Optimal Experimental Design for Generalized Regression Models

Anthony C. Atkinson
The London School of Economics, London UK*
Valerii V. Fedorov
Isaac Newton Institute for Mathematical Sciences, Cambridge, UK
Agnes M. Herzberg
Queen's University, Ontario, Canada
and Rongmei Zhang
FDA, Maryland, USA

December 13, 2011

Abstract

The construction of optimal experimental designs for regression models requires knowledge of the information matrix of a single observation. The latter can be found if the elemental information matrices corresponding to the distribution of the response are known. We present tables of elemental information matrices for distributions that are often used in statistical work. The tables contain matrices for one- and two-parameter distributions. Additionally we describe multivariate normal and multinomial cases. The parameters of response distributions can themselves be parameterized to provide dependence on explanatory variables, thus leading to regression formulations for wide classes of models. We present essential results from optimum experimental design and illustrate our approach with a few examples including bivariate binary responses and gamma regression.

Keywords: adaptive design; convex optimal design; elemental information matrix; equivalence theorem.

1 Introduction

We are concerned with optimal experimental design for a rather general class of regression models. The observations are not constrained to be normally distributed, but might, for example, follow a beta or binomial distribution, or, indeed, any distribution for which the Fisher information matrix exists, together with regular maximum likelihood estimators. By regression we intend that the parameters

*e-mail: a.c.atkinson@lse.ac.uk

of these distributions are themselves functions of further parameters and of explanatory variables, at least some of which can be chosen and controlled in the experiment. Frequently a link function will be required, for example in the parameterization of the variance of a normal distribution or the probability in a Bernoulli model.

Our main objective (see §§2 and 4) is to provide a collection of "elemental" information matrices for specific distributions which are crucial for the solution of optimal design problems. In this way we bring together results that are repetitively scattered throughout the statistical literature. These matrices are essential for constructing information matrices for "single" observations. The latter may in many cases, for example for mixed-effect models, consist of a series of observations on a single subject. For applications to experimental design we require the possibility of independent replications of these series. Combining the concept of a single observation (supported by tables of information matrices) with a collection of sensitivity functions for various popular optimality criteria (see §3) the reader can address numerous tasks in the optimal design of experiments.

We begin in §2 with an introduction to the generalized regression model, continuing in §3 with some basic ideas about optimal experimental design. The core of our paper is §4 where we provide tables of elemental information matrices for one and two parameter univariate distributions. In §4.2 we indicate how further information matrices, excluded from the tables, may be found for members of the location-scale and exponential families. The multivariate normal distribution, including a parameterized variance, and multinomial distributions receive special attention in §4.3 and §4.4. Examples in §5 include the equivalence theorem for D-optimality for normal observations with a particular parameterized variance. Example 3 includes the family of designs for univariate generalized linear models, whereas, in Example 4 there are two binary responses. Example 5 is gamma regression when both the parameters are functions of explanatory variables.

2 Generalized Regression

2.1 Main model

We assume that the observed response Y is distributed as

$$Y \sim p(y|\eta), \tag{1}$$

where Y and y are k-dimensional vectors, with η of dimension l. The parameters η depend on controls $x \in \mathscr{X}$ and $\mathscr{X}$ is a design set (region). Usually $\mathscr{X} \in \mathbb{R}^s$, but in general $\mathscr{X}$ could be a compact set of more complicated structure, for instance, part of a functional space (Fedorov and Hackl 1997, §5.7). In what follows we assume that $\mathscr{X} \in \mathbb{R}^s$ unless otherwise stated. In the standard setting for regression models it is assumed that the expected values of responses Y are parameterized, compare with Pázman (1986), Pukelsheim (1993), Fedorov and Hackl (1997) or

Atkinson, Donev, and Tobias (2007); see Example 1. The most popular examples are "normal" regression, binary regression and Poisson regression. Actually, all regression models with response belonging to a one-parameter family can be treated in a similar way. Generalizations for cases when the components of η are of the same type (e.g. expectations in the multivariate normal) are straightforward. However, in the case of multi-parameter distributions (such as the gamma or Weibull), it is very natural to assume that two or more components of η of different types may depend on controls and unknown parameters. As far as we know, in the experimental design literature only in the "normal" case have different types of components of η been parameterized, namely, the expectation(s) of the response(s) and its variance (variance-covariance matrix). See, for instance, Atkinson and Cook (1995), Dette and Wong (1999) or Fedorov and Leonov (2004). A rare exception is the beta regression considered by Wu, Fedorov, and Propert (2005).

2.2 Likelihood estimators and Fisher information matrices

Before proceeding with the design problem for generalized regression we recall a few commonly known facts and introduce the required notation. Let

$$\eta = \eta(x, \theta), \tag{2}$$

where $\eta^T(x,\theta) = \{\eta_1(x,\theta), \ldots, \eta_k(x,\theta)\}$ are given functions of controls x and unknown regression parameters θ. Further, let n independent observations $\{Y_i\}_1^n$ be made at $\{x_i\}_1^n$. Equations (1) and (2) constitute a generalized regression model.

The maximum-likelihood estimator (MLE) $\hat{\theta}_n$ of θ can be defined as

$$\begin{aligned} \hat{\theta}_n &= \arg\max_{\theta\in\Theta} \prod_{i=1}^{n} p(y_i|\eta(x_i,\theta)) \\ &= \arg\max_{\theta\in\Theta} L(\{y_i\}_1^n, \{x_i\}_1^n, \theta), \end{aligned} \tag{3}$$

where Θ is compact and the true value θ_t of θ is an internal point of Θ. Under rather mild conditions on $p(y|\eta)$, $\eta(x,\theta)$ and on the sequence $\{x_i\}_1^n$ (see, for example, Lehmann and Casella 1998, Chapter 2) the MLE $\hat{\theta}_n$ is strongly consistent and its normalized asymptotic variance-covariance matrix is

$$nD(\theta_t, \{x_i\}_1^n) = M^{-1}(\theta_t, \{x_i\}_1^n), \tag{4}$$

where

$$M(\theta, \{x_i\}_1^n) = \sum_{i=1}^{n} \mu(x_i, \theta), \tag{5}$$

$$\text{and} \quad \mu(x,\theta) = \text{Var}\left[\frac{\partial}{\partial\theta} \ln p(y|\eta(x,\theta)\right]. \tag{6}$$

Introducing

$$F(x,\theta) = \frac{\partial \eta^T(x,\theta)}{\partial\theta}$$

and observing that

$$\frac{\partial}{\partial\theta}\ln p(y|\eta(x,\theta)=\frac{\partial\eta^T(x,\theta)}{\partial\theta}\left[\frac{\partial}{\partial\eta}\ln p(y|\eta)\right]_{\eta=\eta(x,\theta)},$$

one may conclude from (6) that

Lemma 1

$$\mu(x,\theta)=F(x,\theta)\nu(\eta)F^T(x,\theta), \tag{7}$$

where

$$\nu(\eta)=\mathrm{Var}\left[\frac{\partial}{\partial\eta}\ln p(y|\eta)\right]. \tag{8}$$

Note that in (7) and (8) the vector η depends on x and θ but we have omitted the compound subscript $\eta=\eta(x,\theta)$ and will continue to do so if this does not lead to ambiguity.

In what follows we call $\nu(\eta)$ defined by (8) the "*elemental information matrix*" for model (1). It plays a central role in this article.

If there are repeated observations (5) should be replaced by

$$M(\theta,\xi_N)=\sum_{i=1}^{n}r_i\mu(x_i,\theta), \tag{9}$$

or, moving to the normalized information matrix,

$$NM(\theta,\xi)=\sum_{i=1}^{n}w_i\mu(x_i,\theta), \tag{10}$$

where $N=\sum_{i=1}^{n}r_i$ and $\xi=\{w_i,x_i\}_1^n$. If there are no repeated observations then n and N coincide.

More generally,

$$\begin{aligned}M(\theta,\xi) &= \int\mu(x_i,\theta)\xi(dx) \\ &= \int F(x,\theta)\nu(\eta)F^T(x,\theta)\xi(dx),\end{aligned} \tag{11}$$

where ξ could be any probabilistic measure defined on $\mathscr{X}$. Thus the information matrix (11) is completely defined by the design ξ, by the derivatives $F(x,\theta)$ of the regression functions $\eta(x,\theta)$ and by the elemental information matrix $\nu(\eta)$.

3 A Few Basic Facts from Optimal Design Theory

Optimality criteria play a fundamental role in the design of experiments. We follow the well-accepted paradigm of convex design theory (Fedorov 1972, Silvey 1980,

Pázman 1986, Pukelsheim 1993, Fedorov and Hackl 1997, Atkinson et al. 2007) and define an optimal (continuous/approximate) design as

$$\xi^* = \arg\min_{\xi \in \Xi} \Psi\left[M(\xi, \theta)\right], \tag{12}$$

where $\Psi[M]$ is a convex and homogeneous function of matrix M (see Fedorov and Hackl 1997, §2.2). An introduction to optimal design, emphasizing the importance of Fisher information, is in Chapter 7 of Cox and Reid (2000). For the sake of simplicity we confine ourselves throughout this paper to cases when optimal designs defined by (12) have regular information matrices, e.g. non-zero determinants.

3.1 Equivalence theorem

The following theorem, which is often called the equivalence theorem after the "equivalence theorem" of Kiefer and Wolfowitz (1960) for the D-criterion, is the theoretical basis for the development of many numerical methods and for the construction of adaptive designs.

Theorem 1 *A necessary and sufficient condition for ξ^* to be optimal is fulfillment of the inequality*

$$\min_{x \in \mathscr{X}} \psi(x, \theta) \geq 0, \tag{13}$$

where $\psi(x, \xi^*)$ is the directional derivative:

$$\psi(x, \xi^*) = \lim_{\alpha \to 0} \frac{\partial}{\partial \alpha} \Psi\left[M((1-\alpha)\xi^* + \alpha\xi(x))\right], \tag{14}$$

and $\xi(x)$ is a design atomized at x.

Frequently the function $\psi(x, \xi)$ is called a sensitivity function. A table of sensitivity functions for the most popular criteria is presented in Table 1. All functions $\psi(x, \xi)$ are defined by the information matrices of a single observation. These, in turn, are defined by the elemental information matrices (8) and by the derivatives of the regression functions $F(x, \theta)$. Thus Table 1, together with tables from the next section that contain elemental matrices, leads straightforwardly to the specific versions of the equivalence theorem for a wide variety of models; the necessity of special calculations for each individual case is avoided.

Note that Table 1 is built under the assumption that the optimal design ξ^* is regular. In some special cases in rows 2 and 4, for example, it can happen that ξ^* is singular, that is that rank $M(\xi^*) < m$. Such cases are well treated in Pukelsheim (1993), but are beyond the scope of this paper.

All the criteria in Table 1 operate in parameter space, that is they provide a scalar measure of precision, or uncertainty, of the MLE of θ or of the linear transformation $\vartheta = A^T\theta$. In the case of arbitrary, but smooth, transformations $\vartheta(\theta)$ the results of the table still hold with A^T replaced by the vector of derivatives $\partial\vartheta^T/\partial\theta$.

Often a practitioner may be interested in estimation of response functions $\eta(x, \theta)$ or of a utility function $\zeta(x, \theta)$ on some set $\mathscr{Z}$ that may or may not coincide with $\mathscr{X}$. The corresponding criteria are not explicitly included in Table 1. However, the table provides the requisite information for some criteria. For instance, let interest be in

$$\begin{aligned}\int_{\mathscr{Z}} \operatorname{Var}[\zeta(x, \hat{\theta})]dx &\cong \int_{\mathscr{Z}} \frac{\partial \zeta(x,\theta)}{\partial \theta^T} \operatorname{Var}\hat{\theta}\, \frac{\partial \zeta(x,\theta)}{\partial \theta} dx \\ &= \operatorname{tr} \int_{\mathscr{Z}} \frac{\partial \zeta(x,\theta)}{\partial \theta} \frac{\partial \zeta(x,\theta)}{\partial \theta^T} dx \operatorname{Var}\hat{\theta}.\end{aligned}$$

Defining

$$A = \int_{\mathscr{Z}} \frac{\partial \zeta(x,\theta)}{\partial \theta} \frac{\partial \zeta(x,\theta)}{\partial \theta^T} dx,$$

one can use the fourth row of Table 1 to find the necessary sensitivity function.

3.2 Computing optimal designs

Many widely-used numerical algorithms for the construction of optimal designs are based on directional derivatives (Wynn 1970; Fedorov 1972; Fedorov and Hackl 1997).

For instance, in first-order algorithms, forward excursions add weight to the design at

$$x_{s+1}^{+} = \arg\max_{x \in \mathscr{X}} \psi(x; \xi_s, \theta), \tag{15}$$

whereas backward excursions delete weight from

$$x_{s+1}^{-} = \arg\min_{x \in \mathscr{X}} \psi(x; \xi_s, \theta). \tag{16}$$

Again, as in the previous section, Tables 1-4 and Table 5 provide all the components that are needed, now for (15) and (16).

Fedorov and Hackl (1997, §3.2) and Atkinson et al. (2007, §9.5) describe the use of general-purpose second-order algorithms for the construction of optimal designs. These algorithms can be built using elemental information matrices and sensitivity functions, i.e. the first-order directional derivatives, which are presented in Table 1. However, second-order directional derivatives are also needed. Again they are completely defined by the elemental information matrices and by $F(x, \theta)$ (Fedorov and Hackl 1997, §3.2).

Adaptive designs in the optimal design setting are driven by placing the $(n+1)$-th observation at the point that is viewed as the most informative (with respect to the selected criterion) after n observations. The approximate location of this point is defined as

$$x_{n+1} = \arg\max_{x \in \mathscr{X}} \psi(x; \xi_s, \hat{\theta}_n), \tag{17}$$

where $\hat{\theta}_n$ is the maximum likelihood estimate after n observations. Note that (17) is identical to (15) if, instead of the unknown θ, we substitute its estimate $\hat{\theta}_n$. Thus,

knowledge of the sensitivity function - i.e. the knowledge of $F(x,\theta)$ together with the elemental matrix $\nu(\eta)$ - is sufficient to develop adaptive design procedures.

Table 1: Sensitivity functions $\psi(x,\xi)$, $D = M^{-1}$; see (4)

$\Psi(\xi)$	$\psi(x,\xi)$
$\log\|D\|,\ \ \|D\|^{1/m},\ \ \prod_{\alpha=1}^{m}\lambda_\alpha(D)$	$\operatorname{tr}\nu(x)F^T(x)DF(x) - m$
$\log\|A^TDA\|,$ $\dim A = k\times m,\ \operatorname{rank}A = k < m$	$\operatorname{tr}\nu(x)F^T(x)DA(A^TDA)^{-1}A^TDF(x) - k$
$\operatorname{tr} D,\ \ \sum_{\alpha=1}^{m}\lambda_\alpha(D)$	$\operatorname{tr}\nu(x)F^T(x)D^2F(x) - \operatorname{tr} D$
$\operatorname{tr} A^TDA$	$\operatorname{tr}\nu(x)\left(F^T(x)DA\right)^2 - \operatorname{tr} A^TDA$
$\operatorname{tr} D^\gamma,\ \ \sum_{\alpha=1}^{m}\lambda_\alpha^\gamma(D)$	$\operatorname{tr}\nu(x)F^T(x)D^{\gamma+1}F(x) - \operatorname{tr} D^\gamma$
$\lambda_{\min} = \lambda_{\min}(M)$ $= \lambda_{\max}^{-1}(D)$	$\sum_{i=1}^{a}\pi_iP_i^TF(x)\nu(x)F(x)P_i - \lambda_{\min}$ $\lambda_{\min}P_i = MP_i,$ a is the multiplicity of $\lambda_{\min}$, $\sum_{i=1}^{a}\pi_i = 1,\ 0\le\pi_i\le 1$

In many cases the calculation of $\nu(\eta)$ is simpler if, instead of (8) one uses (Lehmann and Casella 1998)

$$\nu(\eta) = -\mathrm{E}\left[\frac{\partial^2}{\partial\eta\partial\eta^{\mathrm{T}}}\ln \mathrm{p}\{\mathrm{y}|\eta(\mathrm{x},\theta)\}\right].$$

In survival analysis for families of continuous distributions on the real line it is common (see, for example, Cox and Oakes 1984) to use the hazard function

$$h(y|\eta) = \frac{p(y|\eta)}{\int_y^\infty p(u|\eta)du} = \frac{p(y|\eta)}{\text{Prob}\,(Y\ge y|\eta)}\ .$$

Efron and Johnstone (1990) recommend representing the information matrix in terms of the hazard function as

$$\nu(\eta) = \mathrm{E}\left[\frac{\partial}{\partial\eta}\ln \mathrm{h}(\mathrm{y}|\eta)\frac{\partial}{\partial\eta^{\mathrm{T}}}\ln \mathrm{h}(\mathrm{y}|\eta)\right]\ .$$

Note that, in general, unlike (8)

$$\nu(\eta) \ne \text{Var}\left[\frac{\partial}{\partial\eta}\ln h(y|\eta)\right],$$

and, equivalently,

$$\mathrm{E}\left[\frac{\partial}{\partial\eta}\ln \mathrm{h}(\mathrm{y}|\eta)\right] \ne 0.$$

4 Elemental Information Matrices

In this section we provide a collection of elemental information matrices for popular distributions together with recommendations for the derivation of matrices for any that we have omitted. In the following examples we show how this collection helps to explore the optimal design problem for almost any plausible generalized regression.

4.1 Univariate distributions

Almost all of the reported elemental information matrices or, at least the corresponding references, can be found in Johnson, Kotz, and Balakrishnan (1994, 1995) and Johnson, Kemp, and Balakrishnan (2005). We have also found useful results in Lehmann and Casella (1998) and in Bernardo and Smith (1994). Table 2 contains elemental information expressions for single parameter distributions with elemental information matrices for two parameter distributions in Table 4.

The distributions in Table 4 include the Weibull and Pareto which are often used for modelling survival or lifetime data. We note (Harris 1968) that a mixture of exponential distributions with a gamma distributed parameter leads to a form of Pareto distribution, which makes it appropriate for modelling heterogeneity in survival times. These distributions are frequently elaborated by the introduction of extra parameters to add flexibility in shape and to allow for times that do not start at zero. Brazauskas (2003) presents information matrices for these more complicated Pareto distributions. Escobar and Meeker (1994), Gertsbakh (1995) and Gupta and Kundu (2006) discuss information matrices for censored survival distributions. Information matrices for bivariate and multivariate Pareto distributions are presented by Yari and Jafari (2006) ,Gupta and Nadarajah (2007) and by Kotz (2008). Elaborations of the Laplace distribution are in Kotz, Kozubowski, and Podgórski (2001). Ali and Nadarajah (2007) cover normal Laplace mixtures with the Dirichelet-multinomial distribution in Paul, Balasooriya, and Banerjee (2005).

If one of these two parameters is assumed to be known then the elemental information for the other (unknown) parameter equals the corresponding diagonal element of the elemental information matrix. In many settings it is helpful to work with parameters ϑ which are often log or logit functions of the original parameters η and can vary between $-\infty$ and ∞. We use $\nu(\vartheta)$ when the parameters ϑ, not η, will be considered as functions of controls x and regression parameters θ. All results of Section 3 stay valid with the obvious replacement of $\nu(\eta)$ by $\nu(\vartheta)$. Table 3 and Table 5 contain information or information matrices for popular choices of those new parameters. The multivariate normal distribution and multinomial distribution are described in the corresponding subsections. Many further references on multivariate distributions can be found in Johnson, Kotz, and Balakrishnan (1997) and Kotz, Balakrishnan, and Johnson (2000)

Table 2: Elemental information for single parameter distributions

Distribution	Density	Mean Variance	Information
Bernoulli(p) $0 \leq p \leq 1$	$p^y(1-p)^{1-y}$	p $p(1-p)$	$1/[p(1-p)]$
Geometric(p) $0 \leq p \leq 1$	$(1-p)^y p$	$(1-p)/p$ $(1-p)/p^2$	$1/[p^2(1-p)]$
Binomial(p,n) $0 \leq p \leq 1$	$\binom{n}{y}p^y(1-p)^{n-y}$	np $np(1-p)$	$n/[p(1-p)]$
Neg. Bin. (p,m) $0 \leq p \leq 1$	$\binom{m+y-1}{m-1}p^m(1-p)^y$	$m(1-p)/p$ $m(1-p)/p^2$	$m/[p^2(1-p)]$
Hypergeometric (p,N,n) $0 \leq p \leq 1$	$\frac{\binom{n}{y}\binom{N-n}{Np-y}}{\binom{N}{Np}}$	np $np(1-p)(N-n)/(N-1)$	$\frac{(N-1)n}{p(1-p)(N-n)}$
Poisson(λ) $\lambda > 0$	$\frac{\lambda^y e^{-\lambda}}{y!}$	λ λ	$1/\lambda$

Table 3: Transformed elemental information for single parameter distributions

Distribution	New parameter	Information
Bernoulli(p)	$\vartheta = \ln \frac{p}{1-p}$	$e^{\vartheta}/(1+e^{\vartheta})^2$
Binomial(p,n)	$\vartheta = \ln \frac{p}{1-p}$	$ne^{\vartheta}/(1+e^{\vartheta})^2$
Poisson(λ)	$\vartheta = \ln \lambda$	e^{ϑ}
Geometric(p)	$\vartheta = \ln \frac{p}{1-p}$	$1/(1+e^{\vartheta})$
Neg. Bin. (p,m)	$\vartheta = \ln \frac{p}{1-p}$	$m/(1+e^{\vartheta})$

4.2 Families of Distributions

If the reader does not find the needed distribution in Table 2 or in Table 4, the following classical results may help (c.f. Lehmann and Casella 1998).

4.2.1 Location-scale families

The density of the location-scale family has the form

$$p(y,\eta) \;=\; \frac{1}{b}\,\pi\left(\frac{y-a}{b}\right)\;, \tag{18}$$

i.e., in our notation, $\eta^T = (a,b)$. The elements of the matrix $\nu(\eta)$ can be found by calculating the integrals

$$\nu_{aa} \;=\; \frac{1}{b^2}\int s^2(u)\pi(u)du, \tag{19}$$

$$\nu_{bb} \;=\; \frac{1}{b^2}\int [us^2(u)+1]^2\pi(u)du, \tag{20}$$

$$\nu_{ab}=\nu_{ba} \;=\; \frac{1}{b^2}\int us^2(u)\pi(u)du, \tag{21}$$

where $s(u) \;=\; \partial \ln \pi(u)/\partial u$. The non-diagonal elements equal zero whenever $\pi(u)$ is symmetric about the origin. The normal, logistic, Cauchy, and double exponential distributions belong to the location-scale family and are included in our tables.

4.2.2 Exponential families.

Most of the common distributions that are in our tables, such as the normal, exponential, gamma, beta, Bernoulli, binomial, and Poisson belong to the exponential

Table 4: Elemental information matrices for two parameter distributions

Distribution	Density	Mean Variance	Support	Information matrix
Normal(a, σ^2) $-\infty < a < \infty$ $\sigma > 0$	$\frac{1}{\sqrt{2\pi\sigma^2}}\, e^{\frac{-(y-a)^2}{2\sigma^2}}$	a σ^2	$-\infty, \infty$	$\sigma^{-2}\begin{pmatrix} 1 & 0 \\ 0 & \frac{1}{2}\sigma^{-2} \end{pmatrix}$
Beta(α, β) $\alpha > 0$ $\beta > 0$	$B(\alpha,\beta)y^{\alpha-1}(1-y)^{\beta-1}$	$\alpha/(\alpha+\beta)$ $\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$	$0 < y < 1$	$\begin{pmatrix} \psi'(\alpha)-\psi'(\alpha+\beta) & -\psi'(\alpha+\beta) \\ -\psi'(\alpha+\beta) & \psi'(\beta)-\psi'(\alpha+\beta) \end{pmatrix}$
Gamma(α, β) $\alpha > 0$ $\beta > 0$	$\frac{1}{\Gamma(\alpha)\beta^\alpha}\, y^{\alpha-1}e^{-y/\beta}$	$\alpha\beta$ $\alpha\beta^2$	$0, \infty$	$\begin{pmatrix} \psi'(\alpha) & 1/\beta \\ 1/\beta & \alpha/\beta^2 \end{pmatrix}$
Logistic(a, b) $-\infty < a < \infty$ $b > 0$	$\frac{e^{-(y-a)/b}}{b[1+e^{-(y-a)/b}]^2}$	a $b^2\pi^2/3$	$-\infty, \infty$	$\frac{1}{3}\begin{pmatrix} b^{-2} & 0 \\ 0 & 1+b^{-2} \end{pmatrix}$
Cauchy(a, b) $-\infty < a < \infty$ $b > 0$	$\frac{b}{\pi[b^2+(y-a)^2]}$	Do not exist	$(-\infty, \infty)$	$\frac{1}{2b^2}I$
Weibull(α, β) $\alpha > 0$ $\beta > 0$	$\frac{\alpha}{\beta}(\frac{y}{\beta})^{\alpha-1}e^{-(y/\beta)^\alpha}$	$\beta\Gamma(1+\alpha^{-1})$ $\beta^2[\Gamma(1+2\alpha^{-1})$ $-\Gamma^2(1+\alpha^{-1})]$	$0, \infty$	$\begin{pmatrix} \frac{\pi^2}{6}+\frac{(1-\gamma)^2}{\alpha^2} & \frac{\gamma-1}{\beta} \\ \frac{\gamma-1}{\beta} & \frac{\alpha^2}{\beta^2} \end{pmatrix}$
Pareto(α, σ) $\alpha > 2$ (for variance) $\sigma > 0$	$\frac{\alpha}{x}(\frac{y}{\sigma})^{-\alpha}$	$\sigma\left(\frac{\alpha}{\alpha-1}\right)$ $\frac{\alpha\sigma^2}{(\alpha-1)^2(\alpha-2)}$	$0, \infty$	$\begin{pmatrix} \frac{\alpha}{\sigma^2(\alpha+2)} & -\frac{1}{\sigma(\alpha+1)} \\ -\frac{1}{\sigma(\alpha+1)} & \frac{1}{\alpha^2)} \end{pmatrix}$
Double exponential (Laplace)(a, b) $-\infty < a < \infty$ $b > 0$	$\frac{e^{-\lvert y-a\rvert/b}}{2b}$	a $2/b^2$	$-\infty, \infty$	$\frac{1}{b^2}I$

$\psi(\alpha) = \frac{\Gamma'(\alpha)}{\Gamma(\alpha)}$ and $\psi'(\alpha) = \frac{d\psi(\alpha)}{d\alpha)}$ are the digamma and trigamma functions; $\gamma = 0.5772$ is Euler's constant, see Abramowitz and Stegun (1965)

Table 5: Transformed elemental information matrices for two-parameter distributions after transformation

Distribution	New parameters	Information matrix
$\mathrm{N}(a, \sigma^2)$	$\vartheta_1 = a$ $\vartheta_2 = \ln \sigma^2$	$\begin{pmatrix} 1/e^{\vartheta_2} & 0 \\ 0 & 1/2 \end{pmatrix}$
$\mathrm{N}(a, a^2)$	$\vartheta = \ln a$	3
$\mathrm{N}(a, k^2 a^2)$	$\vartheta_1 = \ln a$ $\vartheta_2 = k^2$	$\begin{pmatrix} 2 + 1/\vartheta_2 & 1/\vartheta_2 \\ 1/\vartheta_2 & 1/(2\vartheta_2^2) \end{pmatrix}$
$\mathrm{Beta}(\alpha, \beta)$	$\vartheta_1 = \ln \alpha$ $\vartheta_2 = \ln \beta$	$\begin{pmatrix} e^{2\vartheta_1}\psi'(e^{\vartheta_1}) & 0 \\ 0 & e^{2\vartheta_2}\psi'(e^{\vartheta_2}) \end{pmatrix}$ $-\psi'(e^{\vartheta_1} + e^{\vartheta_2}) \begin{pmatrix} e^{2\vartheta_1} & 1 \\ 1 & e^{2\vartheta_2} \end{pmatrix}$
$\mathrm{Gamma}(\alpha, \beta)$	$\vartheta_1 = \ln \alpha$ $\vartheta_2 = \ln \beta$	$e^{\vartheta_1} \begin{pmatrix} e^{\vartheta_1}\psi'(e^{\vartheta_1}) & 1 \\ 1 & 1 \end{pmatrix}$
$\mathrm{Logistic}(a, b)$	$\vartheta_1 = a$ $\vartheta_2 = \ln b$	$\frac{1}{3} \begin{pmatrix} e^{-2\vartheta_2} & 0 \\ 0 & 1 + e^{2\vartheta_2} \end{pmatrix}$
$\mathrm{Cauchy}(a, b)$	$\vartheta_1 = a$ $\vartheta_2 = \ln b$	$\frac{1}{2} \begin{pmatrix} e^{-2\vartheta_2} & 0 \\ 0 & 1 \end{pmatrix}$
$\mathrm{Weibull}(\alpha, \beta)$	$\vartheta_1 = \ln \alpha$ $\vartheta_2 = \ln \beta$	$\begin{pmatrix} \frac{\pi^2}{6} + (1-\gamma)^2 & \frac{\gamma - 1}{e^{\vartheta_2}} \\ \frac{\gamma - 1}{e^{\vartheta_2}} & e^{2\vartheta_1} \end{pmatrix}$

family. For this family the density is written as

$$p(y, \vartheta) = h(y) \exp\left[\eta^T(\vartheta) T(y) - B(\vartheta)\right]. \tag{22}$$

One can also write the density in the canonical form

$$p(y, \vartheta) = h(y) \exp\left[\eta^T T(y) - A(\eta)\right], \tag{23}$$

where η is often called the natural or canonical parameter. Sometimes it is useful to use the so-called mean-value parameter $\tau = \mathrm{E}[T(y)]$. Information matrices for distributions in the canonical form (23) can easily be found. For canonical parameters the elemental information matrix is

$$\nu(\eta) \;=\; \frac{\partial^2 A(\eta)}{\partial \eta \partial \eta^T} \;=\; \mathrm{Var}[T(y)] \;=\; \Sigma, \tag{24}$$

and interestingly enough, for the mean-value parametrization, the elemental information matrix is the inverse matrix of Σ, i.e. $\nu(\tau) = \Sigma^{-1}$.

Note that the multivariate normal and multinomial distributions considered in the next two sections are both members of the exponential family. More about the multivariate exponential family can be found in Kotz, Balakrishnan, and Johnson (2000).

4.3 Multivariate normal distribution

Let $Y \in R^l$ have a normal distribution, i.e.,

$$p(y|a, \Sigma) = (2\pi)^{-l/2}|\Sigma|^{-1/2} exp\{-\frac{1}{2}(y-a)\Sigma^{-1}(y-a)^T\} .$$

We let θ represent the unknown parameters defining the mean $a(x, \theta)$ and the variance-covariance matrix $\Sigma(x, \theta)$. Then the (α, β) element of the information matrix for θ and for a single observation is (Magnus and Neudecker 1988, p. 325):

$$\mu_{\alpha\beta} = \frac{\partial a}{\partial \theta_\alpha}\Sigma^{-1}\frac{\partial a^T}{\partial \theta_\beta} + \frac{1}{2}\mathrm{tr}\left(\Sigma^{-1}\frac{\partial \Sigma}{\partial \theta_\alpha}\Sigma^{-1}\frac{\partial \Sigma}{\partial \theta_\beta}\right). \tag{25}$$

The information matrix for the $l + (l+1)l/2$ parameters a and Σ appears complicated, so we introduce notation to express it in a more compact form (Magnus and Neudecker 1988, §2.4; Harville 2002, §16.4). Let vecΣ be a vector that is constructed from Σ and consists of l^2 components. To build it, stack the columns of Σ beneath each other so that the first column of Σ is on top, followed by the second column of Σ, etc.; the l-th column of Σ is therefore at the bottom of the stack. Because Σ is symmetric, this vector vecΣ contains considerable redundancy. To obtain a parsimonious column vector with the same information as is in Σ, eliminate all elements that come from the super-diagonal elements of Σ. The resulting vector, with only $l(l+1)/2$ elements is denoted vechΣ.

The *duplication matrix* D_l (Magnus and Neudecker 1988, §3.8) is a linear transform that links vecΣ and vechΣ:

$$D_l \mathrm{vech}\Sigma = \mathrm{vec}\Sigma .$$

D_l is a unique matrix of dimension $l^2 \times l(l+1)/2$. We use D_l to express an elemental information matrix with parameters $\vartheta = \{a^T, (\mathrm{vech}\Sigma)^T\}$) in a relatively compact format:

$$\mu(\theta) = \begin{pmatrix} \Sigma^{-1} & 0 \\ 0 & \frac{1}{2}D_m^T(\Sigma^{-1} \otimes \Sigma^{-1})D_m \end{pmatrix} . \tag{26}$$

For example, for the bivariate normal distribution with parameters $\theta = (a_1, a_2, \sigma_1^2, \rho\sigma_1\sigma_2, \sigma_2^2)^T$,

$$D_2 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix},$$

which inserted in (26), yields the compact re-expression of the elemental information matrix:

$$\mu(\theta) = \frac{1}{1-\rho^2} \begin{pmatrix} \Sigma^{-1} & 0 \\ 0 & B \end{pmatrix} , \tag{27}$$

where

$$\Sigma^{-1} = \begin{pmatrix} \frac{1}{\sigma_1^2} & -\frac{\rho}{\sigma_1\sigma_2} \\ -\frac{\rho}{\sigma_1\sigma_2} & \frac{1}{\sigma_2^2} \end{pmatrix} ,$$

$$B = \frac{1}{1-\rho^2} \begin{pmatrix} \frac{2-\rho^2}{4\sigma_1^4} & -\frac{\rho}{\sigma_1^3\sigma_2} & \frac{\rho^2}{2\sigma_1^2\sigma_2^2} \\ -\frac{\rho}{\sigma_1^3\sigma_2} & \frac{1+\rho^2}{\sigma_1^2\sigma_2^2} & -\frac{\rho}{\sigma_1\sigma_2^3} \\ \frac{\rho^2}{2\sigma_1^2\sigma_2^2} & -\frac{\rho}{\sigma_1\sigma_2^3} & \frac{2-\rho^2}{4\sigma_2^4} \end{pmatrix} .$$

4.4 Multinomial distribution

For multinomial observations $y = (y_1, \ldots, y_l, y_{l+1})$ (Bernardo and Smith 1994, §5.4)

$$p(y|\vartheta, n) = \frac{n!}{y_1! \cdots y_l! y_{l+1}!} \vartheta_1^{y_1} \cdots \vartheta_l^{y_l} \vartheta_{l+1}^{y_{l+1}} ,$$

where

$$\sum_{i=1}^{l+1} y_i = n \text{ and } \sum_{i=1}^{l+1} \vartheta_i = 1 .$$

There are actually only l independent parameters. We take, as is common, the elemental parameters to be $\vartheta = (\vartheta_1, \ldots, \vartheta_l)^T$, noting that $\vartheta_{l+1} = 1 - \sum_{i=1}^{l} \vartheta_i$.

The elemental information matrix for ϑ is

$$\mu(\vartheta) = \frac{n}{\vartheta_{l+1}} \begin{pmatrix} \frac{\vartheta_1+\vartheta_{l+1}}{\vartheta_1} & 1 & \cdots & 1 \\ 1 & \frac{\vartheta_2+\vartheta_{l+1}}{\vartheta_2} & \cdots & 1 \\ \cdots & \cdots & \cdots & \cdots \\ 1 & 1 & \cdots & \frac{\theta_l+\vartheta_{l+1}}{\vartheta_1} \end{pmatrix} . \tag{28}$$

The latter can be rewritten as

$$\mu(\vartheta) = n \begin{pmatrix} \vartheta_1 & 0 & \cdots & 0 \\ 0 & \vartheta_2 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & \vartheta_l \end{pmatrix}^{-1} + \frac{n}{\vartheta_{l+1}} ll^T , \tag{29}$$

where $l^T = (1 \cdots 1)$. Recalling that

$$(A + ll^T)^{-1} = A^{-1} - \frac{A^{-1}ll^TA^{-1}}{1 + lA^{-1}l^T} ,$$

we obtain

$$\mu^{-1}(\vartheta) = \frac{1}{n} \begin{pmatrix} \vartheta_1 & 0 & \cdots & 0 \\ 0 & \vartheta_2 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & \vartheta_l \end{pmatrix} - \frac{1}{n}\vartheta\vartheta^T \,. \tag{30}$$

5 Examples

Example 1. Linear regression with normal errors and constant variance.

For normally distributed observations there are two elemental parameters, the mean a and the variance σ^2, see Table 4. Let the variance be constant and

$$\eta(x,\theta) = \begin{pmatrix} a(x,\vartheta) \\ \sigma^2 \end{pmatrix} = \begin{pmatrix} f^T(x) & 0 \\ 0,\ldots,0, & 1 \end{pmatrix} \begin{pmatrix} \vartheta \\ \sigma^2 \end{pmatrix} \,. \tag{31}$$

Often it is assumed that σ^2 does not depend on x. Then from (7), (11) and Table 1 it follows that

$$\mu(x) = \frac{1}{\sigma^2} \begin{pmatrix} f(x)f^T(x) & 0 \\ 0 & 1/2\sigma^2 \end{pmatrix}. \tag{32}$$

Note that the information matrix is block diagonal, i.e. the first m components $\hat{\vartheta}_p$ of the MLE are independent of the last one, $\hat{\vartheta}_v = \hat{\sigma}^2$.

The information matrix for the design ξ is

$$M(\xi) = \begin{pmatrix} M_p(\xi) & 0 \\ 0 & 1/2\sigma^4 \end{pmatrix} \,, \tag{33}$$

where $M_p(\xi) = \int f(x)f^T(x)\xi(dx)$.

We now apply the above to the D-criterion. From Theorem 1 and Table 1 it follows that the sensitivity function is $\psi(x) = f(x){M_p}^{-1}(\xi)f^T(x) - m$, and a necessary and sufficient condition for ξ to be optimal is fulfillment of the inequality

$$f^T(x){M_p}^{-1}(\xi)f(x) \le m \,.$$

The later statement is of course the major part of Kiefer-Wolfowitz equivalence theorem. Note that this inequality does not contain any unknown parameters.

Example 2. Normal linear regression with independently parameterized variance.

Let us continue the previous example with

$$\eta(x,\theta) = \begin{pmatrix} a(x,\vartheta_p) \\ \ln\sigma^2(x,\vartheta_v) \end{pmatrix} = \begin{pmatrix} f^T(x) & 0 \\ 0, & \varphi^T(x) \end{pmatrix} \begin{pmatrix} \vartheta_p \\ \vartheta_v \end{pmatrix} \,. \tag{34}$$

From (7) and the first line of Table 5 it follows that

$$M(\xi) = \begin{pmatrix} M_p(\xi) & 0 \\ 0 & 1/2M_v(\xi) \end{pmatrix} , \tag{35}$$

where $M_p(\xi) = \int f(x)f^T(x)\xi(dx)$ and $M_v(\xi) = \int \varphi(x)\varphi^T(x)\xi(dx)$. Applying Theorem 1 and Table 1 we have for the D-criterion that (compare with Atkinson and Cook 1995)

$$e^{-\vartheta_v^T\varphi(x)} f^T(x){M_p}^{-1}(\xi)f(x) + \frac{1}{2}\varphi(x){M_v}^{-1}(\xi)\varphi^T(x) \leq m_p + m_v ,$$

where $m_p = \dim \vartheta_p$ and $m_v = \dim \vartheta_v$.

Example 3. One parameter families, linear predictor function and D-optimality.

Let $\eta(x,\theta) = h(\theta^T f(x))$, where the range of the inverse link function h coincides with the domain of the corresponding parameter (e.g. between 0 and 1 for p from Table 2). From (7) and Table 1 it can immediately be seen that a necessary and sufficient condition for ξ to be optimal is fulfillment of the inequality

$$\lambda(\theta^T f(x))f^T(x)M^{-1}(\xi)f(x) \leq m ,$$

where $\lambda(u) = \nu(u)\phi^2(u)$, $\phi(u) = \partial h(u)/\partial u$ and $M(\xi) = \int f(x)f^T(x)\xi(dx)$. Compare our results with the special cases presented by Wu (1985) and Torsney and Gunduz (2001) for binary regression or of Ford, Torsney, and Wu (1992) and Atkinson et al. (2007, Cap. 22) who considered generalized linear models.

Example 4. Bivariate binary response model.

Consider two binary outcomes, efficacy and toxicity, from a clinical trial. The possible outcomes are $y = (y_{00}, y_{01}, y_{10}, y_{11})$ with probabilities $\vartheta^T = (\vartheta_1, \ldots, \vartheta_4)$. It is more intuitive to re-express these probabilities respectively as p_{00}, p_{01}, p_{10} and p_{11}. The interpretation of these probabilities is: "probability of no efficacy, no toxicity"; "probability of no efficacy, toxicity"; etc. Let a "single" observation be an observation performed on a cohort of size n. Then

$$p(y|p,n) = \frac{n!}{y_{00}!y_{01}!y_{10},!y_{11}!} p_{00}^{y_{00}} p_{01}^{y_{01}} p_{10}^{y_{10}} p_{11}^{y_{11}} , \tag{36}$$

where $\sum_{i=1}^{2}\sum_{j=1}^{2} y_{ij} = n$ and $\sum_{i=1}^{2}\sum_{j=1}^{2} p_{ij} = 1$. Define $p = (p_{00}, p_{01}, p_{10})$. From (29) it follows that the elemental information matrix for a bivariate binary random variable and a cohort of size n is

$$\mu(\vartheta) = n \begin{pmatrix} p_{00} & 0 & 0 \\ 0 & p_{01} & 0 \\ 0 & 0 & p_{10} \end{pmatrix}^{-1} + \frac{n\, ll^T}{1 - p_{00} - p_{01} - p_{10}} . \tag{37}$$

This formula was derived and used in a number of publications on dose-response studies, see for instance, Dragalin and Fedorov (2006).

Example 5. Gamma regression.

In the case of gamma distributed observations there are two intuitively attractive ways to define the link function:

1. Model the parameters α and β directly:

$$\begin{aligned} \ln \alpha &= f^T(x)\vartheta_\alpha \,, \\ \ln \beta &= \varphi^T(x)\vartheta_\beta \,. \end{aligned}$$

2. Model the mean $a = \alpha\beta$ and variance $b = \alpha\beta^2$:

$$\begin{aligned} \ln a = \ln(\alpha\beta) &= \tilde{f}^T(x)\tilde{\vartheta}_a \,, \\ \ln b = \ln(\alpha\beta^2) &= \tilde{\varphi}^T(x)\tilde{\vartheta}_b \,. \end{aligned}$$

The formulae are more compact for the first case, which is the reason why we proceed with it. Following Sections 2 and 3 we have

$$F(x,\theta) = \begin{pmatrix} \alpha f(x) & 0 \\ 0 & \beta\varphi(x) \end{pmatrix} , \tag{38}$$

where $\theta^T = (\vartheta_\alpha^T, \vartheta_\beta^T)$ and the information matrix of a single observation made at x is (see Lemma 1 and Table 5):

$$\mu(x,\theta) = \alpha \begin{pmatrix} \alpha\psi'(\alpha) f(x) f^T(x)) & f(x)\varphi^T(x) \\ \varphi(x) f^T(x) & \varphi(x)\varphi^T(x) \end{pmatrix} , \tag{39}$$

where $\alpha = e^{f^T(x)\vartheta_\alpha}$ and $\beta = e^{\varphi^T(x)\vartheta_\beta}$. The matrix (39) allows us to build the total information matrix $M(\vartheta, \xi)$. A necessary and sufficient condition for D-optimality follows immediately from (38) and Table 1 :

$$\psi'(\alpha) f^T(x) \left(M^{-1}\right)_{\alpha\alpha} f(x) + \frac{2}{\beta} f^T(x) \left(M^{-1}\right)_{\alpha\beta} \varphi(x) + \frac{\alpha}{\beta^2} \varphi^T(x) \left(M^{-1}\right)_{\beta\beta} \varphi(x)$$

$$\leq m_\alpha + m_\beta \,,$$

where the matrices $\left(M^{-1}\right)_{\alpha\alpha}$, $\left(M^{-1}\right)_{\alpha\beta}$, $\left(M^{-1}\right)_{\beta\beta}$ are the blocks of the inverse information matrix corresponding to parameters ϑ_α and ϑ_β respectively. Of course, one has to remember that α and β are functions of x and of unknown parameters.

6 Conclusions

We have provided a set of tools that makes the optimal design of experiments as routine as possible for the most popular distributions of responses. A key is that the parameters of these distributions may depend on controllable, and perhaps uncontrollable, variables. Once a model is selected, that is the distribution of the responses and the predictive functions relating y and x, the design procedure consists of almost identical steps for all the alternatives enumerated and discussed in this article. We trust that this collection of results will not only streamline the practical aspects of experimental design, but that it will also lead to the development of rather simple software that can incorporate all the cases we have considered (and hopefully some we haven't) in one menu driven toolkit.

It should be clearly understood that there is a wealth of challenging problems that lies beyond the scope of one paper, even if we have surveyed much material. One such example is problems with transformed responses when the transformations depend on unknown parameters as in the Box and Cox (1964) transformation; see Atkinson (2005) for some design aspects in this case.

In various areas of biostatistics it may be challenging to build information matrices when the correlated multivariate responses consist of both continuous and discrete variables. See Tate (1955) (with a correction in Hannan and Tate 1965) and Fedorov, Wu, and Zhang (2012).

These remarks are an explicit call for joint efforts with other statisticians to build a collection of elemental information matrices to make experimental design more attractive and readily available for a wider population of practitioners.

Acknowledgment This paper was written at the Isaac Newton Institute for Mathematical Sciences in Cambridge, England, during the 2011 programme on the Design and Analysis of Experiments.

References

Abramowitz, M. and I. A. Stegun (1965). *Handbook of Mathematical Functions*. New York: Dover Publications.

Ali, M. M. and S. Nadarajah (2007). Information matrices for normal and Laplace mixtures. *Information Sciences 117*, 947–955.

Atkinson, A. C. (2005). Robust optimum designs for transformation of the response in a multivariate chemical kinetic model. *Technometrics 47*, 478–487.

Atkinson, A. C. and R. D. Cook (1995). D-optimum designs for heteroscedastic linear models. *Journal of the American Statistical Association 90*, 204–212.

Atkinson, A. C., A. N. Donev, and R. D. Tobias (2007). *Optimum Experimental Designs, with SAS*. Oxford: Oxford University Press.

Bernardo, J. M. and A. F. M. Smith (1994). *Bayesian Theory.* New York: Wiley.

Box, G. E. P. and D. R. Cox (1964). An analysis of transformations (with discussion). *Journal of the Royal Statistical Society, Series B 26*, 211–246.

Brazauskas, V. (2003). Information matrix for Pareto(IV), Burr, and related distributions. *Communications in Statistics - Theory and Methods 32*(2), 315–325. doi: 10.1081/STA-120018188.

Cox, D. R. and D. Oakes (1984). *Analysis of Survival Data*. Boca Raton: Chapman and Hall/ CRC Press.

Cox, D. R. and R. Reid (2000). *The Theory of the Design of Experiments*. Boca Raton: Chapman and Hall/ CRC Press.

Dette, H. and W. K. Wong (1999). Optimal designs when the variance is a function of the mean. *Biometrics 55*, 925–929.

Dragalin, V. and V. Fedorov (2006). Adaptive designs for dose-finding based on efficacy-toxicity response. *Journal of Statistical Planning and Inference 136*(6), 1800 – 1823. doi: 10.1016/j.jspi.2005.08.005.

Efron, B. and I. M. Johnstone (1990). Fisher's information in terms of the hazard rate. *The Annals of Statistics 18*, 38–62.

Escobar, L. A. and W. Q. Meeker (1994). Algorithm AS 292: Fisher information matrix for the extreme value, normal and logistic distributions and censored data. *Applied Statistics 43*, 533–540.

Fedorov, V. V. (1972). *Theory of Optimal Experiments*. New York: Academic Press.

Fedorov, V. V. and P. Hackl (1997). *Model-Oriented Design of Experiments*. Lecture Notes in Statistics 125. New York: Springer Verlag.

Fedorov, V. V. and S. L. Leonov (2004). Parameter estimation for models with unknown parameters in variance. *Communications in Statistics - Theory and Methods 33*, 2627–2657. doi:10.1081/STA-200037917.

Fedorov, V. V., Y. Wu, and R. Zhang (2012). Optimal dose-finding designs with correlated continuous and discrete responses. *Statistics in Medicine 31*. (To appear).

Ford, I., B. Torsney, and C. F. J. Wu (1992). The use of a canonical form in the construction of locally optimal designs for non-linear problems. *Journal of the Royal Statistical Society, Series B 54*, 569–583.

Gertsbakh, I. (1995). On the Fisher information in type-I censored and quantal response data. *Statistics and Probability Letters 23*, 297–306.

Gupta, A. K. and S. Nadarajah (2007). Information matrices for some bivariate Pareto distributions. *Applied Mathematics and Computation 184*, 1069–1079.

Gupta, R. D. and D. Kundu (2006). On the comparison of Fisher information of the Weibull and Generalizd Exponential distributions. *Journal of Statistical Planning and Inference 136*, 3130 – 3144. doi: 10.1016/j.jspi.2004.11.013.

Hannan, J. F. and R. F. Tate (1965). Estimation of the parameters for a multivariate normal distribution when one variable is dichotomized. *Biometrika 52*, 664–668.

Harris, C. M. (1968). The Pareto distribution as a queue server discipline. *Operations Research 16*, 307–313.

Harville, D. A. (2002). *Matrix Algebra From a Statistician's Perspective*. New York: Springer-Verlag.

Johnson, N. L., A. W. Kemp, and N. Balakrishnan (2005). *Univariate Discrete Distributions, 3rd edition*. New York: Wiley.

Johnson, N. L., S. Kotz, and N. Balakrishnan (1994). *Continuous Univariate Distributions - 1, 2nd edition*. New York: Wiley.

Johnson, N. L., S. Kotz, and N. Balakrishnan (1995). *Continuous Univariate Distributions - 2, 2nd edition*. New York: Wiley.

Johnson, N. L., S. Kotz, and N. Balakrishnan (1997). *Discrete Multivariate Distributions*. New York: Wiley.

Kiefer, J. and J. Wolfowitz (1960). The equivalence of two extremum problems. *Canadian Journal of Mathematics 12*, 363–366.

Kotz, S. (2008). Information matrices for some bivariate Pareto distributions. In G. Betti and A. Lemmi (Eds.), *Advances on Income Inequality and Concentration Measures*, pp. 105–116. London: Routledge.

Kotz, S., N. Balakrishnan, and N. L. Johnson (2000). *Continuous Multivariate Distributions - 1, 2nd edition*. New York: Wiley.

Kotz, S., T. J. Kozubowski, and K. Podgórski (2001). *The Laplace Distribution and Generalizations: a revisit with applications to communications, economics, engineering and finance*. Boston: Birkhäuser-Verlag.

Lehmann, E. and G. Casella (1998). *Theory of Point Estimation, 2nd edition*. New York: Sringer-Verlag.

Magnus, J. R. and H. Neudecker (1988). *Matrix Differential Calculus with Applications in Statistics and Econometrics*. Chichester: Wiley.

Paul, S. R., U. Balasooriya, and T. Banerjee (2005). Fisher information matrix of the Dirichlet-multinomial distribution. *Biometrical Journal 47*, 230236. DOI:10.1002/bimj.200410103.

Pázman, A. (1986). *Foundations of Optimum Experimental Design*. Dordrecht: Reidel.

Pukelsheim, F. (1993). *Optimal Design of Experiments*. New York: Wiley.

Silvey, S. D. (1980). *Optimum Design*. London: Chapman and Hall.

Tate, R. F. (1955). The theory of correlation between two continuous variables when one is dichotomized. *Biometrika 42*, 205–216.

Torsney, B. and N. Gunduz (2001). On optimal designs for high dimensional binary regression models. In A. C. Atkinson, B. Bogacka, and A. Zhigljavsky (Eds.), *Optimal Design 2000*, pp. 275–285. Dordrecht: Kluwer.

Wu, C. F. J. (1985). Efficient sequential designs with binary data. *Journal of the American Statistical Association 80*, 974–984.

Wu, Y. H., V. V. Fedorov, and K. J. Propert (2005). Optimal design for dose response using beta distributed responses. *Journal of Biopharmaceutical Statistics 15*, 753–771. doi: 10.1081/BIP-200067760.

Wynn, H. P. (1970). The sequential generation of D-optimal experimental designs. *Annals of Mathematical Statistics 41*, 1055–1064.

Yari, G. and A. M. D. Jafari (2006). Information and covariance matrices for multivariate Pareto (IV), Burr, and related distributions. *International Journal of Engineering Science 17*, 61–69.