Use of an empirical Bayesian method to analyse data from saturated factorial designs

Marta Yukie Baba
Jorge Alberto Achcar
Sergio Minoru Oikawa

Departamento de Matemática
Estatística e Computação
UNESP, Universidade Estadual Paulista
Presidente Prudente, SP, Brazil

Abstract

“The use of saturated two-levels designs is very popular, especially in industrial applications where the cost of experiments is too high. Standard classical approaches are not appropriate to analyze data from saturated designs, since we only could get the estimates of main factor effects and we do not have degrees of freedom to estimate the variance of the error. In this paper, we propose the use of empirical Bayesian procedures to get inferences for data obtained from saturated designs. The proposed methodology is illustrated assuming a simulated data set”.

Keywords: saturated designs, Plackett-Burman designs, empirical Bayesian methods.
1. Introduction

Factorial designs have been extensively used by experimental researchers in many areas of interest, as agriculture, engineering, medical research, industrial research among many others since the pioneering work of Fisher (1926, 1935) and Yates (1935, 1937). The systematic exploration of factorial designs in industrial applications is described by Box, Hunter and Hunter (1978). Daniel (1959) introduced the graphical method (Q-Q plots) of the half-normal plot to explore the important factors on a response variable of interest; fractional replication was first discussed by Finney (1945). When we have primary interest only in the main effects, we could use saturated factorial designs (see for example, Plackett and Burman, 1946).

The use of saturated designs has been very popular for screening factors, especially in industrial applications where the observations usually are very expensive to be obtained (see for example, Box, Hunter and Hunter, 1978; Daniel, 1976; Wu and Hamada, 2000; Cox and Reid, 2000). This is the case when the number of factors is too large or when we have destructive tests.

A saturated design is a fractional factorial design in which the number of parameters in the main effect models is equal to the number of runs. In this paper, we consider saturated designs in which \( k = n - 1 \) main effects are considered in \( n \) experimental units or experiments without replicates. In such designs all information is used to estimate the main effect parameter, leaving no degrees of freedom to estimate the error variance. The classical analysis allows only the estimation of main effects under the assumption that interactions are negligible.

A well known two-level saturated design is based in the paper of Plackett and Burman (1946), where the constructed designs used Hadamard matrices of order \( n \), where \( n \) is a multiple of 4.

In this paper, we propose the use of empirical Bayesian methods to analyse data from a saturated design, since the use of saturated classical approach based on least squares estimation usually only permits the estimation of the main effects.

The paper is organized as follows: in section 2, we introduce a simulated data set from a saturated two-level design to motivate our approach; in section 3, we introduce a classical analysis for data from a saturated design; in section 4, we introduce a Bayesian analysis assuming conjugated or other priors for the parameters of the model; in section 5, we introduce an
empirical Bayesian approach; in section 6, we analyse the data from a saturated design introduced in section 2; finally, in section 7, we present some concluding remarks.

2. A simulated example

Consider a 12-run Plackett-Burman design and a simulated data set from the model
\[ Y = 2x_1 + x_2 + 1.5x_3 + \varepsilon, \]
where \( \varepsilon \sim N(0,0.25^2) \) and,

\[ x_i = \begin{cases} 
-1 & \text{when factor A uses "low" level (−)} \\
1 & \text{when factor A uses "high" level (+)} 
\end{cases} \]

In the same way, we obtain the codes for factors B to K. From this model used to simulate the data, we observe that the factors A, B and C are active in the experiment. The simulated data are given in Table 1 (data set introduced by Baba and Gilmour, 2006).

<table>
<thead>
<tr>
<th>Factor</th>
<th>Response</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>B</td>
</tr>
<tr>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>+</td>
<td>-</td>
</tr>
<tr>
<td>-</td>
<td>+</td>
</tr>
<tr>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>+</td>
<td>-</td>
</tr>
<tr>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>-</td>
<td>+</td>
</tr>
<tr>
<td>+</td>
<td>-</td>
</tr>
<tr>
<td>-</td>
<td>+</td>
</tr>
<tr>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>
The general model for the data from a two-level saturated design is given by,

\[ Y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + \varepsilon \]  \hspace{1cm} (1)

where \( x_i \) denotes the \( i \)-th factor main effect and \( \varepsilon \sim N(0, \sigma^2) \). The model can be expressed in matrix notation by,

\[ E[\mathbf{y}] = \mathbf{X}\beta \]  \hspace{1cm} (2)

where \( \mathbf{X} \) is a \( n \times p \) matrix showing the levels at which the factors are fixed, \( \beta \) is a \( p \times 1 \) vector of parameters and \( \mathbf{y} \) is \( n \times 1 \) vector of observations.

### 3. A classical analysis

The use of least square estimation permits to obtain estimates for the main effects, but there are no degrees of freedom to estimate the error. This is a great difficult to get inferences from data of a saturated design, that is, the usual analysis of variance (ANOVA) can not be used.

Assuming that the matrix \( \mathbf{X}'\mathbf{X} \) is nonsingular, as is the case of Plackett-Burman design, the least squares estimates of the main effects are given by,

\[ \hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{y} \]  \hspace{1cm} (3)

From the simulated data of Table 1, we get: \( \hat{\beta}_0 = 0.057; \hat{\beta}_1 = 2.005; \hat{\beta}_2 = 1.028; \hat{\beta}_3 = 1.548; \hat{\beta}_4 = 0.031; \hat{\beta}_5 = 0.091; \hat{\beta}_6 = 0.071; \hat{\beta}_7 = 0.021; \hat{\beta}_8 = -0.069; \hat{\beta}_9 = -0.010; \hat{\beta}_{10} = -0.011 \) and \( \hat{\beta}_{11} = 0.031 \). The least squares estimates were obtained using the MINITAB® software.

In practical work, usually the researches consider the use of normal plots to decide by the important factors, but the interpretation of the normal plots depends on how strongly the experimenter believes in factor sparsity.
A possible alternative to analyse data from saturated two-level designs is the use of Bayesian methods.

4. A Bayesian analysis

The likelihood function for \( \beta \) and \( \sigma^2 \) is given by,

\[
f(y | \beta, \sigma^2) = (2\pi\sigma^2)^{-n/2} \exp\left\{ -\frac{1}{2\sigma^2} (y - X\beta)'(y - X\beta) \right\} \tag{4}
\]

Expanding the quadratic form \((y - X\beta)'(y - X\beta)\) in (4), we have,

\[
f(y | \beta, \sigma^2) = (2\pi\sigma^2)^{-n/2} \exp\left\{ -\frac{1}{2\sigma^2} \left[ (\hat{\beta} - \beta)'X'X(\beta - \hat{\beta}) + Q \right] \right\} \tag{5}
\]

where \( \hat{\beta} \) is given by (3) and \( Q = (y - X\hat{\beta})'(y - X\hat{\beta}) \) is the residual sum of squares.

Observe that for saturated designs, we can estimate the vector of parameter \( \beta \), but the quantity \( Q \) is always zero, which means that the error variance cannot be estimated from the likelihood and the analysis used in the general linear model is not appropriate for saturated designs.

In this way, we assume a Bayesian approach to analyse data from a saturated design.

Different priors could be considered to analyse data from saturated designs (see Baba and Gilmour, 2006), but since data from saturated designs provide only limited information, the interpretation of these data depends heavily on the prior assumptions.

Among the different priors considered by these authors, a conjugate prior is given (from (5)) by,

\[
f(\beta, \sigma^2) = \frac{\left( \frac{a}{2} \right)^{d/2}}{(2\pi)^{d/2} \left| V \right|^{d/2} \Gamma(d/2)} \times \frac{\left( \frac{a}{2} \right)^{(d + p + 2)/2}}{(2\pi)^{(d + p + 2)/2} \left| V \right|^{(d + p + 2)/2} \Gamma(d/2)} \tag{5}
\]
\[
\times \exp \left\{ -\frac{1}{2\sigma^2} \left[ (\beta - m)'V^{-1}(\beta - m) \right] + a \right\}
\] (6)

with hyperparameters \(a > 0\), \(d > 0\), \(m \in \mathbb{R}^p\) and \(V\) is a \(p \times p\) positive definite matrix (a Normal - Inverse Gamma distribution).

Assuming different values for the hyperparameters of the prior (6), the use of conjugate priors are very inflexible since very informative priors can lead to the posterior being very vague and centered in the wrong place (see Baba and Gilmour, 2006).

Other priors as a finite mixture of densities or non-conjugate priors are also considered in the literature, but the obtained posterior summaries are heavily dependent of the choice of the prior for \(\beta\) and \(\sigma^2\).

Observe that in a saturated design, we are using \(n\) observations to estimate \(n + 1\) parameters, that is, in absence of any prior knowledge, the data do not provide any information.

5. Use of empirical Bayesian methods

To get some information for the variance \(\sigma^2\) of the error for data of a saturated design, let us assume the following procedure: among the \(g\) experiments considered in the saturated two-level design, we have for each factor, \(g/2\) values in the “high” level (+) and \(g/2\) values in the “low” level (−). From these \(g/2\) values in each level “+” or “−”, we get its standard deviations denoted by \(s_{\ell+}\) and \(s_{\ell-}\), \(l = 1, 2, \ldots, K\).

Denoting by \(y_{\ell+}\) and the \(g/2\) values in the “high” level + for factor \(l\), we have,

\[
s_{\ell+}^2 = \frac{1}{(g/2 - 1)} \sum_{i=1}^{g/2} \left( y_{\ell+} - \bar{y}_{\ell+} \right)^2
\] (7)
where $\bar{y}_{\ell*}^+ = \frac{1}{(g/2)} \sum_{i=1}^{g/2} y_{\ell i}^+, l = 1, 2, \ldots, K.$

In the same way, denoting by $y_{\ell i}^-$ the $g/2$ values in the “low” level (-) of factor $l$,

$$s_{\ell*}^2 = \frac{1}{(g/2 - 1)} \sum_{i=1}^{g/2} (y_{\ell i}^- - \bar{y}_{\ell*}^-)^2$$

where $\bar{y}_{\ell*}^- = \frac{1}{(g/2)} \sum_{i=1}^{g/2} y_{\ell i}^-, l = 1, 2, \ldots, K.$

In this way, we have $K$ values $S_{\ell*}^2$ and $S_{\ell*}^2$, that is, a total of $2K$ sample variances, or $2K$ quantities.

From these $2K$ quantities, we get the sample mean denoted by $\bar{A}$ and the sample variance denoted by $\bar{B}^2$, which can be used to find appropriated values for the hiperparameters of the prior distribution for the variance $\sigma^2$ of the error in (1).

Assuming a gamma prior distribution for $\sigma^2$, that is,

$$\sigma^2 \sim \text{Gamma}(a, b)$$

where $E[\sigma^2] = a/b$ and $\text{Var}[\sigma^2] = a/b^2$, we get values for the hyperparameters $a$ and $b$ by solving the following equations:

$$\begin{cases} \frac{a}{b} = A \\ \frac{a}{b^2} = B^2 \end{cases}$$
6. Analysis of the data of Table 1

Let us assume the simulated data of Table 1 considering a 12-run Plackett-Burman design with model \( Y = 2x_1 + x_2 + 1.5x_3 + \varepsilon \), where \( \varepsilon \sim N(0,0.25^2) \).

Considering the \( g = 12 \) observations in factor A, we have 6 observations in level “high” or “+” given by 1.7502, 2.7130, 4.4825, 1.4302, -0.5429 and 2.5350, from where we get \( s_{+1} = 1.661 \) (see (7)).

In the same way, considering the 6 observations in factor A in the level “low” or “-” we have 0.4377, -4.4129, -1.4384, -2.4942, 0.8998 and -4.6797, and \( s_{-1} = 2.362 \) (see (8)).

With the same approach, we get for other factors B to K the values \( s_{+l} \) and \( s_{-l} \) for the 11 factors.

In table 2, we have these 22 quantities \( (s_{+l}, s_{-l}) \) for the 11 factors.

Table 2. Estimated quantities \( s_{+l} \) and \( s_{-l} \)

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
<th>H</th>
<th>I</th>
<th>J</th>
<th>K</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.661</td>
<td>2.252</td>
<td>2.073</td>
<td>2.920</td>
<td>1.984</td>
<td>2.600</td>
<td>3.260</td>
<td>1.857</td>
<td>3.380</td>
<td>3.010</td>
<td>2.450</td>
</tr>
</tbody>
</table>

Considering the squares for the 22 quantities given in Table 2, we get a sample mean \( A = 8.311 \) and a sample variance \( B^2 = 3.260 \). From equation (10), we get the values of the hyperparameters of the gamma prior distribution (9) for the variance \( \sigma^2 \) of the error in (1), given by \( a = 6.4994 \) and \( b = 0.7820 \).

Also assuming normal \( N(0, 10^6) \) priors for the regression parameters \( \beta_1 = 0, \ 1 = 1, \ldots, 11 \) in (1) (that is, very non-informative priors), we use Markov Chain Monte Carlo (MCMC)
methods to get the posterior summaries of interest (see for example, Chib and Greenberg, 1995; or Gelfand and Smith, 1990).

Using the Winbugs software (Spiegelhalter et al., 2004), we simulated 5000 Gibbs samples (taking every 10th sample) of the joint posterior distribution for $\beta$ and $\sigma^2$, after a “burn-in period” of size 5000.

Convergence of the Gibbs sampling algorithm was monitored by observing the traceplots of the simulated samples for each parameter.

In Table 3, we have the posterior summaries of interest.

In Table 4, we have the observed and predicted values $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1x_1 + \ldots + \hat{\beta}_lx_l$, where $\hat{\beta}_l$ are the Monte Carlo estimates of the posterior means for $\beta_l$, $l = 0, 2, \ldots, 11$ based on the 5000 simulated Gibbs samples given in Table 3, assuming a Gamma (6.4994; 0.7820) prior for $\sigma^2$ and normal $N(0, 10^6)$ priors for $\beta_l$, $l = 0, 2, \ldots, 11$ (use of the empirical Bayesian method). Let us denote this model as “model 1”.

From the results of Table 4, we observe very good predictions assuming “model 1”, that is, good inferences for the data from a saturated two-level design.

In Table 4, we also have the predicted values considering other priors for $\sigma^2$ and normal $N(0, 10^6)$ priors for $\beta_l$ also considering the use of the Winbugs software (5000 Gibbs samples after a “burn-in period” of size 5000).
Table 3. Posterior summaries (use for empirical Bayesian methods)

<table>
<thead>
<tr>
<th>parameter</th>
<th>Posterior mean</th>
<th>S.D</th>
<th>95% credible Interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\beta_0$</td>
<td>0.0694</td>
<td>0.8030</td>
<td>(-1.5270 ; 1.6810)</td>
</tr>
<tr>
<td>$\beta_1$</td>
<td>2.0180</td>
<td>0.8290</td>
<td>(0.3696 ; 3.7140)</td>
</tr>
<tr>
<td>$\beta_2$</td>
<td>1.0270</td>
<td>0.8262</td>
<td>(-0.6176 ; 2.6250)</td>
</tr>
<tr>
<td>$\beta_3$</td>
<td>1.5490</td>
<td>0.8387</td>
<td>(-0.1182 ; 3.2030)</td>
</tr>
<tr>
<td>$\beta_4$</td>
<td>0.0291</td>
<td>0.8174</td>
<td>(-1.5900 ; 1.6920)</td>
</tr>
<tr>
<td>$\beta_5$</td>
<td>0.1135</td>
<td>0.8308</td>
<td>(-1.5220 ; 1.7120)</td>
</tr>
<tr>
<td>$\beta_6$</td>
<td>0.0488</td>
<td>0.8408</td>
<td>(-1.6250 ; 1.6880)</td>
</tr>
<tr>
<td>$\beta_7$</td>
<td>0.0153</td>
<td>0.8458</td>
<td>(-1.6420 ; 1.7150)</td>
</tr>
<tr>
<td>$\beta_8$</td>
<td>-0.0729</td>
<td>0.8198</td>
<td>(-1.7390 ; 1.5540)</td>
</tr>
<tr>
<td>$\beta_9$</td>
<td>-0.0248</td>
<td>0.8231</td>
<td>(-1.6570 ; 1.6090)</td>
</tr>
<tr>
<td>$\beta_{10}$</td>
<td>-0.0320</td>
<td>0.8441</td>
<td>(-1.7390 ; 1.5870)</td>
</tr>
<tr>
<td>$\beta_{11}$</td>
<td>0.02595</td>
<td>0.8245</td>
<td>(-1.6180 ; 1.6650)</td>
</tr>
<tr>
<td>$\sigma^2$</td>
<td>0.14370</td>
<td>0.0701</td>
<td>(0.0634 ; 0.3151)</td>
</tr>
</tbody>
</table>

Table 4. Observed and predicted values

<table>
<thead>
<tr>
<th>observed</th>
<th>Predicted “model 1”</th>
<th>Predicted “model 2”</th>
<th>Predicted “model 3”</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.7502</td>
<td>1.7810</td>
<td>-4.3830</td>
<td>-4.6797</td>
</tr>
<tr>
<td>2.7180</td>
<td>2.7940</td>
<td>6.7420</td>
<td>-4.4129</td>
</tr>
<tr>
<td>0.4377</td>
<td>0.4249</td>
<td>-0.5821</td>
<td>-2.4942</td>
</tr>
<tr>
<td>4.4825</td>
<td>4.4770</td>
<td>11.7900</td>
<td>-1.4384</td>
</tr>
<tr>
<td>1.4302</td>
<td>1.4160</td>
<td>11.7500</td>
<td>-0.5429</td>
</tr>
<tr>
<td></td>
<td>-0.5429</td>
<td>-0.4920</td>
<td>-15.5700</td>
</tr>
<tr>
<td>4.4129</td>
<td>-4.5030</td>
<td>-11.0800</td>
<td>0.8998</td>
</tr>
<tr>
<td>-1.4384</td>
<td>-1.4640</td>
<td>12.0600</td>
<td>1.4302</td>
</tr>
<tr>
<td>-2.4942</td>
<td>-2.4530</td>
<td>12.1200</td>
<td>1.7502</td>
</tr>
<tr>
<td>2.5350</td>
<td>2.5470</td>
<td>8.6140</td>
<td>2.5350</td>
</tr>
<tr>
<td>0.8998</td>
<td>0.9318</td>
<td>-13.9600</td>
<td>2.7130</td>
</tr>
<tr>
<td>-4.6792</td>
<td>-4.6270</td>
<td>10.1800</td>
<td>4.4825</td>
</tr>
</tbody>
</table>

Another possibility is to assume an uniform U(0, 1000) prior for $\sigma^2$ and normal $N(0, 10^6)$ priors for $\beta_i$. Let us denote this model as “model 2”.
A third model denoted as “model 3”, is considered assuming an uniform \( U(0, 1000) \) prior for \( \sigma^2 \) and \( N(\tilde{\beta}_1, 0.01) \) priors for \( \beta \), where \( \tilde{\beta}_1 \) are the least squares estimates for \( \beta_1 \), \( 1 = 0, 2, \ldots, 11 \) (see section 3), that is, very informative priors for the regression parameters, but very non-informative prior for \( \sigma^2 \).

From the results of Table 4, we observe that the predicted values assuming “model 2” and “model 3” are very different of the observed values, that is, we get bad predictions.

From these results, we observe that the use of the proposed empirical Bayesian method could be a powerful methodology to be used in applications of saturated two-level designs.

It is also important to point out that the obtained 95% credible intervals considering “model 1” (empirical Bayesian model) show that the factors \( x_1, x_2 \) and \( x_3 \) are active in the experiment (in agreement to the true values of \( \beta_1, \beta_2 \) and \( \beta_3 \) given by the model \( E[Y] = 2x_1 + x_2 + 1.5x_3 \) used to simulate the data) and very accurate posterior means for \( \beta_1, \beta_2 \) and \( \beta_3 \).

7. Concluding Remarks

The use of saturated factorial designs has been extensively used by industrial researchers and engineers as a powerful methodology for screening factors, especially in the presence of a great number of factors. Usually, we get some information about important factors on the response variable of interest using normal Q-Q plots. The use of these normal plots in the interpretation of factorial two-level experiments usually could be very subjective and in many cases we could be in difficulties to find the active factors on the response of interest.

The use of the empirical Bayesian approach introduced in this paper, could be of great interest in applications.

We also observed very good predictions and inferences for the parameters of interest using our proposed methodology.
References


DANIEL, C. (1959) Use of half-normal plot in interpreting factorial two-level experiments. Technometrics, 1, 149.


