ON THE IDENTIFIABILITY
OF A SUPERSATURATED DESIGN

Jiahua Chen, University of Waterloo
and
Dennis K.J. Lin, University of Tennessee

IIQP Research Report
RR-95-01

January 1995

ON THE IDENTIFIABILITY OF A SUPERSATURATED DESIGN

Jiahua Chen
Department of Statistics and Actuarial Science
University of Waterloo
Waterloo, ON N2L 3G1 Canada

and

Dennis K.J. Lin
Department of Statistics
University of Tennessee
Knoxville, TN 37996, USA

ABSTRACT

A supersaturated design investigates k factors in only $n(< k + 1)$ experimental runs. The goal here is to identify, presumably only a few, relatively dominant effects with a minimum cost. While the construction of supersaturated designs has been widely explored, the data analysis aspect of such designs remains primitive. Obviously, if a large number of the factors under investigation have significant effects, their estimation will be masked, making it, in general, not possible to identify any individual effects. We study the following problem: How many dominant effects are allowed to assure a meaningful data analysis for a supersaturated design with the maximum correlation ρ? The results obtained support the fundamental concern of the $E(s^2)$ criterion introduced by Booth and Cox (1962). Furthermore, under normality assumption, we also obtained a lower bound of the probability that the factor with the largest estimated effect has indeed the largest true effect. This bound depends on the relative size of the large effect and the maximum correlation of the underlying design. Under some mild assumptions, we show that this probability is satisfactorily large. Consequently, by carefully constructing supersaturated designs, we not only save the cost of the experiment, but also make reliable inferences.

Key words: Alias; Bias; Correlation; Estimability; Projection.

1 Introduction

Many preliminary industrial screening experiments typically contain a large number of potentially relevant factors. Among them, only a few are believed to be active. The goal here is to identify those (relatively) few dominant active factors with the least possible number of experimental runs. A (two-level) supersaturated design is a matrix of ± 1's with n rows and typically, a large number of columns k which is much larger than n in general. Hence, it studies a large number (k) factors with a few (n) runs.

First constructed systematically by Booth and Cox (1962), supersaturated designs have received a great deal of attention in the recent literature (see, Lin, 1991, 1993, 1995; Wu, 1993; Nguyen, 1994; Deng, Lin and Wang, 1994; and Seewald, 1994). While the construction of supersaturated designs has been widely explored, the data analysis aspect of such designs remains primitive. This is the major interest of this paper.

Naturally, such a design will not allow us to estimate main effects of all k factors. This is, however, not always needed if we believe that only a small number (p) of them are active. The underlying assumption is that if there are only a few active factors, we should be able to identify them. Once they are identified, the design can then be projected onto a lower dimension space. The resulting design is then an under-saturated design, and the ordinary data analysis can then be applied.

To assure that all active factors can be properly estimated after the projection, we must carefully select the supersaturated design. Suppose, for example, it is known that there are at most four active effects among all the factors. To assure identifiability, we need to select a supersaturated design such that any four columns of the design are linearly independent among others. We shall show that the correlations among the columns of the design are crucial to guarantee linear independence. This is, in fact, the fundamental concern in Booth and Cox (1962)'s E(s^2) criterion.

Furthermore, in order to project the design into the set of active factors, it is necessary to identify those active factors correctly. Due to random noises and the partial aliasing structure of the design, we may not always get the correct set of active factors. However, we shall show that if the effects of the active factors are reasonably large as compared with inert factors and random noise, the probability of obtaining the correct set of active factors is satisfactorily high.

2 Estimability and Correlation Structure

Consider a supersaturated design in n experimental runs to investigate k ($\geq n-1$) factors. If $\mathbf{X}$ denotes the $n \times k$ design matrix (without intercept column), our model is

$$Y = \mu\mathbf{1} + \mathbf{X}\beta + \epsilon$$

where $\mathbf{Y}$ is the $(n \times 1)$ observable data vector; μ is the intercept term and $\mathbf{1}$ is an n-vector of 1's; β is a $(k \times 1)$ fixed parameter vector for the unknown factor effects; and ϵ is a vector assumed to be distributed as N(0, $\sigma^2\mathbf{I_n}$). Because k is larger than $n-1$, it is clear that the $\mathbf{X}$ matrix cannot be of full rank and orthogonality is only possible for certain pairs of design columns.

Note that once those active factors are identified, the whole design $\mathbf{X}$ is then projected into a much lower dimension. Hence, the estimability of the effects of these factors depends on whether or not the projected design has full rank. We will show that the largest number of active factors, which can be identified from a supersaturated design, depends on the correlations between columns of $\mathbf{X}$. Let $\mathbf{N} = \{i_1, i_2, \ldots, i_p\}$ and $\mathbf{A} = \{i_{p+1}, i_{p+2}, \ldots, i_k\}$ denote indexes of inert and active factors, respectively, so that $\mathbf{N} \cup \mathbf{A} = \{\mathit{1}, \mathit{2}, \ldots, k\}$. Also, denote the projective design matrix as $\mathbf{X}_p$, and ξ_i the columns of $\mathbf{X}$, $i = 1, 2, \ldots, k$. Defining $Corr(\xi_i, \xi_j) = \xi_i'\xi_j/n$ for any $1 \leq i, j \leq k$, we have:

THEOREM 1 *If* $|Corr(\xi_i, \xi_j)| < \rho = \frac{1}{p-1}$ *for all* $i \neq j$, *then* $\mathbf{X}_p$ *is of full rank.*

Proof: Write $\mathbf{X}_p'\mathbf{X}_p = (x_{ij})$. It is easy to see that $x_{ij} = nCorr(\xi_i, \xi_j)$. Hence, (i) $x_{ii} = n$ and (ii) $|x_{ij}| \leq \rho n$, for all $i \neq j$. Therefore $x_{ii} > \sum_{j \neq i} |x_{ij}|$. This implies that $\mathbf{X}_p'\mathbf{X}_p$ is positively definite. Hence, X_p is nonsingular. Q.E.D.

Note that if $X = (\xi_1, \ldots, \xi_k)$ is a supersaturated design, then $X = (\pm\xi_{\tau(1)}, \ldots, \pm\xi_{\tau(k)})$ for any choice of ± signs and the permutation function τ is an equivalent design. We will not distinguish equivalent designs in this paper.

When $\max|Corr(\xi_i, \xi_j)| = \rho (= \frac{1}{p-1})$ for $i \neq j$, there is no definite answer to this problem. A simple counter-example is the supersaturated design

$$X = \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}.$$

Let $p = 2$, we have $Corr(\xi_1, \xi_2) = 1 = 1/(p-1)$. However, $X_2 = X$ is singular.

Despite this counter-example, X_p of most useful supersaturated designs has full rank when $|Corr(\xi_i, \xi_j)| \leq 1/(p-1)$. For the 12-run supersaturated designs given by Lin (1991, 1993, 1995) and Wu(1993), it can be verified that any submatrice consisting of four columns has full rank.

LEMMA 1 *Let $X = (\xi_1, \ldots, \xi_k)$ of size $n \times k$ be a supersaturated design with entries ± 1. If $|Corr(\xi_i, \xi_j)| \leq \rho = 1/(p-1)$ for all $i \neq j$ and a submatrix $X_p = (\xi_1, \ldots, \xi_p)$ is singular, then there is an equivalent submatrix of X_p such that $X_p^t X_p/n = (1+\rho)I - \rho\mathbf{1}\mathbf{1}^t$.*

Proof: If X_p is singular and $|Corr(\xi_i, \xi_j)| \leq \rho = 1/(p-1)$ for all $i \neq j$, we must be able to find a ξ_i, say ξ_1, such that $|Corr(\xi_1, \xi_j)| = \rho$ for all $j \neq 1$, else X_p is not singular by using the same proof of Theorem 1. Clearly, there is an equivalent design such that $Corr(\xi_1, \xi_j) = -\rho$ for all $j \neq 1$. So

$$X_p^t X_p/n = [\rho_{ij}] = \begin{bmatrix} 1 & -\rho\mathbf{1}^t \\ -\rho\mathbf{1} & Y \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ -\rho\mathbf{1} & Y - \rho^2\mathbf{1}\mathbf{1}^t \end{bmatrix} \begin{bmatrix} 1 & -\rho\mathbf{1}^t \\ 0 & I \end{bmatrix}.$$

Now we make an induction assumption that X_p is singular only if $Corr(\xi_i, \xi_j) = \pm 1/(p-1)$ for $1 \leq i \neq j \leq p$. This assumption is obviously true for $p = 2$. Let us assume that it is also true for $p-1$. Then, the above X_p is singular if and only if $Y - \rho^2\mathbf{1}\mathbf{1}^t$ is singular. Note the order of $Y - \rho^2\mathbf{1}\mathbf{1}^t$ is $(p-1) \times (p-1)$. By the induction assumption, we have, for $i \neq j$,

$$(\rho_{ij} - \rho^2)/(1 - \rho^2) = \pm 1/(p-2).$$

Solving this equation, we get $\rho_{ij} = -1/(p-1)$. Q.E.D.

So, for a 12-run supersaturated design with the maximum absolute correlation being 1/3, if the submatrix $X_4 = (\xi_1, \ldots, \xi_4)$ is singular, then $\sum_{i=1}^{4} \xi_i = 0$ for an equivalent X_4. Clearly, designs in Lin (1991, 1993, 1995) and Wu (1993) do not contain such structures. Therefore, the maximum estimable number of factors is at least four.

For a 20-run supersaturated design with maximum correlation 1/4, if the submatrix $X_5 = (\xi_1, \ldots, \xi_5)$ is singular, then $\sum_{i=1}^{5} \xi_i = 0$ after choosing a proper equivalent design. However, this is impossible since the summation of five ±1's cannot be zero. We can summarize this example into the following theorem.

THEOREM 2 *If* $|Corr(\xi_i, \xi_j)| \leq \rho = \frac{1}{p-1}$ *for all* $i \neq j$*, then* $\mathbf{X}_p$ *is nonsingular when* p *is odd.*

Theorems 1 and 2 ensure that if there are at most p active factors, when using a supersaturated design with maximum correlation less than or equal to $1/(p-1)$ for odd p, it is always possible to estimate all of them. However, in doing so, we have to first identify these active factors. Thus, an important question that needs to be addressed is: Assuming there are only p active factors, what is the probability these p factors have the largest estimated effects? The answer to this question depends on the relative size of their effects and also partially on the estimation method. We will discuss this problem in the next section.

3 Identifiability

The conventional point estimation for β_i's is

$$\hat{\beta}_i = (\bar{y}_{(i)}^{+} - \bar{y}_{(i)}^{-})/2 = x_i'Y/n \tag{1}$$

where $\bar{y}_{(i)}^{+}$ and $\bar{y}_{(i)}^{-}$ are the averages of responses y_i's for factor x_i being high-(+) and low-(−) level respectively. Consequently we have $E(\hat{\beta}_i) = \beta_i + \sum_{j \neq i} \rho_{ij}\beta_j$, where ρ_{ij} is the correlation between columns ξ_i and ξ_j. Suppose β_m is the maximal effect among all β_i's. We are interested in the probability that the estimate of β_i, $\hat{\beta}_m$, will remain the largest among all estimates. To compute ϕ_m=Prob$[\hat{\beta}_m > \max_{i \neq m} \hat{\beta}_i]$, the following two lemmas are needed.

LEMMA 2 *Assume $Y_1, \ldots, Y_n$ are independent and normally distributed with the same variance and the design satisfies condition $\rho \leq 1/3$. Let $\hat{\beta}_i, i = 1, \ldots, k$ be defined as in (1), ϕ_{ma} = Prob$[\hat{\beta}_m > \hat{\beta}_a]$, and ϕ_{mb} = Prob$[\hat{\beta}_m > \hat{\beta}_b\ ]$, then*

$$\text{Prob}[\hat{\beta}_m > max(\hat{\beta}_a, \hat{\beta}_b)] \geq \phi_{ma} \cdot \phi_{mb}.$$

Proof. Note that

$$\text{Prob}[\hat{\beta}_m > max(\hat{\beta}_a, \hat{\beta}_b)] = \text{Prob}[\hat{\beta}_m > \hat{\beta}_a | \hat{\beta}_m > \hat{\beta}_b]\text{Prob}[\hat{\beta}_m > \hat{\beta}_b].$$

Hence, we need only to show

$$\text{Prob}[\hat{\beta}_m > \hat{\beta}_a | \hat{\beta}_m > \hat{\beta}_b] \geq \text{Prob}[\hat{\beta}_m > \hat{\beta}_a].$$

Let $Z_1 = \hat{\beta}_m - \hat{\beta}_a$ and $Z_2 = \hat{\beta}_m - \hat{\beta}_b$. By the conditions of the lemma, the correlation between two estimates is no larger than 1/3, and we have $Cov(Z_1, Z_2) \geq 0$. So,

$$\text{Prob}[\hat{\beta}_m > \hat{\beta}_a | \hat{\beta}_m > \hat{\beta}_b] = \text{Prob}[Z_1 > 0 | Z_2 > 0] \geq \text{Prob}[Z_1 > 0]$$

as required. Q.E.D.

Following Lemma 2, by induction we have $\phi_m = \prod_{i \neq m, i=1}^{k} \phi_{mi}$ where ϕ_{mi} = Prob$[\hat{\beta}_m \geq \hat{\beta}_i]$.

LEMMA 3 *Given m and i, $\phi_{mi} \geq \Phi\left(\sqrt{\frac{n}{2(1-\rho_{mi})}} \cdot \delta_{mi}\right)$, where $\delta_{mi} = (E\hat{\beta}_m - E\hat{\beta}_i)/\sigma$ and Φ is the standard normal cumulative distribution function.*

Proof: First note that $\hat{\beta}_m - \hat{\beta}_i \sim N(E(\hat{\beta}_m - \hat{\beta}_i), \frac{2(1-\rho_{mi})}{n}\sigma)$. Thus,

$$\begin{aligned}
\text{Prob}[\hat{\beta}_m - \hat{\beta}_i \geq 0] &= P[\frac{(\hat{\beta}_m - \hat{\beta}_i) - E(\hat{\beta}_m - \hat{\beta}_i)}{\sqrt{\frac{2}{n}}(1 - \rho_{mi})\sigma} \geq \frac{-E(\hat{\beta}_m - \hat{\beta}_i)}{\sqrt{\frac{2}{n}}(1 - \rho_{mi})\sigma}] \\
&= \Phi\left[\frac{\sqrt{n}E(\hat{\beta}_m - \hat{\beta}_i)}{\sqrt{2(1 - \rho_{mi})}\sigma}\right] = \Phi\left(\sqrt{\frac{n}{2(1 - \rho_{mi})}} \cdot \delta_{mi}\right).
\end{aligned}$$

Q.E.D.

Note that whether $\hat{\beta}_m$ is larger than $\hat{\beta}_i$ depends on several factors. One important factor is the value of ρ_{mi}. A positive ρ_{mi} makes $n/[2(1-\rho_{mi})]$ larger and hence improves the chance to identify β_m. On the other hand, a positive ρ_{mi} also brings $E\hat{\beta}_i$ closer to $E\hat{\beta}_m$ which makes δ_{mi} smaller and hence reduces the chance to identify β_m.

It is straightforward then from lemmas 2 and 3, that:

THEOREM 3 *The probability that the $\hat{\beta}_m$ is the largest estimated effect is*

$$\phi_m \geq \prod_{i \neq m, i=1}^{k} \Phi\left(\sqrt{\frac{n}{2(1-\rho_{mi})}} \cdot \delta_{mi}\right).$$

The lower bounds given in Theorem 3 can be calculated directly for any specific design. Table 1 shows some simulation results based on the supersaturated designs constructed by Lin (1991, 1993 and 1995) and Wu (1993). For each case, given the design and the number of factors, the simulations were conducted in the following way:

(a) Randomly select a number m from 1 to k. Let all $\beta_j = 0$ when $j \neq m$, and β_m=1 or 2 separately.

(b) Generate n of ϵ_i's from N(0,1) to construct the responses $y_i = \sum_{u=1}^{k} x_{iu}\beta_u + \epsilon_i$, $i = 1, 2, \ldots, n$.

(c) Obtain $\hat{\beta}_j$ by Equation (1) for all j; and record whether $\hat{\beta}_m$ is indeed the maximum.

(d) Repeat (a) to (c) 5000 times.

Note that the supersaturated designs constructed by half-fraction Hadamard matrices (Lin, 1993) can only examine $k = N - 2$ factors in $n = N/2$ runs, while the 12-run supersaturated design using interaction columns (Lin, 1991 and 1995; Wu, 1993) can study as many as 66 factors. The case k=10 is not supersaturated, but is a reference benchmark to be compared with the performance of supersaturated designs. It is clear that in all the cases, the lower bounds are satisfactorily large.

Table 1 about here

Supersaturated designs with $|\rho_{ij}| \leq 1/3$ were recommended by Lin (1995). In fact, all designs discussed in Table 1 have such a property. In this case, we can extend Theorem 3 to:

COROLLARY 1 *If $|\rho_{ij}| \leq 1/3$, then $\phi_m \geq \prod_{i \neq m, i=1}^{k} \Phi\left(\sqrt{\frac{3n}{8}} \cdot \delta_{mi}\right)$.*

Table 2 shows lower bound probabilities as given in Corollary 1 for various combinations of (n, k, δ_{mi}). Note that these probabilities do not depend on the design. Also, the probabilities given here are smaller than the probabilities given in Table 1, as expected. In general, if $\delta_{mi} \geq 2$, the largest effect can always be correctly identified, a similar observation made by Lin (1995).

Table 2 about here

If there is a set of factors which are active, a bound can be found as follows. Recall that for any two events E_1 and E_2,

$$\text{Prob}(E_1 E_2) \geq \text{Prob}(E_1) + \text{Prob}(E_2) - 1.$$

Let A be the set of active factors. For any $m \in A$, define $\psi_m = \text{Prob}(\hat{\beta}_m \geq \hat{\beta}_i, \text{for } i \notin A)$. Then, the probability that $\hat{\beta}_{m'}$ and the $\hat{\beta}_{m''}$ are the largest estimated effects is $\psi_{m'} + \psi_{m''} - 1$. If both $\psi_{m'}$ and $\psi_{m''}$ are larger than 99%, this bound is 98%. Obviously, we have $\psi_m \geq \phi_m$ for any $m \in A$. Hence, a lower bound of ψ_m can be obtained from the last theorem.

More generally, we have

THEOREM 4 *Let $A = 1, 2, \ldots, p$ correspond to p active factors in the design. Assume the conditions in Theorem 1 are satisfied. Then, the probability that $\hat{\beta}_i, i = 1, \ldots, p$ are the p largest estimated factors is no less than*

$$\psi_1 + \psi_2 + \cdots + \psi_p - (p-1).$$

For example, if $p = 4$ and each of $\psi_i, i = 1, \ldots, p$ is larger than 99%, this bound becomes 96%– a very satisfactorily large lower-bound probability. However, the bound decreases rapidly when p increases or the ψ_i's become smaller. Of course, for such cases, a supersaturated design is not recommended.

Note that the results given above can be straightforwardly extended to the reverse case to find Prob$[\hat{\beta}_w \leq \hat{\beta}_i]$ where β_w is the minimal effect among all β_i's.

4 Remarks

Another possible point estimate for β is via the least squares method. Namely, $\hat{\beta}_G = (\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'y$. Because of the singularity of $\mathbf{X}'\mathbf{X}$ in a supersaturated design setting, the least squares estimate $\hat{\beta}_G = (\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'y$ is not unique. However, one can show that the $\hat{\beta}$ defined in (1) is not one of the least square estimates. They are intrinsically different from $\hat{\beta}_G$.

LEMMA 4 *The conventional estimator $\hat{\beta}$ is not a least squares estimator.*

Proof. Note that $E(\hat{\beta}) = \mathbf{X}'\mathbf{X}\beta/n$ and $Var(\hat{\beta}) = (\mathbf{X}'\mathbf{X})\sigma^2/n^2$; while $E(\hat{\beta}_G) = (\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\mathbf{X}\beta$ and $Var(\hat{\beta}_G) = (\mathbf{X}'\mathbf{X})^{-}\sigma^2$. Hence, if $\hat{\beta}$ is one of the least squares estimate of β, there must be a $(\mathbf{X}'\mathbf{X})^{-}$, such that $\mathbf{X}'\mathbf{X}/n = (\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\mathbf{X}$ and $(\mathbf{X}'\mathbf{X})/n^2 = (\mathbf{X}'\mathbf{X})^{-}$. However, these two identities imply $\frac{\mathbf{X}'\mathbf{X}}{n}$ is idempotent. On the other hand, the trace (sum of diagonal elements) of $\frac{\mathbf{X}'\mathbf{X}}{n}$ is obviously k from the structure of the supersaturated designs, hence it too has full rank. A $k \times k$ matrix which is *idempotent* and of full rank at the same time, has to be the identity matrix. This is impossible for a supersaturated design since $k > n$. Q.E.D.

Note that we have implicitly assumed that all columns of $\mathbf{X}$ have half $+1$'s and half -1's. This is necessary for ignoring the intercept term in the model.

We prefer to use $\hat{\beta}$ here, because: (1) $\hat{\beta}_G$ is not unique, and more important, (2) the screening process typically has only a vague idea of the model, in this case, the criterion of minimizing the overall residual sum squares (as in obtaining $\hat{\beta}_G$) does not seem very appropriate.

Acknowledgements

This research was partially supported by the Natural Sciences and Engineering Research Council of Canada.

References

Booth, K. H. V. and Cox, D. R. (1962), "Some Systematic Supersaturated Designs," *Technometrics*, **4** 489-495.

Deng, L.Y., Lin, D.K.J. and Wang, J.N. (1994), "Supersaturated Designs Using Hadama Matrices," *IBM Research Report*, **No. 19470** IBM T.J. Watson Research Center.

Lin, D.K.J. (1991), "Supersaturated Designs," *Working Papers*, **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*, to appear.

Nguyen, Nam (1994), "An Algorithmic Approach to Constructing Supersaturated Designs," preprint.

Seewald, W. (1994), "Analysis of Experimental Data When Only One Factor is Important," *Technical Report*, **No. 9405**, Ciba-Geigy Ltd, Mathematical Applications, Switzerland.

Wu, C.F.J. (1993), "Construction of Supersaturated Design Through Partially Aliased Interactions," *Biometrika*, **80** 661-669.

Table 1: Successful identification probabilities in 5000 simulations

Design	Run	β	Number of factors(k)					
	size		10	20	30	40	50	60
HFHM	12	1	0.9462	0.9062				
		2	1	1				
	18	1	0.9888	0.9794	0.9694			
		2	1	1	1			
	24	1		0.9950	0.9934	0.9912		
		2		1	1	1		
IntCol	12	1	0.9560	0.9026	0.8556	0.8092	0.7842	0.7530
		2	0.9999	1.0000	1.0000	1.0000	0.9999	0.9999

HFHM=Supersaturated designs using half Fraction Hadamard matrices (Lin, 1993)

IntCol=Supersaturated designs using interaction columns (Lin, 1991 and Wu, 1993)

Table 2: Lower bound probabilities given by Corollary 1

Run	δ_{mi}	Number of factors(k)				
size		10	20	30	40	50
12	1	0.8574	0.7237	0.6092	0.5134	0.4327
	2	0.9999	0.9998	0.9997	0.9996	0.9995
16	1	0.9374	0.8725	0.8121	0.7558	0.7034
	2	1.000	1.000	1.000	1.000	1.000
20	1	0.9726	0.9430	0.9143	0.8865	0.8595
	2	1.000	1.000	1.000	1.000	1.000
24	1	0.9879	0.9747	0.9616	0.9487	0.9360
	2	1.000	1.000	1.000	1.000	1.000