Canonical moments, orthogonal polynomials with application to statistics

Holger Dette
Ruhr-Universität Bochum
Fakultät für Mathematik
44780 Bochum
Germany
email: holger.dette@ruhr-uni-bochum.de

April 16, 2002

Abstract

In this paper we describe the special role of moment theory for the construction of optimal designs in statistical regression models. A careful introduction in the problem of designing experiments for certain polynomial regression models is given, and it is demonstrated that the maximization of certain Hankel determinants over the moment space plays a particular role for the construction of optimal designs in these models. We introduce the theoy of canonical moments, which povide a powerful tool for the maximization of functionals of Hankel determinants and illustrate its application in several statistical problems. On the other hand these results can be used for the derivation of several new results in approximation theory. As examples we give simple proofs for the asymptotic distribution of the zeros of classical orthogonal polynomials, generalize the trigonometric identity $\sin^2\theta + \cos^2\theta = 1$ to abitrary systems of polynomials orthogonal with respect to a measure with compact support and give a solution of a nonlinear extremal problem for polynomials.

AMS 1991 Subject Classifications: 33C45; 62K05

Key words and phrases: orthogonal polynomials, Hankel matrix, canonical moments, trigonometric identity, D-optimal design, model robust designs, weak asymptotics, extremal problems for polynomials

Contents

1 Optimal designs for polynomial regression

1.1 Introduction to regression models

In the following section we give a careful explanation of the application of moment theory in the construction of optimum designs for polynomial regression models. For a more general description of optimum experimental designs we refer to the monographs of Fedorov (1972), Silvey (1980) and Pukelsheim (1993). For the sake of brevity we will only mention the results which are relevant for the discussion presented in this paper.

An important model in statistics is the *univariate polynomial regression model* of degree $d \in \mathbb{N}_0$

$$Y \;=\; \sum_{j=0}^{d} \theta_j x^j + \varepsilon \;=\; f(x)^T\theta + \varepsilon \ , \tag{1.1}$$

where $\theta = (\theta_0, \theta_1, \ldots, \theta_d)^T$ is a vector of unknown parameters, $f(x) = (1, x, \ldots, x^d)^T$ is the vector of monomials up to the order d and ε is a random error with mean $E(\varepsilon) = 0$ and variance $\operatorname{Var}(\varepsilon) = \sigma^2 > 0$. The interpretation of the model (1.1) is that Y is the result of a measurement at a point $x \in \mathcal{X}$ which is the sum of the expectation, *the deterministic mean effect* $f(x)^T\theta$, and an additive error term ε. Y is called the *response* at the point $x \in \mathcal{X}$. In general the relationship between x and Y would have $f(x)^T\theta$ replaced by some arbitrary unknown function $g(x)$. For convenience this function $g(x)$ is assumed to be a polynomial of degree d.

The set $\mathcal{X}$ of all possible points where observations are assumed to be located is the interval $[-1, 1]$ (if not stated otherwise) and is called the *design space*. The variance of the random term ε in (1.1) (which subsumes quite different sources of error) is assumed to be independent of the specific point x, where the response Y is observed. This assumption is referred to in the literature as the *homoscedastic assumption*. The goal of the experiment is to estimate the unknown parameters $\theta_0, \ldots, \theta_d$ in the polynomial regression model, where n observations

$$Y_j = f(x_j)^T\theta + \varepsilon_j \qquad\qquad (j = 1, \ldots, n) \tag{1.2}$$

at experimental conditions $x_1, \ldots, x_n \in \mathcal{X}$ are available. The x_i values are not necessarily distinct, i.e. repeated observations at some x_i are allowed, however all observations are assumed to be *uncorrelated*, i.e.

$$E(\varepsilon_i\varepsilon_j) = \begin{cases} \sigma^2 & \text{if } i = j \\ 0 & \text{else.} \end{cases} \tag{1.3}$$

If the different responses and errors are collected in vectors $Y = (Y_1, \ldots, Y_n)^T$ and $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)^T$, then (1.2) and (1.3) can be conveniently written in matrix form

$$Y = X\theta + \varepsilon \ ,$$

where

$$X \;=\; \begin{bmatrix} 1 & x_1 & \cdots & x_1^d \\ 1 & x_2 & \cdots & x_2^d \\ \vdots & \vdots & & \vdots \\ 1 & x_n & \cdots & x_n^d \end{bmatrix} \in \mathbb{R}^{n\times(d+1)}$$

denotes the $n \times (d+1)$ *design matrix*. The expectation and the dispersion (matrix) of the random vector Y are given by (note (1.3))

$$E(Y) = X\theta \qquad\qquad D(Y) = \sigma^2 I_n \tag{1.4}$$

where I_n denotes the $n \times n$ identity matrix.
For the estimation of the unknown parameters θ from the observed data $Y = (Y_1, \dots, Y_n)^T$ we restrict our considerations to *linear unbiased estimates* for θ, which are estimators of the form

$$\hat{\theta}_L = LY \tag{1.5}$$

where $L \in \mathbb{R}^{(d+1)\times n}$ is a given $(d+1) \times n$ matrix such that

$$E[\hat{\theta}] = LX\theta = \theta \tag{1.6}$$

is satisfied for all $\theta \in \mathbb{R}^{d+1}$. Obviously the condtion (1.6) is equivalent to the condition that the matrix L is a left inverse of the matrix X, that is $LX = I_{d+1}$. Note that the dispersion matrix of a linear estimator (1.5) is nonnegative definite, i.e.

$$D(\hat{\theta}_L) = D(LY) = \sigma^2 LL^T \;\geq\; 0,$$

and different linear unbiased estimators (specified by different matrices) can be compared by a partial ordering. To be precise we define for symmetric matrices $A,\, B \in \mathbb{R}^{(d+1)\times(d+1)}$

$$\begin{array}{lll} A \geq B & \text{if and only if} & A - B \text{ is nonngegative definite} \\ A > B & \text{if and only if} & A - B \text{ is positive definite} \end{array} \tag{1.7}$$

The partial ordering defined on the set of symmetric matrices is called the *Loewner ordering*. It is a well known fact in statistics that this dispersion matrix can be minimized (in the Loewner ordering) with respect to all linear unbiased estimators $\hat{\theta}_L$ for θ.

Theorem 1.1. *For the linear model with moment assumptions (1.4) wtih rank $(X) = d+1$, the estimator*

$$\hat{\theta}^{GM} \;=\; (X^T X)^{-1} X^T Y \tag{1.8}$$

is the best linear unbiased estimator (BLUE) with respect to the Loewner ordering; that is,

$$\sigma^2 (X^T X)^{-1} \;=\; D(\hat{\theta}^{GM}) \;\leq\; D(\hat{\theta}_L) \tag{1.9}$$

for all linear unbiased estimators $\hat{\theta}_L$ for the parameter θ.

Proof. From (1.6) we obtain $LX = I$ for any L with $E[LY] = \theta$. Then, since,

$$((X^T X)^{-1} X^T - L)((X^T X)^{-1} X^T - L)^T \;\geq\; 0$$

it follows that

$$(X^T X)^{-1} - LX(X^T X)^{-1} - (X^T X)^{-1} X^T L^T + LL^T \;\geq\; 0.$$

Since $LX = I$ we have

$$D(\hat{\theta}_L) \ = \ \sigma^2 LL^T \ \geq \ \sigma^2 (X^T X)^{-1} \ .$$

□

Theorem 1.1. is usually called the *Gauss Markov Theorem* and the estimator $\hat{\theta}^{GM}$ is called the *Gauss Markov estimator* for the full parameter vector θ. We point out here that optimal linear estimators for linear combinations of the components of θ are simply obtained by taking the corresponding linear combinations of the components of the estimator $\hat{\theta}^{GM}$. Note also that $\hat{\theta}^{GM}$ is the well known *least squares estimator*, which is obtained by minimzing the function

$$\sum_{i=1}^{n} \Big[Y_i - \sum_{j=0}^{d} \theta_j x_i^j \Big]^2$$

with respect to the choice of the parameters $\theta_0, \ldots, \theta_d$.

1.2 Optimal designs for regression models

Note that Theorem 1.1 gives a lower bound for the smallest possible variance of an estimator of the form LY, which is given by

$$\sigma^2 (X^T X)^{-1},$$

where the matrix on the right hand side is defined by

$$X^T X \ = \ n \begin{bmatrix} 1 & c_1 & c_2 & \cdots & c_d \\ c_1 & c_2 & c_3 & \cdots & c_{d+1} \\ \vdots & \vdots & \vdots & & \vdots \\ c_d & c_{d+1} & c_{d+2} & \cdots & c_{2d} \end{bmatrix} . \tag{1.10}$$

and

$$c_j = \frac{1}{n} \sum_{i=1}^{n} x_i^j \ , \quad j = 0, \ldots, 2d.$$

Moreover the matrix $X^T X$ depends on the design points $x_1, \ldots, x_n$ chosen by the experimenter and a reasonable question is, if the matrix $(X^T X)^{-1}$ can be further minimized (with respect to the Loewner ordering) by an appropriate choice of the experimental conditions $x_1 \ldots, x_n$. Equivalently one could try to maximize $X^T X$ as a function of the design points $x_1 \ldots, x_n$. However it can be proved [see Pukelsheim (1993), Chapter 4] that such a minimization or maximization is not possible except in the case $d = 0$ of a constant polynomial, which is of course not interesting from a practical point of view. The reason for these difficulties is that the Loewner ordering on the set of symmetric matrices is not complete. Therefore it is common practice to maximize real valued functionals of the matrix $X^T X$, where the functionals have a particular statistical meaning. These functions are usually called *optimality criteria* in the literature and we recall the most commonly used criteria here for the sake of completeness.

For a statistical interpretation of these criteria we refer to the books of Fedorov (1972), Silvey (1980) and Pukelsheim (1993). The *D-optimality criterion* determines the points $x_1, \ldots, x_n$ such that the determinant

$$|X^T X| \longrightarrow \max \tag{1.11}$$

becomes maximal. Similary, the A- and E- optimality look for arrangements of the design points such that

$$\left[\operatorname{tr}(X^T X)^{-1}\right]^{-1} \longrightarrow \max$$

$$\lambda_{\min}(X^T X) \longrightarrow \max$$

are maximal, respectively, where $\lambda_{\min}(A)$ denotes the minimal eigenvalue of a symmetric matrix A. For later purposes we finally mention the *D_1-optimality criterion*, which determines the designs points $x_1, \ldots, x_n$ such that

$$\left[e_d (X^T X)^{-1} e_d\right]^{-1} \longrightarrow \max \tag{1.12}$$

is maximal, where $e_d = (0, \ldots, 0, 1) \in \mathbb{R}^{d+1}$ denote the $(d+1)$th unit vector in $\mathbb{R}^{d+1}$. We begin with a careful discussion of the D-optimality criterion and the particular example of the linear and quadratic regression model.

Example 1.2. Consider the case $d = 1$ in (1.1), for which the polynomial regression model reduces to the well known model of *linear regression.* The matrix X is given by

$$X = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \in \mathbb{R}^{n \times 2} ,$$

which gives

$$X^T X = \begin{bmatrix} n & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 \end{bmatrix} .$$

The Gauss Markov estimator for the parameters θ_0 and θ_1 is obtained from Theorem 1.1.

$$\hat{\theta}_0 = \bar{Y}_n - \hat{\theta}_1 \bar{x}_n$$

$$\hat{\theta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x}_n)(Y_i - \bar{Y}_n)}{\sum_{i=1}^n (x_i - \bar{x}_n)^2}$$

where $\bar{Y}_n = \frac{1}{n}\sum_{i=1}^n Y_i$ and $\bar{x}_n = \frac{1}{n}\sum_{i=1}^n x_i$ denote the mean of the observations and design points, respectively. The covariance matrix is given by $\sigma^2 (X^T X)^{-1}$ and the D-criterion advises the experimenter to choose observations at the points $x_1, \ldots, x_n$ such that the determinant

$$|X^T X| = n \sum_{i=1}^n (x_i - \bar{x}_n)^2$$

is maximal. The D-optimal designs for the linear regression model have been determined by Hofmann and Jung (1975). If the number of observations is an even number, say $n = 2m$, then it is easy to see that the best choice for maximizing this determinant is to take half of the total observations at each of the extreme points of the designs space $\mathcal{X} = [-1, 1]$, i.e.

$$\begin{aligned} x_1 &= \ldots = x_m = -1 \\ x_{m+1} &= \ldots = x_{2m} = 1. \end{aligned}$$

If an odd number of observations, say $n = 2m + 1$ is available, the situation is slightly more complicated, but it can be shown that the best allocation is to take m observations at one and the other $m + 1$ observations at the other extreme point of the design space, i.e.

$$\begin{aligned} x_1 &= \ldots = x_{m+1} = -1 \\ x_{m+2} &= \ldots = x_{2m+1} = 1. \end{aligned}$$

Thus for the D-optimality criterion and the linear regression model the determination of an optimal design (in other words an allocation of the points $x_1, \ldots, x_n$ such that the determinant of the matrix $X^T X$ becomes maximal) is fairly simple. Note that the optimal designs are unique subject to a reflection at the origin.

We now consider the *quadratic polynomial regression* model, for which the situation is more complicated. The matrix X is given by

$$X = \begin{bmatrix} 1 & x_1 & x_1^2 \\ \vdots & \vdots & \vdots \\ 1 & x_n & x_n^2 \end{bmatrix} \in \mathbb{R}^{n \times 3},$$

which yields

$$X^T X = \begin{bmatrix} n & \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 \\ \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 & \sum_{i=1}^n x_i^3 \\ \sum_{i=1}^n x_i^2 & \sum_{i=1}^n x_i^3 & \sum_{i=1}^n x_i^4 \end{bmatrix}.$$

An application of the Cauchy Binet formula shows that the determinant of this matrix is given by

$$|X^T X| = \sum_{1 \le i < j < k \le n} (x_k - x_j)^2 (x_k - x_i)^2 (x_j - x_i)^2,$$

which has to be maximized with respect to the choice of the design points $x_1, \ldots, x_n \in [-1, 1]$. We will not give the explicit details of this maximization, but refer to the work of Gaffke and Krafft (1982). For the solution of the optimization problem three cases have to be considered. In all cases the optimum allocation is to take only observations at the points -1, 0 and 0 and to allocate the observations at these points as equal as possible. More precisely, if $n = 3m$, we use

$$\begin{aligned} x_1 &= \ldots = x_m = -1 \\ x_{m+1} &= \ldots = x_{2m} = 0 \\ x_{2m+1} &= \ldots = x_{3m} = 1 \end{aligned}$$

for $n = 3m + 1$ and $n = 3m + 2$ the optimal allocations are given by

$$\begin{aligned} x_1 &= \ldots = x_m = -1 \\ x_{m+1} &= \ldots = x_{2m+1} = 0 \\ x_{2m+2} &= \ldots = x_{3m+1} = 1 \end{aligned}$$

and

$$\begin{aligned} x_1 &= \ldots = x_{m+1} = -1 \\ x_{m+2} &= \ldots = x_{2m+1} = 0 \\ x_{2m+2} &= \ldots = x_{3m+2} = 1 \end{aligned}$$

respectively. We finally mention that the optimal designs are unique subject to a reflection at the origin.

We are now ready to formalize the illustrated optimization problems. To this end assume that the distinct points among $x_1, \ldots, x_n$ are the points $x_1, \ldots, x_l$ $(l \leq n)$ and let n_i denote the number of times the particular point x_i occurs among $x_1, \ldots, x_n$ $(i = 1, \ldots, l)$. By this procedure one obtains a probability measure $\xi_{(n)}$ on the design space $\mathcal{X} = [-1, 1]$ with finite support $\{x_1, \ldots, x_l\}$ and mass n_i/n at the point x_i $(i = 1, \ldots, l)$. We call any probability measure with finite support and masses which are multiples of $1/n$ an *exact design* for sample size n and summarize the information of such a measure in the matrix

$$\xi_{(n)} = \begin{pmatrix} x_1 & \cdots & x_l \\ \frac{n_1}{n} & \cdots & \frac{n_l}{n} \end{pmatrix} .$$

The first row of this matrix gives the points in the design space $\mathcal{X}$ where observations have to be taken and the second row tells the experimenter how many observations have to be taken at these points.

Example 1.3. Consider the quadratic regression model in Example 1.2 and assume that we can take $n = 17$ observations. The design

$$\xi_{(17)} = \begin{pmatrix} -1 & 0 & 1 \\ \frac{6}{17} & \frac{5}{17} & \frac{6}{17} \end{pmatrix}$$

is the (exact) D-optimal design. On the other hand the design

$$\xi_{(17)} = \begin{pmatrix} -1 & 0 & 1 \\ \frac{5}{17} & \frac{7}{17} & \frac{5}{17} \end{pmatrix}$$

is also exact for sample size $n = 17$, but it is not D-optimal.

Example 1.4. Consider the D_1-optimality criterion defined in (1.12) for the quadratic regression model. Using Cramers rule it is easy to see that the function in (1.12), which has to be maximized with respect to the choice of the points $x_1, \ldots, x_n \in [-1, 1]$ is given by

$$\left[e_2^T (X^T X)^{-1} e_2\right]^{-1} \;=\; \frac{\sum_{1\le i<j<k\le n}(x_k - x_j)^2 (x_k - x_i)^2 (x_j - x_i)^2}{n \sum_{i=1}^n (x_i - \bar{x}_n)^2}$$

The maximization of this expression is somewhat complicated and we only state the result which is due to Krafft and Schaefer (1995) in order to illustrate the difficulty of this concept of optimization. If $n = 4p + q$, $q \in \{0, 1, 3\}$ and $p \ge 1$ (or $p = 0$ and $q = 3$) the D_1-optimal design $\xi^*_{(n)}$ is unique and given by

$$\begin{aligned}
\xi^*_{(4p)} &= \begin{pmatrix} -1 & 0 & 1 \\ \frac{1}{4} & \frac{1}{2} & \frac{1}{4} \end{pmatrix} && (n = 4p) \\
\xi^*_{(4p+1)} &= \begin{pmatrix} -1 & 0 & 1 \\ \frac{p}{4p+1} & \frac{2p+1}{4p+1} & \frac{p}{4p+1} \end{pmatrix} && (n = 4p+1) \\
\xi^*_{(4p+3)} &= \begin{pmatrix} -1 & 0 & 1 \\ \frac{p+1}{4p+3} & \frac{2p+1}{4p+3} & \frac{p+1}{4p+3} \end{pmatrix} && (n = 4p+3).
\end{aligned}$$

In the case $n = 4p + 2$ the situation is substantially more complicated and there are two (exact) optimal designs, namely

$$\xi^*_{(4p+2)} = \begin{pmatrix} -1 & x_0(p) & 1 \\ \frac{p}{4p+2} & \frac{2p+1}{4p+2} & \frac{p+1}{4p+2} \end{pmatrix} \qquad (n = 4p+2) \tag{1.13}$$

and its reflection at the point 0. Here $x_0(p)$ is the real root of the cubic polynomial

$$(2p+1)^2 x^3 - 3(2p+1)x^2 + (20p^2 + 20p + 3)x - 2p - 1. \tag{1.14}$$

With the notation of an exact design, the matrix $X^T X$ can be written as a Stieltjes integral

$$\begin{aligned}
X^T X &= \sum_{j=1}^n f(x_j) f^T(x_j) \;=\; \sum_{j=1}^l n_j f(x_j) f^T(x_j) \;=\; n \sum_{j=1}^l \frac{n_j}{n} f(x_j) f^T(x_j) \\
&= n \int_{\mathcal{X}} f(x) f^T(x) d\xi_{(n)}(x) \;=\; n M(\xi_{(n)}),
\end{aligned} \tag{1.15}$$

where $f^T(x) = (1, x, \ldots, x^d)$ denotes the vector of monomials up to the order d and the last equality defines the $(d+1) \times (d+1)$ matrix $M(\xi_{(n)})$. Note that if $c_i = \int_{\mathcal{X}} x^i d\xi_{(n)}(x)$ denotes the ith moment of the exact design $\xi_{(n)}$, then the matrix

$$M(\xi_{(n)}) = (c_{i+j})_{i,j=0}^d$$

is the *Hankel matrix* of the design $\xi_{(n)}$. In general the maximization of a function of $M(\xi_{(n)})$ over the set of all exact designs is a highly nonlinear discrete optimization problem, which can only be solved in rare circumstances similar to the examples presented above. For these reasons the concept of optimization introduced so far has to be modified appropriately. One main difficulty is that for a fixed sample size n the set of all exact designs for this sample size is not convex. In the following we will slightly modify the definition of a design in order to make the set of all designs convex.

Definition 1.5. *An approximate design is a probability measure on the design space $\mathcal{X}$ with finite support and an approximate design will usually be represented in the matrix form*

$$\xi = \begin{pmatrix} x_1 & \cdots & x_l \\ w_1 & \cdots & w_l \end{pmatrix} . \tag{1.16}$$

The set of all approximate designs is denoted by Ξ and the matrix

$$\begin{aligned} M(\xi) &= \int_{\mathcal{X}} f(x) f^T(x) d\xi(x) = \sum_{j=1}^{l} w_j f(x_j) f^T(x_j) \\ &= \begin{bmatrix} c_0 & c_1 & \cdots & c_d \\ c_1 & c_2 & \cdots & c_{d+1} \\ \vdots & \vdots & & \vdots \\ c_d & c_{d+1} & \cdots & c_{2d} \end{bmatrix} \end{aligned} \tag{1.17}$$

is called *a moment matrix, information matrix or Hankel matrix*, where $c_i = \int_{\mathcal{X}} x^i d\xi(x)$ denotes the ith moment of the design ξ.

Note that the support points of the design ξ, say $x_1, \ldots, x_l$, give the locations where observations have to be taken and the masses $w_1, \ldots, w_l$ give the proportions of the total observations taken at the corresponding points. Obviously, an exact design for the sample size n is also an approximate one but the converse is in general not true, because the weights in (1.16) are not necessarily multiples of $1/n$. Very often an approximate design is called a *design for an infinite sample size*, because it arises from the exact design of sample size n when n tends to infinity. However, for a finite sample size n the numbers $w_j n$ are not necessarily integers and an optimal approximate design has to be approximated by an exact design for sample size n using appropriate rounding procedures.

Example 1.6. Consider the quadratic regression ($d = 2$) for the sample size $n = 17$. By the above discussion the approximate design arises from the exact design if the sample size tends to infinity. Therefore observing the discussion in Example 1.2 the D-optimal approximate design for the quadratic regression model on the interval $[-1, 1]$ is given by

$$\xi^* = \begin{pmatrix} -1 & 0 & 1 \\ 1/3 & 1/3 & 1/3 \end{pmatrix}$$

(note that this is not the common way of determining optimal approximate designs, because in general the exact designs are not known and the concept of optimal approximate designs is introduced in order to deal with the discrete optimization problem). From the approximate optimal design we get by an appropriate rounding procedure an exact design for the required sample size, e.g.

$$\tilde{\xi}_{(17)} = \begin{pmatrix} -1 & 0 & 1 \\ \frac{5}{17} & \frac{7}{17} & \frac{5}{17} \end{pmatrix}$$

By the same Example 1.2 the D-optimal exact design for sample size 17 is

$$\xi_{(17)} = \begin{pmatrix} -1 & 0 & 1 \\ \frac{6}{17} & \frac{5}{17} & \frac{6}{17} \end{pmatrix}$$

The performance of the design obtained from an approximate design and a rounding procedure with respect to the D-optimal exact design (for the given sample size) is usally measured in terms of the *D-efficiency*

$$\left(\frac{|M(\tilde{\xi}_{(17)})|}{|M(\xi_{(17)})|} \right)^{1/3} \approx 99,07\% \ ,$$

where $1/3$ in the exponent corresponds to the number of unknown parameters in the quadratic regression model. We note that the design obtained from the approximate design using a rounding procedure is very efficient in the sense that the determinant of its information matrix is close to the determinant of the information matrix of the D-optimal exact design.

Example 1.7. Consider the quadratic regression Example 1.3 where we are interested in finding the D_1-optimal design, which maximizes

$$(e_2{}^T M^{-1}(\xi) e_2)^{-1} = \frac{|M(\xi)|}{c_2 - c_1^2} \tag{1.18}$$

where $e_2 = (0, 0, 1)^T$,

$$M(\xi) = \begin{bmatrix} c_0 & c_1 & c_2 \\ c_1 & c_2 & c_3 \\ c_2 & c_3 & c_4 \end{bmatrix}$$

and $c_j = \int_{-1}^{1} x^j d\xi(x)$ denotes the jth moment of the design ξ. We will show in Example 3.2 (in a more general context) that the optimum approximate D_1-optimal design is given by the measure

$$\xi^* = \begin{pmatrix} -1 & 0 & 1 \\ \frac{1}{4} & \frac{1}{2} & \frac{1}{4} \end{pmatrix}.$$

This means that the experimenter should take 1/4 of the observations at the points -1 and 1 and 1/2 of the observations at the point 0. If the sample size n is not a multiple of 4, a rounding procedure is applied to produce an exact design for the sample size n. In order to compare (exact) designs obtained by this procedure with the optimal exact designs $\xi^*_{(n)}$ of Example 1.4 we apply the following (simple) apportionment method. If n_0 is the closest integer to $n/4$ we use the exact design

$$\tilde{\xi}_{(n)} \;=\; \begin{pmatrix} -1 & 0 & 1 \\ \frac{n_0}{n} & 1 - \frac{2n_0}{n} & \frac{n_0}{n} \end{pmatrix}$$

as approximation of the optimal design ξ^* (if there are two integers with the same distance to $n/4$ we define n_0 as the smaller one). Whenever $n \neq 4p+2$ the design $\tilde{\xi}_{(n)}$ coincides with the optimal exact design $\xi^*_{(n)}$ of Example 1.4.

It is reasonable to compare the performance of the two designs $\xi^*_{(4p+2)}$ and $\tilde{\xi}_{(4p+2)}$ by the ratio

$$r(\tilde{\xi}_{(4p+2)}, \xi^*_{(4p+2)}) \;=\; \frac{(e_2{}^T M^{-1}(\tilde{\xi}_{(4p+2)})e_2)^{-1}}{(e_2{}^T M^{-1}(\xi^*_{(4p+2)})e_2)^{-1}}.$$

The following table contains these ratios and the solution $x_0(p)$ of the equation (1.14) for different values of p

p	1	2	3	4	5
n	6	10	14	18	22
$x_0(p)$	0.0707	0.0408	0.0289	0.0224	0.0183
$r(\tilde{\xi}_{(4p+2)}, \xi^*_{(4p+2)})$	0.9327	0.9759	0.9877	0.9925	0.9950

We note that there appear only minor differences between the exact design $\tilde{\xi}_{(n)}$ (constructed by an approximation to the optimal approximate design) and the optimum exact design $\xi^*_{(n)}$ (constructed by integer optimization). Thus the approximate design approach provides an efficient solution of the exact design problem.

A general result in statistical design theory [see Pukelsheim and Rieder (1993)] shows that under assumptions of differentiability the loss of efficiency by using an exact design obtained from an optimal approximate design by an appropriate rounding procedure is of order $O(\frac{1}{n^2})$, where n denotes the sample size. Thus for reasonable sample sizes the concept of approximate designs seems to be justified in the sense that the application of an appropriate rounding procedure to the optimal approximate design yields efficient designs for the given sample size. We will use this concept thoughout the remaining part of this paper.

1.3 D-optimal approximate designs for polynomial regression

Comparing the formulas (1.15) and (1.17) it seems to be appropriate to call an approximate design D-optimal if it maximizes the determinant

$$|M(\xi)| = \left| \Big(\int_{-1}^{1} x^{i+j} d\xi(x) \Big)_{i,j=0}^{d} \right| \to \max_{\xi} \tag{1.19}$$

in the class of all designs Ξ. This problem was solved simultaneously by Guest (1958) and Hoel (1958). Note that the optimization problem is now convex, which makes the optimization a little easier. However, the maximization problem in (1.19) is still an infinite dimensional one, because we do not have any information regarding the number of the support of a D-optimal approximate design (we only know that it has finite support). The following result gives a characterization of the D-optimal approximate design and is due to Kiefer and Wolfowitz (1960). As a by-product it provides an upper bound for the number of support points of the D-optimal approximate design in a polynomial regression model. As a consequence we are able to reduce the infinite dimensional optimization problem to a finite dimensional one.

Theorem 1.8. equivalence Theorem for D-optimal designs*) An approximate design ξ^* is D-optimal for the polynomial regression model if and only if the inequality*

$$d(x,\xi^*) = f^T(x)M^{-1}(\xi^*)f(x) \leq d+1 \tag{1.20}$$

holds for all $x \in [-1,1]$. Moreover, there is equality in (1.20) at all support points of the D-optimal design ξ^*.

Proof. For a concave function (in other words an optimality criterion) $\Phi : \xi \to \Phi(\xi) \in \mathbb{R}$ on the set Ξ of all approximate designs define the *Fréchet derivative* of Φ at the design ξ_1 in direction of ξ_2 by

$$F_\Phi(\xi_1,\xi_2) = \lim_{\varepsilon\to 0^+} \frac{1}{\varepsilon}\{\Phi((1-\varepsilon)\xi_1+\varepsilon\xi_2) - \Phi(\xi_1)\}.$$

Note that the limit exists, because the concavity of Φ implies that the expression

$$H_{\xi_1,\xi_2}(\varepsilon) \;=\; \frac{1}{\varepsilon}\{\Phi((1-\varepsilon)\xi_1+\varepsilon\xi_2) - \Phi(\xi_1)\}$$

is decreasing with $\varepsilon > 0$. Now, if ξ^* maximizes the function Φ, then we obviously have

$$\Phi((1-\varepsilon)\xi^*+\varepsilon\xi) - \Phi(\xi^*) \leq 0 \quad \forall\, \xi \in \Xi,$$

which implies

$$F_\Phi(\xi^*,\xi) \leq 0 \quad \forall\, \xi \in \Xi. \tag{1.21}$$

On the other hand, if (1.21) is satsified, it follows observing that $H_{\xi_1,\xi_2}(\varepsilon)$ is decreasing with ε that

$$\Phi(\xi) - \Phi(\xi^*) \leq F_\Phi(\xi^*,\xi) \leq 0 \quad \forall\, \xi \in \Xi,$$

which means that ξ^* maximizes Φ. In other words ξ^* maximizes the function Φ if and only if (1.21) holds. If we choose $\xi = \delta_x$ as a Dirac measure at the point $x \in [-1,1]$, then the optimality of ξ^* implies that

$$F_\Phi(\xi^*,\delta_x) \leq 0 \quad \forall\, x \in [-1,1] \tag{1.22}$$

On the other hand observe that the Fréchet derivative is linear with respect to convex combinations of the second argument [see e.g. Silvey (1980)], i.e.

$$F_\Phi\Big(\xi^*, \sum_{i=1}^k \lambda_i \delta_{x_i}\Big) = \sum_{i=1}^k \lambda_i F_\Phi(\xi^*, \delta_{x_i}); \quad \forall\ \lambda_1, \ldots, \lambda_k > 0; \sum_{i=1}^k \lambda_i = 1;$$

then we obtain from (1.22) the relation (1.21). This shows that (1.21) and (1.22) are equivalent, and consequently a designs ξ^* maximizes the function Φ if and only if the inequality (1.22) holds. All what remains is the calculation of Fréchet derivative for the D-optimality criterion. It actually turns out that the function $\xi \to |M(\xi)|$ is not concave. However the function $\Phi(\xi) = \log |M(\xi)|$ is concave on the set of all approximate designs [see Fedorov (1972)] and taking the logarithm does obviously not change the optimization problem. For this function we obtain by a straightforward calculation for all $x \in [-1, 1]$

$$\begin{aligned} F_\Phi(\xi_1, \delta_x) &= \operatorname{tr}(M(\delta_x) M^{-1}(\xi_1)) - (d+1) \\ &= \operatorname{tr}(f(x) f^T(x) M^{-1}(\xi_1)) - (d+1) \\ &= f^T(x) M^{-1}(\xi_1) f(x) - (d+1), \end{aligned}$$

which completes the proof of the first part of Theorem 1.8. For the proof of the second assertion regarding the support points of the D-optimal design, let $\xi^* = \sum_{i=1}^k \lambda_i \delta_{x_i}$ denote the D-optimal design. Then it is easy to see that

$$0 = F_\Phi(\xi^*, \xi^*) = \sum_{i=1}^k \lambda_i\, F_\Phi(\xi^*, \delta_{x_i}) \leq 0\ ,$$

where the last inequality follows from the inequality (1.22). But this implies

$$F_\Phi(\xi^*, \delta_{x_i}) = 0 \quad \forall\ x_i\ ,$$

which is equivalent to the equation

$$f^T(x_i) M^{-1}(\xi^*) f(x_i) = d+1$$

for all support points x_i of the D-optimal design ξ^*. □

It is worthwhile to mention that the characterization of the optimal design given in the previous theorem does neither depend on the particular regression model (here the polynomials) nor on the specific optimality criterion (here the- D-optimality criterion). All what is required is Fréchet differentiablity of the (concave) function Φ and a few regularity assumptions on the regression model. For more general characterizations avoiding differentiability assumptions for the optimality criterion we refer to the monograph of Pukelsheim (1993). We will now illustrate the application of Theorem 1.8 in the quadratic regression model.

Example 1.9. Consider the quadratic regression model ($d = 2$) and the two designs

$$\xi^* = \begin{pmatrix} -1 & 0 & 1 \\ 1/3 & 1/3 & 1/3 \end{pmatrix}, \quad \xi^{**} = \begin{pmatrix} -1 & 0 & 1 \\ 1/4 & 1/2 & 1/4 \end{pmatrix}. \tag{1.23}$$

The corresponding moment matrices are given by

$$M(\xi^*) = \begin{pmatrix} 1 & 0 & 2/3 \\ 0 & 2/3 & 0 \\ 2/3 & 0 & 2/3 \end{pmatrix}, \quad M(\xi^{**}) = \begin{pmatrix} 1 & 0 & 1/2 \\ 0 & 1/2 & 0 \\ 1/2 & 0 & 1/2 \end{pmatrix};$$

and the functions $d(\cdot, \cdot)$ are easily calculated as

$$d(x, \xi) = \tfrac{3}{2}(2 - 3x^2 + 3x^4) \ ; \qquad d(x, \xi^{**}) = 2 - 2x^2 + 4x^4$$

These functions are depicted in Figure 1. By Theorem 1.8 the design ξ^* is in fact D-optimal and the design ξ^{**} is not D-optimal (we will show later that the design ξ^{**} is in fact the D_1-optimal design, see Example 3.2 below).

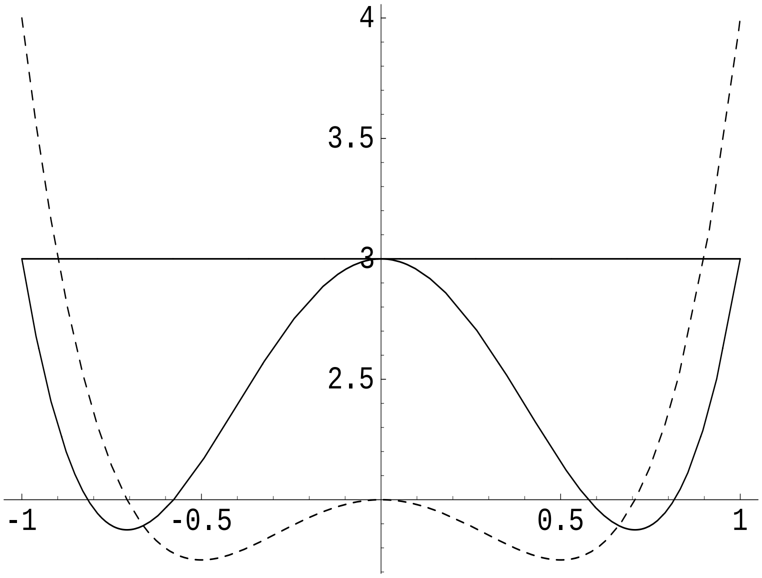

Figure 1: *The function $d(x, \xi)$ defined in (1.20) in the quadratic regression model for the designs ξ^* and ξ^{**} given in (1.23). The design is D-optimal if and only if the curve stays below the line $y \equiv 3$. Solid line: design ξ^*; dashed line: design ξ^{**}.*

We now have all tools for determining the approximate D-optimal design for the general polynomial regression model of degree d. Recall that for the general model the function d in Theorem 1.8 is given by

$$d(x, \xi) = (1, \ldots, x^d) M^{-1}(\xi) (1, \ldots, x^d)^T \ ,$$

which is obviously a polynomial of degree $2d$. By Theorem 1.8 a design ξ^* is D-optimal for the polynomial regression model of degree d on the interval $[-1,1]$ if and only if

- $d(x,\xi^*) \leq d+1 \quad \forall\ x \in [-1,1]$
- $d(x,\xi^*) = d+1 \quad \forall\ x \in \text{ supp}(\xi^*)$

Moreover, the matrix $M(\xi^*)$ is positive definite and this implies that the leading coefficient of the polynomial $d(x,\xi^*)$ is also positive. Now a careful counting of the zeros with corresponding multiplicities shows that the D-optimal design has at most $d+1$ support points and if its support has $d+1$ points it must contain the extreme points of the design space, i.e. -1 and 1. On the other hand we need at least $d+1$ support points in order to have a nonsingular matrix $M(\xi^*)$, which implies

$$\begin{aligned} \#\,\text{supp}(\xi^*) &= d+1 \\ \{-1,1\} &\subset \text{supp}(\xi^*). \end{aligned}$$

Now let

$$\xi^* = \begin{pmatrix} x_0 & \dots & x_d \\ w_0 & \dots & w_d \end{pmatrix}; x_0 = -1; x_d = 1$$

denote the D-optimal design and observe the representation

$$\begin{aligned} M(\xi^*) &= \int_{-1}^{1} f(x) f^T(x) d\xi^*(x) \\ &= \begin{bmatrix} 1 & \sum_i w_i x_i & \dots & \sum_i w_i x_i^d \\ \sum_i w_i x_i & \sum_i w_i x_i^2 & \dots & \sum_i w_i x_i^{d+1} \\ \vdots & \vdots & \ddots & \vdots \\ \sum_i w_i x_i^d & \sum_i w_i x_i^{d+1} & \dots & \sum_i w_i x_i^{2d} \end{bmatrix} = X^T W X, \end{aligned}$$

where the matrices $X \in \mathbb{R}^{(d+1)\times(d+1)}$ and $W \in \mathbb{R}^{(d+1)\times(d+1)}$ are defined by

$$X^T = \begin{bmatrix} 1 & \dots & 1 \\ x_0 & \dots & x_d \\ \vdots & \ddots & \vdots \\ x_0^d & \dots & x_d^d \end{bmatrix}, \quad W = \begin{bmatrix} w_0 & & & \\ & w_1 & & \\ & & \ddots & \\ & & & w_d \end{bmatrix}$$

(all other entries in the matrix W are 0). A straightforward calculation shows that

$$|M(\xi^*)| = |X|^2 |W| = \prod_{i=0}^{d} w_i \prod_{0 \leq i < j \leq d} (x_i - x_j)^2, \tag{1.24}$$

and this expression has to be maximized with respect to the choice of the weights $w_0, \dots, w_n \in (0,1)$ (subject to the constraint $\sum_{j=0}^{d} w_j = 1$) and the support points $-1 = x_0 < x_1 < \dots x_{d-1} <$

$x_d = 1$. A straightforward optimization with respect to the weights w_i shows that these have to be all equal, i.e.

$$w_i = \frac{1}{d+1} \qquad i = 0, 1, \dots, d.$$

The determination of the optimal support points $x_1, \dots, x_{d-1}$ is more complicated but can be performed following the arguments given in Szegö (1959). Taking partial derivatives of the logarithm we obtain from (1.24) the system of equations

$$\begin{aligned} 0 &= \frac{\partial}{\partial x_k} \log \prod_{i=1}^{d-1} (1 - x_i^2)^2 \prod_{1 \le i < j \le d-1} (x_i - x_j)^2 \\ &= \sum_{i \ne k} \frac{2}{x_k - x_i} - \frac{4x_k}{1 - x_k^2} \qquad k = 1, \dots, d-1. \end{aligned} \tag{1.25}$$

Let $f(x) = \prod_{j=1}^{d-1}(x - x_j)$ denote the polynomial of degree $d - 1$, which vanishes precisely at the points $x_1, \dots, x_{d-1}$ then it is easy to see that (1.25) gives

$$0 = \frac{f''(x_k)}{f'(x_k)} - \frac{4x_k}{1 - x_k^2} \qquad k = 1, \dots, d-1 \ .$$

Because f is a polynomial of degree $d - 1$ these equations provide a differential equation for the polynomial f, that is

$$(1 - x^2) f''(x) - 4x f'(x) + (d-1)(d+2) f(x) = 0,$$

where the factor of $f(x)$ is obtained by comparing leading coefficients. It is well known that the only polynomial solution of this equation is given by the Jacobi polynomial

$$f(x) = P_{d-1}^{(1,1)}(x)$$

which is proportional to the derivative of the Legendre polynomial $P_d'(x)$ [see Szegő (1959)]. We summarize these results in the following Theorem.

Theorem 1.10. *The D-optimal design for the polynomial regression model of degree d on the interval* $[-1, 1]$ *has equal masses at the roots of the polynomial*

$$(1 - x^2) P_d'(x),$$

where P_d *denotes the dth Legendre polynomial orthogonal with respect to the Lebesgue measure on the interval* $[-1, 1]$.

2 Canonical moments: simple properties and first applications

In the previous section we used the equivalence theorem to reduce an infinite dimensional optimization problem to a finite one. The solution of the resulting optimization problem could be characterized by a differential equation for a polynomial, which had the support points of the D-optimal design as its roots. In this section we will use a more direct approach, which does not require the reduction of the optimization to a finite dimensional maximization problem. Note that the information matrix in the polynomial regression model of degree d is given by the *Hankel matrix*

$$M(\xi) = \begin{bmatrix} c_0 & c_1 & \cdots & c_d \\ c_1 & c_2 & \cdots & c_{d+1} \\ \vdots & \vdots & & \vdots \\ c_d & c_{d+1} & \cdots & c_{2d} \end{bmatrix},$$

where $c_i = \int_{\mathcal{X}} x^i d\xi(x)$ is the ith moment of the design ξ. Moreover, because the interval under consideration is compact, any design is determined by its moments. Thus formally any function of the matrix $M(\xi)$ defined on the set of all approximate designs Ξ can be written as a function on the *moment space*

$$\mathcal{M}_{2d} = \Big\{(c_1, \ldots, c_{2d}) \mid c_i = \int_{\mathcal{X}} x^i d\xi(x) \ , i = 1, \ldots, 2d; \quad \xi \in \Xi\Big\};$$

that is

$$\Phi(M(\xi)) = \tilde{\Phi}(c_1, \ldots, c_{2d}) \ , \tag{2.1}$$

for an appropriate function $\tilde{\Phi} : \mathcal{M}_{2d} \to \mathbb{R}$. In other words, the determination of the D-optimal design corresponds to a constrained non-linear $2d$-dimensional maximization problem.

2.1 Canonical moments

The problem of characterizing the moment points in the moment space $\mathcal{M}_{2d}$ is known as the *Hausdorff moment problem* and can be solved by using the *Hankel determinants*

$$\underline{H}_{2m} = \begin{vmatrix} c_0 & \cdots & c_m \\ \vdots & & \vdots \\ c_m & \cdots & c_{2m} \end{vmatrix} \quad \overline{H}_{2m+1} = \begin{vmatrix} c_0 - c_1 & \cdots & c_m - c_{m+1} \\ \vdots & & \vdots \\ c_m - c_{m+1} & \cdots & c_{2m} - c_{2m+1} \end{vmatrix}$$
$$\underline{H}_{2m+1} = \begin{vmatrix} c_0 + c_1 & \cdots & c_m + c_{m+1} \\ \vdots & & \vdots \\ c_m + c_{m+1} & \cdots & c_{2m} + c_{2m+1} \end{vmatrix} \quad \overline{H}_{2m} = \begin{vmatrix} c_0 - c_2 & \cdots & c_{m-1} - c_{m+1} \\ \vdots & & \vdots \\ c_{m-1} - c_{m+1} & \cdots & c_{2m-2} - c_{2m} \end{vmatrix} \tag{2.2}$$

$(m = 0, \ldots, d)$. We will use the following characterization to define a one to one mapping from the moment space $\mathcal{M}_{2d}$ onto the unit cube $[0, 1]^{2d}$. For a proof of the following theorem we refer to Shohat and Tamarkin (1943) or Dette and Studden (1997).

Theorem 2.1.

(i) $(c_1, \ldots c_n) \in \mathcal{M}_n$ *if and only if* $\underline{H}_i$ *and* $\overline{H}_i$ *are nonnegative for* $i = 1, \ldots, n$.

(ii) $(c_1, \ldots c_n) \in Int\mathcal{M}_n$ *if and only if* $\underline{H}_i$ *and* $\overline{H}_i$ *are positive for* $i = 1, \ldots, n$.

For each sequence of moments $c = (c_1, c_2, \ldots)$ let

$$N = N(c) = \min\{n \in \mathbb{N} \mid (c_1 \ldots c_n) \in \partial M_n\} \tag{2.3}$$

denote the minimum integer such that $(c_1, \ldots, c_N)$ is on the boundary of the Nth moment space $\mathcal{M}_N$. If $(c_1, \ldots, c_n) \in \mathrm{Int}\mathcal{M}_n$ for all $n \geq 1$, define $N(c) = \infty$ while $N(c) = 1$ if $c_1 \in \partial\mathcal{M}_1$. Thus $(c_1, \ldots, c_k) \in \mathrm{Int}\mathcal{M}_k$ for $k < N(c)$ and $(c_1, \ldots, c_N) \in \partial\mathcal{M}_N$ and, of course, $(c_1, \ldots c_k) \in \partial\mathcal{M}_k$ for $k \geq N + 1$ (see the previous theorem). For a given sequence of moments $c = (c_0, c_1, c_2, \ldots)$ of a probability measure μ on the interval $[-1, 1]$ we now define for each $n \in \mathbb{N}_0$

$$\begin{aligned} c_{n+1}^{+} &= \max\Big\{ \int_{-1}^{1} x^{n+1} d\eta(x) \,\Big|\, \eta \text{ prob. measure with } \ c_j = \int_{-1}^{1} x^j d\eta(x) \ \forall j = 1, \ldots, n \Big\} \\ c_{n+1}^{+} &= \min\Big\{ \int_{-1}^{1} x^{n+1} d\eta(x) \,\Big|\, \eta \text{ prob. measure with } \ c_j = \int_{-1}^{1} x^j d\eta(x) \ \forall j = 1, \ldots, n \Big\} \end{aligned} \tag{2.4}$$

as the maximum and minimum of the $(n+1)$th moment over the set of all probability measures η whose moments up to the order n coincide with the given moments $(c_1, \ldots, c_n)$. The *canonical moment sequence* is then defined for $k \leq N(c)$ by

$$p_k \ = \ p_k(c) \ = \ \frac{c_k - c_k^-}{c_k^+ - c_k^-}, \tag{2.5}$$

where c_k^- and c_k^+ are defined in (2.4). Note that the canonical moments vary in the interval $[0, 1]$. Moreover, if $N = N(c) < \infty$, then $p_j \in (0, 1)$, $1 \leq j < N$ and p_N is either 0 or 1. It is easy to see that this mapping is one to one [see Dette and Studden (1997)] and consequently any probability measure on the interval $[0, 1]$ is uniquely determined by its canonical moments. We finally mention that canonical moments were introduced in a series of papers by Skibinsky (1967, 1968, 1969, 1976, 1986) and are also implicitly mentioned in the work of Karlin and Shapeley (1953).

Example 2.2. We briefly discuss the calculation of the first two canonical moments. For the first canonical moment we observe that $c_1^+ = 1$, $c_1^- = -1$ and obtain by definition (2.5) that

$$p_1 = \frac{c_1 + 1}{2}.$$

The calculation of the second canonical moment is slightly more complicated. Note that $c_1 \in \mathrm{Int}\mathcal{M}_1$ if and only if $c_1 \in (-1, 1)$. Because the variance of a random variable is nonnegative and the second moment is bounded by 1 we have $c_2^+ = 1$ and $c_2^- = c_1^2$, which gives for $c_1 \in (0, 1)$

$$p_2 = \frac{c_2 - c_1^2}{1 - c_1^2} \ .$$

Conversely, we can express the second moment c_2 in terms of the first two canonical moments p_1, p_2 and obtain $c_1 = 2p_1 - 1$

$$c_2 = 4p_1q_1p_2 + (2p_1 - 1)^2, \tag{2.6}$$

where $q_1 = 1 - p_1$.

2.2 Simple properties

The above definition provides a one-to-one map, say T, from the moment space

$$\mathcal{M} = \left\{(c_1, c_2 \ldots) \mid c_i = \int_{\mathcal{X}} x^i d\xi(x) \ , i = 1, 2, \ldots; \quad \xi \in \Xi\right\}$$

onto a set $\mathcal{S}$ defined by

$$\mathcal{S} \ = \ \Big(\bigcup_{k=0}^{\infty} \mathcal{S}_k\Big) \bigcup \mathcal{S}_\infty, \tag{2.7}$$

where

$$\mathcal{S}_\infty = \{(p_1, p_2, \ldots) | 0 < p_i < 1, \text{ for all } i \geq 1\}$$

and for $k \geq 0$

$$\mathcal{S}_k = \{(p_1, \ldots, p_k, p_{k+1}) \mid 0 < p_i < 1, \ \ 1 \leq i \leq k, \ \ p_{k+1} = 0 \text{ or } 1\} \ .$$

Any $c \in \text{Int}\mathcal{M}$ corresponds to some point in $\mathcal{S}_\infty$, while for any sequence of canonical moments $p = (p_1, p_2, \ldots) \in \mathcal{S}_\infty$ the corresponding sequence $c = (c_1, c_2, \ldots)$ can be defined successively from (2.5). Therefore it is evident that T maps Int$M$ onto $\mathcal{S}_\infty$ in a one-to-one manner. Similarly, each $(p_1, \ldots, p_n)$ is uniquely determined by $(c_1, c_2, \ldots, c_n)$. In the following we list a few interesting properties of canonical moments. For a proof see Dette and Studden (1997).

Simple properties 2.3.

- *the canonical are invariant under a linear transformation of the corresponding measure and interval.*
- *the design ξ is symmetric if and only if $p_{2k-1} = 1/2$ for $k \geq 1$ and $2k - 1 \leq N(c)$.*
- *$p_k = 0$ if and only if $c_k = c_k^-$; $p_k = 1$ if and only if $c_k = c_k^+$; in both cases $(c_1, \ldots, c_j) \in \partial\mathcal{M}_j$ for all $j \geq k$ and the corresponding design ξ has finite support.*
- *$p_{2d} = 1$ if and only if #supp$(\xi) = d + 1$ and $\{-1, 1\} \subset$ supp(ξ)*
- *$p_{2d+2} = 0$ if and only if #supp$(\xi) = d + 1$ and supp$(\xi) \subset (-1, 1)$*
- *$p_{2d+1} = 1$ if and only if #supp$(\xi) = d + 1$ and $1 \in$ supp(ξ), $-1 \notin$ supp(ξ)*
- *$p_{2d+1} = 0$ if and only if #supp$(\xi) = d + 1$ and $-1 \in$ supp(ξ), $1 \notin$ supp(ξ)*

In general the problem of calculating the canonical moments of a given measure ξ is very complicated if definition (2.5) is used directly. As more efficient method we present a representation of the canonical moments in terms of Hankel determinants.

Theorem 2.4. *For all $1 \leq n \leq N(c)$*

$$p_n = \frac{\underline{H}_n \overline{H}_{n-2}}{\underline{H}_{n-1} \overline{H}_{n-1}} \alpha_n, \quad q_n = 1 - p_n = \frac{\underline{H}_{n-2} \overline{H}_n}{\underline{H}_{n-1} \overline{H}_{n-1}} \alpha_n \ ,$$

where $\alpha_n = 1$ if n is even and $\alpha = \frac{1}{2}$ if n is odd.

Proof. We consider only the case $n = 2d$ even and the representation of p_{2d}. We show

$$c_{2d} - c_{2d}^- \ = \ \underline{H}_{2d} / \underline{H}_{2d-2}, \quad c_{2d}^+ - c_{2d} \ = \ \overline{H}_{2d} / \overline{H}_{2d-2} \ , \tag{2.8}$$

which implies by the definiton of p_{2d}

$$p_{2d} \ = \ \frac{\underline{H}_{2d} \overline{H}_{2d-2}}{\underline{H}_{2d} \overline{H}_{2d-2} + \overline{H}_{2d} \underline{H}_{2d-2}}, \tag{2.9}$$

where $\underline{H}_{-1} = \overline{H}_{-1} = \underline{H}_0 = \overline{H}_0 = 1$. The assertion now follows from the identity

$$\underline{H}_{2d-1} \overline{H}_{2d-1} = \underline{H}_{2d-2} \overline{H}_{2d} + \overline{H}_{2d-2} \underline{H}_{2d} \ ,$$

where the proof of this identity is complicated and can be found in Dette and Studden (1997). To obtain the expression for $c_{2d} - c_{2d}^-$ in (2.8) we note again that $\underline{H}_{2d}$ would be zero if we replace c_{2d} by c_{2d}^- (note that $(c_1, \ldots, c_{2d-1}, c_{2d}^-) \in \partial \mathcal{M}_{2d}$ which implies $\underline{H}_{2d} = 0$, by Theorem 2.1). Then writing $c_{2d} = c_{2d}^- + (c_{2d} - c_{2d}^-)$ for the last element in the determinant $\underline{H}_{2d}$ gives $\underline{H}_{2d} = (c_{2d} - c_{2d}^-) \underline{H}_{2d-2}$. The value of $c_{2d}^+ - c_{2d}$ in (2.8) is verified in a similar manner. □

Example 2.5. Let $\xi_{\alpha\beta}$ denote the Beta distribution on the interval $(0,1)$ with density

$$w_{(\alpha,\beta)}(x) \ = \ \frac{1}{B(\beta+1, \alpha+1)} x^{\beta} (1-x)^{\alpha} \qquad 0 < x < 1 \tag{2.10}$$

where $\alpha, \beta > -1$ and

$$B(p,q) \ = \ \int_0^1 x^{p-1} (1-x)^{q-1} dx \ = \ \frac{\Gamma(p)\Gamma(q)}{\Gamma(p+q)} \qquad (p, q > 0) \tag{2.11}$$

denotes the *Beta-integral* and $\Gamma(\cdot)$ the *Gamma function* (see Johnson and Kotz, 1970). The ordinary moment of $\xi_{\alpha\beta}$ are

$$c_j \ = \ \frac{B(\beta+1+j, \alpha+1)}{B(\beta+1, \alpha+1)} \ = \ \frac{\Gamma(\beta+j+1)}{\Gamma(\beta+1)} \frac{\Gamma(\alpha+\beta+2)}{\Gamma(\alpha+\beta+2+j)} \qquad j \geq 1.$$

The canonical moments of the Beta distribution on the interval $[0,1]$ were first determined by Skibinsky (1969). This author showed that the canonical moments of the measure $\xi_{\alpha\beta}$ are given by

$$p_{2j} \ = \ \frac{j}{2j+1+\alpha+\beta} \qquad p_{2j-1} \ = \ \frac{\beta+j}{2j+\alpha+\beta} \qquad j \geq 1 \ . \tag{2.12}$$

An alternative approach using the hypergeometric series of Gauss (1813) can be found in Dette and Studden (1997). Because canonical moments are invariant with respect to linear transformations of the underlying interval, the Beta distribution on the interval $[-1,1]$ has the same canonical moments.

Note that $p_{2j-1} = 1/2$ if and only if $\alpha = \beta$ which means that $\xi_{\alpha\beta}$ is symmetric with respect to the midpoint $x = 1/2$. Two special cases should be mentioned. For $\alpha = \beta = 0$, $\xi_{\alpha\beta}$ is the uniform distribution on the interval $[0,1]$ and the canonical moments are given by

$$p_{2k} \ = \ \frac{k}{2k+1}, \ \ p_{2k-1} \ = \ \frac{1}{2} \qquad k \geq 1 \ .$$

For $\alpha = \beta = -1/2$, $\xi_{\alpha\beta}$ gives the *arc-sine distribution* with canonical moments $p_k \equiv 1/2$ for all k. This indicates that the arc-sine distribution with density

$$w_{(-1/2,-1/2)}(x) \ = \ \frac{1}{\pi\sqrt{x(1-x)}} \qquad 0 < x < 1$$

has moments in the *center* of the moment space. We mention once again that the uniform and arc-sine distribution on the interval $[-1,1]$ have the same canonical moments as the corresponding measures on the interval $[0,1]$.

Example 2.6. The Binomial distribution ξ_B is given by the mass distribution

$$b(x;N,p) \ = \ \binom{N}{x} p^x(1-p)^{N-x} \qquad x = 0,1,\ldots,N$$

where $p \in (0,1)$ and $N \in \mathbb{N}$. The ordinary moments of ξ_B are somewhat complicated and are given by

$$c_r \ = \ \sum_{x=0}^{N} \binom{N}{x} p^x(1-p)^{N-x}x^r \ = \ N! \sum_{j=0}^{r} \frac{S(r,j)}{(N-j)!} p^j \qquad (r \geq 1)$$

where $S(r,j)$ denote the *Stirling numbers* of the second kind defined by

$$S(r,j) \ = \ \frac{1}{j!} \sum_{k=0}^{j} (-1)^{j-k} \binom{j}{k} k^r \quad (j \leq r)$$

(see Johnson, Kotz and Kemp, 1992). The canonical moments of the Binomial distribution have a much simpler form and were obtained by Skibinsky (1969) as

$$p_{2j-1} = p, \qquad p_{2j} = \frac{j}{N} \qquad j = 1,2,\ldots,N \ .$$

They are calculated with reference to the interval $[0, N]$ or alternatively on $[0, 1]$ by moving the mass $b(x; N, p)$ to x/N for $x = 0, 1, \ldots, N$. Note that the sequence of canonical moments terminates at $p_{2N} = 1$, which reflects the fact that the Binomial distribution is supported on a finite number of points.

We finally note that the canonical moments also appear in the sequence of orthogonal polynomials with respect to the measures $d\xi(x)$, $(1+x)d\xi(x)$, $(1-x)d\xi(x)$ and $(1-x^2)d\xi(x)$ [see Dette and Studden (1997)].

Theorem 2.7.
(i) *The monic orthogonal polynomials on the interval* $[-1, 1]$ *with respect to the measure* $d\xi(x)$ *satisfy the following recursion formula* $\underline{R}_{-1}(y) = 0$, $\underline{R}_0(y) = 1)$

$$\begin{aligned} \underline{R}_1(y) &= y + 1 - 2p_1 \ , \\ \underline{R}_{m+1}(y) &= (y + 1 - 2q_{m-1}p_{2m} - 2q_{2m}p_{2m+1})\underline{R}_m(y) \\ &\qquad -4q_{2m-2}p_{2m-1}q_{2m-1}p_{2m}\underline{R}_{m-1}(y) \qquad m \geq 1 \end{aligned}$$

(ii) *The monic orthogonal polynomials on the interval* $[-1, 1]$ *with respect to the measure* $(1-x^2)d\xi(x)$ *satisfy the following recursion formula* $\overline{S}_{-1}(y) = 0$, $\overline{S}_0(y) = 1$,

$$\begin{aligned} \overline{S}_{m+1}(y) &= (y + 1 - 2p_{2m+1}q_{2m+2} - 2p_{2m+2}q_{2m+3})\overline{S}_m(y) \\ &\qquad -4p_{2m}q_{2m+1}p_{2m+1}q_{2m+2}\overline{S}_{m-1}(y) \qquad m \geq 0 \end{aligned}$$

Note that if the measure ξ is symmetric about zero then we have $p_{2i+1} = 1/2$, $i \geq 0$. The polynomials $\underline{R}_m$ and $\overline{S}_m$ orthogonal with respect to the measures $d\xi$ and $(1-y^2)d\xi$ are even or odd functions according as m is even or odd. The corresponding recursion equations are particularly simple and given by

$$\begin{aligned} \underline{R}_0(y) &= 1, \ \underline{R}_1(y) = y, \\ \underline{R}_{m+1}(y) &= y\underline{R}_m(y) - q_{2m-2}p_{2m}\underline{R}_{m-1}(y) \qquad m \geq 1 \end{aligned}$$

$$\begin{aligned} \overline{S}_0(y) &= 1, \ \overline{S}_1(y) = y, \\ \overline{S}_{m+1}(y) &= y\overline{S}_m(y) - p_{2m}q_{2m+2}\overline{S}_{m-1}(y) \qquad m \geq 1 \end{aligned}$$

2.3 Canonical moments and D-optimal approximate designs

The following result shows that the function $\tilde{\Phi}$ in (2.1) has a surprisingly simple representation in terms of canonical moments. This transfers the constrained optimization problem on the space $\mathcal{M}_{2d}$ to an elementary maximization problem on the unit cube $[0, 1]^{2d}$.

Theorem 2.8. *If ξ is a design on the interval $[-1,1]$ with canonical moments $p_1, p_2, \ldots$, then*

$$|M_{2d}(\xi)| = \underline{H}_{2d} = \begin{vmatrix} c_0 & c_1 & \cdots & c_d \\ c_1 & c_2 & \cdots & c_{d+1} \\ \vdots & \vdots & & \vdots \\ c_d & c_{d+1} & \cdots & c_{2d} \end{vmatrix} = 2^{d(d+1)} \prod_{j=1}^{d} (q_{2j-2} p_{2j-1} q_{2j-1} p_{2j})^{d-j+1}$$

$(q_j = 1 - p_j; q_0 = 1)$.

Proof. The assertion follows from Theorem 2.4, which shows that

$$\begin{aligned} \frac{\underline{H}_{2d}}{\underline{H}_{2d-2}} &= p_{2d} \frac{\underline{H}_{2d-1} \overline{H}_{2d-1}}{\underline{H}_{2d-2}} \cdot \frac{\overline{H}_{2d-3}}{\overline{H}_{2d-2} \overline{H}_{2d-3}} \\ &= 2 q_{2d-1} p_{2d} \frac{\underline{H}_{2d-1}}{\underline{H}_{2d-3}} = 4 q_{2d-2} p_{2d-1} q_{2d-1} p_{2d} \frac{\underline{H}_{2d-2}}{\underline{H}_{2d-4}} \\ &= \ldots = 2^{2d} \prod_{j=1}^{d} q_{2j-2} p_{2j-1} q_{2j-1} p_{2j} . \end{aligned}$$

Observing $\underline{H}_0 = 1$ and

$$\underline{H}_{2d} = \prod_{j=1}^{d} \frac{\underline{H}_{2j}}{\underline{H}_{2j-2}}$$

we obtain the assertion of Theorem 2.8 by a straightforward calculation. □

The maximization of the determinant of $M(\xi)$ in terms of canonical moments is now straightforward. Observing that the canoncial moments vary independently in the interval $[0,1]$, we obtain from Theorem 2.8 the following corollary by a direct calculation.

Corollary 2.9. *The D-optimal design for the polynomial regression model of degree d has canonical moments*

$$p_{2j-1} = \frac{1}{2}, \quad p_{2j} = \frac{d-j+1}{2(d-j)+1}, \quad j = 1, 2, \ldots, d \ . \tag{2.13}$$

Note that in principal Corollary 2.9 solves the D-optimal design problem by characterizing the optimal design in terms of its canonical moments. However, for applications it is necessary to know the support points and weights corresponding to the measure determined by the sequence (2.13), because these give the locations where the observations are taken and the proportions of the total observations to be taken at these locations. For the determination of these quantities we consider the *Stieltjes transform* of the measure ξ and its corresponding power series and continued fraction expansion, i.e.

$$\int_{-1}^{1} \frac{d\mu(x)}{z-x} = \sum_{j=0}^{\infty} \frac{c_j}{z^{j+1}} = \frac{1\,|}{|\,z+1} - \frac{2p_1\,|}{|\,1} - \frac{2q_1p_2\,|}{|\,z+1} - \frac{2q_2p_3\,|}{|\,1} - \ \ldots \tag{2.14}$$

The first equality follows from the theorem of dominated convergence and a series expansion of the integrand on the left hand side. The second expression can be derived by the correspondence between power series and continued fractions [see Wall (1948) or Perron (1954a,b)] and some generalizations of Theorem 2.8 for the determinants $\underline{H}_{2d-1}$, $\overline{H}_{2d-1}$ and $\overline{H}_{2d}$. An alternative proof can be found in Dette and Studden (1997). The continued fraction in (2.14) converges uniformly on compact sets $K \subset \mathbb{C}$ with positive distance from the interval $[-1,1]$. However, in the case $p_{2d} = 1$ which is of interest here, the Stieltjes transform is in fact a rational function

$$\begin{aligned} H(z) &= \int_{-1}^{1} \frac{d\mu(x)}{z-x} \\ &= \frac{1\,|}{|\,z+1} - \frac{2p_1\,|}{|\,1} - \frac{2q_1p_2\,|}{|\,z+1} - \frac{2q_2p_3\,|}{|\,1} - \ \ldots \ - \frac{2q_{2d-1}p_{2d}\,|}{|\,z+1} \\ &= \frac{1\,|}{|\,z+1} - \frac{1|}{|1} - \frac{p_2\,|}{|\,z+1} - \frac{q_2|}{|1} - \ \ldots \ - \frac{2q_{2d-2}\,|}{|\,1} - \frac{p_{2d}\,|}{|\,z+1} \\ &= \frac{A_d(z)}{B_{d+1}(z)} \ , \end{aligned} \tag{2.15}$$

where the second equality is derived under the assumption of symmetry (i.e. $p_{2i-1} = \frac{1}{2}$ for all $i = 1, \ldots, d$) and A_d and B_{d+1} are polynomials of degree d and $d+1$, respectively. Consequently we obtain for the support and the weights of the D-optimal design

$$\begin{aligned} \operatorname{supp}(\xi) &= \{\ z\ \in \mathbb{C} \mid B_{d+1}(z) = 0\ \}, \\ \xi(x) &= \lim_{z \to x} H(z)(z-x) = \frac{A_d(x)}{B'_{d+1}(x)} \quad \forall x \in \operatorname{supp}(\xi), \end{aligned} \tag{2.16}$$

and all what remains is the calculation of the polynomials A_d and B_{d+1}. For that purpose we use the represenation of the partial numerators and denominators of a continued fraction

$$b_0 + \frac{a_1\,|}{|\,b_1} + \frac{a_2\,|}{|\,b_2} + \ \ldots + \frac{a_n\,|}{|\,b_n} = \frac{A_n}{B_n} \tag{2.17}$$

in terms of *continuants* defined by

$$A_n = K\begin{pmatrix} & a_1 & \ldots & & a_n \\ b_0 & b_1 & \ldots & & b_n \end{pmatrix} := \begin{vmatrix} b_0 & -1 & 0 & \cdots & 0 & 0 \\ a_1 & b_1 & -1 & \cdots & 0 & 0 \\ 0 & a_2 & b_2 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & b_{n-1} & -1 \\ 0 & 0 & 0 & \cdots & a_n & b_n \end{vmatrix}$$

$$B_n \;=\; K\left(\begin{array}{cccccc} & a_2 & & \dots & a_n & \\ b_1 & & b_2 & \dots & & b_n \end{array}\right) \;:=\; \left|\begin{array}{cccccc} b_1 & -1 & 0 & \cdots & 0 & 0 \\ a_2 & b_2 & -1 & \cdots & 0 & 0 \\ 0 & a_3 & b_3 & \cdots & 0 & 0 \\ \vdots & \vdots & & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & b_{n-1} & -1 \\ 0 & 0 & 0 & \cdots & a_n & b_n \end{array}\right| .$$

For the polynomial B_{d+1} in the denominator of (2.15) we therefore obtain

$$\begin{aligned} B_{d+1} \;&=\; K\left(\begin{array}{ccccccccccc} & -1 & & -p_2 & & -q_2 & & \dots & & -q_{2d-2} & & -p_{2d} & \\ z+1 & & 1 & & z+1 & & 1 & & \dots & & 1 & & z+1 \end{array}\right) \\ &=\; (z^2-1)K\left(\begin{array}{ccccccccccc} & -q_2 & & -p_2 & & -q_4 & & \dots & & -q_{2d-2} & & -p_{2d-2} & \\ 1 & & z+1 & & 1 & & z+1 & & \dots & & 1 & & z+1 \end{array}\right) \\ &=\; (z^2-1)Q_{d-1}(z), \end{aligned}$$

where the last line defines the polynomial Q_{d-1} and the second identity follows by a tedious calcluation of the corresponding determinants and an induction argument [see Dette and Studden (1997), Section 2.5]. An expansion of the last determinant now give a recursive relation for the polynomals $Q_j(z)$, i.e. $Q_0(z) = 1$, $Q_1(z) = z$ and

$$\begin{aligned} Q_{j+1}(z) \;&=\; zQ_j(z) - q_{2d-2j}p_{2d-2j-2}Q_{j-1}(z) \\ &=\; zQ_j(z) - \frac{j(j+2)}{(2j+1)(2j+3)}Q_{j-1}(z) \;, \end{aligned}$$

where we used the representation in Corollary 2.9 for the canonical moments of the D-optimal design. Comparing this recursion with the recursive relation for the monic Jacobi polynomials $\hat{P}_j^{(1,1)}(z)$ [see Chihara (1978)]we obtain that

$$Q_{d-1}(z) \;=\; \hat{P}_{d-1}^{(1,1)}(z) \;.$$

A similar calculation shows for the polynomial A_d in the numerator

$$A_d(z) \;=\; \hat{P}_d^{(0,0)}(z) \;,$$

where $\hat{P}_d^{(0,0)}(z)$ is the monic version of the Jacobi polynomial $P_d^{(0,0)}(z)$ on the interval $[0,1]$ (in other words the monic version of the Legendre polynomial P_d). Observing that $\hat{P}_{d-1}^{(1,1)}(z)$ is proportional to the derivative of P_d we obtain for the support points of the D-optimal design

$$\text{supp}(\xi) \;=\; \{\; z \mid (z^2-1)P_d'(z) = 0 \;\}$$

and for the weights at the support points $x_0, \dots, x_d$

$$\begin{aligned} \xi(x_j) &= \frac{A_d(x)}{\frac{d}{dz}B'_{d+1}(z)|_{z=x_j}} = \frac{\hat{P}_d^{(0,0)}(x_j)}{(z^2-1)\hat{P}_{d-1}^{(1,1)}(z)|_{z=x_j}} \\ &= \frac{\hat{P}_d^{(0,0)}(x_j)}{(d+1)\hat{P}_d^{(0,0)}(z)|_{z=x_j}} = \frac{1}{d+1} \qquad j = 0, \dots, d \end{aligned}$$

where the third identity follows from a standard identity for the Jacobi polynomials [see e.g. Szegö (1959)]. Note that this provides an alternative proof of Theorem 1.10 based on the theory of canonical moments.

2.4 Asymptotic distribution of zeros of orthogonal polynomials

A further application of the theory of canonical moments consists in the maximization of generalized Hankel determinants, which are determinants of matrices of the form

$$M^{(\alpha,\beta)}(\xi) = \Big[\int_{-1}^{1} (1-x)^{\alpha+1}(1+x)^{\beta+1}x^{i+j}d\xi(x)\Big]_{i,j=0}^{d-1}, \tag{2.18}$$

where $\alpha, \beta > -1$ are given constants. In the statistical context this corresponds to the covariance matrix in a polynomial regression model with an heteroscedastic error structure, where the variance at the point x is proportional to $(1-x)^{-\alpha-1}(1+x)^{-\beta-1}$ [see Fedorov (1972)]. The following theorem was proved by Studden (1982) using an extension of the theory described in the previous section.

Theorem 2.10. *The canonical moments of the design maximizing the determinant of the matrix $M^{(\alpha,\beta)}(\xi)$ defined in (2.18) are given by*

$$\begin{aligned} p_{2j}^* &= \frac{d-j}{2(d-j)+1+\alpha+\beta} \qquad j = 1, \dots, d \\ p_{2j-1}^* &= \frac{\beta+d-j+1}{2(d-j)+2+\alpha+\beta} \qquad j = 1, \dots, d \end{aligned} \tag{2.19}$$

and the corresponding design ξ_d^* satisfies

$$\begin{aligned} supp(\xi_d^*) &= \{x \mid P_d^{(\alpha,\beta)}(x) = 0\} \\ \xi_d^*(x) &= \frac{1}{d} \quad \forall\, x \in supp(\xi^*) \end{aligned}$$

where $P_d^{(\alpha,\beta)}(x)$ denotes the dth Jacobi polynomial on the interval $[-1, 1]$.

Note that the solution of the maximization problem for the determinant of the matrix in (2.18) yields to a uniform distribution on the zeros of the dth Jacobi polynomial $P_d^{(\alpha,\beta)}(x)$. The asymptotic behviour of this distribution with increasing degree has been of some interest in approximation theory [see e.g. Van Assche (1987), Gawronski (1993), Bosbach and Gawronski (1998), Faldey and Gawronski (1995), Dette and Studden (1992, 1995), Dette (1995c), Kuijlaars and Van Assche (1999)]. In the next theorem we state a typical result in this area and present an elmentary proof based on the theory of canonical moments. For a motivation note that for $d \to \infty$ the canonical moments defined in (2.19) satisfy

$$\lim_{d\to\infty} p_j^* \;=\; \frac{1}{2} \;\;,\;\; \forall j \in \mathbb{N} \;.$$

By Example 2.5 the measure on the interval $[-1,1]$ corresponding to the limit sequence of canonical moments is the arc-sine distribution with densitiy

$$\frac{1}{\pi}\frac{1}{\sqrt{1-x^2}} I_{(-1,1)}(x).$$

Because this distribution is determined by its moments and the mapping between canonical and ordinary moments is one to one and continuous, it follows that the uniform distribution on the roots of the Jaobi polynomials converges weakly to the arc-sine distribution, that is

$$\lim_{d\to\infty} N_d^{(\alpha,\beta)}(x) \;:=\; \lim_{d\to\infty} \frac{1}{d}\Big\{ z \le x \mid P_d^{(\alpha,\beta)}(x) \Big\} \;=\; \frac{1}{\pi}\int_{-1}^{x} \frac{dt}{\sqrt{1-t^2}}$$

for all $x \in [-1,1]$. The following result generalizes this statement to Jacobi polynomials with parameters α_d, β_d depnding on the degree d of the polynomial.

Theorem 2.11. *Assume that*

$$d \to \infty \;,\quad \lim_{d\to\infty} \frac{\alpha_d}{d} \to a \;,\quad \lim_{d\to\infty} \frac{\beta_d}{d} \to b,$$

where $a, b \ge 0$, *then*

$$\begin{aligned} \lim_{d\to\infty} N_d^{(\alpha_d,\beta_d)}(x) \;&:=\; \lim_{d\to\infty} \frac{1}{d}\Big\{ z \le x \mid P_d^{(\alpha_d,\beta_d)}(z) = 0 \Big\} \\ &=\; \frac{2+a+b}{2\pi}\int_{r_1}^{x} \frac{\sqrt{(r_2-t)(t-r_1)}}{1-t^2} dt, \end{aligned}$$

where

$$r_{1,2} = \frac{b^2 - a^2 \pm 4\sqrt{(a+1)(b+1)(a+b+1)}}{(2+a+b)^2}$$

Proof. By the preceding discussion the canonical moments of the uniform distribution on the set

$$\{z \mid P_d^{(\alpha_d,\beta_d)}(z) = 0\}$$

satisfy

$$\begin{aligned}\lim_{d\to\infty} p^*_{2i} &= \lim_{d\to\infty} \frac{d-i}{2(d-i)+\alpha_d+\beta_d+1} = \frac{1}{2+a+b} = h \\ \lim_{d\to\infty} p^*_{2i-1} &= \lim_{d\to\infty} \frac{\beta_d+d-i+1}{2(d-i)+2+2_d+\beta_d} = \frac{b+1}{2+a+b} = g\end{aligned} \tag{2.20}$$

The identification of the corresponding measure is a little complicated. To be precise consider at first the measure $\xi^{g,h}$ on the interval $[-1,1]$ corresponding to the sequence

$$\begin{aligned} p_{4j-2} &= g \in (0,1) \quad j \in I\!N \\ p_{4j} &= h \in (0,1) \quad j \in I\!N \\ p_{2j-1} &= 1/2 \qquad\quad\; j \in I\!N \end{aligned} \tag{2.21}$$

The continued fraction expansion of the Stieltjes transform

$$S(z,\xi^{g,h}) = \int_{-1}^{1} \frac{d\xi_{g,h}(x)}{z-x}$$

is obtained by an even contraction [see Perron (1954a)] from (2.14) and is given by

$$\begin{aligned} S(z,\xi^{g,h}) &= \frac{1 \rvert}{\lvert z} - \frac{g \rvert}{\lvert z} - \frac{(1-g)h \rvert}{\lvert z} - \frac{(1-h)g \rvert}{\lvert z} - \frac{(1-g)h \rvert}{\lvert z} - \ \ldots \\ &= \frac{z \rvert}{\lvert z^2-g} - \frac{g(1-g)h \rvert}{\lvert z^2-\eta} - \frac{\mu \rvert}{\lvert z^2-\eta} - \frac{\mu \rvert}{\lvert z^2-\eta} - \ \ldots \end{aligned}$$

where second identity follows from a further even contraction [see Perron (1954a)] and the constants η and μ are defined by

$$\begin{aligned} \eta &= g(1-h)+h(1-g) \\ \mu &= g(1-g)h(1-h) \end{aligned} \tag{2.22}$$

We find that

$$\begin{aligned} H(z) &= \frac{1 \rvert}{\lvert z^2-\eta} - \frac{\mu \rvert}{\lvert z^2-\eta} - \frac{\mu \rvert}{\lvert z^2-\eta} - \ \ldots \\ &= \frac{1/2\sqrt{\mu} \rvert}{\lvert (z^2-\eta)/2\sqrt{\mu}} - \frac{1/4 \rvert}{\lvert (z^2-\eta)/2\sqrt{\mu}} - \frac{1/4 \rvert}{\lvert (z^2-\eta)/2\sqrt{\mu}} - \ \ldots \\ &= \frac{1}{\sqrt{\mu}} \left(\frac{z^2-\eta}{2\sqrt{\mu}} - \sqrt{\frac{(z^2-\eta)^2}{4\mu} - 1} \right), \end{aligned}$$

where the branch of the square root is defined by

$$\left| \frac{z^2-\eta}{2\sqrt{\mu}} - \sqrt{\frac{(z^2-\eta)^2}{4\mu} - 1} \right| < 1 \ . \tag{2.23}$$

Thus it follows that

$$\begin{aligned} S(z,\xi^{g,h}) &= \frac{z}{z^2-g-g(1-g)hH(z)} \\ &= \frac{1}{2h}\frac{(1-2h)z^2+(h-g)-\sqrt{(z^2-\eta)^2-4\mu}}{z(1-z^2)}, \end{aligned} \tag{2.24}$$

where the branch of the square root is defined by (2.23). For the identification of the corresponding probability measure we use the inversion formula for the Stieltjes transfom [see e.g. Dette and Studden (1997), Chapter 3].

Because $S(z,\xi^{g,h})$ can be extended from the lower half plane to a continuous function in a neighborhood of any $u_0 \in (-1,1)\setminus\{0\}$ it follows that the absolute continuous part of $\xi^{g,h}$ is given by

$$\frac{1}{2\pi h}\frac{\sqrt{4\mu-(x^2-\eta)^2}}{|x|(1-x^2)}I\left\{|x^2-\eta|<2\sqrt{\mu}\right\}\ , \tag{2.25}$$

where $\eta = g(1-h)+h(1-g)$, $\mu = g(1-g)h(1-h)$. Jumps of $\xi^{g,h}$ are only possible at the poles of $S(z,\xi^{g,h})$. We investigate the situation at $z=0$, the other cases are treated similarly. If $g=h$ it is straightforward to show that $S(z,\xi^{g,h})$ has in fact no pole at $z=0$ and consequently $\xi^{g,h}(\{0\})=0$ in this case. Observing the definition of η and μ in (2.22) we see that $h\neq g$ if and only if $\eta^2>4\mu$. For $z=-iv$ and sufficiently small v this determines the sign of the square root in (2.23) to satisfy $\mathcal{R}(\sqrt{\dots})<0$. Now Theorem 3.6.5 in Dette and Studden (1997) yields

$$\begin{aligned} \xi^{g,h}(\{0\}) &= \lim_{v\to 0}\operatorname{Im}\, S\{v\ S(-iv,\xi^{g,h})\} \\ &= \lim_{v\to 0}\operatorname{Im}\left\{\frac{(2h-1)v^2+(h-g)+\sqrt{(v^2+\eta)^2-4\mu}}{-i2h(1+v^2)}\right\} \\ &= \frac{h-g+|h-g|}{2h} \;=\; \begin{cases} (h-g)/h & \text{if} \quad h>g \\ 0 & \text{if} \quad h<g \end{cases} \end{aligned}$$

Similary, it can be shown that in the case $g+h>1$ the measure $\xi^{g,h}$ has additional masses $(g+h-1)/2$ at the points -1 and 1. Now note that for the specific choice in (2.20) we have $h\leq g$ and $g+h\leq 1$. Consequently, there is in fact only an absolute continuous part of $\xi_{g,h}$ given by (2.25). In other words, the measure $\xi^{g,h}$ corresponding to the sequence

$$\frac{1}{2},g,\frac{1}{2},h,\frac{1}{2},g,\frac{1}{2},\dots$$

on the interval $[-1,1]$ is absolute continuous with density given in (2.25). By Theorem 1.3.5 in Dette and Studden (1997) the measure $\tilde{\xi}^{g,h}$ corresponding to the sequence in (2.20) on the interval $[0,1]$ is related to $\xi^{g,h}$ by $\tilde{\xi}^{g,h}([0,x])=\xi^{g,h}([-\sqrt{x},\sqrt{x}])$. Therefore the density of $\tilde{\xi}^{g,h}$ is given by

$$h(x) = \frac{1}{2h\pi}\frac{\sqrt{4\mu-(x-\eta)^2}}{x(1-x)}I\{|x-\eta|<2\sqrt{\mu}\}\ . \tag{2.26}$$

Because all zeros of $P_n^{(\alpha_n,\beta_n)}(x)$ are located in the interval $(-1,1)$ the limit distribution ξ satisfies supp$(\xi)\subset (-1,1)$. Now ξ is induced through $\tilde{\xi}^{g,h}$ by the linear transformation $y = 2x - 1$ and the assertion follows by transforming the density in (2.26) onto the interval $[-1,1]$

$$\frac{d\xi}{dx} \;=\; \frac{1}{2}h(\frac{y+1}{2}) \;=\; \frac{1}{2h\pi}\frac{\sqrt{16\mu - (y+1-2\eta)^2}}{(1-y^2)} I\{|y+1-2\eta| < 4\sqrt{\mu}\}\ ,$$

and observing that $2\eta - 1 \pm 4\sqrt{\mu} = r_{1,2}$. □

3 Discrimination designs and extremal problems for polynomials

3.1 Discrimination designs

So far it has been assumed that the linear model (1.1) is known by the experimenter. As pointed out by Anderson (1962), Atkinson and Cox (1974) or Spruill (1990) there are many applications, where precise knowledge about the form of the regression function is not available and the analysis of the data is performed in two steps. In the first step the data is used to identify an appropriate regression model and the second step might consist of performing some statistical analysis in the determined model. For example, if a cubic regression model is assumed by the experimenter, the results of the experiments will typically be used to test whether a quadratic model would be more appropriate. In this case "good" designs have to address at least three different tasks: 1) the problem of testing the hypothesis $H_0 : \theta_3 = 0$ for the "highest" coefficient in the cubic model, 2) the problem of estimating the parameters in the full cubic polynomial if the test rejects the hypothesis H_0, 3) the problem of estimating the parameters in the reduced quadratic regression model if the test does not reject the hypothesis H_0. In this section we illustrate the application of canonical moments in this field and discuss further applications of our approach in approximation theory. We assume that the main interest of the experimenter is the identification of the degree of the underlying polynomial regression and an optimal design for this task has to be constructed. Optimal designs for this problem are called *optimal discrimination designs*. Because the degree of the polynomial regression is only known to be less or equal than d we have to use a further index in our notation, namely the degree $l \in \{1,\ldots,d\}$ for the polynomial model under consideration. To be precise let

$$h_l(x) \;=\; \sum_{i=1}^{l} \theta_{li}x^i \;=\; \theta_l^T f_l(x) \qquad l = 1,\ldots d$$

denote a polynomial regression model of degree l, where $f_l(x) = (1, x, \ldots, x^l)^T$ denotes the vector of monomials up to the order l and $\theta_l = (\theta_{l0},\ldots,\theta_{ll})^T \in \mathbb{R}^{l+1}$ is the vector of unknown parameters in the polynomial model of degree $l = 1,\ldots,d$. It can be shown [see Pukelsheim (1993) or Dette and Studden (1997)] that a "good" choice of a design for model discrimination should make the quantities

$$\delta_l^2(\xi) \;=\; (e_l^T M_l^{-1}(\xi)e_l)^{-1} \;=\; \frac{|M_l(\xi)|}{|M_{l-1}(\xi)|} \quad (l = 1,\ldots,d)$$

as large as possible, where $e_l = (0, \ldots, 0, 1)^T \in \mathbb{R}^{l+1}$ is the $(l+1)$th unit vector and

$$M_l(\xi) \;=\; \int_{-1}^{1} f_l(x) f_l(x)^T d\xi(x) \;=\; (c_{i+j})_{i,j=0}^{l} \tag{3.1}$$

denotes the moment matrix of the design ξ in the polynomial regression of degree l. As expected, a simultaneous maximization of these quantities is impossible and we have to restrict ourselves again to the maximization of real valued functions of these quantities [see Dette (1994, 1995c]. As a first function we consider the geometric mean of $\delta_1^2, \ldots, \delta_d^2$ defined by

$$\Psi^\beta(\xi) \;=\; \prod_{l=1}^{d} \left(\delta_l^2(\xi)\right)^{\beta_l} \;=\; \prod_{l=1}^{d} \left(\frac{|M_l(\xi)|}{|M_{l-1}(\xi)|} \right)^{\beta_l} \;, \tag{3.2}$$

where $\beta_1, \ldots, \beta_d$ are given nonnegative weights with $\sum_{j=1}^{d} \beta_j = 1$. A design ξ_β is called a *optimal discriminating design* with respect to the the *prior* $\beta = (\beta_1, \ldots, \beta_d)$ if and only if ξ_β maximizes the weighted geometric mean defined in (3.2). Note that the weight β_l reflects the experimenter's belief about the adequacy of the polynomial of degree l. The optimal design maximizing the function in (3.2) can be easily characterized in terms of its canonical moments.

Theorem 3.1. *The optimal discriminating design with respect to the prior* $\beta = (\beta_1, \ldots, \beta_d)$ $(\beta_d > 0)$ *is uniquely determined by its canonical moments*

$$\begin{aligned} p_{2i} &= \frac{\sigma_i}{\sigma_i + \sigma_{i+1}} \qquad i = 1, \ldots, d-1\;, \qquad p_{2d} = 1 \\ p_{2i-1} &= \frac{1}{2} \qquad\qquad\;\; i = 1, \ldots, d \end{aligned}$$

where $\sigma_i = \sum_{l=i}^{d} \beta_l$, $i = 1, \ldots, d$.

Proof. By definition of the criterion Ψ^β we have to maximize the function in (3.2) which reduces by Theorem 2.8 to

$$\begin{aligned} \Psi^\beta(\xi) &= C \prod_{l=1}^{d} \prod_{j=1}^{l} (q_{2j-2} p_{2j-1} q_{2j-1} p_{2j})^{\beta_l} \\ &= C \prod_{j=1}^{d} \prod_{l=j}^{d} (q_{2j-2} p_{2j-1} q_{2j-1} p_{2j})^{\beta_l} \\ &= C \prod_{j=1}^{d} (q_{2j-1} p_{2j-1})^{\sigma_j} \prod_{j=1}^{d-1} q_{2j}^{\sigma_{j+1}} p_{2j}^{\sigma_j} \; p_{2d}^{\sigma_d} \;, \end{aligned}$$

where $p_1, p_2, \ldots$ denote the canonical moments of the design ξ $(q_0 = 1)$ and the constant C does not depend on the design ξ. The assertion now follows by a straightforward maximization of this function in terms of the canonical moments. □

Example 3.2. Consider the uniform prior $\beta_l = 1/d$ $(l = 1, \dots, d)$, then it is easy to see that the criterion (3.2) reduces to the D-optimality criterion

$$\Psi^\beta(\xi) \;=\; |M_d(\xi)|^{\frac{1}{d}}$$

and Theorem 3.1 gives the canoncial moments of the D-optimal design derived in Section 2 [see formula (2.13)]. As a further application consider the prior $\beta_1 = \ldots = \beta_{d-1} = 0$, $\beta_d = 1$, which corresponds to a discrimination between a polynomial of degree $d-1$ and d. In this case the criterion (3.2) reduces to the D_1-optimality criterion

$$\Psi^\beta(\xi) \;=\; \frac{|M_d(\xi)|}{|M_{d-1}(\xi)|}$$

and we obtain from Theorem 3.1 by a straightforward calculation that the optimal canonical moments are given by

$$p_i = \frac{1}{2} \;,\; i = 1, \dots, 2d-1 \;\;;\;\; p_{2d} = 1 \;.$$

It can be shown by similar techniques as illustrated in Chapter 2 [see e.g. Studden (1980a)] that the design ξ^* corresponding to this sequence is supported at

$$\{\; x \;\mid (x^2-1)T_d'(x) = 0 \;\} \;=\; \left\{\; \cos(\frac{j\pi}{d}) \mid j = 0, \dots, d \right\}$$

with masses given by

$$\xi^*\Big(\cos(\frac{j\pi}{d})\Big) \;=\; \begin{cases} \frac{1}{d} & \text{if } 1 \le j \le d-1 \\ \frac{1}{2d} & \text{if } \quad\; j = 0, d \end{cases}$$

It is interesting to note that there exists a converse of Theorem 3.1, which shows that any symmetric design maximizes a function of the form (3.2). Although on the first glance this result is not too helpful from a statistical point of view, it will be a very useful tool for deriving new identities for orthogonal polynomials in the next section. The proof is a straightforward application of Theorem 3.1, solving for the corresponding weights.

Theorem 3.3. *Let ξ denote a symmetric design on the interval $[-1,1]$ with canonical moments $p_j \in (0,1)$ for all $1 \le j \le 2d-1$ and $p_{2d} = 1$, then ξ maximizes the function Ψ^β defined in (3.2), where the weights $\beta_1, \dots, \beta_d$ are given by*

$$\beta_l \;=\; \prod_{j=1}^{l-1} \frac{q_{2j}}{p_{2j}}(1 - \frac{q_{2l}}{p_{2l}}) \qquad\qquad l = 1, \dots, d \;. \tag{3.3}$$

Note that the weights in (3.3) can become negative and in this case there is no statistical interpretation of the criterion Ψ^β. Consider for example the class of polynomial models up to degree 5 on the interval $[-1,1]$ and the distribution ξ_B with masses proportional to $1 : 5 : 10 : 10 : 5 : 1$ at the points -1, $-3/5$, $-1/5$, $1/5$, $3/5$ and 1, respectively. This is the Binomial distribution with parameters $p = 1/2$ and $n = 5$ transformed to the interval $[-1,1]$. By Example 2.6 it follows that the canonical moments of even order of ξ_B are given by $p_{2i} = i/5$ $(i = 1, \dots, 5)$ while the canonical moments of odd order are $1/2$. By Theorem 3.3 the design ξ_B maximizes the function Ψ^β in (3.2) where the vector of weights is given by

$$\beta \;=\; (-3, -2, 2, 3, 1),$$

which does not define a prior on the class of polynomials up to degree 5.

3.2 Identities for orthogonal polynomials

Throughout this section we assume that the weights in the criterion (3.2) are arbitrary (not necessarily nonnegative) numbers with sum 1 and $\beta_d \neq 0$. In this case the function Ψ^β is not necessarily concave. But nevertheless we can give a necessary condition for a design maximizing the function Ψ_β, which is of similar structure as the equivalence theorem for the D-optimality criterion stated in Theorem 1.8.

Lemma 3.4. *If the design ξ^* maximizes the function $\Psi^\beta(\xi)$ in (3.2) over the class of all probability measures on the interval $[-1,1]$, then the inequality*

$$\sum_{l=1}^{d} \beta_l \frac{\left(e_l^T M_l^{-1}(\xi^*) f_l(x)\right)^2}{e_l^T M_l^{-1}(\xi^*) e_l} \;\leq\; 1 \tag{3.4}$$

holds for all $x \in [-1,1]$ with equality for the support points of ξ^*.

Proof. For a probability measure ξ on the interval $[-1,1]$ with $|M_d(\xi)| \neq 0$ define

$$\Phi\,(\xi) \;=\; \log \Psi^\beta(\xi) \;=\; -\sum_{l=1}^{m} \beta_l \log e_l^T M_l^{-1}(\xi) e_l.$$

Let

$$F_\Phi\,(\xi, \eta) \;=\; \frac{d}{d\alpha} \Phi\,((1-\alpha)\xi + \alpha\eta)\,|_{\alpha=0+} \tag{3.5}$$

denote the Frechét derivative of the function Φ at ξ in the direction of η. For a matrix $A = (a_{ij})$ we define its derivative by differentiating the elements, that is

$$\frac{\partial}{\partial t} A = \left(\frac{\partial}{\partial t} a_{ij}\right)_{ij},$$

then it follows for a nonsingular square matrix

$$\frac{\partial}{\partial t} A^{-1} \;=\; -\,A^{-1} \frac{\partial}{\partial t} A\; A^{-1}\;.$$

This implies that for $l = 1, \dots, d$

$$\frac{d}{d\alpha} \log \Big[e_l^T \left\{ (1-\alpha) M_l(\xi) + \alpha M_l(\eta) \right\}^{-1} e_l \Big] \Big|_{\alpha=0+}$$
$$= \ 1 \ - \ \frac{e_l^T M_l^{-1}(\xi) M_l(\eta) M_l^{-1}(\xi) e_l}{e_l^T M_l^{-1}(\xi) e_l}$$

and consequently the directional derivative of Φ at ξ in the direction of η is given by

$$F_\Phi \left(\xi, \eta \right) \ = \ -1 + \sum_{l=1}^{m} \beta_l \frac{e_l^T M_l^{-1}(\xi) M_l(\eta) M_l^{-1}(\xi) e_l}{e_l^T M_l^{-1}(\xi) e_l}. \tag{3.6}$$

If ξ^* maximizes Ψ^β or equivalently Φ, then $F_\Phi \left(\xi^*, \eta \right) \leq 0$ for all η. If $\eta = \eta_x$ concentrates mass one at $x \in [-1,1]$, then

$$0 \ \geq \ F_\Phi \left(\xi^*, \eta_x \right) \ = \ -1 \ + \ \sum_{l=1}^{d} \beta_l \frac{\left(e_l^T M_l^{-1}(\xi^*) f_l(x) \right)^2}{e_l^T M_l^{-}(\xi^*) e_l}, \tag{3.7}$$

which is equivalent to (3.4). Moreover, integrating this inequality with respect to the measure $d\xi^*(x)$ gives

$$\int_{-1}^{1} F_\Phi \left(\xi^*, \eta_x \right) d\xi^*(x) \ = \ 0$$

and shows that $F_\Phi \left(\xi^*, \eta_x \right)$ vanishes on the support of the design ξ^*. This proves the second assertion of the Lemma. □

It is worthwhile to demonstrate at this point how concavity is used in the proof of the converse of Lemma 3.4. Integrating (3.4) and observing (3.6) and (3.7) it follows that

$$F_\Phi \left(\xi^*, \eta \right) \ \leq \ 0$$

for all probability measures η on the interval $[-1,1]$. Now the concavity of the function Φ implies that

$$\Phi \left(\eta \right) - \Phi \left(\xi^* \right) \ \leq \ F_\Phi \left(\xi^*, \eta \right) \ \leq \ 0$$

[see the proof of Theorem 1.8] proving that ξ^* maximizes Φ (or equivalently Ψ^β). A sufficient condition for the concavity of Φ is that all weights β_l in the function Ψ^β are nonnegative.

Lemma 3.5. *Let ξ denote a measure on the interval $[-1,1]$ such that $|M_d(\xi)| \neq 0$. The polynomials*

$$P_l(x,\xi) \ = \ \left(e_l^T M_l^{-1}(\xi) e_l \right)^{-1/2} e_l^T M_l^{-1}(\xi) f_l(x) \qquad l = 0, \dots, d \tag{3.8}$$

are orthonormal with respect to the measure $d\xi(x)$.

Proof. Obviously, the function $\hat{P}_d(x) = e_d^T M_d^{-1}(\xi) f_d(x)$ defines a polynomial of degree d and the identity

$$\int_{-1}^{1} \hat{P}_d(x) f_d^T(x) d\xi(x) \;=\; e_d^T M_d^{-1}(\xi) \int_{-1}^{1} f_d(x) f_d^T(x) d\xi(x) \;=\; e_d^T$$

shows that $\hat{P}_d(x)$ is the dth orthogonal polynomial with respect to the measure $d\xi(x)$ with L^2-norm $e_d^T M^{-1}(\xi) e_d$. □

Theorem 3.6. *Let ξ^* denote a symmetric probability measure on the interval $[-1, 1]$, with canonical moments of even order $p_2, \ldots, p_{2d} > 0$. The orthonormal polynomials $\{P_j(x, \xi^*)\}_{j=0}^{d}$ and $\{Q_j(x, \xi^*)\}_{j=0}^{d-1}$ with respect to the measures $d\xi^*(x)$ and $(1 - x^2)d\xi^*(x)$ satisfy the identity*

$$\sum_{l=1}^{d} \beta_l^* P_l^2(x, \xi^*) \;=\; 1 - (1 - x^2)\delta_{d-1}^* Q_{d-1}^2(x, \xi^*), \tag{3.9}$$

where the constants β_l^* and δ_l^* are defined by

$$\begin{aligned} \beta_l^* &= \prod_{j=1}^{l-1} \frac{q_{2j}}{p_{2j}} \left(1 - \frac{q_{2l}}{p_{2l}}\right) \qquad l = 1, \ldots, d-1 \\ \beta_d^* &= \prod_{j=1}^{d-1} \frac{q_{2j}}{p_{2j}} p_{2d} \;, \qquad \delta_{d-1}^* = \prod_{j=1}^{d-1} \frac{q_{2j}}{p_{2j}} q_{2d}. \end{aligned} \tag{3.10}$$

Proof. Let $\bar{\xi}$ denote the symmetric probability measure with the same canonical moments $\bar{p}_j = p_j$ as ξ^* up to the order $2d - 1$ and $\bar{p}_{2d} = 1$. It follows from Theorem 2.8 and the proof of Lemma 3.5 that the L^2-norm of the monic orthogonal polynomials $\underline{R}_d(x, \xi^*)$ with respect to the measure $d\xi^*(x)$ is given by

$$\begin{aligned} (e_d^T M_d^{-1}(\xi^*) e_d)^{-1} \;&=\; \frac{\underline{H}_{2d}}{\underline{H}_{2d-2}} \;=\; \int_{-1}^{1} \underline{R}_d^2(x, \xi^*) d\xi^*(x) \\ &=\; 2^{2d} \prod_{j=1}^{d} q_{2j-2} p_{2j-1} q_{2j-1} p_{2j} \;=\; \prod_{j=1}^{d} q_{2j-2} p_{2j} \;. \end{aligned} \tag{3.11}$$

A similar identity yields for the L^2-norm of the monic orthogonal polynomial $\overline{S}_d(x, \xi^*)$ with respect to the measure $(1 - x^2)d\xi^*(x)$

$$\begin{aligned} \frac{\overline{H}_{2d}}{\overline{H}_{2d-2}} \;&=\; \int_{-1}^{1} \underline{S}_{d-1}^2(x, \xi^*)(1 - x^2) d\xi^*(x) \\ &=\; 2^{2d} \prod_{j=1}^{d} p_{2j-2} q_{2j-1} p_{2j-1} q_{2j} \;=\; \prod_{j=1}^{d} p_{2j-2} q_{2j} \end{aligned} \tag{3.12}$$

[see Dette and Studden, Remark 2.3.7]. Observing (3.11) and Theorem 2.7 we obtain for the orthonormal polynomials with respect to the measures $\bar{\xi}$ and ξ^* satisfy

$$\begin{aligned} P_l(x,\bar{\xi}) &= P_l(x,\xi^*) \qquad l=1,\dots,d-1 \\ P_d(x,\bar{\xi}) &= \sqrt{p_{2d}}P_d(x,\xi^*). \end{aligned} \tag{3.13}$$

Now Theorem 3.3 shows that the probability measure $\bar{\xi}$ maximizes the function Ψ^β in (3.2) for the weights $\beta=(\beta_1,\dots,\beta_d)$ given by

$$\beta_l = \prod_{j=1}^{l-1}\frac{\bar{q}_{2j}}{\bar{p}_{2j}}\left(1-\frac{\bar{q}_{2l}}{\bar{p}_{2l}}\right) = \begin{cases} \beta_l^* & \text{if } 1\le l\le d-1 \\ \frac{1}{p_{2d}}\beta_d^* & \text{if } l=d. \end{cases} \tag{3.14}$$

Here the last equality is a consequence of the definition (3.10) and the fact that the canonical moments of the measure ξ^* and $\bar{\xi}$ up to the order $2d-1$ are identical. By Lemma 3.4 and 3.5 it therefore follows for the orthonormal polynomials $P_n(x,\bar{\xi})$ with respect to the measure $d\bar{\xi}(x)$

$$1 \ge \sum_{l=1}^d \beta_l \frac{\left(e_l^T M_l^{-1}(\bar{\xi}) f_l(x)\right)^2}{e_l^T M_l^{-1}(\bar{\xi}) e_l} = \sum_{l=1}^d \beta_l P_l^2(x,\bar{\xi}) = \sum_{l=1}^d \beta_l^* P_l^2(x,\xi^*)$$

whenever $x\in[-1,1]$. Here we have used (3.13) and (3.14) in the last equality. Moreover, the second part of Lemma 3.4 shows that there is equality on the support of the measure $\bar{\xi}$ which contains $d+1$ points including -1 and 1 (note that $\bar{p}_{2d}=1$ and recall the simple properties in 2.3). In can be shown [see Dette and Studden (1997), Theorem 2.2.3, transferred to the interval $[-1,1]$] that these support points are given by the zeros of the polynomial $(x^2-1)\bar{S}_{d-1}^2(x,\bar{\xi})$ where $\bar{S}_{d-1}(x,\bar{\xi})$ is the $(d-1)$th monic orthogonal polynomial with respect to the measure $(1-x^2)d\bar{\xi}(x)$. By Theorem 2.7 $\bar{S}_{d-1}(x,\bar{\xi})$ is proportional to the $(d-1)$th orthonormal polynomial $Q_{d-1}(x,\xi^*)$ with respect to the measure $(1-x^2)d\xi^*(x)$. Therefore the polynomials $\sum_{l=1}^d \beta_l^* P_l^2(x,\xi^*)-1$ and $(x^2-1)Q_{d-1}^2(x,\xi^*)$ are of degree $2d$, nonpositive on the interval $[-1,1]$ and equal to 0 at the $d+1$ support points of $\bar{\xi}$. Counting zeros with multiplicities shows that the polynomials must be proportional and a comparison of the leading coefficients shows

$$\begin{aligned} \sum_{l=1}^d \beta_l^* P_l^2(x,\xi^*) - 1 &= \left(\prod_{j=1}^{d-1} p_{2j}^{-2}\right)x^{2d} + \dots \\ &= \delta_{d-1}^*(x^2-1)Q_{d-1}^2(x,\xi^*) \end{aligned}$$

where we used (3.11) and (3.12) in both equalities. This is equivalent to (3.9) and proves the assertion. □

Note that Theorem 3.6 provides an identity for the sum of squares of orthogonal polynomials with respect to an arbitrary (symmetric) measure on the interval $[-1,1]$. For further more general identities of this type for polynomials orthogonal with respect to a not necessarily symmetric measure on a compact interval we refer to Dette (1993). In the remaining part of this section we illustrate the identity of Theorem 3.6 in two examples.

Example 3.7. Let μ^T denote the arc-sine distribution on the interval $[-1,1]$ which has canonical moments $p_j = 1/2$ ($j \in \mathbb{R}$) [see Example 2.5]. The orthonormal polynomials $Q_{d-1}(x,\mu^T)$ with respect to the measure $(1-x^2)d\mu^T(x) = \sqrt{1-x^2}dx/\pi$ are proportional to the Chebyshev polynomials of the second kind $U_{d-1}(x)$. From (3.12) it follows that the leading coefficient is $\sqrt{2}2^{d-1}$ which shows that

$$Q_{d-1}(x,\mu^T) = \sqrt{2}U_{d-1}(x).$$

Similarly (3.11) shows that the orthonormal polynomials with respect to the arc-sine measure are given by

$$P_d(x,\mu^T) = \sqrt{2}T_d(x) \ .$$

Now $\beta_l^* = 0$ ($l = 1,\dots,d-1$), $\beta_d^* = \delta_{d-1}^* = 1/2$ and (3.9) reduces to the well known "trigonometric identity"

$$(1-x^2)U_{d-1}^2(x) + T_d^2(x) = 1$$

for the Chebyshev polynomial of the first and second kind.

Example 3.8. Let $\mu^{(\alpha)}$ denote the probability measure on the interval $[-1,1]$ with density

$$c_\alpha(1-x^2)^{\alpha-1/2} \qquad \alpha > -\frac{1}{2},\ \alpha \neq 0 \ . \tag{3.15}$$

The constant in (3.15) is given by

$$c_\alpha \ = \ \frac{\Gamma(2\alpha+1)}{2^{2\alpha}[\Gamma(\alpha+1/2)]^2} \ = \ \frac{\Gamma(\alpha+1)}{\sqrt{\pi}\Gamma(\alpha+1/2)} \ ,$$

which can be obtained from Example 2.5 (and a transformation to the interval $[-1,1]$) and the duplication formula for the Gamma function. The orthonormal polynomials with respect to the measure $d\mu^{(\alpha)}(x)$ are proportional to the ultraspherical polynomials $C_m^{(\alpha)}(x)$ [see Szegö (1959)]. The constant of proportionality can be obtained from the coefficient of x^m in $C_m^{(\alpha)}(x)$. It follows from Example 2.5 that the canonical moments of the measure $\mu^{(\alpha)}$ are given by

$$p_{2i}^{(\alpha)} = \frac{i}{2(i+\alpha)} \qquad p_{2i-1} = \frac{1}{2} \qquad (i \in \mathbb{N}). \tag{3.16}$$

If $\hat{C}_m^{(\alpha)}(x)$ denotes the monic version of $C_m^{(\alpha)}(x)$, this gives the representation [see (3.11)]

$$\begin{aligned} P_l^2(x,\mu^{(\alpha)}) &= \prod_{j=1}^{l} \left(q_{2j-2}^{(\alpha)}p_{2j}^{(\alpha)}\right)^{-1} \left[\hat{C}_l^{(\alpha)}(x)\right]^2 \\ &= 2^{2l}\frac{\Gamma(2\alpha)\Gamma(l+1+\alpha)\Gamma(l+\alpha)}{\Gamma(l+1)\Gamma(l+2\alpha)\Gamma(\alpha)\Gamma(\alpha+1)} \left[\hat{C}_l^{(\alpha)}(x)\right]^2 \qquad (3.17) \\ &= \frac{\Gamma(2\alpha)\Gamma(l+1)(l+\alpha)}{\alpha\Gamma(l+2\alpha)} \left[C_l^{(\alpha)}(x)\right]^2 , \end{aligned}$$

where the last equality follows from the duplication formula for the Gamma function and the fact that the leading coefficient of the polynomial $C_l^{(\alpha)}(x)$ is given by

$$\frac{2^l \Gamma(l+\alpha)}{\Gamma(\alpha)\Gamma(l+1)}.$$

Similarly, the polynomial $Q_{d-1}(x, \mu^{(\alpha)})$ orthonormal with respect to the measure

$$(1-x^2)d\mu^{(\alpha)}(x) \; = \; c_\alpha (1-x^2)^{\alpha+1/2} dx$$

is proportional to $C_{d-1}^{(\alpha+1)}(x)$ and given by

$$Q_{d-1}^2(x, \mu^{(\alpha)}) \; = \; \frac{2\Gamma(d)\Gamma(2\alpha+1)(d+\alpha)}{\Gamma(d+2\alpha+1)} \left[C_{d-1}^{(\alpha+1)}(x) \right]^2 .$$

Finally, the constants β_l^* and δ_{m-1}^* in (3.10) can be calculated as

$$\beta_l^* \; = \; \begin{cases} -\dfrac{\Gamma(l+2\alpha)}{\Gamma(l+1)\Gamma(2\alpha)} & l = 1, \ldots, d-1 \\ \dfrac{d\Gamma(d+2\alpha)}{2(d+\alpha)\Gamma(d)\Gamma(2\alpha+1)} & l = d \end{cases}$$

and

$$\delta_{d-1}^* \; = \; \frac{\Gamma(d+2\alpha+1)}{2(d+\alpha)\Gamma(d)\Gamma(2\alpha+1)}.$$

Consequently we obtain from (3.9) the following identity for the sum of squares of ultraspherical polynomials

$$\left[\frac{d}{2\alpha} C_d^{(\alpha)}(x) \right]^2 - \sum_{l=0}^{d-1} \frac{l+\alpha}{\alpha} \left[C_l^{(\alpha)}(x) \right]^2 \; = \; (x^2-1) \left[C_{d-1}^{(\alpha+1)}(x) \right]^2 , \qquad (3.18)$$

which has a nice application in mathematical physics (see Dehesa, Van Assche and Yanez (1997)).

3.3 Maximin discrimination designs

In this Section we go back to the statistical problem of determining optimal discrimination designs for the degree of a polynomial regression. In Section 3.1 we used a geometric mean to discriminate between competing designs and in Section 3.2 we related the design problem to identities for orthogonal polynomials. In the present section we concentrate on a different criterion, which relates the design problem to a nonlinear extremal problem for polynomials [see the following section]. More precisely, we consider the function

$$\Psi(\xi) \; = \; \min \left\{ 2^{2l-2} \frac{|M_l(\xi)|}{|M_{l-1}(\xi)|} \;\middle|\; l = 1, \ldots, d \right\} \qquad (3.19)$$

and call a design maximizing Ψ a *maximin-optimal discrimination design.* The factors 2^{2l-2} in (3.19) are introduced because the determinants

$$\frac{|M_l(\xi)|}{|M_{l-1}(\xi)|} \tag{3.20}$$

are of quite different order for different values of the index l. Observing Example 3.2 we see that the design maximizing the ratio in (3.20) has canonical moments $p_i = 1/2$, $(i = 1, \ldots, 2l-1)$, $p_{2l} = 1$ and we obtain from Theorem 2.8 that

$$\sup_\xi \frac{|M_l(\xi)|}{|M_{l-1}(\xi)|} \ = \ 2^{-2l+2} \ .$$

This means that we standardized each term in (3.19) by the maximum value obtainable by maximzing $|M_l(\xi)|/|M_{l-1}(\xi)|$ seperately.

Theorem 3.9. *For a design ξ with $|M_d(\xi)| \neq 0$ define*

$$\mathcal{N}(\xi) = \left\{ j \in \{1, \ldots, d\} \ \middle| \ 2^{2j-2} \frac{|M_j(\xi)|}{|M_{j-1}(\xi)|} = \Psi(\xi) \ \right\} . \tag{3.21}$$

A design ξ^* is a maximin-optimal discriminating design if and only if $|M_d(\xi^*)| \neq 0$ and for any $l \in \mathcal{N}(\xi^*)$ there exist a nonnegative number α_l such that

$$\sum_{l \in \mathcal{N}(\xi^*)} \alpha_l \ = \ 1 \tag{3.22}$$

and such that the inequality

$$\sum_{l \in \mathcal{N}(\xi^*)} \alpha_l \frac{(e_l^T M_l^{-1}(\xi^*) f_l(x))^2}{e_l^T M_l^{-1}(\xi^*) e_l} \ \leq \ 1 \tag{3.23}$$

holds for all $x \in [-1, 1]$. Moreover there is equality in (3.23) for all support points of the maximin-optimal discriminating design.

Note that this result provides a similar characterization of the the optimal design as given for the D-optimality criterion in Theorem 1.8 or in Lemma 3.4 for the geometric mean. However, there is an important difference which should be pointed out here. While Theorem 1.8 and Lemma 3.4 are directly applicable to check the optimality of a given design, this is not possbile for Theorem 3.9. The reason is that it is not clear how to choose the weight α_l (except in the case where $\#\mathcal{N}(\xi^*) = 1$) for a given design ξ^*. These quantities appear because of the non-differentiablity of the criterion (3.19) and represent certain subgradients of a concave function. For the same reason a proof of Theorem 3.9 is based on general arguments of convex analysis [see Pukelsheim (1993)] and is not given here [see Dette (1995) or Dette and Studden (1997) for more details]. Nevertheless this result provides one of the main tools in identifying the design maximizing the function Ψ in (3.19).

Theorem 3.10. *The design ξ^* maximizing the function $\Psi(\cdot)$ in (3.19) is uniquely determined and has canonical moments*

$$\begin{aligned} p_{2j-1} &= \frac{1}{2} \qquad j = 1, \ldots, d \\ p_{2j} &= \frac{d-j+2}{2(d-j)+2} \quad j = 1, \ldots, d\ . \end{aligned} \tag{3.24}$$

Moreover, the support points of ξ^* are obtained as the roots of the polynomial

$$(x^2-1)U_d'(x)$$

and the weights are given by

$$\xi^*(x) = \begin{cases} \dfrac{1}{d+2} & \text{if } U_d'(x) = 0 \\ \dfrac{3}{2}\dfrac{1}{d+2} & \text{if} \quad x = \mp 1 \end{cases}$$

Proof. Assume that ξ^* has the canonical moments as specified in Theorem 3.10. Observing Theorem 2.8 it then straightforward to show that

$$\frac{|M_l(\xi^*)|}{|M_{l-1}(\xi^*)|} \cdot 2^{2l-2} = \frac{|M_d(\xi^*)|}{|M_{d-1}(\xi)|} 2^{2d-2}$$

for all $l = 1, \ldots, d$. By definition of the set $\mathcal{N}(\xi^*)$ in (3.21) this implies

$$\mathcal{N}(\xi^*) = \{1, \ldots, d\}\ .$$

For the application of Theorem 3.9 we have to identify the weights α_l and we will use Theorem 3.3 for this purpose. This result shows that ξ^* also maximizes a geometric mean

$$\Psi^{\beta^*}(\xi) = \prod_{l=1}^{d} \Big(\frac{|M_l(\xi)|}{|M_{l-1}(\xi)|} \Big)^{\beta_l^*}$$

if the weights β_l^* are choosen appropriately, that is

$$\beta_l^* = \prod_{j=1}^{l-1} \frac{q_{2j}}{p_{2j}} \Big(1 - \frac{q_{2l}}{p_{2l}}\Big) = \frac{2}{d(d+1)}(d-l+1) \tag{3.25}$$

($l = 1, \ldots, d$), where the last equality follows from the represenation for the canonical moments (of even order) of the design ξ^*, i.e.

$$p_{2l} = \frac{d-l+2}{2(d-l)+2}, \quad l = 1, \ldots, d.$$

Now Lemma 3.4 shows that the inequality

$$\sum_{l=1}^{d} \beta_l^* \frac{(e_l^T M_l^{-1}(\xi^*) f_l(x))^2}{e_l^T M_l^{-1}(\xi^*) e_l} \;\leq\; 1 \tag{3.26}$$

holds for all $x \in [-1,1]$. Moreover, from (3.25) it is easy to see that $\sum_{l=1}^{d} \beta_l^* = 1$ and that $\beta_l^* \geq 0$ for all $l = 1, \ldots, d$. Oberving $\mathcal{N}(\xi^*) = \{1, \ldots, d\}$ it therefore follows that we can use the weights $\alpha_l = \beta_l^*$ in Theorem 3.9 and obtain from (3.26)

$$\sum_{l \in \mathcal{N}(\xi^*)} \alpha_l \frac{(e_l^T M_l^{-1}(\xi^*) f_l(x))^2}{e_l^T M_l^{-1}(\xi^*) e_l} \;\leq\; 1$$

for all $x \in [-1,1]$. From Theorem 3.9 it therefore follows that ξ^* defined by the canoncial moments in (3.24) maximzes the function

$$\Psi(\xi) = \min\Big\{ \frac{|M_l(\xi)|}{|M_{l-1}(\xi)|} 2^{2l-2} \mid l = 1, \ldots, d \Big\} ,$$

which proves the first assertion of Theorem 3.10. The representation of the support points and weights is now obtained by similar arguments as given in Section 2.3 and therefore omitted. □

Example 3.11. In this example we discuss the problem of constructing a maximin optimal discriminating design for the discrimination between a linear, quadratic and cubic polynomial regression model. Thus we have $d = 3$

$$U_3'(x) \;=\; (8x^3 - 4x)' = 24x^2 - 4$$

and we obtain from Theorem 3.9 for the maximin optimal discriminating design

$$\xi^* \;=\; \begin{pmatrix} -1 & -1/\sqrt{6} & 1/\sqrt{6} & 1 \\ 0.3 & 0.2 & 0.2 & 0.3 \end{pmatrix} .$$

3.4 Extremal problems for polynomials

In this section we discuss some extensions of an extremal property of the Chebyshev polynomials of the first kind [see Chebyshev (1959)]. More precisely we consider the problem

$$\min_{a_0, \ldots, a_{d-1}} \sup_{x \in [-1,1]} \Big| x^d - \sum_{j=0}^{d-1} a_j x^j \Big| \tag{3.27}$$

of best approximation of the power x^d by a (real) polynomial of degree $d - 1$. It is well known (see Natanson (1955) or Rivlin (1990)) that the minimum value in (3.27) is given by

$1/2^{d-1}$ and the "best" polynomial $x^d - \sum_{j=0}^{d-1} a_j x^j$ is given by $1/2^{d-1}T_d(x)$ where $T_d(x)$ is the Chebyshev polynomial of the first kind. We will present a new proof of this result, which was proposed by Studden (1980b) and is based on a game theoretic argument. The main idea is to relate the extremal problem to the D_1-optimal design problem which can be solved with canonical moments [see Example 3.2]. This duality allows the treatment of more general extremal problems. With the notation of the previous sections (3.27) can be rewritten as ($a \in \mathbb{R}^{d+1}$)

$$\begin{aligned}
\inf_{|a^T e_d|^2=1} \sup_{x\in[-1,1]} (a^T f_d(x))^2 &= \inf_{|a^T e_d|^2=1} \sup_{\xi} \int_{-1}^{1} (a^T f_d(x))^2 d\xi(x) \\
&= \inf_{|a^T e_d|^2=1} \sup_{\xi} a^T M_d(\xi) a \\
&= \sup_{\xi} \inf_{|a^T e_d|^2=1} a^T M_d(\xi) a && (3.28) \\
&= \sup_{\xi} \inf_{a\in\mathbb{R}^{d+1}} \frac{a^T M_d(\xi) a}{(a^T e_d)^2} && (3.29) \\
&= \sup_{\xi} (e_d^T M_d^{-1}(\xi) e_d)^{-1} && (3.30) \\
&= \sup_{\xi} \frac{|M_d(\xi)|}{|M_{d-1}(\xi)|}.
\end{aligned}$$

with the convention that $(e_d^T M_d^{-1}(\xi)e_d)^{-1}$ is zero if $M_d(\xi)$ is singular. Here the equality in (3.28) follows from a game theoretic argument and the fact that the kernel $w(a,\xi) = a^T M_d(\xi) a$ is convex in a and concave (even linear) in ξ. Note that the calculation of the supremum in (3.29) can be restricted to the set of designs with nonsingular moment matrix of order $2d$ and that the equality between (3.29) and (3.30) is a consequence of Cauchy's inequality

$$(a^T e_d)^2 \leq a^T M(\xi) a \cdot e_d^T M^{-1}(\xi) e_d. \tag{3.31}$$

Now (3.30) is the D_1-optimal design problem in a polynomial regression of degree d which was solved in Example 3.2. If $\xi_d^{D_1}$ denotes the D_1-optimal design and

$$P_d(x) = \hat{a}^T f_d(x) \quad , \quad (\hat{a}^T e_d)^2 = 1 \tag{3.32}$$

is a optimal solution of (3.27), then there must be equality in (3.31), i.e.

$$\hat{a} = cM^{-1}(\xi_d^{D_1})e_d, \tag{3.33}$$

where the constant c is determined by $(\hat{a}^T e_d)^2 = 1$. Combining (3.32), (3.33) with Lemma 3.5 it follows that $P_d(x)$ is the dth monic orthogonal polynomial with respect to the measure $d\xi_d^{D_1}(x)$ (up to the sign). Now the canonical moments of $\xi_d^{D_1}$ up to the order $2d-1$ are $1/2$ and coincide with the canonical moment of the arc-sine distribution. Therefore Theorem 2.7 shows that $P_d(x)$ is the dth monic orthogonal polynomial with respect to the arc-sine measure, i.e.

$$P_d(x) = \frac{1}{2^{d-1}} T_d(x) \ ,$$

which determines the solution of the extremal problem (3.27).

In the remaining part of this article we discuss a more general extremal problem which cannot be solved by the "classical" methods of approximation theory. To be precise let $I = \{i_1, \dots, i_n\}$ denote a subset of $\{1, \dots, d\}$ containing d and define

$$P_I \ := \ \Big\{(P_j)_{j \in I} \mid P_j \in I\!P_j, \ j \in I, \ \sup_{x \in [-1,1]} \sum_{j \in I} P_j^2(x) \leq 1\Big\}$$

as the set of all polynomials of degree $i_1, \dots, i_n$ such that the sup-norm of the sum of squares is bounded by 1 on the interval $[-1, 1]$. In the following $m_l(P_l)$ denotes the leading coefficient of the polynomial $P_l \in I\!P_l$. We are interested in the (nonlinear) extremal problem

$$(\mathcal{P}_I) \qquad \max\Big\{\sum_{l \in I} \beta_l m_l^2(P_l) \mid (P_l)_{l \in I} \in P_I\Big\}$$

where $\beta = (\beta_{i_1}, \dots, \beta_{i_n})$ denotes a vector of positive weights with sum 1. Note that for $I = \{d\}$ the extremal problem $(\mathcal{P}_I)$ reduces to the problem of maximizing the highest coefficient among all polynomials of (precise) degree d with sup-norm bounded by 1. This is an alternative formulation of the "classical" Chebyshev approximation problem (3.27). Throughout this section the orthogonal polynomials with leading coefficient 1 corresponding to a probability measure will be denoted by $R_j(x, \xi)$ and their (squared) L_2-norm by

$$k_j(\xi) \ = \ \int_{-1}^{1} R_j^2(x, \xi) d\xi(x) \ = \ \frac{|M_j(\xi)|}{|M_{j-1}(\xi)|} \ = \ (e_j^T M_j^{-1}(\xi) e_j)^{-1}. \tag{3.34}$$

The main step for solving the extremal problem $(\mathcal{P}_I)$ is the following duality which is the analogue in the game theoretic argument in (3.28). A proof is based on Fenchel's duality theorem in convex analysis and can be found in Dette (1995b) or Dette and Studden (1997).

Theorem 3.12. *If $\Xi := \{\xi \in \Xi \mid |M_d(\xi)| > 0\}$ denotes the set of all probability measures with nonsingular Hankel matrix $M_d(\xi)$, then the following duality holds*

$$(\mathcal{P}_I) \qquad \max\Big\{\sum_{l \in I} \beta_l m_l^2(P_l) \mid (P_j)_{j \in I} \ \in \ P_I\Big\} \ = \ \min_{\xi \in \Xi} \max_{j \in I} \{\beta_j k_j^{-1}(\xi)\} \qquad (\mathcal{D}_I)$$

and solutions of $(\mathcal{P}_I)$ and $(\mathcal{D}_I)$ exist.
Moreover, let ξ^* be a solution of the problem $(\mathcal{D}_I)$,

$$\mathcal{M}(\xi^*) = \{j \in I \mid \beta_j^{-1} k_j(\xi^*) = \min_{i \in I} \beta_i^{-1} k_i(\xi^*)\},$$

and $\sqrt{1/k_j(\xi^*)} R_j(x, \xi^*)$ denote the jth orthonormal polynomial with respect to the measure $d\xi^*(x)$. Then there exist constants $\alpha_j \geq 0$ with sum 1 satisfying

$$\alpha_j = 0 \qquad \text{if} \ \ j \in I \setminus \mathcal{M}(\xi^*) \tag{3.35}$$

$$\sum_{j\in I} \alpha_j k_j^{-1}(\xi^*) R_j^2(x,\xi^*) \ \le \ 1 \quad \textit{for all} \ \ x \in [-1,1] \ . \tag{3.36}$$

With this choice $\{\sqrt{\alpha_j/k_j(\xi^*)}R_j(x,\xi^*)\}_{j\in I}$ *is a solution of the extremal problem* $(\mathcal{P}_I)$.

Note that the dual problem $(\mathcal{D}_I)$ contains as a special case the optimization problem considered in (3.19) ($n = d$, $\beta_j = 2^{-2j+2}$, $j = 1,\dots,n$). While from a statistical point of view the main interest are the support points and weights of the solution ξ^* of $(\mathcal{D}_I)$ Theorem 3.12 shows that the orthogonal polynomials with respect to the optimal design $d\xi^*(x)$ are needed for the solution of the extremal problem $(\mathcal{P}_I)$. These polynomials can be calculated by the recurrence relations given in Theorem 2.7 and the L_2-norm is given

$$k_d(\xi^*) \ = \ \int_{-1}^{1} R_d^2(x,\xi^*) d\xi^*(x) \ = \ 2^{2d} \prod_{j=1}^{d} q_{2j-2}^* p_{2j-1}^* q_{2j-1}^* p_{2j}^* \ , \tag{3.37}$$

where $p_1^*, p_2^*, \dots$ denote the canonical moments of the optimal design ξ^*. The following result specifies these canonical moments in explicit form.

Theorem 3.13. *The solution* ξ^* *of the dual problem* $(\mathcal{D}_I)$ *is uniquely determined by its canonical moments* $(p_1^*,\dots,p_{2d}^*)$ *where* $p_{2j-1}^* = 1/2$ $(j = 1,\dots,d)$, $p_{2d}^* = 1$, $p_{2(d-j)}^* = 1/2$ *if* $d-j \notin I$ *and*

$$p_{2(d-j)}^* \ = \ \max\Big\{1 - \frac{\beta_d}{\beta_{d-j}} \prod_{i=d-j+1}^{d-1} (q_{2i}^* p_{2i}^*)^{-1}, \ \frac{1}{2}\Big\}. \tag{3.38}$$

if $d-j \in I$.

Proof. The result can be proved by similar arguments as given in the proof of Theorem 3.10. We provide an alternative proof, which is directly based on the duality result of Theorem 3.12 and uses the identities for orthogonal polynomials derived in Section 3.2. For $d-j \in I$ let

$$\gamma_{d-j} = 1 - \beta_d/\beta_{d-j} \prod_{i=d-j+1}^{d-1} (q_{2i}^* p_{2i}^*)^{-1} \, (\gamma_d = 1);$$

then it is easy to see (observing (3.37) and (3.38)) that for $d-j \in I$

$$\gamma_{d-j} \ \ge \ \frac{1}{2} \quad \text{if and only if} \quad \beta_d^{-1} k_d(\xi^*) = \beta_{d-j}^{-1} k_{d-j}(\xi^*)$$
$$\gamma_{d-j} \ < \ \frac{1}{2} \quad \text{if and only if} \quad \beta_d^{-1} k_d(\xi^*) < \beta_{d-j}^{-1} k_{d-j}(\xi^*)$$

Consequently, we have for the set $\mathcal{M}(\xi^*)$ in Theorem 3.12

$$d \in \mathcal{M}(\xi^*) \ = \ \{j \in I \mid \gamma_j \ge \frac{1}{2}\} \tag{3.39}$$
$$p_{2j}^* \ = \ \frac{1}{2} \quad \text{if } j \notin \mathcal{M}(\xi^*). \tag{3.40}$$

In the following define weights $\alpha_1, \ldots, \alpha_d$ by

$$\alpha_j \;=\; \prod_{i=1}^{j-1} \frac{q_{2i}^*}{p_{2i}^*} \left(1 - \frac{q_{2j}^*}{p_{2j}^*}\right) . \tag{3.41}$$

These have sum 1 and are nonnegative, by the definition of p_{2j}^* in (3.38). Additionally we have by (3.40) $\alpha_j = 0$ whenever $j \notin \mathcal{M}(\xi^*)$. From (3.37) it follows that the polynomials $k_l^{-1/2}(\xi^*)R_l(x,\xi^*)$ are orthonormal with respect to the measure $d\xi^*(x)$ and Theorem 3.6 shows (note that $p_{2d}^* = 1$)

$$\sum_{j=1}^{d} \alpha_j k_j^{-1}(\xi^*) R_j^2(x,\xi^*) \;=\; \sum_{j \in \mathcal{M}(\xi^*)} \alpha_j k_j^{-1}(\xi^*) R_j^2(x,\xi^*) \;\le\; 1 \tag{3.42}$$

for all $x \in [-1,1]$. In other words

$$\{R_j^*(x)\}_{j\in I} \;:=\; \left\{ \sqrt{\alpha_j / k_j(\xi^*)} R_j(x,\xi^*) \right\}_{j \in I} \;\in\; P_I \tag{3.43}$$

and by the definition of $\mathcal{M}(\xi^*)$ and $\sum_{j\in\mathcal{M}(\xi^*)} \alpha_j = 1$ it follows

$$\sum_{j\in I} \beta_j m_j^2(R_j^*) \;=\; \sum_{j\in\mathcal{M}(\xi^*)} \beta_j m_j^2(R_j^*) \;=\; \frac{\beta_d}{k_d(\xi^*)} \;=\; \max\{\beta_j k_j^{-1}(\xi^*) \mid j \in I\} \; .$$

Therefore we have equality in Theorem 3.12 for $\{R_j^*\}_{j\in I} \in P_I$, $\xi^* \in \Xi$ and the assertion of the theorem follows. □

Example 3.14. We will conclude with two examples illustrating the application of Theorem 3.12 and 3.13. Consider at first the weights $\beta_1 = \ldots = \beta_d = \frac{1}{d}$, then the problem is to maximize

$$\sum_{l=1}^{d} m_l^2(P_l)$$

subject to the constaints

$$\sum_{l=1}^{d} P_l^2(x) \;\le\; 1 \quad \forall \;\; x \in [-1,1] \; . \tag{3.44}$$

In this case we have from Theorem 3.13 $p_i^* = 1/2$ for all $i = 1, \ldots, 2d-1$, $p_{2d}^* = 1$, the corresponding optimal design is the D_1-optimal design (see Example 3.2) and

$$\mathcal{M}(\xi^*) = \{d\}.$$

¿From Theorem 3.12 and the proof of Theorem 3.13 we obtain for the extremal polynomials

$$P_l(x) = 0 \; , \quad l = 1, \ldots, d-1 \; , \quad P_d(x) = T_d(x) \; .$$

On the other hand, if $\beta_l = 2^{-2l+2}$, $l = 1, \ldots, d$ we are maximizing

$$\sum_{l=1}^{d} 2^{-2l+2} m_l^2(P_l)$$

subject to the constrained (3.44) and the dual problem is the maximin discrimination design problem discussed in Section 3.3. Thus the canonical moments of the solution ξ^* of the corresponding dual problem ξ^* can either be obtained from Theorem 3.9 or 3.13 and are given by (3.24). Consequently we have

$$\mathcal{M}(\xi^*) = \{1, \ldots, d\}$$

and all polynomials are needed for the solution of the extremal problem. From Theorem 2.7 we obtain for the monic orthogonal polynomials with respect to the measure $d\xi^*(x)$ the recursive relation $R_1(x, \xi^*) \;=\; x$,

$$R_2(x, \xi^*) \;=\; x^2 - \frac{d+1}{2d}$$

$$R_{l+1}(x, \xi^*) \;=\; x R_l(x, \xi^*) \;-\; \frac{1}{4} R_{l-1}(x, \xi^*)\ , \quad l = 1, \ldots, d-1$$

and $R_l(x, \xi^*)$ can be identified as a difference of a Chebyshev polynomial of the first and second kind [see Chihara (1978), Chapter VI - 13], that is

$$R_l(x, \xi^*) \;=\; 2^{1-l} \Big[T_l(x) - \frac{1}{d} U_{l-2}(x) \Big]\ , \quad l = 1, \ldots, d\ .$$

Finally, we obtain from (3.24) and (3.41) and a straightforward calculation

$$\alpha_l \;=\; \frac{2(d-l+1)}{d(d+1)}\ , \quad l = 1, \ldots, d$$

$$k_l(\xi^*) \;=\; \Big(\frac{1}{4}\Big)^{l-1} \frac{d+1}{2d}$$

and Theorem 3.12 yields for the corresponding extremal polynomials

$$P_l(x) \;=\; \frac{2\sqrt{d-l+1}}{d+1} \Big[T_l(x) - \frac{1}{d} U_{l-2}(x) \Big]\ , \quad l = 1, \ldots, d\ .$$

Acknowledgements. This paper summarizes a series of lectures of the auhtor given in Inzell at the *Summer-School on Orthogonal Polynomials, Harmonic Analysis and Applications* in September 2001. The author would like to thank the organizers for the invitation to this interesting workshop. Parts of this paper were written, during the second author's visit at Purdue University and the author would like to thank the Department of Statistics for its hospitality. The financial support of the Deutsche Forschungsgemeinschaft (SFB 475, Reduction of complexity in multivariate data structures) is gratefully acknowledged.

4 References

Anderson, T. W. (1962). The choice of the degree of a polynomial regression as a multiple decision problem. *Annals of Mathematical Statistics*, **33**, 255–265.

Atkinson, A. C. and Cox, D. R. (1974). Planning experiments for discriminating between models. *Journal of the Royal Statistical Society, Ser. B*, **36**, 321–334.

C. Bosbach, W. Gawronski (1998). Strong asymptotics for Laguerre polynomials with varying weights. *J. Comput. Appl. Math.* **99**, 77-89.

Chebyshev, P. L. (1859). Sur les questions de minima qui se retachent à la représentation approximative de fonctions. *Mémoires de l'Académie Impériale des Sciences de St-Pétersbourg. Sixième Série Sciences Mathématiques et Physiques*, **7**, 199–291.

Chihara, T. S. (1978). *Introduction to Orthogonal Polynomials.* Gordon and Breach, NY.

Dehesa, J. S., Van Assche, W., Yanez, A. (1997). Information entropy of classical orthogonal polynomials and their application to harmonic oscillator and Coulomb potentials. Methods and Applications in Analysis 4, 91-110.

Dette, H. (1993). New Identities for orthogonal polynomials on a compact interval. *Journal of Mathematical Analysis and its Applications*, **179**, 547–573.

Dette, H. (1994). Discrimination designs for polynomial regression on a compact interval. *Annals of Statistics*, **22**, 890-904.

Dette, H. (1995a). Optimal designs for identifying the degree of a polynomial regression. *Annals of Statistics*, **23**, 1248–1267.

Dette, H. (1995b). A note on some peculiar extremal phenomena of the Chebyshev polynomials. *Proceedings of the Edinburgh Mathematical Society*, **38**, 343–355.

Dette, H. (1995c). Characterizations for generalized Hermite and sieved ultraspherical polynomials. *Transactions of the American Mathematical Society*, **346**, 691–712.

Dette, H. and Studden, W. J. (1992). On a new characterization of the classical orthogonal polynomials. *Journal of Approximation Theory*, **71**, 3–17.

Dette, H. and Studden, W. J. (1995). Some new asymptotic properties for the zeros of Jacobi, Laguerre and Hermite polynomials. *Constructive Approximation*, **11**, 227–238.

Dette, H. and Studden, W.J. (1997). The Theory of Canonical Moments with Applications in Statistics. *Probability and Analysis*, Wiley, N.Y.

J. Faldey, W. Gawronski (1995). On the limit distribution of the zeros of Jonquiere polynomials and generalized classical orthogonal polynomials. *J. Approx. Theory* **81**, 231-249.

Fedorov, V. V. (1972). *Theory of optimal experiments.* Academic Press, New York.

Gaffke, N., Krafft, O. (1982). Exact D-optimum designs for quadratic regression. J. Roy. Statist. Soc., Ser. B, 44, 394-397.

Gauss, C. F. (1813). Disquisitiones gererales circa seriem infinitam $1+\frac{\alpha\beta}{1\cdot\gamma}+\frac{\alpha(\alpha+1)\beta(\beta+1)}{1\cdot2\cdot\gamma(\gamma+1)}+\frac{\alpha(\alpha+1)(\alpha+2)\beta(\beta+1)(\beta+2)}{1\cdot2\cdot3\cdot\gamma(\gamma+1)(\gamma+2)}+etc.$. *Commentationes Societatis Regiae Scientiarum Goettingensis Recentiones*, Vol **2**, 1–46; Werke, Band **3**, Königliche Gesellschaft der Wissenschaften, Göttingen (1876), 123–162.

W. Gawronski (1993). Strong asymptotics and the asymptotic zero distributions of Laguerre polynomials $L_n^{(an+\alpha)}$ and Hermite polynomials $H_n^{(an+\alpha)}$. *Analysis* **13**, 29-67.

Guest, P. G. (1958). The spacing of observations in polynomial regression. *Annals of Mathematical Statistics*, **29**, 294–299.

Hoel, P. G. (1958). Efficiency problems in polynomial estimation. *Annals of Mathematical Statistics*, **29**, 1134–1145.

Hofmann, G., Jung, W. (1975). On sequential and non-sequential D-optimal experimental design. Biometrische Z. **17**, 329-336.

Johnson, N. L. and Kotz, S. (1970). *Continuous univariate distributions.* Houghton Mifflin, New York.

Johnson, N. L., Kotz, S. and Kemp, A. W. (1992). *Univariate discrete distributions.* John Wiley & Sons, New York.

Karlin, S. and Shapeley, L. S. (1953). Geometry of Moment Spaces, *Amer. Math. Soc. Memoir No.* **12**, Amer. Math. Soc., Providence, Rhode Island.

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

Krafft, O. and Schaefer, M. (1995). Exact Elfving-minimax designs for quadratic regression. *Statistica Sinica*, **5**, 475–485.

A.B.J. Kuijlaars, W. Van Assche (1999). The asymptotic zero distribution of orthogonal polynomials with varying recurrence coefficients. *J. Approx. Theory* **99**, 167-197.

Natanson, T. P. (1955). *Konstruktive Funktionentheorie.* Akademie Verlag, Berlin.

Perron, O. (1954a). *Die Lehre von den Kettenbrüchen.* Band I, Teubner, Stuttgart.

Perron, O. (1954b). *Die Lehre von den Kettenbruüchen.* Band II, Teubner, Stuttgart.

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

Pukelsheim, F. and Rieder, S. (1992). Efficient rounding of approximate designs. *Biometrika*, **79**, 763–770.

Rivlin, T. J. (1990). *Chebyshev polynomials.* Wiley & Sons, New York.

Shohat, J. A. and Tamarkin, J. D. (1943). *The Problem of Moments.* Mathematical Surveys No. 1, Amer. Math. Soc., Providence, RI

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

Skibinsky, M. (1967). The range of the $(n+1)$th moment for distributions on $[0, 1]$. *Journal of Applied Probability*, **4**, 543–552.

Skibinsky, M. (1968). Extreme nth moments on $[0, 1]$ and the inverse of a moment space map. *Journal of Applied Probability*, **5**, 693–701.

Skibinsky, M. (1968). Minimax estimation of a random probability whose first N moments are known. *Annals of Mathematical Statistics*, **39**, 492–501.

Skibinsky, M. (1969). Some striking properties of binomial and beta moments. *Annals of Mathematical Statistics*, **40**, 1753–1764.

Skibinsky, M. (1976). Sharp upper bounds for probability on an interval when the first three moments are known. *Annals of Statistics*, **4**, 187–213.

Skibinsky, M. (1986). Principal representations and canonical moment sequences for distributions on an interval. *Journal of Mathematical Analysis and its Applications*, **120**, 95–120.

Spruill, M. G. (1990). Good designs for testing the degree of a polynomial mean. *Sankhyā, The Indian Journal of Statistics, Ser. B*, **52**, 67–74.

Studden, W. J. (1980a). D_s-optimal designs for polynomial regression using continued fractions. *Annals of Statistics*, **8**, 1132–1141.

Studden, W. J. (1980b). On a problem of Chebyshev. *Journal of Approximation Theory*, **29**, 253–260.

Studden, W. J. (1982). Optimal designs for weighted polynomial regression using canonical moments. *Third Purdue Symposium on Decision Theory and Related Topics (S. S. Gupta, J. O. Berger eds.)*, Vol. 2, 335–350.

Szegö, G. (1959). *Orthogonal Polynomials*. Amer. Math. Soc. Colloq. Publ., Vol. 23, Amer. Math. Soc., NY.

Van Assche, W. (1987). Asymptotics for orthogonal polynomials. *Lecture Notes in Mathematics*, No. 1265, Springer-Verlag, Berlin/New York.

Wall, H. S. (1948). *Analytic Theory of Continued Fractions*. Van Nostrand, New York.