Journal of Statistical Planning and Inference 143 (2013) 1195–1202

Contents lists available at SciVerse ScienceDirect

Journal of Statistical Planning and Inference

journal homepage: www.elsevier.com/locate/jspi

A note on robustness of D-optimal block designs for two-colour microarray experiments

R.A. Bailey [a,*], Katharina Schiffl [b,c], Ralf-Dieter Hilgers [c]

[a] School of Mathematical Sciences, Queen Mary University of London, Mile End Road, London E1 4NS, UK
[b] Für Medizinsche Statistik, Aachen, Germany
[c] Aachen University, Department of Medical Statistics, 52074 Aachen, Germany

ARTICLE INFO

Article history:
Received 26 February 2012
Received in revised form
4 January 2013
Accepted 16 January 2013
Available online 29 January 2013

Keywords:
Breakdown number
Design of microarray experiments
D-optimality
Graph theory
Robustness

ABSTRACT

Two-colour microarray experiments form an important tool in gene expression analysis. Due to the high risk of missing observations in microarray experiments, it is fundamental to concentrate not only on optimal designs but also on designs which are robust against missing observations. As an extension of Latif et al. (2009), we define the *optimal* breakdown number for a collection of designs to describe the robustness, and we calculate the breakdown number for various D-optimal block designs. We show that, for certain values of the numbers of treatments and arrays, the designs which are D-optimal have the highest breakdown number. Our calculations use methods from graph theory.

1. Introduction

Microarrays play a key role in modern molecular biology since they enable simultaneous monitoring of the expression levels of thousands of genes: see, for example, Brown and Botstein (1999). The main goal of cDNA microarray experiments is to identify significantly up- or down-regulated genes. These genes serve as possible targets for therapy for severe diseases, such as malignant tumours. Since two samples of different treatments are coloured green and red and are applied (hybridized) onto one microarray, designs for microarray experiments can be considered as row–column designs with two rows corresponding to the two dyes. Ignoring the two different dyes, the designs can be considered as incomplete-block designs with block size 2, provided that the number of treatments exceeds two.

Microarray experiments have been widely studied in the literature. For instance, Kerr et al. (2000) first recommended analysing microarray data with analysis-of-variance models. Most papers focus on the derivation of optimal designs in specific scenarios, but only a few authors address the problem of missing values. Missing values often occur in microarray experiments: for example, due to insufficient resolution, image corruption, or simply dust or scratches on the slide (Latif et al., 2009). Thus, this data cannot be involved in the analysis of the experiment (Troyanskaya et al., 2001) and so it is important to use robust experimental designs, which ensure estimability of the treatment effects even if a few observations are missing. Latif et al. (2009) investigated specific robustness properties of commonly used microarray designs. They proposed two robustness criteria and calculated these criteria for the commonly used designs. However, to date no attempts have been made to investigate these robustness criteria analytically. We will derive an upper bound for

* Corresponding author. Tel.: +44 20 7882 5517; fax: +44 20 7882 7684.
E-mail addresses: r.a.bailey@qmul.ac.uk (R.A. Bailey), kschiffl@web.de (K. Schiffl), rhilgers@ukaachen.de (R.-D. Hilgers).

0378-3758/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.jspi.2013.01.005

the breakdown number, which enables us to define an optimal breakdown number and then investigate some published optimal designs with respect to the breakdown number.

This paper is structured as follows. Section 2 introduces the statistical model which is used to describe microarray experiments. Robustness criteria are defined in Section 3, where optimal robustness properties are derived. Section 4 shows that several families of published D-optimal designs achieve the optimal breakdown number, under a slight simplification of the original model. Implications for the full model are discussed in Section 5, and a short conclusion is given in Section 6.

2. Preliminaries

Suppose that there are t treatments and a arrays. The statistical analysis is based on the gene-specific model

$$\log_2(y_{ij\ell}) = \tau_i + \alpha_j + \delta_\ell + \epsilon_{ij\ell}, \tag{1}$$

where $y_{ij\ell}$ describes the intensity of treatment i coloured in dye ℓ on array j, for $i \in \{1,\ldots,t\}$, $j \in \{1,\ldots,a\}$ and $\ell \in \{\text{green,red}\}$, and $\epsilon_{ij\ell}$ are the error terms.

Suppose that array j has treatments i and k coloured green and red respectively. For analysis using intra-array information only, model (1) can be replaced by

$$\log_2\left(\frac{y_{ijgreen}}{y_{kjred}}\right) = \tau_i - \tau_k + \delta_{green} - \delta_{red} + \epsilon_{ijgreen} - \epsilon_{kjred}. \tag{2}$$

As in Bailey (2007, Sections 2–6), we begin by ignoring the dye effect in the consideration of robustness and optimality; that is, we assume that $\delta_{green} = \delta_{red}$. For $j=1,\ldots,a$, put $z_j = \log_2(y_{ijgreen}) - \log_2(y_{kjred})$. Then, written in matrix notation, model (2) simplifies to

$$z = X\tau + \eta, \tag{3}$$

where z is the a-dimensional vector $[z_1,\ldots,z_a]^\top$, τ is the t-dimensional vector $[\tau_1,\ldots,\tau_t]^\top$ of unknown treatment effects, and X is the $a \times t$ design matrix, with each row containing exactly one 1 and one -1, all other entries being equal to zero. The term η is the random error vector with independent identically distributed components having expectation zero and variance σ^2.

This approach can always be used for blocks of size 2, whether or not the treatments are coloured: see Cox and Snell (1981, Example I). We return to the full model (2) in Section 5.

In most situations one is interested in estimating all linear contrasts of the parameter vector τ. A design is called *connected* if all linear contrasts in τ are estimable. If the design matrix is X then the matrix $X^\top X$ is called the *information matrix* of the design. Let $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_{t-1} \geq \lambda_t$ be the eigenvalues of $X^\top X$; these are non-negative. The entries in each row of $X^\top X$ sum to zero, so $\lambda_t = 0$. It can be shown (Shah and Sinha, 1989) that the design is connected if and only if the remaining $t-1$ eigenvalues are non-zero. In this case, the vector τ is estimable in the hyperplane $\sum_i \tau_i = 0$, and the volume of the confidence ellipsoid for τ is inversely proportional to $\sqrt{\prod_{i=1}^{t-1} \lambda_i}$. Thus, a design is called *D-optimal* if it maximizes the value of $\prod_{i=1}^{t-1} \lambda_i$.

3. Optimal breakdown number

Latif et al. (2009) introduced the *breakdown number* for microarray experiments, but they did not derive designs with optimal breakdown numbers for given values of t and a. Adapting their definition to the case where all linear contrasts are to be estimated gives the following.

Definition. Assume the model (3), with $a \times t$ design matrix X. Given any subset S of $\{1,\ldots,a\}$, let X_S be the design matrix obtained from X by deleting the rows corresponding to the arrays in S. The *breakdown number* of the design is equal to m if all contrasts are estimable with reduced design matrix X_S for all subsets S of size $m-1$ (that is, with $m-1$ arrays in which one or both observations are missing) but there exists at least one subset S of size m for which not all contrasts are estimable.

Note that, for $1 \leq n \leq a$, every design matrix X with a rows gives $\binom{a}{n}$ matrices X_S.

Since designs with large breakdown numbers can be considered robust, we aim to search for designs which maximize the breakdown number. Let $\Omega_{t,a,2}$ be the collection of all binary designs for t treatments using a arrays of size 2.

Definition. A design in $\Omega_{t,a,2}$ has *optimal breakdown number* if it maximizes the breakdown number over all designs in $\Omega_{t,a,2}$.

Each design in $\Omega_{t,a,2}$ can be considered as a graph with vertices $1,\ldots,t$. The number of edges joining distinct vertices i and k is equal to the number of arrays where treatments i and k are applied. If the edge e joins vertices i and k, then i and e are said to be *incident* with each other, and i and k are said to be *adjacent* to each other. Two edges are adjacent if they have a vertex in common. The degree of vertex i, denoted $\deg(i)$, is the number of edges incident with vertex i.