1 - \Phi_d(\rho\sqrt{(n-x)/2}) - \Phi_d(\rho\sqrt{(n+x)/2}) + 2\Phi_d(\rho\sqrt{(n-x)/2})\Phi_d(\rho\sqrt{(n+x)/2}) \end{aligned}$$

where $x = t(\zeta_0)'t(\zeta)$, $\rho = \zeta_0/\sigma$ and Φ is the distribution function of $N(0, 1)$.

For any given design d , we get from (7)

$$P(d, \rho) = \min_{\zeta_0 \subset \xi_2} \min_{\zeta \in \mathcal{A}(\xi_2, \zeta_0)} G_d(x, \rho). \tag{10}$$

Table 1 provides the searching probabilities of the available supersaturated designs having the equal occurrence property that every column of the design matrix has equal number of plus and minus ones. The designs considered here are due to Booth and Cox (1962, for $(n, m) = (12, 16)$ and $(12, 24)$), Bulutoglu and Cheng (2004, for $(n, m) = (10, 15)$ and $(14, 19)$), Lin (1993 and 1995 for $(n, m) = (6, 10)$ and $(12, 66)$) and Liu and Dean (2004, for $(n, m) = (6, 10)$, $(8, 21)$ and $(12, 22)$).

It is evident from the above table that the design given by Lin (1995) for $n = 12, m = 66$ performs better than the designs proposed by other authors with respect to probability of correct searching and also the number of factors

Table 1
Searching probabilities for available supersaturated designs

n	m	$\rho = \zeta_0/\sigma$					
		1	1.2	1.4	1.6	1.8	2
6	10	0.9022	0.9477	0.9737	0.9875	0.9944	0.9976
8	21	0.9153	0.9537	0.9759	0.9881	0.9945	0.9977
10	15	0.9194	0.9548	0.9761	0.9882	0.9945	0.9977
12	16	0.9750	0.9915	0.9974	0.9993	0.9998	0.9999
12	22	0.9207	0.9551	0.9761	0.9882	0.9945	0.9977
12	24	0.9207	0.9551	0.9761	0.9882	0.9945	0.9977
12	66	0.9750	0.9915	0.9974	0.9993	0.9998	0.9999
14	19	0.9765	0.9917	0.9974	0.9993	0.9998	0.9999

included in the design. It is also to be noted that for $\rho(=\zeta_0/\sigma) \geq 1.6$, the searching probability of all the designs is close to unity.

4. Searching probabilities for two nonzero effects

Consider the case $k = 2$, i.e., there is at most two nonzero main effects in ξ_2 . Then model (3) becomes

$$\mathbf{y} = \mathbf{1}\mu + t_1(\zeta_1)\zeta_1 + t_2(\zeta_2)\zeta_2 + \mathbf{e}, \tag{11}$$

where $t_1(\zeta_1)$ and $t_2(\zeta_2)$ are the $n \times 1$ columns of T corresponding to the two nonzero effects ζ_1 and ζ_2 in ξ_2 . Let $\zeta = (\zeta_1, \zeta_2)$. We assume that the components of error vector \mathbf{e} are distributed independently with a normal distribution $N(0, \sigma^2)$. Also the true model (6) becomes

$$\mathbf{y} = \mathbf{1}\mu + t_1(\zeta_{10})\zeta_{10} + t_2(\zeta_{20})\zeta_{20} + \mathbf{e}. \tag{12}$$

Consider a design d and define the event

$$P(d, \zeta_0, \rho_1, \rho_2) = P\left(\bigcap_{\zeta \in \mathcal{A}(\xi_2, \zeta_0)} \{s(\zeta_0)^2 \leq s(\zeta)^2\}\right),$$

where $\zeta_0 = (\zeta_{10}, \zeta_{20})$, $\rho_1 = \zeta_{10}/\sigma$, $\rho_2 = \zeta_{20}/\sigma$, $s(\zeta_0)^2$ and $s(\zeta)^2$ are as defined in (4) and $\mathcal{A}(\xi_2, \zeta_0)$ is the set of all possible ζ in ξ_2 of which at least one element is not in ζ_0 . Following Srivastava (1975), the probability of correct searching for the design d is then given by

$$P(d, \rho_1, \rho_2) = \min_{\zeta_0 \subset \xi_2} P(d, \zeta_0, \rho_1, \rho_2), \tag{13}$$

where $\mathcal{A}(\xi_2, \zeta_0)$ denotes the set of all possible ζ of ξ_2 of which at least one element is not in ζ_0 .

We next present the probabilities given in (13) for two ($k = 2$) nonzero effects through simulation. The designs given by Booth and Cox (1962) with $(n, m) = (12, 16)$, Bulutoglu and Cheng (2004) with $(n, m) = (14, 19)$, Liu and Dean (2004) with $(n, m) = (12, 22)$ and Lin (1995) with $(n, m) = (12, 66)$ are found to satisfy the condition given in (2) with $k = 2$. The simulation procedure is described below.

Based on the above true model and for a given design d , we have generated \mathbf{y} for different values of ρ_1 and ρ_2 . For each of those \mathbf{y} , the value of $s(\zeta_0)^2$ is calculated under the true model. Using the same \mathbf{y} we have calculated the values of $s(\zeta)^2$ for all possible $\binom{m}{2} - 1$ choices of $\zeta(=\zeta_1, \zeta_2) \subset \xi_2$ assuming ζ_1 and ζ_2 as the nonzero effects. This procedure is repeated 10 000 times and the proportion of repetition in which $s(\zeta_0)^2$ is less than $s(\zeta)^2$ is recorded. This entire procedure is then repeated for $\binom{m}{2}$ choices of ζ_0 . The searching probabilities for a design is then obtained, following the definition given in (13). The results are summarized in Table 2.

The search probabilities in Table 2 are in general lower than those probabilities in Table 1. Furthermore, such search probability is higher when both ρ_1 and ρ_2 are large (greater than 1.8, say). All designs have a similar search probabilities,

Table 2
The probabilities of correct searching

	$\rho_1 = 1$	$\rho_1 = 1.2$	$\rho_1 = 1.4$	$\rho_1 = 1.6$	$\rho_1 = 1.8$	$\rho_1 = 2$
Booth and Cox (1962) with $(n, m) = (12, 16)$						
$\rho_2 = 1$	0.4823	0.5436	0.6657	0.7233	0.7744	0.8239
$\rho_2 = 1.2$	0.6156	0.6325	0.6745	0.7356	0.7761	0.8535
$\rho_2 = 1.4$	0.639	0.6576	0.6899	0.7521	0.7965	0.8963
$\rho_2 = 1.6$	0.6887	0.7254	0.7432	0.8534	0.8732	0.9542
$\rho_2 = 1.8$	0.7503	0.7651	0.8852	0.9256	0.9571	0.9751
$\rho_2 = 2$	0.7738	0.8089	0.9031	0.9371	0.9744	0.9846
Bulutoglu and Cheng (2004) with $(n, m) = (14, 19)$						
$\rho_2 = 1$	0.5355	0.6326	0.7678	0.7831	0.7901	0.7995
$\rho_2 = 1.2$	0.7025	0.7631	0.7956	0.8256	0.8654	0.8845
$\rho_2 = 1.4$	0.7554	0.8012	0.8241	0.8574	0.8957	0.9132
$\rho_2 = 1.6$	0.7965	0.8254	0.8854	0.9248	0.9541	0.9788
$\rho_2 = 1.8$	0.8254	0.8574	0.9152	0.9546	0.9785	0.9901
$\rho_2 = 2$	0.7637	0.8678	0.9345	0.9771	0.9801	0.9965
Liu and Dean (2004) with $(n, m) = (12, 22)$						
$\rho_2 = 1$	0.1637	0.2345	0.3051	0.4056	0.4301	0.4377
$\rho_2 = 1.2$	0.2468	0.3341	0.3687	0.4236	0.4965	0.5936
$\rho_2 = 1.4$	0.3561	0.3788	0.4251	0.4932	0.5967	0.6885
$\rho_2 = 1.6$	0.4894	0.5012	0.5133	0.5932	0.6789	0.7331
$\rho_2 = 1.8$	0.5324	0.5562	0.5789	0.6589	0.7521	0.7922
$\rho_2 = 2$	0.5781	0.6012	0.6321	0.7341	0.8433	0.8637
Lin (1995) with $(n, m) = (12, 66)$						
$\rho_2 = 1$	0.4300	0.5426	0.6547	0.6987	0.7781	0.8000
$\rho_2 = 1.2$	0.5912	0.6000	0.6789	0.7451	0.8100	0.8561
$\rho_2 = 1.4$	0.6891	0.7131	0.7400	0.7935	0.8565	0.9131
$\rho_2 = 1.6$	0.7121	0.7156	0.8475	0.8900	0.9200	0.9600
$\rho_2 = 1.8$	0.7535	0.7658	0.8956	0.9323	0.9900	0.9943
$\rho_2 = 2$	0.7589	0.7700	0.9600	0.9700	0.9985	1.0000

although the design of Lin (1995) with $(n, m) = (12, 66)$ is slightly higher while the design of Liu and Dean (2004) with $(n, m) = (12, 22)$ is slightly lower.

5. An algorithm for constructing supersaturated designs with high searching probability

This section provides the construction algorithm for obtaining a supersaturated design capable of searching and estimating a single factor with high probability of correct searching. The algorithm is presented below.

- (1) Consider a Hadamard matrix of order n such that $n \geq 96$ and $n = 0 \pmod{12}$. Delete the first column of all ones.
- (2) Select the last column as a branching column and consider the rows which has +1 in the branching column. So we have a resulting matrix of order $(n/2) \times (n - 2)$.
- (3) Delete the columns for which $k = 0$ and $11 \pmod{12}$ where k is the column number.

For example, when $n = 96$, we have a balanced supersaturated design of $n/2 = 48$ runs and $n - 2 = 94$ columns. According to (3) above, delete the columns 11, 12, 23, 24, 35, 36, 47, 48, 59, 60, 71, 72, 83, and 84 (a total of 14 columns) and result in a supersaturated design with 80 factors in 48 runs. In general, we have a supersaturated design of $n/2$ row (runs) with for $m = n - 2 - [(n - 2)/12] \times 2$ columns (factors). The resulting design is displayed on the website <http://www.smeal.psu.edu/faculty/dk15/ProbSSD>. The probability of correct searching is found to be close to 1.0000 for $\rho = 1$. Since the probability of correct searching is a monotone non-decreasing function of ρ , as proved in Shirakura et al. (1996), the probability of correct searching will be 1 for any value of $\rho \geq 1$.

6. Conclusion

In this paper, we have applied the probability of correct searching to the area of supersaturated designs. Some general properties, especially for $k = 1$ and 2, have been revealed. There are many existing criteria for supersaturated design. The linkages between such a search probability to other supersaturated design criteria are currently under investigation.

References

- Booth, K.H.V., Cox, D.R., 1962. Some systematic supersaturated designs. *Technometrics* 4, 489–495.
- Box, G.E.P., Meyer, R.D., 1986. An analysis for unreplicated fractional factorials. *Technometrics* 28, 11–18.
- Bulutoglu, A.D., Cheng, C.S., 2004. Construction of $E(s^2)$ -optimal supersaturated designs. *Ann. Statist.* 32 (4), 1662–1678.
- Chatterjee, K., Gupta, S., 2003. Construction of supersaturated designs involving s -level factors. *J. Statist. Plann. Inference* 113, 589–595.
- 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. *Statistica Sinica* 7, 929–939.
- Deng, L.Y., Lin, D.K.J., Wang, J., 1999. A resolution rank criterion for supersaturated designs. *Statistica 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.
- Li, W.W., Wu, C.F.J., 1997. Columnwise-pairwise algorithms with applications to the construction of supersaturated designs. *Technometrics* 39, 171–179.
- Lin, D.K.J., 1991. Systematic Supersaturated Designs. Working Paper No. 264, College of Business Administration, 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. *Quality Eng.* 11 (1), 133–139.
- Lin, D.K.J., 2000. Supersaturated design: theory and application. In: Park, S.H., Vining, G.G. (Eds.), *Statistical Process Monitoring and Optimization*. Marcel Dekker, New York. (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.
- Satterthwaite, F., 1959. Random balance experimentation. *Technometrics* 1, 111–137 (with discussion).
- Shirakura, T., Tkahashi, T., Srivastava, J.N., 1996. Searching probabilities for nonzero effects in search designs for the noisy case. *Ann. Statist.* 24, 2560–2568.
- Srivastava, J.N., 1975. Designs for searching nonnegligible effects. In: Srivastava, J.N. (Ed.), *A Survey of Statistical Design and Linear Models*. North-Holland, Amsterdam, pp. 507–519.
- Tang, B., Wu, C.F.J., 1997. A method for constructing supersaturated designs and its $E(s^2)$ optimality. *Canad. J. Statist.* 25, 191–201.
- Wu, C.F.J., 1993. Construction of supersaturated designs through partially aliased interactions. *Biometrika* 80, 661–669.
- Yamada, S., Lin, D.K.J., 1997. Supersaturated designs including orthogonal base. *Canad. J. Statist.* 25, 203–213.
- 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.