Journal of Statistical Planning and Inference 92 (2001) 267–282

journal of statistical planning and inference

www.elsevier.com/locate/jspi

An application of information theory and error-correcting codes to fractional factorial experiments

Irad Ben-Gal[a,*], Lev B. Levitin[b]

[a] *Department of Industrial Engineering, Tel-Aviv University, Ramat-Aviv, Tel-Aviv 69978, Israel*
[b] *Department of Electrical & Computer Engineering, Boston University, 8 St. Mary's Street, Boston, MA 02215, USA*

Received 1 September 1999; received in revised form 7 May 2000; accepted 16 June 2000

Abstract

The objective of *design of experiments* (DOE) is addressed by introducing an information optimality criterion, which is based on concepts adopted from *information theory*. In particular, experiments are specified to maximize the information in the system responses about estimators of the system parameters. It is shown that one has to maintain certain *resolution* of the design matrix to maximize the information, obtainable by a design, about a system described by a linear model with interactions. The correspondence between *error-correcting codes* and *fractional factorial experiments* provides a method to attain the required resolution with a smaller fractional factorial experiment by increasing the number of levels associated with each factor – a result that in the context of experimental design seems counterintuitive. In particular, the Gilbert–Varshamov and the Singleton bounds are employed to obtain bounds on the size of the fractional experiment. Analytical approximations and numerical results are given and illustrated by examples. © 2001 Elsevier Science B.V. All rights reserved.

MSC: 62K15; 62B10

Keywords: Design of experiments; Information in experiments; H-optimal designs; Coding theory

1. Introduction

Fractional factorial experiments (FFE) are often applied to *screening experiments* in which many factors are considered with the purpose of identifying those that have a significant effect on the *response*. Screening experiments are often used as building blocks to fit empirical response models, seeking to relate a *response* Y to the values of

[*] Corresponding author.
E-mail addresses: bengal@eng.tau.ac.il (I. Ben-Gal), levitin@enga.bu.edu (L.B. Levitin).

PII: S0378-3758(00)00165-8

control factors $x_1,\dots,x_n$ where the underlying relationship is unknown. The empirical model can be written as

$$Y = g(x_1,x_2,\dots,x_n;B_1,\dots,B_p) + \varepsilon, \tag{1.1}$$

where g approximates an unknown function by a first- or second-order polynomial in $x_1,\dots,x_n$ with coefficients $B_1,B_2,\dots,B_p$, which are the estimators of the unknown system parameters $\beta_1,\beta_2,\dots,\beta_p$ and ε represents the observation error (noise). In practice, estimators are usually obtained by the method of least squares or maximum likelihood from a set of m experiments. Then, the experiments are represented by the $m \times p$ design matrix, $\boldsymbol{x}$, whose rows correspond to the experiments and columns are associated with the system parameters (untraditionally, the design matrix is denoted here by small letter to distinguish it from random variables that are denoted by capital letters).

In this paper, the task of defining a fraction of a factorial design, out of all possible combinations of factor levels, is considered and analyzed from an information–theoretical perspective. Section 2 gives some background on Shannon's information measure, the construction of linear error-correcting codes and its correspondence to the construction of FFE. In Section 3, we introduce an information optimality criterion, which is based on Shannon's information measure. The new criterion seeks to maximize the information gained from experiments about estimators of the system parameters. It is shown that, for systems described by a multiple linear regression model, this criterion requires to maintain a certain resolution of the design matrix. The problem of obtaining a desired resolution with limited number of experiments is then analyzed in Section 4, by using the isomorphism between the construction of FFE and linear error-correcting codes. In particular, the desired resolution is achieved with an FFE of a smaller size by increasing the number of levels associated with each factor – a result that seems somewhat paradoxical. Bounds on the size of the fractional factorial designs are given by use of the Gilbert–Varshamov and the Singleton bounds. Section 5 concludes the paper.

2. Background

2.1. Shannon's information measure

Let Y and Λ be two continuous random variables (r.v.'s) with marginal probability density functions (pdfs) denoted, respectively, by $f_Y(y)$ and $f_\Lambda(\lambda)$ and a joint pdf $f(y,\lambda)$ (random variables, other than ε, are denoted here by capitals, and their values by small letters). The *information in Y about Λ*, denoted by $I(Y;\Lambda)$, has been introduced by Shannon (1948) and defined as

$$I(Y;\Lambda) = H(\Lambda) - H(\Lambda \mid Y) = \int_{\{y,\lambda\}} f(y,\lambda)\log\frac{f(y,\lambda)}{f_Y(y)f_\Lambda(\lambda)}\,\mathrm{d}y\,\mathrm{d}\lambda, \tag{2.1}$$

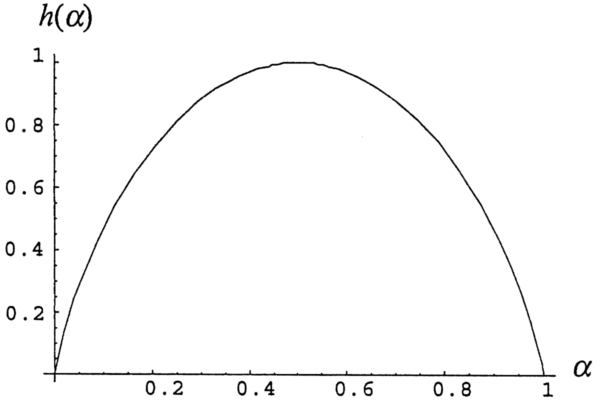

Fig. 1. The binary entropy of a discrete random variable.

where $H(\Lambda)$ is the *differential entropy* of Λ defined as

$$H(\Lambda) = -\int_{\{\lambda\}} f_\Lambda(\lambda) \log f_\Lambda(\lambda)\, \mathrm{d}\lambda \tag{2.2}$$

and $H(\Lambda \mid Y)$ is the *conditional differential entropy* of Λ given Y which is defined as the expected value of the entropy of the conditional distribution, averaged over the conditioning random variable, i.e.,

$$H(\Lambda \mid Y) = -\int_{\{y,\lambda\}} f(\lambda, y) \log f_{\Lambda|Y}(\lambda \mid y)\, \mathrm{d}y\, \mathrm{d}\lambda. \tag{2.3}$$

Thus, information is interpreted as the reduction of the entropy of one r.v. conditioned by another r.v. and entropy is used as a measure of uncertainty. For example, in Fig. 1 the entropy of a discrete r.v. taking one value with probability α and another value with probability $1-\alpha$, is plotted. The *binary entropy function*, given by $h(\alpha) = -\alpha \log \alpha - (1-\alpha)\log(1-\alpha)$, is measured in shannons (or bits) if the log is to the base 2. Note that the entropy depends only on the probability distribution of the r.v. and not on the values taken by the r.v. It is a concave function of the probability α and equals 0 when $\alpha = 0$ or 1, i.e., when the variable is not random and no uncertainty is associated with its outcome. Moreover, it can be shown that Shannon's definition of entropy is closely related to the notion of entropy in thermodynamics (e.g., see Cover and Thomas, 1991).

2.2. Linear error-correcting codes (ECC) and fractional factorial experiments (FFE)

Block codes introduce controlled amounts of redundancy into transmitted data stream, providing the receiver with the ability to detect and to correct errors caused by noise in the communication channel. A *q-ary block code* $\boldsymbol{C}$ of *length* n and *size* M consists of

M codewords $\{\boldsymbol{c}_0, \boldsymbol{c}_1, \ldots, \boldsymbol{c}_{M-1}\}$, each codeword being a vector from $\boldsymbol{Z}_q^n$: $\boldsymbol{c}_i \in \boldsymbol{Z}_q^n$ $(i = 0, 1, \ldots, M-1)$ where $\boldsymbol{Z}_q = \{0, 1, \ldots, q-1\}$. If any error that results in a change of values of no more than t symbols in a codeword can be corrected with code $\boldsymbol{C}$, the code is called *t-error-correcting code*.

The *Hamming distance* between two vectors from $\boldsymbol{Z}_q^n$ is the number of coordinates in which these words differ. The *distance d* of a block code $\boldsymbol{C}$ is the minimum Hamming distance between any two distinct codewords from $\boldsymbol{C}$. It can be seen that $d = 2t + 1$. If $M = q^\kappa$ and $\boldsymbol{C}$ is a linear subspace of $\boldsymbol{Z}_q^n$, $\boldsymbol{C}$ is called a *linear* (n, κ) code of *dimension* κ. Since every κ-digit q-ary message can be encoded by a codeword of $\boldsymbol{C}$, κ is the number of *information digits*, $r = n - \kappa$ is the number of *redundant digits*, or *redundancy* of the code (see, e.g., Wicket, 1995) for more details on error-correcting codes).

To construct a linear q-ary (n, κ) code with distance d, one has to find a linear subspace $\boldsymbol{C}$ in $\boldsymbol{Z}_q^n$ of dimension κ such that any vector in $\boldsymbol{C}$ has a *weight* (number of nonzero components) not less than d. The null space of $\boldsymbol{C}$ (called also the *dual subspace* or *orthogonal subspace*) is a linear subspace of $\boldsymbol{Z}_q^n$ of order $q^{n-\kappa}$ (dimension r) that is interpreted as the *dual code* $\boldsymbol{C}^\perp$. Any vector from $\boldsymbol{C}$ is orthogonal to any vector from $\boldsymbol{C}^\perp$. $\boldsymbol{C}^\perp$ can be represented by a matrix whose rows are the q^r vectors (codewords) of $\boldsymbol{C}^\perp$. Then, the n column vectors of $\boldsymbol{C}^\perp$ form a column space of rank r, which is a linear subspace of $\boldsymbol{Z}_q^\zeta$, $\zeta = q^r$.

It has been known that there exists a correspondence between error-correcting codes and fractional factorial designs. Bose (1950, 1961) has presented a mathematical isomorphism between those structures and showed how both the problem of determining the alphabet of a code and the problem the selection of experiments to be included in an FFE design (the subset selection problem) can be reduced to the κ-surjective matrix problem. Delsarte (1973) has obtained a linear programming bound for orthogonal arrays. Sloane and Stufken (1996) extended this bound for orthogonal arrays with mixed levels. Hardin and Sloane (1993) suggested a numerical algorithm to construct optimal designs for several performance criteria. We follow Kishen (1948) and Bose (1961) in the presentation below.

An isomorphism between ECC and FFE can be established (see Bose, 1961) by mapping vectors of $\boldsymbol{Z}_q^n$ to a group of elements, E_i, $i = 1, \ldots, q^n$, where the zero vector is mapped to the identity element of the group. Addition of vectors (component-wise addition modulo q) is preserved and represented by multiplication of elements in the group. Multiplication (modulo q) of a vector by a constant $\boldsymbol{p}$ $(\boldsymbol{p} = 0, 1, \ldots, q-1)$ is preserved and represented by multiplication of the exponents (by $\boldsymbol{p}$ modulo q) of generators of the group forming the image of the vector. Then, the problem of constructing a fractional factorial experiment (FFE) with resolution d is directly related to the problem of constructing a t-error-correcting code (ECC) with distance $d = 2t + 1$. The images of the basis elements of the vector space $\boldsymbol{Z}_q^n$, denoted here by $F_1, F_2, \ldots, F_n$ are the generators of the group that can be identified with the *n factors* of a factorial experiment, in which each factor can be chosen at one of q distinct *levels*. An individual *treatment* in which the factor F_i is experimented at the level q_i $(i = 0, 1, \ldots, n)$ may be written as

$$f_1^{q_1} f_2^{q_2} \ldots f_n^{q_n}, \quad 0 \leqslant q_i \leqslant q-1. \tag{2.4}$$

where each treatment is a row in the design matrix.

Then, an element E of the group can be expressed in the form

$$E = F_1^{a_1} F_2^{a_2} \ldots F_n^{a_n} \quad \text{where } 0 \leqslant a_i \leqslant q-1,\ i = 1,\ldots,n. \tag{2.5}$$

E is a *t-factor interaction*, if t of the numbers $a_1, a_2, \ldots, a_n$ are nonzero (when $t=1$, E is a *main factor* effect). Any κ-independent interactions $E_1, E_2, \ldots, E_\kappa$ (such that their images in the vector space $\boldsymbol{Z}_q^n$ are linearly independent) generate a subgroup G of order q^κ, which corresponds to the linear code $\boldsymbol{C}$. Accordingly, the images of the codewords of the dual code $\boldsymbol{C}^\perp$ (which are the rows of the design matrix) constitute a $q^{-\kappa}$th fraction of the total number of q^n possible experiments. Hence, this subgroup is a *fractional factorial experiment* (FFE). Let L be any interaction not belonging to G. Then the interactions $LE_1^{a_1} E_2^{a_2} \ldots E_\kappa^{a_\kappa}$ $0 \leqslant a_i \leqslant q-1$, $i = 1, 2, \ldots, \kappa$, are *aliases* of L. The set of all such interactions for given L is said to be the *alias set* of L (this alias set is the image of a coset of the linear code in $\boldsymbol{Z}_q^n$). By observing the responses of a fractional factorial experiment, one can estimate the effects of the sum of all the aliases of L, though the effect of L individually cannot be estimated (Bose 1950, 1961; Finney, 1945; Kishen, 1948). Since it is in general more important to estimate lower-order interactions, it is of interest to choose the fundamental subgroup G in such a way that (for a specified t) no t-factor or lower-order interaction is aliased with another t-factor or lower-order interaction, i.e., such that any alias set should not contain more than one t-factor or lower order interaction. A fractional factorial design with such a property is said to have a *resolution d*, where $d = 2t+1$. It is known that in order to achieve such resolution d it is necessary and sufficient that every interaction represented by a nonzero element of G should have $2t+1$ or more factors (Bose, 1961). Thus, the distance d of the primary code $\boldsymbol{C}$ appears to be an important parameter of a fractional factorial experiment, where it is interpreted as the experiment resolution with the properties outlined above.

3. Information optimality criterion in experimental design (H-optimality)

3.1. Measuring the information obtained in sequential experiments

Let us apply Shannon's information measure to the subset selection problem and define a family of information quantities. Consider a system described by an empirical model where Y and Λ are continuous random variables representing, respectively, the experiment response and the estimator of an unknown characteristic of the system. A conceivable formulation of the subset selection task (which seeks to define a subset of experiments out of all possible combinations of factor levels) can now be written as achieving $\max_{\boldsymbol{x}} [I(Y;\Lambda)]$, i.e., as maximizing the information in the experiment responses about a system estimator over a set of feasible designs. Our approach is similar to those suggested by Box and Hill (1967) and Fedorov (1972) where experiments are designed to provide maximum discrimination among various models or hypotheses based on the information measure.

In a sequence of k experiments, one can consider the *conditional information*,

$$I(Y_k;\Lambda\,|\,Y_1,\dots,Y_{k-1}) = H(\Lambda\,|\,Y_1,\dots,Y_{k-1}) - H(\Lambda\,|\,Y_1,\dots,Y_{k-1},Y_k), \tag{3.1}$$

which, in the context of experimental design, is interpreted as the *incremental information* gained from the kth experiment response Y_k, given the responses of previous experiments $Y_1,\dots,Y_{k-1}$. Since information satisfies the chain rule, the *total information*, which is gained from a set of experiments, can be expressed as

$$I(Y_1,\dots,Y_{K-1},Y_K;\Lambda) = \sum_{k=1}^{K} I(Y_k;\Lambda\,|\,Y_1,\dots,Y_{k-2},Y_{k-1}). \tag{3.2}$$

Let us specify an information criterion for the multi-dimensional case, when one considers information in the experiment responses about the parameter estimators. Specifically, we define $\Lambda \equiv \boldsymbol{B}$, where $\boldsymbol{B}$ is a random vector of parameter estimators. The information criterion for choosing the design matrix $\boldsymbol{x}$ is then to achieve $\max_{\boldsymbol{x}}[I(\boldsymbol{Y};\boldsymbol{B})]$, where $\boldsymbol{Y}$ is a random vector of system responses. In particular, consider an experiment that consist of m individual treatments with n control factors represented by the following p-dimensional, multiple linear regression model,

$$\boldsymbol{Y} = \boldsymbol{x}\beta + \varepsilon, \tag{3.3}$$

where, ε is a m-dimensional (m-dim) vector of i.i.d. Gaussian random variables with zero mean and variance σ^2; β is a p-dim vector of unknown parameters; $\boldsymbol{x}$ is a $m \times p$ design matrix of controlled factors, which is to be determined by the designer; and $\boldsymbol{Y}$ is a random vector of experiment responses which is m-variate normally distributed.

The additive Gaussian noise models are considered here for certain reasons. First, the Gaussian distribution maximizes the entropy over all distributions with the same covariance matrix (Cover and Thomas, 1991). Thus, the normal distribution provides us with an upper bound on the uncertainty of a r.v. with an unknown pdf. Second, the normal distribution is widely used in DOE and regression models, and is practically justified in many situations by the Central Limit Theorem. Last, the Gaussian models allow us to obtain general analytic expressions.

Let $\boldsymbol{B}$ be the maximum likelihood estimator of β, which is also the least-squares estimator for the Gaussian case. As well known (e.g., see Myers and Montgomery, 1995), $\boldsymbol{B}$ is p-variate normally distributed, i.e.,

$$\boldsymbol{B} = (\boldsymbol{x}'\boldsymbol{x})^{-1}\boldsymbol{x}'\boldsymbol{Y} \sim N_p(\beta,\sigma^2(\boldsymbol{x}'\boldsymbol{x})^{-1}), \tag{3.4}$$

where $\boldsymbol{x}$ is a design matrix of rank p ($p \leqslant m$), which guarantees that the matrix $\boldsymbol{x}'\boldsymbol{x}$ is nonsingular, and $\boldsymbol{x}'$ is $\boldsymbol{x}$ transposed.

Let $\boldsymbol{Y}_k$ be the response vector of the kth experiment in a multiple linear regression model,

$$\boldsymbol{Y}_k = \boldsymbol{x}_k\beta + \varepsilon, \tag{3.5}$$

where $\boldsymbol{x}_k$ is the design matrix used in the kth experiment. Then, the maximum likelihood estimator $\boldsymbol{B}_k$, whose pdf $f_{\boldsymbol{B}_k}(\boldsymbol{b}_k) = f_{\boldsymbol{B}|\boldsymbol{Y}_1,\dots,\boldsymbol{Y}_k}(\boldsymbol{b}\,|\,\boldsymbol{y}_1,\dots,\boldsymbol{y}_k)$, is distributed as follows:

$$\boldsymbol{B}_k \sim N_p\left[\left(\sum_{i=1}^{k}\boldsymbol{x}_i'\boldsymbol{x}_i\right)^{-1}\left(\sum_{i=1}^{k}\boldsymbol{x}_i'\boldsymbol{y}_i\right);\left(\sum_{i=1}^{k}\boldsymbol{x}_i'\boldsymbol{x}_i\right)^{-1}\sigma^2\right], \tag{3.6}$$

where the pdf of the estimator represents the "partial knowledge" about the system parameters resulting from a series of k experiments.

The conditional marginal distribution of $\boldsymbol{Y}_k$ can be obtained by applying a Bayesian inference approach using the previous $k-1$ responses and designs. Thus,

$$\begin{aligned}&\boldsymbol{Y}_k\,|\,(\boldsymbol{Y}_1=\boldsymbol{y}_1,\dots,\boldsymbol{Y}_{k-1}=\boldsymbol{y}_{k-1})\\ &\quad\sim N_m\left[\boldsymbol{x}_k\left(\sum_{i=1}^{k-1}\boldsymbol{x}'_i\boldsymbol{x}_i\right)^{-1}\left(\sum_{i=1}^{k-1}\boldsymbol{x}_i'\boldsymbol{y}_i\right);\left(\sum_{i=1}^{k-1}\boldsymbol{x}_i'\boldsymbol{x}_i\right)^{-1}\left(\sum_{i=1}^{k}\boldsymbol{x}_i'\boldsymbol{x}_i\right)\sigma^2\right].\end{aligned} \tag{3.7}$$

Recall (e.g. Cover and Thomas, 1991) that the differential entropy of a Gaussian p-dimensional r.v. Λ with a covariance matrix Γ is given by

$$H(\Lambda)=\tfrac{1}{2}\log(2\pi e)^p\det\Gamma. \tag{3.8}$$

Now, the *incremental information* in the responses about the parameter estimators can be obtained.

Theorem 1. *The* incremental information *in the responses about the parameter estimators in a Gaussian multiple linear regression model is given by*

$$I(\boldsymbol{Y}_k;\boldsymbol{B}|\boldsymbol{Y}_1,\dots,\boldsymbol{Y}_{k-1})=\frac{1}{2}\log\det\left[\boldsymbol{I}_p+\boldsymbol{x}_k'\boldsymbol{x}_k\left(\sum_{i=1}^{k-1}\boldsymbol{x}_i'\boldsymbol{x}_i\right)^{-1}\right], \tag{3.9}$$

*where $\boldsymbol{I}_p$ is the p-*dim *identity matrix.*

Proof. The conditional distribution of $\boldsymbol{B}_k$, $f_{\boldsymbol{B}|\boldsymbol{Y}_1,\dots,\boldsymbol{Y}_k}(\boldsymbol{b}|\boldsymbol{y}_1,\dots,\boldsymbol{y}_k)$, is given by (3.6). Then, using (3.8) to express differential entropies of $\boldsymbol{B}_k$ and $\boldsymbol{B}_{k-1}$, one obtains that

$$\begin{aligned}I(\boldsymbol{Y}_k;\boldsymbol{B}|\boldsymbol{Y}_1,\dots,\boldsymbol{Y}_{k-1})&=H(\boldsymbol{B}|\boldsymbol{Y}_1,\dots,\boldsymbol{Y}_{k-1})-H(\boldsymbol{B}|\boldsymbol{Y}_1,\dots,\boldsymbol{Y}_{k-1},\boldsymbol{Y}_k)\\ &=\frac{1}{2}\log\det\left[\left(\sum_{i=1}^{k}\boldsymbol{x}_i'\boldsymbol{x}_i\right)\left(\sum_{i=1}^{k-1}\boldsymbol{x}_i'\boldsymbol{x}_i\right)^{-1}\right]\\ &=\frac{1}{2}\log\det\left[\boldsymbol{I}_p+\boldsymbol{x}_k'\boldsymbol{x}_k\left(\sum_{i=1}^{k-1}\boldsymbol{x}_i'\boldsymbol{x}_i\right)^{-1}\right].\quad\square\end{aligned} \tag{3.10}$$

One can also calculate the total amount of information gained from K sequential experiments as follows.

Corollary 1. *The* total information *gained from K experiment responses about the parameter estimators in a Gaussian multiple linear regression model is given by*

$$I(\boldsymbol{Y}_2,\ldots,\boldsymbol{Y}_{K-1},\boldsymbol{Y}_K;\boldsymbol{B}|\boldsymbol{Y}_1)=\frac{1}{2}\log\det\left[\left(\sum_{k=1}^{K}\boldsymbol{x}_k'\boldsymbol{x}_k\right)(\boldsymbol{x}_1'\boldsymbol{x}_1)^{-1}\right]. \tag{3.11}$$

Proof. The proof follows from (3.2) and (3.10) by summation of $K-1$ incremental information terms:

$$\begin{aligned}I(\boldsymbol{Y}_2,\ldots,\boldsymbol{Y}_{K-1},\boldsymbol{Y}_K;\boldsymbol{B}|\boldsymbol{Y}_1)&=\sum_{k=2}^{K}I(\boldsymbol{Y}_k;\boldsymbol{B}|\boldsymbol{Y}_1,\ldots,\boldsymbol{Y}_{k-2},\boldsymbol{Y}_{k-1})\\&=\sum_{k=2}^{K}\frac{1}{2}\log\det\left[\left(\sum_{i=1}^{k}\boldsymbol{x}_i'\boldsymbol{x}_i\right)\left(\sum_{i=1}^{k-1}\boldsymbol{x}_i'\boldsymbol{x}_i\right)^{-1}\right]\\&=\frac{1}{2}\log\prod_{k=2}^{K}\det\left[\left(\sum_{i=1}^{k}\boldsymbol{x}_i'\boldsymbol{x}_i\right)\left(\sum_{i=1}^{k-1}\boldsymbol{x}_i'\boldsymbol{x}_i\right)^{-1}\right]\\&=\frac{1}{2}\log\det\prod_{k=2}^{K}\left[\left(\sum_{i=1}^{k}\boldsymbol{x}_i'\boldsymbol{x}_i\right)\left(\sum_{i=1}^{k-1}\boldsymbol{x}_i'\boldsymbol{x}_i\right)^{-1}\right]\\&=\frac{1}{2}\log\det\left[\left(\sum_{k=1}^{K}\boldsymbol{x}_k'\boldsymbol{x}_k\right)\left(\boldsymbol{x}_1'\boldsymbol{x}_1\right)^{-1}\right].\qquad\square\end{aligned} \tag{3.12}$$

Here the conditioning over $\boldsymbol{Y}_1$ comes from the use of the Bayesian inference approach in situations when the designer has no advance knowledge about the pdf of $\boldsymbol{B}$. Then, the vector of the first experiment responses $\boldsymbol{Y}_1$ enables the designer to establish a prior distribution of $\boldsymbol{B}$ which is updated by successive experiment responses.

3.2. The H-optimality criterion and orthogonal designs

Observations given below may assist in specifying designs that maximize the total information. We call these designs *H-optimal*, where H stands for entropy. The first observation relates the *H-opt*imality criterion to the well-known D-optimality criterion.

Theorem 2. *In the multiple linear regression model with an additive Gaussian noise the H-*opt*imality criterion and the D-optimality criterion coincide.*

Proof. The D-optimality criterion implies minimization of the determinant of the covariance matrix of the vector of parameter estimators $\boldsymbol{B}_k$. In our case, as seen from (3.6), the D-optimality criterion requires that in a series of K experiments the determinant δ^{-1} of the matrix

$$\left(\sum_{i=1}^{K}\boldsymbol{x}_i'\boldsymbol{x}_i\right)^{-1}$$

is to be minimized. This is equivalent to maximization of the determinant δ of the inverse matrix

$$\sum_{i=1}^{K} \boldsymbol{x}_i'\boldsymbol{x}_i.$$

On the other hand, by (3.11), the total information gained in K experiments is a monotonically increasing function of δ (for given $\boldsymbol{x}_1'\boldsymbol{x}_1$). Thus, maximization of the total information $I(\boldsymbol{Y}_2,\ldots,\boldsymbol{Y}_{K-1},\boldsymbol{Y}_K;\boldsymbol{B}|\boldsymbol{Y}_1)$ is equivalent to minimization of δ^{-1}, which proves the theorem. □

The coincidence of *H-opt*imality and *D*-optimality in the Gaussian case should not be surprising: it is simply due to the one-to-one relation between the differential entropy and the determinant of the covariance matrix, as seen from (3.8). *D*-optimal designs have been extensively investigated in DOE literature (see, e.g., Hardin and Sloane, 1993; Keiefer and Wolfowitz, 1959; St. John and Draper, 1975; Wynn, 1970). In particular, it is known that for multiple linear regression model with coded factors (i.e. factors with level range from -1 to 1), the *D*-optimality criterion (and hence, in Gaussian case, the *H-opt*imality criterion) requires the normalized design matrix to be *orthogonal* so that all off-diagonal elements of $\boldsymbol{x}'\boldsymbol{x}$ are zeros and the diagonal elements of $\boldsymbol{x}'\boldsymbol{x}$ are forced to be as large as possible (Box and Draper, 1971; Montgomery, 1991; Myers and Montgomery, 1995). For a normalized matrix the covariance matrix is the identity.

Orthogonal designs play an important role in DOE. First, since the covariance matrix of $\boldsymbol{B}$ is given by $\sigma^2(\boldsymbol{x}'\boldsymbol{x})^{-1}$, it is clear that by making the off-diagonal elements of $\boldsymbol{x}'\boldsymbol{x}$ zeros, one keeps components of $\boldsymbol{B}$ uncorrelated, and by maximizing the diagonal elements of $\boldsymbol{x}'\boldsymbol{x}$, one minimizes the individual variance of the components of $\boldsymbol{B}$. Second, if columns of the design matrix are orthogonal, then the vectors of levels of the factors (or interactions) associated with these columns, are linearly independent. Thus, their effects are not *aliased* and can be estimated from the experiment independently of each other. The minimum number of columns that are linearly dependent determines the design *resolution*. Thus, orthogonality of the design matrix $\boldsymbol{x}$ requires a resolution level that assures that all the model terms are not aliased with each other. In general, a resolution at least $d=2t+1$ is necessary and sufficient to make all interactions of order t or lower independent (note, however, that for certain models no FFE can provide orthogonality).

Consider now the *H-opt*imality criterion where information is obtained by using orthogonal designs with coded factors. Theorem 3 and Corollary 3 then follows.

Theorem 3. *The* incremental information *gained from the response of the kth orthogonal experiment with coded factors about the parameter estimators in a Gaussian multiple linear regression model is given by*

$$I(\boldsymbol{Y}_k;\boldsymbol{B}|\boldsymbol{Y}_1,\ldots,\boldsymbol{Y}_{k-1})=\frac{p}{2}\log\left(\frac{k}{k-1}\right). \tag{3.13}$$

Proof. Orthogonal designs with coded factors satisfy $\boldsymbol{x}'\boldsymbol{x} = m \cdot \boldsymbol{I}_p$. By applying this result to (3.9) the following incremental information is obtained:

$$I(\boldsymbol{Y}_k; \boldsymbol{B} | \boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_{k-1}) = \frac{1}{2} \log \det[(km\boldsymbol{I}_p)((k-1)m\boldsymbol{I}_p)^{-1}]$$

$$= \frac{p}{2} \log\left(\frac{k}{k-1}\right). \quad \square \tag{3.14}$$

Corollary 3. *The* total information *gained from responses of K orthogonal experiments with coded factors about the parameter estimators in a Gaussian multiple linear regression model is given by*

$$I(\boldsymbol{Y}_2, \ldots, \boldsymbol{Y}_{K-1}, \boldsymbol{Y}_K; \boldsymbol{B} | \boldsymbol{Y}_1) = \frac{p}{2} \log K. \tag{3.15}$$

Proof. The result follows directly by applying (3.13) to (3.2). □

Note that (3.15) is the maximum amount of information obtained from a series of K-coded experiments about the parameters of a multiple linear regression model with Gaussian noise. It is of interest to point out that information per estimator component, which is equal to $\frac{1}{2} \log K$, increases at a slow logarithmic rate with the number of experiments (or equivalently, that asymptotically the incremental information decreases inversely proportionally to k). Thus, as $K \to \infty$ one can obtain infinite amount of information (since $\boldsymbol{B}$ is a continuous r.v.), however, at a decreasing rate. This observation implies that at a certain point of time it is no longer efficient to continue experimenting when considering the cost of the experiment. The answer to the interesting question "*when should one stop experimenting*"? is further complicated when the value of σ^2 is unknown and has to be estimated from the responses. A suggested solution, which implies an optimal stopping rule, is considered in Ben-Gal and Caramanis (1998) by application of a dynamic programming framework to the DOE.

Summarizing the above, one can see that as the order of factor interactions in the linear regression model grows, one requires a design matrix with a higher resolution in order to maintain orthogonality, and hence, in the Gaussian case, *H-opt*imality. The problem of obtaining designs with high resolution for a limited number of experiments is one of the central problems in DOE. It has been pointed out in Section 2.2 that an isomorphic problem can be poised in the field of error-correcting codes. In the next section we use the correspondence between error-correcting codes and fractional factorial designs to maintain the required resolution that assures *H-opt*imality when the number of experiments is limited.

4. Construction of smaller FFE by increasing the number of levels

As described in Section 2.2, the construction of a FFE with n factors, q levels per factor, q^r treatments and resolution d is isomorphic to the construction of a (n, κ, d)

q-ary linear error-correcting code $\boldsymbol{C}$ with $\kappa = n - r$. The FFE matrix is, in fact, the matrix which has all the words of the *dual code* $\boldsymbol{C}^{\perp}$ as rows. Accordingly, the construction problem discussed above can be formulated in both fields as follows. Given *number of factors* (*code length*) n and a required *resolution* (*distance*) d, find a subgroup G with the maximum κ so that the size of the *FFE* (*dual-code*) $q^{n-\kappa}$ is minimized. An equivalent problem is the following. Given *number of factors* (*code length*) n and a *fraction parameter* (*dimension*) κ, find a subgroup G such that the FFE (ECC) attains the maximum *resolution* (*distance*) d.

In many systems the number of levels associated with each factor is large. Specifically, for continuous factors it is a common practice to discretize the level range by a large number of discrete points. Usually, it is beneficial to restraint q to a small number in order to maintain a small size of the FFE. However, for codes with a small q-ary alphabet and a fixed distance d, the number of redundant digits r is large, and vice versa. This dependence between the values of r and q leads to a tradeoff optimization problem. It is possible, therefore, that decreasing the number of levels can, in fact, increase the size of the FFE. It is a challenging problem to find, for a given resolution d, the optimal number of levels q_{opt} for which the size q_{opt}^{r} of the FFE is minimal. Unfortunately, this problem cannot be solved completely at present time since the exact functional relationship between d, q and r is unknown. Yet, some results can be obtained by using the correspondence described above and by applying known coding bounds and constructions to the design of FFE.

The analysis presented below shows that in many cases one can achieve a smaller FFE for a required (fixed) resolution by increasing the number of levels. This is an interesting observation, which in the context of experimental design seems somewhat counterintuitive.

Note that if the number of levels q (the size of the code alphabet) is considerably smaller than the number of the factors n (the length of the code) then the only guaranteed minimum value of r is given, in general, by the Gilbert–Varshamov bound (Gilbert, 1952; Macwilliams and Sloane, 1977):

$$\sum_{i=0}^{d-2} (q-1)^i \binom{n-1}{i} < q^r. \tag{4.1}$$

However, if the number of levels q (the size of the code alphabet) exceeds $n-2$ and q is a power of a prime then the Singleton bound (Singleton, 1964; Wicket, 1995),

$$d \leqslant n - \kappa + 1 = r + 1, \tag{4.2}$$

is attainable and there exist codes that satisfy this bound. These codes are called *maximum distance separable* codes (MDS codes) and their construction is known (Reed and Solomon, 1960). Thus, for $q = n-1$ we can always construct a FFE of size $(n-1)^{d-1}$. This is, for a given resolution, the smallest attainable size of FFE based on MDS codes. Denote the sizes of FFE (in binary logarithmic scale) based on

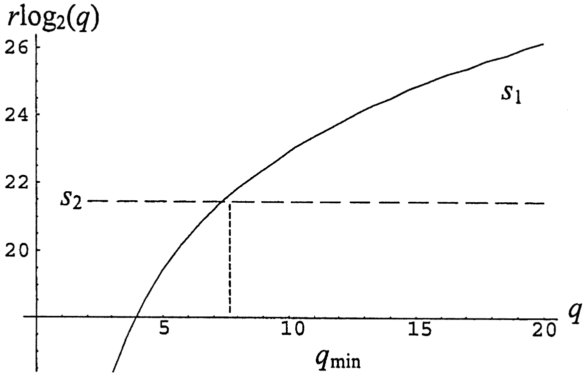

Fig. 2. The FFE size according to the Gilbert–Varshamov bound (s_1) as a function of the number of levels (q), for fixed number of factors ($n=42$) and a fixed resolution ($d=5$). s_2 (dashed) is the size according to the Singleton bound at $q=n-1=41$.

(4.1) and (4.2) at $q=n-1$, respectively, by

$$s_1=\log_2\sum_{i=0}^{d-2}(q-1)^i\binom{n-1}{i}$$

and

$$s_2=(d-1)\log_2(n-1). \tag{4.3}$$

Fig. 2 compares s_1 and s_2 for various values of q, while both the number of factors and the resolution are fixed ($n=42$, $d=5$). It is seen that at approximately $q=7.4$ these bounds intersect. Hence, if the number of levels associated with each factor is 8 or higher, then it is better to increase the number of levels to $q=n-1$ (in this case to $q=41$) in order to achieve a smaller FFE, keeping the same resolution.

Theorem 4. *Let $q_{\min}$ be the positive root of the equation*

$$\sum_{i=0}^{d-2}(q-1)^i\binom{n-1}{i}=(n-1)^{d-1}, \tag{4.4}$$

for given values of n and d (provided that $n-1$ is a power of a prime).

Then, for any $q\geqslant q_{\min}$ the left-hand part of (4.4) is larger than the right-hand part. Thus, for any given n and d, and for q, $q_{\min}\leqslant q\leqslant n-1$, the size of an FFE based on an MDS code with an alphabet of $n-1$ symbols is smaller than the size of an FFE based on a code with q symbols which satisfies the Gilbert–Varshamov bound.

Proof. Note that, obviously, $q_{\min}>1$, and the left-hand part of (4.4) is a monotonically increasing function of q for any $q\geqslant q_{\min}$. Thus, by definition of $q_{\min}$, the left-hand part of (4.4) is larger than the right-hand part for $q\geqslant q_{\min}$. Since the right-hand part of

(4.4) expresses the size of an FFE based on an (n, κ, d) MDS code with an alphabet of $n-1$ symbols, this proves the theorem. □

Theorem 4 implies that if $q_{\min} \leqslant q \leqslant n-1$, it is better to increase the number of levels to $n-1$ and to use a MDS code rather than one that lies on Gilbert–Varshamov bound, as illustrated by Fig. 2.

An explicit approximate expression for $q_{\min}$ can be obtained in the case when $d/n \ll 1$.

Theorem 5. *Let* $q_{\min}$ *be the positive root of* (4.4). *Then for* $d/n \ll 1$, $q_{\min}$ *can be approximated by*

$$q_{\min}(n,d) \approx \frac{(d-2)}{e}(n-1)^{1/(d-2)} + 1. \tag{4.5}$$

Proof. We use the well-known approximation for the sum of terms of a binomial expansion which is valid for any $l/n < \frac{1}{2}$, $a \geqslant 1, n \gg 1$:

$$\sum_{i=0}^{l} a^i \binom{n}{i} \approx a^l \exp\left[n \cdot h\left(\frac{l}{n}\right)\right]. \tag{4.6}$$

Here $h(x) = -x\ln x - (1-x)\ln(1-x)$ is the binary entropy function (to the base e).

Then, from (4.4), we obtain

$$(q-1)^{d-2} \exp\left[(n-1)h\left(\frac{d-2}{n-1}\right)\right] = (n-1)^{d-1}$$

or

$$(d-2)\ln(q-1) + (n-1)h\left(\frac{d-2}{n-1}\right) = (d-1)\ln(n-1). \tag{4.7}$$

For $x \ll 1$, $h(x) \approx x - x\ln x$.

Hence, for $d/n \ll 1$, (4.7) yields

$$\ln(q-1) = \frac{1}{d-2}\ln(n-1) + \ln(d-2) - 1. \tag{4.8}$$

Thus, the positive root $q_{\min}(n,d)$ of (4.4) is approximately equal to

$$q_{\min}(n,d) \approx \frac{(d-2)}{e}(n-1)^{1/(d-2)} + 1. \quad \square \tag{4.9}$$

It is readily seen from (4.5) that $q_{\min}$ is rather sensitive to the value of d and less sensitive to the value of n, as shown in Fig. 3. Moreover, $q_{\min}$ behaves in a nonmonotonic fashion as a function of d for fixed n (see Fig. 4 for various values of n). The same properties of $q_{\min}$ can be seen from Table 1, which gives the exact values of $q_{\min}$ computed numerically from (4.4) for certain values of n and d. The results for small distances (resolutions) ($d = 3, 4, 5$) are presented for illustration purposes (to show the behavior of $q_{\min}$), since for small distances, codes that lie substantially higher than Gilbert–Varshamov bound are known.

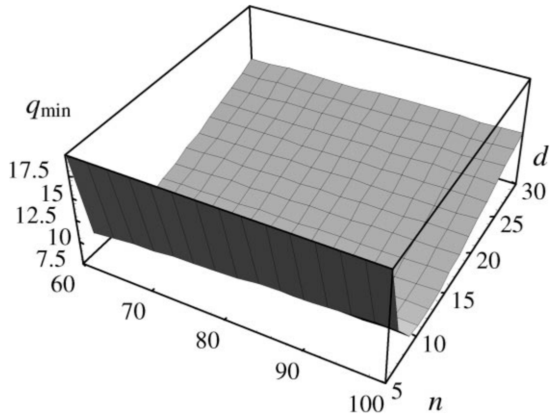

Fig. 3. Approximated values of $q_{\min}$ computed from (4.5) as a function of the number of factors ($n = 60,\dots,100$) and the resolution ($d = 5,\dots,30$).

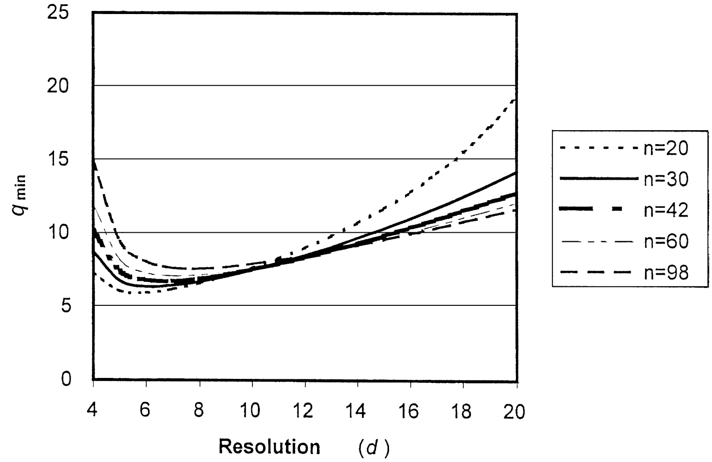

Fig. 4. Exact values of $q_{\min}$, computed numerically from (4.4) as a function of the resolution ($d = 4,\dots,20$) for various values of n.

5. Conclusions

The results presented in this paper can be divided into two parts. In the third section the subset selection problem is analyzed by introducing a new information optimality criterion called *H-optimality*. The coincidence of this criterion to the *D*-optimality criterion is shown for the case when the system is subjected to Gaussian noise. Moreover, it is shown that for the multiple linear regression model, *H-opt*imality requires a certain design resolution, so that all terms in the model can be estimated independently (i.e., such that they are not aliased). This requirement is further addressed in

Table 1
The exact $q_{\min}$ values, computed numerically from (4.4), as a function of the number of factors n and the required resolution d

d/n	10	14	20	30	42	60	98
3	9.89	13.92	19.94	29.96	41.97	59.97	97.98
4	5.37	6.22	7.28	8.71	10.14	11.93	14.99
5	5.13	5.55	6.06	6.75	7.39	8.18	9.42
6	5.47	5.67	5.96	6.38	6.80	7.27	8.05
7	6.09	6.07	6.19	6.45	6.73	7.08	7.63
8	6.93	6.63	6.59	6.71	6.89	7.15	7.58
9	8.02	7.32	7.08	7.07	7.17	7.36	7.69
10	9.46	8.14	7.65	7.50	7.53	7.65	7.91
11	—	9.11	8.30	7.98	7.93	7.99	8.18
12	—	10.27	9.02	8.50	8.36	8.36	8.49
13	—	11.69	9.82	9.07	8.83	8.76	8.84
14	—	13.46	10.72	9.68	9.33	9.19	9.20
15	—	—	11.71	10.32	9.85	9.63	9.58
17	—	—	14.11	11.74	10.96	10.57	10.39
20	—	—	19.46	14.24	12.82	12.10	11.68
23	—	—	—	17.33	14.92	13.76	13.05
25	—	—	—	19.86	16.48	14.96	14.01
27	—	—	—	22.96	18.19	16.19	14.98
30	—	—	—	29.46	21.08	18.91	16.51
35	—	—	—	—	27.17	21.93	19.22
40	—	—	—	—	36.15	26.29	22.13

Section 4, where the problem of maintaining a desired resolution with limited number of experiments is analyzed. The analysis is based on the isomorphism between the construction of fractional factorial experiments and error-correcting codes. The correspondence between these areas was summed up by Sloane (1994): "A good code is a large set of vectors of given length, from a given field, such that the Hamming distance between them is as large as possible; a good orthogonal array is a small set of vectors such that their 'dual distance' is as large as possible". It has been shown that, under certain conditions, one can design a smaller *H-optimal* fractional factorial experiment by increasing the number of levels. This phenomenon, which in the context of experimental design looks paradoxical, is due to the advantage of having a small number of redundant digits for codes with a large alphabet.

An interesting example of such a situation has been pointed out by an anonymous referee for Ben-Gal and Levitin (1998): "suppose $n = 30$, $d = 5$, q at least 4. Then with $q = 4$ one gets $k \leqslant 23$, thus a dual code (FFE) size of $4^7 = 16384$, but with $q = 5$ one has an example where $k = 24$ and a dual code (FFE) size of $5^6 = 15625$".

In practice, this result is applicable for large size experiments which should be conducted fast and within a reasonable cost. These conditions, for example, take place in computer hardware testing where each experiment is essentially a single test pattern generated by the tester.

Acknowledgements

The authors would like to thank the referees for their thorough reviews and useful remarks.

References

Ben-Gal, I., Caramanis, M., 1998. A stochastic dynamic programming framework for an efficient adaptive design of experiments. Mimeograph Series No. 23B-98. ADMS Lab, Boston University, Boston, MA.

Ben-Gal, I., Levitin, L.B., 1998. Bounds on code distance and efficient fractional factorial experiments. Proceedings of the 1998 IEEE International Symposium on Information Theory, Boston, MA, 1998, p. 436.

Bose, R.C., 1950. Mathematics of factorial designs. Proceeding of the International Congress of Mathematicians, Vol. 1, pp. 543–548.

Bose, R.C., 1961. On some connections between the design of experiments and information theory. Bull. Inst. Internat. Statist. 38, 257–271.

Box, G.E.P., Draper, N.R., 1971. Fractional designs, the $|X'X|$ criterion, and some related matters. Technometrics 13 (4), 731–742.

Box, G.E.P., Hill, W.J., 1967. Discrimination among mechanistic models. Technometrics 9 (1), 57–71.

Cover, T.M., Thomas, J.A., 1991. Elements of Information Theory. Wiley, New York.

Delsarte, P., 1973. An algebraic approach to the association scheme of coding theory. Philips Res. Rep. 10 (Suppl.) 1–97.

Fedorov, V.V., 1972. Theory of Optimal Experiments. Academic Press, New York.

Finney, D.J., 1945. The fractional replications of factorial arrangements. Ann. Eugen. Lond. 12, 291–301.

Gilbert, E.N., 1952. A comparison of signaling alphabets. Bell System Tech. J. 31, 504–522.

Hardin, R.H., Sloane, N.J.A., 1993. A new approach to the construction of optimal designs. J. Statist. Plann. Inference 37, 339–369.

Keiefer, J., Wolfowitz, J., 1959. Optimum designs in regression problems. Ann. Math. Statist. 30, 271–294.

Kishen, K., 1948. On fractional replication of the general symmetrical factorial design. J. Indian Soc. Agricultural Statist. 1, 91–106.

Macwilliams, F.J., Sloane, N.J.A., 1977. The Theory of Error-Correcting Codes. North-Holland, Amsterdam.

Montgomery, D.C., 1991. Design and Analysis of Experiments. Wiley, New York.

Myers, R.H., Montgomery, D.C., 1995. Response Surface Methodology. Wiley, New York.

Reed, I.S., Solomon, G., 1960. Polynomial codes over certain finite fields. J. SIAM 8, 300–304.

Shannon, C.E., 1948. A mathematical theory of communication. Bell Systems Tech. J. 27 part I, 379–423; part II, 623–656.

Singleton, R.C., 1964. Maximum distance Q-ary codes. IEEE Trans. Inform. Theory 10, 116–118.

Sloane, N.J.A., 1994. Application of error-correcting codes to orthogonal arrays and the design of experiments. Proceedings of the IEEE International Symposium on Information Theory, p. 104.

Sloane, N.J.A., Stufken, J., 1996. A linear programming bound for orthogonal arrays with mixed levels. J. Statist. Plann. Inference 56 (2), 295–305.

St. John, R.C., Draper, N.R., 1975. D-optimality for regression designs: a review. Technometrics 17, 15–23.

Wicket, S.B., 1995. Error Control Systems for Digital Communication and Storage. Prentice-Hall, Englewood Cliffs, NJ.

Wynn, H.P., 1970. The sequential generation of D-optimum experimental designs. Ann. Math. Statist. 41, 1655–1664.