Probability of correct model identification in supersaturated design

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

\textsuperscript{a} Department of Statistics, Visva-Bharati University, West Bengal 731235, India
\textsuperscript{b} Department of Supply Chain & Information Systems, The Pennsylvania State University, University Park, PA 16802, USA

\textbf{ABSTRACT}

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.

\textcopyright 2009 Elsevier B.V. All rights reserved.

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).


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 discusses 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...
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 \( a_1a_2 \ldots a_n, 0 \leq a_j \leq m_j - 1, 1 \leq j \leq n \). Under factor sparsity situation, we assume that only \( k \) out of \( n \) factors are active.

For \( 1 \leq j \leq n \), let \( I_j \) be the \( m_j \times 1 \) vector with all elements unity, \( I_{m_j-1}^\top \) 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

\[
P_j P_j^\top = I_{m_j-1}, \quad P_j I_j = 0,
\]

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 \( v = \prod m_i \). 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 \( D(m_1 \times m_2 \times \cdots \times m_n) \) 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 \( D(m_1 \times m_2 \times \cdots \times m_n) \). Then the linear model corresponding to \( d \) is given by

\[
y = b_N \mu + Z_1 \theta_1 + \cdots + Z_n \theta_n + \epsilon,
\]

where \( y \) is the \( N \times 1 \) vector of observations, \( b_N \) is the \( N \times 1 \) vector with each element \( N^{-1/2} \), \( \mu \) is the general mean effect, \( \theta_j, 1 \leq j \leq n \), is the \((m_j - 1) \times 1 \) vector of unknown main effect contrasts corresponding to the factor \( F_j \) and \( \epsilon \) is the \( N \times 1 \) vector of random error components. The components of \( \epsilon \) are assumed to be independently and identically distributed with mean \( 0 \) and variance \( \sigma^2 \). 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

\[
y = b_N \mu + Z_{i_1} \theta_{i_1} + \cdots + Z_{i_k} \theta_{i_k} + \epsilon.
\]

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

\[
S(h)^2 = y'(I_{m_{i_k}}^\top - Q_h)y,
\]

where

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

Following Srivastava (1975), choose \( S(h) \) 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 \( D(m_1 \times m_2 \times \cdots \times m_n) \) 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 \( E_{h_1h_2} \) as

\[
E_{h_1h_2} = (S(h_1)^2 \leq S(h_2)^2).
\]

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

\[
P_{h_1}(d) = P\left( \bigcap_{h_2 \neq h_1 \in H(k)} E_{h_1h_2} | M(h_1) \right).
\]

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

\[
P(d) = \frac{1}{v} \sum_{h_1 \in H(k)} P_{h_1}(d).
\]

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 \( P(E_{h_1h_2} | M(h_1)) \) as

\[
P(E_{h_1h_2} | M(h_1)) = P(u_{h_1} > u_{h_2} | M(h_1)), \quad \text{where} \ h_1, h_2 \in H(k), h_1 \neq h_2.
\]

Assuming \( L_h \) to be of full column rank, we can find a non-singular matrix \( R_h \) such that \((L_h' L_h)^{-1} = R_h K_h' \).
Define $z_{h_1} = R_1^L L_1^y$ and $z_{h_2} = R_2^L L_2^y$. It is to be noted that, for $i = 1, 2$, $z_i$ is a random vector of order $r_i \times 1$, where $r_i$ is the number of columns of $L_i$. The following lemma will be helpful in proving Theorem 1.

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

\[
P(E_{h_1, h_2} | M(h_1)) = P \left( \sum_{a=1}^{r_1} z_{h_1 a}^2 > \sum_{a=1}^{r_2} z_{h_2 a}^2 \right),
\]

where, $z_{h_1} = ((z_{h_1 \alpha}) \sim N(r_1, (R_{h_1})^{-1} \theta_{h_1}, \sigma^2 r_1^*)$, $z_{h_2} = ((z_{h_2 \alpha}) \sim N(r_2, (R_{h_2})^{-1} \theta_{h_1}, \sigma^2 r_2^*)$. Moreover, here we have $\text{cov}(z_{h_1}, z_{h_2}) = \sigma^2 R_1^L L_1 \theta_1 R_2$. 

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

This section considers the case $k = 1$. Suppose $m_1 = \cdots = m_\iota = m$. For $1 \leq i \leq \iota$, let us write the matrix $L_i$ as $L_i = [l_{i1}^{(i)} \ldots l_{im_i}^{(i)}]$ and the vector $\theta_i$ as $\theta_i = [\theta_{i1}, \ldots, \theta_{i(m_i-1)}]$. Then, for any $h \in H(1)$ and for a column balanced supersaturated design $d$, the matrix $R_h$ is given by $R_h = \text{diag}(\sqrt{\nu/N}, \ldots, \sqrt{\nu/N})$, where $\nu = m^\iota$. 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

\[
\mu_{h_1 \alpha} = (\sqrt{\nu/N}) \theta_{h_1 \alpha}, \quad \mu_{h_2 \alpha} = (\sqrt{\nu/N}) \sum_{a=1}^{m-1} l_{a}^{(h_2)} l_{a}^{(h_1)} \theta_{h_1 \alpha},
\]

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

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

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

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

where

\[
p_\alpha^{(h_1, h_2)} = 1 - \Phi \left( \frac{\eta_{\iota 1 \alpha}}{\sigma_{1 \alpha}} \right) - \Phi \left( \frac{\eta_{\iota 2 \alpha}}{\sigma_{2 \alpha}} \right) + 2 \Phi \left( \frac{\eta_{\iota 1 \alpha}}{\sigma_{1 \alpha}} \right) \Phi \left( \frac{\eta_{\iota 2 \alpha}}{\sigma_{2 \alpha}} \right),
\]

with

\[
\eta_{\iota 1 \alpha} = \sqrt{\frac{\nu}{N}} \left( \theta_{h_1 \alpha} + \frac{\nu}{N} \sum_{a=1}^{m-1} l_{a}^{(h_2)} l_{a}^{(h_1)} \theta_{h_1 \alpha} \right), \quad \eta_{\iota 2 \alpha} = 2 \sigma^2 \left( 1 + \frac{\nu}{N} \sum_{a=1}^{m-1} l_{a}^{(h_2)} l_{a}^{(h_1)} \right),
\]

\[
\eta_{\iota 2 \alpha} = \sqrt{\frac{\nu}{N}} \left( \theta_{h_1 \alpha} - \frac{\nu}{N} \sum_{a=1}^{m-1} l_{a}^{(h_2)} l_{a}^{(h_1)} \theta_{h_1 \alpha} \right), \quad \sigma_{2 \alpha} = 2 \sigma^2 \left( 1 - \frac{\nu}{N} \sum_{a=1}^{m-1} l_{a}^{(h_2)} l_{a}^{(h_1)} \right).
\]

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

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

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

**Definition 2.** Among two designs $d_1, d_2 \in D(m^n)$, $d_1$ 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 $d_1, d_2 \in D(3^7)$, with 9 runs. $d_1$ is due to Xu and Wu (2005) and $d_2$ has been constructed by permuting the 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 $\theta_{\alpha}/\sigma = \rho$, $1 \leq \alpha \leq m-1$ 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).
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 $1 \leq i \leq n$, define

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

(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

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

(9)

The threshold value $\xi$ 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 $\theta_{ij}$’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 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 $L_n^t L_n$ is singular, the problem of singularity is tackled by setting few effects of few factors at zero level and replacing $(L_n^t L_n)^{-1}$ by a $g$-inverse of $L_n^t L_n$.

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 fitness 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, $\nu = m_1 \times \cdots \times m_r$. It is easy to note that there is a one-to-one correspondence between $T$ and the set $I$. Suppose, in the $i$th generation we have $r$ different chromosomes $C^i_1, C^i_2, \ldots, C^i_r$, 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 $I$.

<table>
<thead>
<tr>
<th>$T$</th>
<th>$I$</th>
</tr>
</thead>
<tbody>
<tr>
<td>000</td>
<td>1</td>
</tr>
<tr>
<td>020</td>
<td>2</td>
</tr>
<tr>
<td>012</td>
<td>3</td>
</tr>
<tr>
<td>100</td>
<td>4</td>
</tr>
<tr>
<td>101</td>
<td>5</td>
</tr>
<tr>
<td>110</td>
<td>6</td>
</tr>
<tr>
<td>111</td>
<td>7</td>
</tr>
<tr>
<td>102</td>
<td>8</td>
</tr>
<tr>
<td>112</td>
<td>9</td>
</tr>
<tr>
<td>120</td>
<td>10</td>
</tr>
<tr>
<td>121</td>
<td>11</td>
</tr>
<tr>
<td>122</td>
<td>12</td>
</tr>
<tr>
<td>200</td>
<td>13</td>
</tr>
<tr>
<td>201</td>
<td>14</td>
</tr>
<tr>
<td>202</td>
<td>15</td>
</tr>
<tr>
<td>210</td>
<td>16</td>
</tr>
<tr>
<td>211</td>
<td>17</td>
</tr>
<tr>
<td>212</td>
<td>18</td>
</tr>
<tr>
<td>220</td>
<td>19</td>
</tr>
<tr>
<td>221</td>
<td>20</td>
</tr>
<tr>
<td>222</td>
<td>21</td>
</tr>
<tr>
<td>27</td>
<td></td>
</tr>
</tbody>
</table>
A typical chromosome and the corresponding design of a $3^3$ factorial experiment for $N = 6$ is shown below.

<table>
<thead>
<tr>
<th>Typical chromosome</th>
<th>Plan (Columns as runs)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 0 1 1 1 1 2 2 2</td>
<td>0 0 1 2 0 0 0 2 1 2</td>
</tr>
<tr>
<td>0 1 0 0 1 2 0 0 0</td>
<td>1 5 12 16 23 27 0 1 0 2</td>
</tr>
<tr>
<td>2 1 0 2 0 2 0 0 1</td>
<td>0 1 2 0 1 2 0 1 2 2</td>
</tr>
</tbody>
</table>

From the above mentioned $r$ chromosomes $C_j^1$, $C_j^2$, ..., $C_j^r$, 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 $n_{ij}^y$ be the number of samples to be drawn in the $j$th stage of the $i$th generation corresponding to the $y$th gene. The values of $n_{ij}^y$ are sufficiently large and grow with stages. Let $P_{ij}^y$ and $S_{ij}^y$ be respectively the mean and standard deviation of the probability of correct model identification of $C_j^y$ at the $j$th stage in the $i$th generation, $1 \leq y \leq r$. On the basis of $P_{ij}^y$ and $S_{ij}^y$ we construct a confidence interval for $y$th chromosome as $(L_{ij}^y, U_{ij}^y)$, where

$$L_{ij}^y = P_{ij}^y - t_{1-(1-\alpha)^{1/2}} \frac{S_{ij}^y}{\sqrt{n_{ij}^y}},$$

$$U_{ij}^y = P_{ij}^y + t_{1-(1-\alpha)^{1/2}} \frac{S_{ij}^y}{\sqrt{n_{ij}^y}}.$$  

$S_{ij}^y$ 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 $S_{ij}^y$. Define $M_{ij}^y = \max_y L_{ij}^y$ and construct the set $F_{ij}^y = \{C_j^y | U_{ij}^y > M_{ij}^y\}$. Then $F_{ij}^y$ is the desired set of elitist chromosomes. In our implementation, the cardinality of $F_{ij}^y$ decreases with an increase in $n_{ij}^y$. 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 $\sigma_1$ and $\xi$. 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>
<thead>
<tr>
<th>Plan (Columns as runs)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 0 1 2 0 0 0 2 1 2</td>
</tr>
<tr>
<td>1 5 12 16 23 27 0 1 0 2</td>
</tr>
<tr>
<td>0 1 2 0 1 2 0 1 2 2</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Typical chromosome</th>
<th>Plan (Columns as runs)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$0 0 1 1 1 1 2 2 2$</td>
<td>$0 0 1 2 0 0 0 2 1 2$</td>
</tr>
<tr>
<td>$0 1 0 0 1 2 0 0 0$</td>
<td>$1 5 12 16 23 27 0 1 0 2$</td>
</tr>
<tr>
<td>$2 1 0 2 0 2 0 0 1$</td>
<td>$0 1 2 0 1 2 0 1 2 2$</td>
</tr>
</tbody>
</table>
It is to be remarked that the probability of correct model identification increases with \( \xi \) and decreases with \( \sigma_1 \). The increase in the values of the probability with that of \( \xi \) can be attributed as logical, as \( \xi \) 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 \( \sigma_1 \) also can be explained by the fact, as the values of \( \sigma_1 \) increases, the absolute values of the coefficients of negligible effects increases causing the difficulties in identifying the true non-negligible effects.

### Appendix. Proof of Theorem 1

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

\[
P(d) \geq \frac{1}{n} \sum_{h_1=1}^{n} \sum_{h_2=1}^{n} \sum_{\alpha=1}^{m-1} P(B_\alpha^{(h_1 h_2)} | M(h_1)) - (n-1)(m-2) - (n-2).
\]

Now, let

\[
p_\alpha^{(h_1 h_2)} = P(B_\alpha^{(h_1 h_2)} | M(h_1)) = P(z_{h_1 \alpha}^2 > z_{h_2 \alpha}^2 | M(h_1)) = P(x_{1\alpha} x_{2\alpha} > 0)
\]

\[
= 1 - \Phi \left( \frac{\eta_{1\alpha}}{\sigma_{1\alpha}} \right) - \Phi \left( \frac{\eta_{2\alpha}}{\sigma_{2\alpha}} \right) + 2 \Phi \left( \frac{\eta_{1\alpha}}{\sigma_{1\alpha}} \right) \Phi \left( \frac{\eta_{2\alpha}}{\sigma_{2\alpha}} \right),
\]

where \( x_{1\alpha} = z_{h_1 \alpha} + z_{h_2 \alpha} \sim N(\eta_{1\alpha}, \sigma_{1\alpha}^2) \) and \( x_{2\alpha} = z_{h_1 \alpha} - z_{h_2 \alpha} \sim N(\eta_{2\alpha}, \sigma_{2\alpha}^2) \).

The proof of Theorem 1 thus follows. \( \Box \)

### References


