Robust screening by edge designs

Stella Stylianou

School of Science
Mathematical Sciences
RMIT University
Melbourne

stella.stylianou@rmit.edu.au

ICORS 2017, University of Wollongong,
3 – 7 July 2017
Outline:

1. Introduction and Preliminaries
   - Motivation and Basic Notation
   - Screening designs
   - Weighing designs

2. Construction of Edge designs
   - The Construction using a conference matrix
   - A construction using weighing design $W(n; n - 2)$
   - A general construction using weighing designs $W(n; k)$
   - Some economical D-optimal edge designs

3. The Analysis of Edge Designs
   - Model independent analysis - robust screening
   - A linearity test using the edge designs

4. Discussion and Future Research

5. Some References
Motivation: Screening Designs

In some industrial experiments:

- A number of quantitative factors has to be studied.
- Only a few of them is expected to be important - effect sparsity assumption (Box and Meyer, 1986).
- Screening designs aim at identifying the few dominant main effects.
- Two level designs are often used as screening designs.
- Active non-linear terms cause linear regression and two level designs to fail.

D.C Montgomery, Designs and Analysis of Experiments, Wiley.
A. Dean and D. Voss, Designs and Analysis of Experiments, Springer.
The non-linear model: \( y = f(x_i, \varepsilon) \),

The linear model: \( y = \beta_0 + \sum_{i=1}^{m} \beta_i x_i + \varepsilon \),

where

- \( y \) is the response vector (dependent variable),
- \( x_i \) are the independent variables (in the linear case these are columns corresponding to the main effects),
- \( \beta_0, \beta_i \), are unknown constant coefficients corresponding to the intercept, and main effects respectively,
- \( \varepsilon \) is the error vector with components \( \varepsilon_j \) being iid \( N(0, \sigma^2) \).
Some traditional two- or three-level screening designs

- Full $s^m$ or fractional factorial designs $s^{m-p}$, $s = 2, 3$.
- Orthogonal arrays $OA(n, m, s, t)$, $s = 2, 3$.
- Two level factorial designs (or other two level arrangements) with center points. Center points indicate curvature but do not allow a separate estimation of quadratic effects.
- L-efficient designs constructed algorithmically by maximizing some defined criterion L. For example, one can use D-criterion [Jones and Nachtsheim (2011)], Q-criterion [Tsai, Gilmour, and Meal (2000, 2004)], and many more.

Elster and Neumaier (1995) defined an alternative class of two-level screening designs called edge designs.

- The measurements are arranged into a set $E$ of pairs that within these pairs the coordinates differ in one component only.
- Can be used for screening out the relevant factors without the assumption of a specific model.
- Constructed easily by using known designs.

Suppose that $W^T$ is the design matrix of a chemical balance weighing experiment.

- Then, $W = W(m; k)$
  - is a square matrix of order $m$ with entries $\{0, \pm 1\}$,
  - has $k$ non-zero entries per row and column,
  - satisfies $WW^T = W^T W = kl_m$,
  - is universally optimal in the class of all $m$-observation chemical balance weighing designs, with $m$ objects, such that at most $k$ objects are used in each weighing (see Kiefer (1975)).

- $k$ is called the weight of $W$.
- A permutation matrix is the identity matrix with its rows re-ordered.

Some Examples

\[
W_{4a} = \begin{pmatrix}
1 & 0 & 1 & 0 \\
0 & 1 & 0 & 1 \\
-1 & 0 & 1 & 0 \\
0 & -1 & 0 & 1
\end{pmatrix}, \quad W_{4b} = \begin{pmatrix}
1 & 1 & -1 & 0 \\
-1 & 1 & 0 & -1 \\
-1 & 0 & -1 & 1 \\
0 & -1 & -1 & -1
\end{pmatrix}
\]

\[
W_{4c} = \begin{pmatrix}
0 & 0 & 1 & 0 \\
0 & 0 & 0 & 1 \\
1 & 0 & 0 & 0 \\
0 & 1 & 0 & 0
\end{pmatrix}, \quad W_{4d} = \begin{pmatrix}
1 & 1 & -1 & -1 \\
-1 & 1 & 1 & -1 \\
-1 & -1 & -1 & -1 \\
1 & -1 & 1 & -1
\end{pmatrix}
\]

- \( W_{4a} = W(4; 2) \) is a weighing design of order 4 and weight 2.
- \( W_{4c} = W(4; 1) \) can be considered as a permutation matrix of order 4.
- \( W_{4b} = W(4; 3), \ W_{4c} = W(4; 1), \ W_{4d} = W(4; 4). \)
A weighing design $W(n; n-1)$ is called **conference matrix**.

\[
W_6 = \begin{pmatrix}
+ & + & - & - & 0 & + \\
+ & - & - & + & - & 0 \\
+ & - & + & 0 & + & + \\
- & - & 0 & - & - & + \\
0 & - & - & - & + & - \\
- & 0 & - & + & + & +
\end{pmatrix},
\]

\[
W_8 = \begin{pmatrix}
0 & - & + & + & - & + & - & - \\
- & - & + & - & - & 0 & + & + \\
+ & 0 & + & + & + & + & + & + \\
+ & - & - & + & - & - & 0 & + \\
- & - & 0 & + & + & - & + & - \\
+ & - & - & - & 0 & + & + & - \\
+ & - & + & - & + & - & - & 0 \\
- & - & - & 0 & + & + & - & +
\end{pmatrix}
\]

- Notation: $-$ stands for $-1$ and $+$ denotes $1$.
- $W_6 = W(6; 5), \ W_8 = W(8; 7)$. 

Stella Stylianou  
Robust screening by edge designs
Let \( C = W(n; n - 1) \) satisfying \( C_{ii} = 0 \). Then the following matrix is a D-optimal edge design with \( 4n \) runs and \( 2n \) edges.

**Proof:** Set \[
X = \begin{bmatrix}
  j & C + I \\
  j & C - I \\
  j & -C + I \\
  j & -C - I \\
\end{bmatrix}.
\]

The \((i, n+i)\) rows and the \((2n+i, 3n+i)\) rows, \( i = 1, \ldots, n \), define edges.

\( X \) is an edge design with \( 4n \)-runs and \( 2n \)-edges (two edges for each variable).

\( X^TX = 4nI_{n+1} \) and thus \( X \) is D-optimal.

Filling the weighing design $W(n; n - 2)$

Let $W = W(n, n - 2)$ and $P, Q$ permutation matrices such that $W + P + Q$ is $\{1, -1\}$ matrix and $W^T P = -P^T W$, then the following matrix is a D-optimal edge design with $4n$ runs and $2n$ edges.

**Proof:** Set

$$X = \begin{bmatrix} j & W + P + Q \\ j & -W - P - Q \\ j & W + P - Q \\ j & -W - P + Q \end{bmatrix}.$$ 

The $(i, 2n + i)$ rows and the $(n + i, 3n + i)$ rows, $i = 1, \ldots, n$, define edges.

$X$ is an edge design with $4n$-runs and $2n$-edges (two edges for each variable).

$X^T X = 4nI_{n+1}$ and thus $X$ is D-optimal.
## An example

\[
W = W(6, 4)
\]

\[
\begin{bmatrix}
1 & 0 & 0 & -1 & 1 & 1 \\
0 & -1 & 1 & 1 & 1 & 0 \\
-1 & 1 & 1 & 0 & 0 & 1 \\
-1 & 0 & 0 & -1 & 1 & -1 \\
0 & -1 & 1 & -1 & -1 & 0 \\
-1 & -1 & -1 & 0 & 0 & 1 \\
\end{bmatrix}
\]

\[
P = \begin{bmatrix}
010000 \\
100000 \\
000100 \\
001000 \\
000011 \\
000010 \\
\end{bmatrix},
\quad
Q = \begin{bmatrix}
001000 \\
000001 \\
000010 \\
010000 \\
100000 \\
000100 \\
\end{bmatrix}.
\]

\[
WP^T = -PW^T, \quad W + P + Q \text{ is } \{1, -1\} \text{ matrix.}
\]

\[
PP^T = P^TP = Q^TQ = QQ^T = I.
\]
An example (cont.)

Following the construction \( X = \begin{bmatrix} j & W + P + Q \\ j & -W - P - Q \\ j & W + P - Q \\ j & -W - P + Q \end{bmatrix} \) we obtain

\[
X^T = \begin{bmatrix}
\cdots & \cdots & \cdots & \cdots & \cdots \\
+ & + & + & + & + \\
- & - & - & - & - \\
+ & + & + & + & + \\
- & - & - & - & - \\
+ & + & + & + & + \\
- & - & - & - & - \\
+ & + & + & + & + \\
+ & + & + & + & + \\
+ & + & + & + & + \\
+ & + & + & + & + \\
+ & + & + & + & + \\
+ & + & + & + & +
\end{bmatrix}
\]

\( X \) is the desirable D-optimal edge design with 24 runs, 6 factors and 12 edges (two for each variable).
Following the construction $X = \begin{bmatrix} j & W + P + Q \\ j & -W - P - Q \\ j & W + P - Q \\ j & -W - P + Q \end{bmatrix}$ we obtain

\[
X^T = \begin{bmatrix}
\end{bmatrix}
\]

$X$ is the desirable D-optimal edge design with 24 runs and 12 edges (two for each variable).
An example (cont.)

Following the construction $X =$

$$
\begin{bmatrix}
  j & W + P + Q \\
  j & -W - P - Q \\
  j & W + P - Q \\
  j & -W - P + Q \\
\end{bmatrix}
$$

we obtain

$X^T =
\begin{bmatrix}
  x_0
  
  x_1
  
  x_2
  
  x_3
  
  x_4
  
  x_5
  
  x_6
\end{bmatrix}
$

\begin{align*}
  &+++
  
  &---
  
  &---
  
  &+++ \\
  &---
  
  &+++ \\
  &---
  
  &--- \\
  &+++
  
  &---
  
  &---
  
  &+++ \\
\end{align*}

$X$ is the desirable D-optimal edge design with 24 runs and 12 edges (two for each variable).
Filling the weighing design $W(n; k)$

- Let $W = W(n; k)$ and $P_1, P_2, \ldots, P_{n-k}$ permutation matrices such that $W + P_1 + P_2 + \ldots + P_{n-k}$ is a $\{1, -1\}$ matrix, then there exists a D-optimal edge design with $n2^{n-k}$ edges and $n2^{n-k}+1$ runs.

**Proof:** Let $Z$ be the $2^{n-k+1} \times (n - k + 1)$ matrix which corresponds to a full $2^{n-k+1}$ factorial design and

$$A = \begin{bmatrix}
W \\
P_1 \\
P_2 \\
\vdots \\
P_{n-k}
\end{bmatrix}, \quad X = \begin{bmatrix} j \\
\vdots \\
j
\end{bmatrix}, \quad Y = (Z \otimes I)A,$$

where $\otimes$ is the Kronecker product and $X$ is the desirable D-optimal edge design with $n2^{n-k+1}$ runs and $n2^{n-k}$ edges.
The proposed $(3n \times n)$ D-optimal edge design is

$$X = \begin{pmatrix}
C_n + I_n \\
-C_n + I_n \\
C_n - I_n
\end{pmatrix},$$

where $C_n$ is a conference matrix of order $n$ and $-C_n$ is its full fold-over design.

- $X$ has a model independent screening ability.
- Provides a linearity test for the underlying model.
- Fewer runs than the known D-optimal edge designs.
General model independent analysis procedure

- Factor sparsity assumption (Lenth (1989)) implies that almost all differences $z_{ij} := y_i - y_j, \ (i,j) \in E$, consist of noise only.
- If we assume that the noise in the data is additive, normally distributed with zero mean and variance $\sigma^2$, the $n - p$ of the $z_i$ are normally distributed with zero mean and variance $2\sigma^2$.
- Because of the unknown number of outliers, the variance must be estimated in a robust way. For example, we can use the median estimate $\hat{\sigma} = \frac{\text{median}\{|z_{ij}| : (i,j) \in E\}}{2^{\frac{1}{2}} \times 0.675}$ which is consistent when $p = 0$ (see Lenth (1989)), and hence can be expected to give reliable results when $p \ll n$.
- Outliers then determine active factors ($|z_{ij}| > 4\hat{\sigma}$).
A simulated example: The settings

- The matrix $X$ with 24 runs, 6 factors and 12 edges.
- $y = h(x_1, x_2, x_3, x_4, x_5, x_6) = 10 + 6x_3 + 16x_4 + \varepsilon$
- $u = f(x_1, x_2, x_3, x_4, x_5, x_6) = -3 + x_3 + x_4 + 3(x_3 + 2x_4)^2 + \varepsilon$
  $\varepsilon \sim N(0, 1)$.

- The response vector $y$ in the case of the noisy linear function is $y^T = (0.22, 30.86, 34.06, 0.31, 0.54, 19.76, 19.98, -11.84, -11.87, 20.89, 17.86, -1.15, -11.60, 33.87, 32.73, 0.29, -0.41, -11.52, 31.20, -11.96, -12.09, 21.46, 20.79, 31.62)$
  while the response vector $u$ in the case of the noisy quadratic function is $u^T = (0.39, 24.53, 25.97, 0.94, 0.54, -0.93, -1.12, 20.56, 22.87, 1.35, -1.11, 0.22, 20.86, 28.06, 26.31, 0.54, -0.24, 21.98, 26.16, 22.13, 22.89, -2.14, -1.15, 26.40)$.
Using linear regression revealed two main effects ($x_3$ and $x_4$) and gave \( \hat{h}(x) = 6.01x_3 + 16.09x_4 + 10.16 + \varepsilon, \varepsilon \) of mean zero and $\hat{\sigma} = 0.89$.

An analysis of the edges shows the same relevant variables ($x_3$ and $x_4$) with $\min\{|z_3|, |z_4|\} > 4\hat{\sigma}$, $\hat{\sigma} = 2.2$.

Note that we may use only one edge for each variable since the results obtained from edges on the same variable are similar.

<table>
<thead>
<tr>
<th>$x_3$</th>
<th>$x_6$</th>
<th>$x_5$</th>
<th>$x_2$</th>
<th>$x_1$</th>
<th>$x_4$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$y_1 - y_{13}$</td>
<td>$y_2 - y_{14}$</td>
<td>$y_3 - y_{15}$</td>
<td>$y_4 - y_{16}$</td>
<td>$y_5 - y_{17}$</td>
<td>$y_6 - y_{18}$</td>
</tr>
<tr>
<td>11.81</td>
<td>-3.01</td>
<td>1.33</td>
<td>0.02</td>
<td>0.95</td>
<td>31.28</td>
</tr>
</tbody>
</table>

Since relevant factors found by both methods are the same, we conclude that these are the correct active factors.

Screening with six columns of a saturated design of order 24 reveals the same relevant factors.
**A simulated example (cont): The noisy quadratic function**

- Using linear regression revealed no main effects and gave $\hat{f}(x) = 11.84 + \varepsilon$, with $\varepsilon$ of mean zero and $\hat{\sigma} = 12.5$.
- An analysis of the edges revealed as relevant the variables ($x_3$ and $x_4$) with $\min\{|z_3|, |z_4|\} > 4\hat{\sigma}$, $\hat{\sigma} = 2.6$.
- Note that we may use only one edge for each variable since the results obtained from edges on the same variable are similar.

<table>
<thead>
<tr>
<th>$x_3$</th>
<th>$x_6$</th>
<th>$x_5$</th>
<th>$x_2$</th>
<th>$x_1$</th>
<th>$x_4$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$u_1 - u_{13}$</td>
<td>$u_2 - u_{14}$</td>
<td>$u_3 - u_{15}$</td>
<td>$u_4 - u_{16}$</td>
<td>$u_5 - u_{17}$</td>
<td>$u_6 - u_{18}$</td>
</tr>
<tr>
<td>-20.47</td>
<td>3.53</td>
<td>0.34</td>
<td>1.48</td>
<td>0.78</td>
<td>22.91</td>
</tr>
</tbody>
</table>

- Different relevant factors, so we conclude that the union of the two subsets may or may not contain the correct active factors but the expected model can not be linear.
- Screening with six columns of a saturated design of order 24 reveals no relevant factors, so no further investigation is possible.
Using the $(3n \times n)$ edge design

\[ X = \begin{pmatrix} C_n + I_n \\ -C_n + I_n \\ C_n - I_n \end{pmatrix}, \]

where $C_n$ is a conference matrix of order $n$ and $-C_n$ is its full fold-over design.

- A model independent screening can be applied as previously using the values $z_i = y_i - y_{2n+i}, \ i = 1, \ldots, n$.
- If the model is linear then the values $w_i = y_i + y_{n+i}, \ i = 1, \ldots, n$ will also provide an estimate of the magnitude of effect of $x_i$. 
Using the \((3n \times n)\) edge design (cont.)

- Robust measure of the variances, in order to select the outliers

\[
\hat{\sigma}_z = \frac{\text{med}\{|z_i| : i = 1, \ldots, n\}}{0.675\sqrt{2}}, \quad \hat{\sigma}_w = \frac{\text{med}\{|w_i| : i = 1, \ldots, n\}}{0.675\sqrt{2}}
\]

- Retain, in set \(A\), active variables \(x_i\) from \(z_i\) (those with \(|z_i| > 4\hat{\sigma}_z\)). No assumptions for the underlying model were made till now and thus \(z_i\) provide a screening method which is model independent (see Elster and Neumaier (1995)).

- To check if the underlying model is linear we use both \(z_i\) and \(w_i\). We retain, in set \(B\), the active variables \(x_i\) using \(w_i\) (those with \(|w_i| > 4\hat{\sigma}_w\)).

- If sets \(A\) and \(B\) are not the same then the model is not linear otherwise the linearity hypothesis for the model cannot be rejected.
An illustrating example

- Let $n = 10$ and construct the $(30 \times 10)$ design matrix as above. Let $y$ be the response vector obtained by the simulated (non-linear) model $y \sim N_{30}(e^{-5x_2}+4x_4-5x_5, I_{30})$. The values of contrasts $z_i = y_i - y_{i+20}, i = 1, 2, \ldots, 10$ identify the variables $x_2, x_4$ and $x_5$ as influential.

- The values of contrasts $w_i = y_i + y_{i+10}, i = 1, 2, \ldots, 10$ identify the variables $x_3, x_6, x_8$ and $x_{10}$ as influential.

- The two sets of active effects are not the same so the underlying model is not linear.

- Using linear regression we obtained that none of the factors were active.

- Even though this example seems to be extreme, one may fall into the same difficulties when screening with real data where things are usually more complex than in the linear case.
In general, the use of edge designs guarantees that
- irrelevant variables are never treated as relevant, in contrast to classical designs, and there is only
- a very small chance that a relevant variable is not correctly recognized, which occurs when the two function values nearly agree on the corresponding edges.
- Thus, screening with edge designs is robust.

It is of interest to construct new edge designs with desirable properties.

How addition or deletion of runs affect edge designs?
Some References

More References


