We discuss the construction of minimax Bayesian T-optimal, D-ratio optimal and sequential G-optimal designs for discrimination between competing wavelet representations of nonparametric regression models. In our examples we use the multiwavelet and Daubechies wavelet systems. We find that symmetry is an important property of designs for multiwavelet models, whereas designs for Daubechies wavelet models are nonsymmetric.

KEY WORDS: Daubechies wavelet; D-optimality; Multiwavelet; Nonparametric regression; T-optimality; Sequential Designs; Weighted least squares.

1. INTRODUCTION

Wavelet representations of nonlinear and nonparametric mean response functions have been discussed in details and very well motivated in the literature by several authors. For example, Antoniadis and Leblanc (2000) show that linear wavelet regression is a simpler and faster alternative to generalized smooth models commonly used in modelling binary response data. The general approach is to approximate a square integrable mean response function, say $E[y|x] = \eta(x)$, by a linear combination of components of a wavelet system

$$\eta(x) = d_0\phi(x) + \sum_{j} \sum_{k} c_{jk}\psi_{jk}(x),$$

where $\phi(x)$ is a scaling function and for $j, k$ integers, $\psi_{jk}(x) = 2^{j/2}\psi(2^j x - k)$ are dilated and translated versions of a primary wavelet $\psi(x)$. The flexible and adaptive properties of wavelets...
make such a linear representation possible and useful, especially, when the exact mathematical
structure of \( \eta(x) \) is unknown. This explains why wavelets have found wide applications in several
areas of statistics and in particular, in nonparametric statistics.

It is possible to represent the function \( \eta(x) \) exactly if the experimenter can use infinite number
of wavelet terms in the expression (1.1). Unfortunately, infinitely many terms cannot be used
in actual computations. Therefore, the question which has to be addressed is: given a nonpara-
metric mean response function \( \eta(x) \), what is the appropriate finite number of wavelet terms that
can be used to adequately represent \( \eta(x) \) or the maximum level \( m \) at which to terminate the
approximation? Once \( m \) has been determined, the expression (1.1) can be written as

\[
E[y|x] = \eta(x) = d_0\phi(x) + \sum_{j=0}^{m} \sum_{k=0}^{2^j-1} c_{jk}\psi_{jk}(x) + f(x),
\]

where \( f(x) \) represent the unused wavelet terms. In vector notations, we have \( E[y|x] = q_m^T(x)\beta_m + f(x) \), where \( q_m(x) = (\phi(x), \psi_{00}(x), \psi_{10}(x), \psi_{11}(x), \ldots, \psi_{m,2^m-1}(x))^T \) and \( \beta_m = (d_0, c_{00}, \ldots, c_{m,2^m-1})^T \) are \( 2^{m+1} \times 1 \) dimensional vectors. Clearly, at each level of approximation \( j = 0, 1, 2, \ldots, m \), there
are \( 2^j \) dilated and translated versions of \( \psi(x) \).

An experimenter may be tempted to chose a very large value for \( m \). However, the consequences
of choosing an unreasonably large value for \( m \) has been highlighted by Härdle, Kerkyacharian,
Picard and Tsybakov (1998). They noted that, “with increasing level the linear wavelet estimates
become rougher”. That is, the wavelet estimate of \( \eta(x) \) will include wiggles and spikes that are
not features of the mean response if \( m \) is unreasonably high. Antoniadis, Gregoire and McKeague
(1994) made an attempt to address this question by recommending that experimenters should
examine only \( m = 3, 4 \) and \( 5 \) for sample sizes between 100 and 200. One of the shortcomings of
this recommendation is that if the mean response curve for the experiment is a simple smooth
curve but the experimenter was able to make more than 100 observations, this approach will lead to
an inefficient estimate of the mean response. This implies that the structure of the mean response
curve has to be considered when choosing a value for \( m \). Since the “true” mathematical structure
of \( \eta(x) \) is unknown, it seems reasonable to take this problem into consideration at the design
stage by measuring observations at design points \( \{x_i\}_{i=1}^N \) selected in such a way as to maximize
the differences between competing values of \( m \). Using these observations, a simple test can then
be applied to select the appropriate value of \( m \). The focus of this paper is on constructing such
designs.
In contrast to the common approach of constructing approximate continuous designs, we will construct exact integer-valued designs for discrimination based on a simulated annealing algorithm described in §3.3. The simulated annealing algorithm has previously been used by other authors such as Fang and Wiens (2000), and Haines (1987) for constructing exact optimal designs. Aside from the considerable simplification in the mathematical problem, both analytically and numerically that the annealing algorithm provides, we are now able to exhibit designs for Daubechies wavelets which has no closed mathematical expression. Typically, experimenters use approximate design theory to construct designs which are continuous probability functions $\xi(x)$ on a given design space. A disadvantage of this approach is that the number of observations, out of a total of $n$, allocated to a particular design point $x_i$ given by $n_\xi(x_i)$, is in general, not an integer. Thus, the experimenter has to approximate this quantity when it is not an integer and hope that the approximate design will be at least near optimal. See for instance, Box and Draper (1959), Atkinson and Fedorov (1975), and Wiens (1992) for applications of approximate design theory to linear regression models. Particular applications to wavelet models can be found in Herzberg and Traves (1994) and Oyet and Wiens (2000).

Our discussion will be organized in the following way. In §2, we outline the general wavelet model, discuss the problem of parameter estimation and also establish some notations. The criteria for choosing our nonsequential and sequential designs are discussed in §3 and §4 respectively. In order to account for the uncertainty in the true structure of $\eta(x)$, our design criteria will be functions of the mean squared error (MSE) of the estimated parameter vector rather than the variance. We shall see in §3 that using MSE based criteria will require that we adopt a minimax approach in constructing our designs. Our designs will therefore be said to be robust, in the minimax sense, against small deviations from the assumed value of $m$. References on the subject of designs for discrimination that are not listed in this paper can be found in Biswas and Chaudhuri (2002) and in Dette and Kwiecien (2004).

2. GENERAL WAVELET MODEL AND PARAMETER ESTIMATION

Following the discussion in §1, linear wavelet models often considered in the literature on non-parametric regression can be written as

$$y_j(x_i) = q_m^T(x_i)\beta_m + f(x_i) + \varepsilon_{ij}, \quad i = 1, 2, \ldots, N, \ j = 1, 2, \ldots, n_i,$$

where $x_i \in [0,1]$ are given by a prespecified grid, $n_i$ is the number of observations at $x_i$ and $n = \prod_{i=1}^{N} n_i$ is the total number of observations. The error terms $\varepsilon_{ij}$ are assumed to be independent and identically distributed normal random variables with zero mean and constant variance $\sigma^2$. 

3
We introduce the methodology we have used with a discussion on two competing wavelet models. A brief discussion on how to construct designs for a sequence of nested wavelet models is provided at the end of Theorem 3.1. Suppose that the two competing wavelet models have maximum level of approximations \( m - 1 \) and \( m \). Then, we embed the \((m - 1)\)-th order model into the \( m \)-th order model and write

\[
y_j(x_i) = q^T_{(m-1)}(x_i)\beta_{(m-1)} + z^T(x_i)\gamma + f(x_i) + \varepsilon_{ij}, \quad i = 1, 2, \ldots, N, \quad j = 1, 2, \ldots, n_i, \quad (2.2)
\]

where \( q_{(m-1)}(x_i) \) and \( z(x_i) \) are both \( 2^m \times 1 \) dimensional vectors. Clearly, \( q_m(x) = (q^T_{(m-1)}(x), z^T(x))^T \). If the wavelet system used in the approximation is orthonormal, then \( \int_0^1 q_m(s) f(s) ds = 0 \). This result implies that \( \beta_m = (\beta^T_{(m-1)}, \gamma^T) \) in (2.2) is identifiable, provided the matrix \( \int_0^1 q_m(s)q^T_m(s) ds \) is invertible. In §3, we shall use the discrete analogue of the identifiability condition

\[
\frac{1}{N} \sum_{i=1}^{N} q_m(x_i)f(x_i) = 0, \quad (2.3)
\]

to derive an expression for the contamination term \( f(x) \). Two examples of orthonormal wavelet systems, which we have also used in this paper, are the multiwavelet and the Daubechies wavelet systems. Properties of these wavelets are briefly described in Oyet and Wiens (2003, §2, pp. 113-114).

2.1. Estimation

Some methods for estimating \( \beta_m \) that have been widely studied include a wavelet version of the Gasser-Müller estimator (GM) (see Antoniadis et al. 1994), a modified wavelet version of the Gasser-Müller estimator (MGM) and weighted least squares (WLS) (see Oyet and Sutradhar 2002), all of which can be combined with the nonlinear method of thresholding (see Donoho and Johnstone 1995). In Oyet and Sutradhar (2002), it was shown that both the MGM and WLS estimators are more efficient than the GM estimator through a simulation study. Clearly, the method used to estimate the parameters of a model has some influence on the choice of design. It is therefore essential for an experimenter to decide on the estimation method to be adopted prior to constructing designs. In this paper, we will assume that upon making independent observations \( y_{ij} \) at design points \( x_i \), the experimenter estimates \( \beta_m \) by weighted least squares

\[
\hat{\beta}_m = \frac{1}{n} \sum_{i=1}^{n} w_i q_m(x_i)q^T_m(x_i) \frac{1}{n} \sum_{i=1}^{n} w_i q_m(x_i)y_{ij}, \quad (2.4)
\]

where \( w_i \) are given by Oyet and Wiens (2000), as \( w(x_i) = \int_0^1 ||q_m(s)|| ds/||q_m(x_i)|| \).
We end this section by establishing some notations we shall use throughout this paper. Let \( J_{r,r} = \text{diag}(1(\|\hat{\beta}_1\| > \lambda), \ldots, 1(\|\hat{\beta}_r\| > \lambda)) \). Two common threshold estimators (see Vidakovic 1999) correspond to “hard thresholding”: \( \delta^h(\hat{\beta}_m, \lambda) = J\hat{\beta}_m \) and “soft thresholding”: \( \delta^s(\hat{\beta}_m, \lambda) = J(\hat{\beta}_m - \text{sgn}(\hat{\beta}_m) \cdot \lambda) \). In either case, we observe that the properties of \( \hat{\eta}(x;m) = q^T(x)\delta(\hat{\beta}_m, \lambda) \) will be more difficult to study since \( \delta(\hat{\beta}_m, \lambda) \) is a nonlinear function of \( \hat{\beta}_m \). We however note that since thresholding is carried out after the parameter vector \( \beta_m \) has been estimated, the threshold estimators will have no influence on the design for estimating \( \beta_m \). Therefore our design criteria will be functions of the MSE of \( \hat{\beta}_m \) and not of \( \delta(\hat{\beta}_m, \lambda) \).

It is clear that if the wavelet terms which have not been used in the actual approximation of the mean response function \( \eta(x) \) denoted by \( f(x) \) is too large, so also will the bias in (2.4) and the wavelet representation will be completely wrong. Thus, we impose a bound,

\[
\frac{1}{N} \sum_{i=1}^{N} f^2(x_i) \leq \tau^2, \tag{2.5}
\]

on \( f(x) \) for a known constant \( \tau \). In what follows, we shall see that our designs will not depend on \( \tau \) but on a parameter \( \nu = \sigma^2/(n\tau^2) \) which can be interpreted as a measure of the relative importance of bias versus variance to the experimenter. As \( \nu \to \infty \), the variance component of the MSE become dominant. Our study show that the optimal wavelet design points for different values of \( \nu \) are the same and differ only in the number of observations allocated to each design point (see §3.3). Therefore, the choice of \( \nu \) should not be a source of difficulty.

We end this section by establishing some notations we shall use throughout this paper. Let \( p_i = n_i/n \) and \( m_i = p_iw_i \). Define \( f = (f(x_1), f(x_2), \ldots, f(x_N))^T \), the \( N \times N \) matrices \( W = \text{diag}(w_1, \ldots, w_N) \) and \( M = \text{diag}(m_1, \ldots, m_N) \), the \( N \times 2^m \) matrices \( Q_* = [q_{(m-1)}(x_1), q_{(m-1)}(x_2), \ldots, q_{(m-1)}(x_N)]^T \), \( Z = (z(x_1), z(x_2), \ldots, z(x_N))^T \) and the \( N \times 2^m+1 \) matrix \( Q = (Q_* \mid Z) \).

Then, in terms of \( B = Q^TMQ \), \( b = Q^Tf \), \( D = Q^TMWQ \), and \( H = B^TDB^{-1} \), we find that \( d = \text{bias} \hat{\beta}_m = B^{-1}b \) and \( K = \text{cov} \hat{\beta}_m = \sigma^2 H^{-1} \). Thus, the MSE of \( \hat{\beta}_m \) in an \( m \)th order model is \( M(\xi_N, f) = B^{-1}bb^TB^{-1} + \sigma^2 H^{-1}/n \), where \( \xi_N \) is the \( N \)-point design

\[
\xi_N = \frac{x_1, x_2, \ldots, x_N}{p_1, p_2, \ldots, p_N}.
\]

Similarly, we can write the MSE of \( \hat{\beta}_{(m-1)} \) in an \((m-1)\)th order model as \( M_*(\xi_N, f_*) = B_*^{-1}b_*b_*^TB_*^{-1} + \sigma^2 H_*^{-1}/n \), where \( f_* \), \( b_* \), \( B_* \) and \( H_* \) are the equivalents of \( f \), \( b \), \( B \) and \( H \). Throughout this paper,
matrices or any parameter obtained from the \((m - 1)\)th model will have a "*" subscript. Such a matrix or parameter will correspond to the same matrix or parameter without an asterisk obtained from the \(m\)th order model.

We assume that \(Q\), \(Q\) and \(Z\) are of full rank, and define the singular value decomposition of \(Q\) by \(Q = U_{N \times r} \Lambda_{r \times r} V_{r \times r}^T\), where \(U^T U = V^T V = I_r\) and \(\Lambda\) is a diagonal matrix of singular values \(\lambda_i(Q) (i = 1, \ldots, r)\) of \(Q\). Observe from (2.3) that \(V A U^T f = 0\). This implies that \(f\) belongs to the vector space \([col(U)]^\perp\) which is the orthogonal complement of the column space of \(U\). Now, let \(\tilde{U}_{N \times N - r}\) be a matrix whose columns form an orthogonal basis of \([col(U)]^\perp\). Then, \(U = [U : \tilde{U}]\) satisfies \(U U^T = UU^T + \tilde{U}\tilde{U}^T = I_N\), and any vector \(f\) satisfying (2.3) and (2.5) is representable as \(f = \alpha \tilde{U} e\), where \(||e|| = 1\) and \(\alpha\) is a normalizing constant. The choice \(\alpha = \tau \sqrt{N}\) ensures equality in (2.3). That is,

\[
f = \tau \sqrt{N} \tilde{U} e, \quad ||e|| = 1. \tag{2.6}
\]

3. NONSEQUENTIAL DESIGNS

Clearly, in order to compare the \((m - 1)\)th order wavelet model to the \(m\)th order model, we have to fit (2.2), without the term \(f(x)\) and then test the null hypothesis \(H_0: \beta_{2m + 1} = \beta_{2m + 2} = \cdots = \beta_{2m + 1} = 0\). Let \(O_{2m}\) be the \(2m \times 2m\) zero matrix and \(I_{2m}\) be the \(2m \times 2m\) identity matrix. Now, by defining the \(2m \times 2m + 1\) partitioned matrix \(C = (O_{2m} \mid I_{2m})\) we can write \(H_0\) as \(H_0: \gamma = C\beta_m = 0\), where \(0\) is the \(2m \times 1\) zero vector. By fitting (2.2) without the term \(f(x)\), we have assumed that the \(m\)th order wavelet approximation is exact, when in fact it is not. Under this assumption, the non-centrality parameter for the F-test under the alternative hypothesis can be written as (see Pukelsheim 1993, pp. 70)

\[
\Delta(\xi_N) = (C\beta_m)^T [CKC^T]^{-1} (C\beta_m) = \gamma^T [\text{cov}(\gamma)]^{-1} \gamma. \tag{3.1}
\]

Now, the larger the values of \(\Delta(\xi_N)\), the larger values of the \(F\)-statistic we expect and the clearer the test detects a significant deviation from the hypothesis \(H_0\). Thus, in order to detect departures from the \((m - 1)\)th model we will construct designs to maximize \(\Delta(\xi_N)\). Since \(\Delta(\xi_N)\) is a function of the unknown parameter vector \(\beta_m\), an experimenter may wish to define an alternative criterion which does not depend on \(\beta_m\) (see §3.2) or construct the designs sequentially (see §4).

3.1. T-optimality

In view of the fact that (2.1) is only an approximation, we modify the definition of \(\Delta(\xi_N)\) by replacing \(\text{cov}(\gamma)\) in (3.1) by the MSE matrix \(M(\xi_N, f)\) to obtain the T-optimality criterion

\[
\Delta(\xi_N, f) = (C\beta_m)^T [CM(\xi_N, f)C^T]^{-1} (C\beta_m). \tag{3.2}
\]
The dependence of $\Delta(\xi_N, f)$ on $\beta_m$ can be resolved by adopting a Bayesian approach. Thus, we define a prior distribution $p_0(\beta_m)$ on $\beta_m$ (see Atkinson and Fedorov 1975) and then take the expectation $\delta(\xi_N, f) = \int p_0(\beta_m)\Delta(\xi_N, f)d\beta_m$ to eliminate $\beta_m$. Define $\beta_0 = \int \beta_m p_0(\beta_m)d\beta_m$, $\Gamma_0 = C(\beta_m - \beta_0)(\beta_m - \beta_0)^TCP0(\beta_m)d\beta_m$ and simplify $\delta(\xi_N, f)$ to obtain
\[
\delta(\xi_N, f) = \text{tr} \left[CM(\xi_N, f)C^T\right]^{-1}[\Gamma_0 + (C\beta_0)(C\beta_0)^T]^{\frac{a}{b}}.
\] (3.3)

The result (3.4) which is now independent of $\beta_0$ follow directly from Beckenbach and Bellman (1965) and is obtained by minimizing (3.3) with respect to $\beta_0$.

**Lemma 3.1.** Let $\delta(\xi_N, f)$ be defined by (3.3). Then,
\[
\min_{\beta_0} \delta(\xi_N, f) = \kappa^{2-(m+1)}|CM(\xi_N, f)C^T|^{-2-(m+1)},
\] (3.4)
where $\kappa = |\Gamma_0|$ and $\cdot$ is the determinant function.

It is clear, from Lemma 3.1 that the design $\xi_N^*$ which maximizes $\min_{\beta_0} \delta(\xi_N, f)$ is equivalent to the design that minimizes $|CM(\xi_N, f)C^T|$. We now apply the minimax approach since $|CM(\xi_N, f)C^T|$ is a function of $f$. That is, the minimax T-optimal design is a solution to $\min_{\xi_N} \max_f |CM(\xi_N, f)C^T|$, for every $f(x)$ satisfying (2.6). Now, $|CM(\xi_N, f)C^T| = \frac{\sigma^2}{n} 2^{(m+1)}|\mathcal{H}|^{\frac{1}{2}} 1 + \frac{\nu}{\sigma^2}b^TDb$, where $\mathcal{H} = \mathbf{C}^{-1}\mathbf{C}^T$ and $\mathcal{D} = \mathbf{B}^{-1}\mathbf{C}^T\mathbf{H}^{-1}\mathbf{C}\mathbf{B}^{-1}$. Define $\mathbf{M}_w = \mathbf{U}^T\mathbf{M}\mathbf{W}\mathbf{U}$, $\mathbf{M}_l = \mathbf{U}^T\mathbf{M}\mathbf{U}$, for $l = 1, 2$ and $\mathbf{F} = \Lambda^{-1}\mathbf{V}^{-1}\mathbf{C}^T$. By using the expression $\mathbf{Q} = \mathbf{U}\Lambda\mathbf{V}^T$ from 8.2.1, and (2.6), we can rewrite $\mathcal{H}$ and $b^TDb$ as
\[
\mathcal{H} = \mathbf{F}^T\mathbf{M}_1^{-1}\mathbf{M}_w\mathbf{M}_1^{-1}\mathbf{F} \quad \text{and} \quad b^TDb = \tau^2\mathbf{N}\mathbf{e}^T\tilde{\mathbf{U}}^T\mathbf{M}\mathbf{U}\mathbf{M}_{1}^{-1}\mathbf{F}\mathbf{H}^{-1}\mathbf{F}^T\mathbf{M}_{1}^{-1}\mathbf{U}^T\tilde{\mathbf{M}}\tilde{\mathbf{U}}e. \tag{3.5}
\]

The above simplifications lead to the result outlined in Theorem 3.1. See the appendix for an outline of the proof.

**Theorem 3.1.** Let $\mathbf{E} = \mathcal{H}^{-1/2}$ and $\mathbf{G} = \|\mathbf{M}\mathbf{U}_1^{-1}\mathbf{F}\mathbf{E}\|^2 - \|\mathbf{F}\mathbf{E}\|^2$, where $\|\mathbf{A}\|^2 = \mathbf{A}^T\mathbf{A}$. Then, for fixed $\nu = \frac{\sigma^2}{n\tau^2}$,
\[
\max_{\|\mathbf{e}\|=1} |CM(\xi_N, \mathbf{e})C^T| = \frac{1}{\sigma^2} \frac{\|\mathbf{M}\|_{\mathcal{F}}}{n} \|\mathcal{H}\|^{\frac{1}{2}} 1 + \nu^{-1}N\lambda_{\max}(\mathbf{G}) \tag{3.6}
\]

Any $N$-point design $\xi_N^*$ which minimizes (3.6) will be called a minimax Bayesian T-optimal design. We note that Theorem 3.1 can only be used to construct designs for discriminating between two nested wavelet models. The problem is to construct designs for discriminating between $k > 2$ nested wavelet models of orders $m_1, m_2, \ldots, m_k$ where $m_1 < m_2 < \cdots < m_k$. Let $\mathbf{C}_j = \bigcup_{j=1}^{k-1} \mathbf{C}_j \mathcal{M}(\xi_N, f)\mathbf{C}_j^T$. Then, following Lauter (1974) the design can be chosen to maximize $\mathbf{P}_{j=1}^{k-1} |\mathbf{C}_j \mathcal{M}(\xi_N, f)\mathbf{C}_j^T|^{-2m_1 + 1}$ or equivalently to minimize $\mathbf{P}_{j=1}^{k-1} |\mathbf{C}_j \mathcal{M}(\xi_N, f)\mathbf{C}_j^T|$ in this case, expression (3.6) becomes $\mathbf{P}_{j=1}^{k-1} |\mathbf{C}_j \mathcal{M}(\xi_N, f)\mathbf{C}_j^T|^{-2m_1 + 1/2} \{1 + \nu^{-1}N\lambda_{\max}(\mathbf{G}_j)\}$. 

7
3.2. D-ratio Optimal Designs

Some experimenters may object to using T-optimal designs due to the dependence of the criterion on the unknown parameter vector $\beta_m$. The experimenter then has to consider designs constructed under an alternative criterion such as the D-optimality criterion. We recall that under ordinary least squares, $\|\text{cov}(\hat{\gamma})\|^{-1} = \frac{f}{|C K C^T|} = |Z^T M Z - Z^T M Q_s Q_s^T M Z|^{-1} = \frac{|Q_s^T M Q_s|}{|Z^T M Q_s|}$. That is, $\|\text{cov}(\hat{\gamma})\|^{-1} = \frac{|\text{cov}(\hat{\beta}_m)|}{|\text{cov}(\hat{\beta}_{m-1})|}$. Thus, designs which minimize $\|\text{cov}(\hat{\gamma})\|$ will also minimize $|\text{cov}(\hat{\beta}_m)|/|\text{cov}(\hat{\beta}_{m-1})|$. It is clear that $|\text{cov}(\hat{\beta}_m)|$ is the classical D-optimality criterion under the full $m$-th order model and $|\text{cov}(\hat{\beta}_{m-1})|$ is the D-optimality criterion when $H_0$ is true. Thus, the ratio $|\text{cov}(\hat{\beta}_m)|/|\text{cov}(\hat{\beta}_{m-1})|$ is in some ways similar to a likelihood ratio test statistic $H_0$. Now, due to the uncertainty in the true structure of the mean response, we prefer to use the mean squared error matrix in place of the covariance matrix. Consequently, we define the D-ratio optimality criterion as

$$
\mathcal{L}(\xi_N) = \max_f \left| \frac{M(\xi_N, f)}{M_s(\xi_N, f)} \right|^{1/2m+1}, \quad (3.7)
$$

and any $N$-point design $\xi_N$ which minimizes $\mathcal{L}(\xi_N)$ will be said to be a D-ratio optimal design.

We have not used the term standardized or D-efficient to refer to the criterion (3.7) because these terms have been used by Dette (1997) and Pukelsheim and Rosenberger (1993) to define similar but not exactly the same criterion.

**Theorem 3.2.** Let $M_l$, $l = 1, 2$ be as defined in Theorem 3.1 and $\mathcal{L}(\xi_N)$ be defined by (3.7). Let $G = M_w^{-1/2} (M_2 - M_1^2) M_w^{-1/2}$ and $G_s = M_w^{-1/2} (M_2 - M_1^2) M_w^{-1/2}$. Then, for fixed $\nu = \frac{\sigma^2}{n \tau^2}$ and $\nu_s = \frac{\sigma^2}{n \tau_s^2}$, an $N$-point standardized D-optimal design is any design $\xi_N^*$ which minimizes

$$
\mathcal{L}(\xi_N) = \frac{\mu}{\sigma^2} \frac{n}{\nu} \frac{n}{\nu_s} \frac{|1 + (N/\nu) \lambda_{max}(G)|^{1/2m+1}}{|1 + (N/\nu_s) \lambda_{max}(G_s)|^{1/2m}} \cdot \frac{||H_s||^{1/2m+1}}{||H||^{1/2m+1}}. \quad (3.8)
$$

We note that the relative importance of bias versus variance to the experimenter does not depend on whether the $m$-th order model is selected or whether the $(m-1)$-th order model is best. It is therefore reasonable to take $\nu = \nu_s$ when constructing designs (see §3.3 and §4). Interested readers should refer to the proof of Theorem 3.1 for the technique for deriving the terms containing $\lambda_{max}(G)$ and $\lambda_{max}(G_s)$.

3.3. Minimax T- and D-ratio Optimal Designs

In this section, we discuss the simulated annealing algorithm and use it to construct $N$-point minimax robust T-optimal and D-ratio optimal designs under Daubechies and multiwavelet models.
the chance of accepting a move that will increase the loss function becomes vanishingly small.

The probability of accepting a move that will increase the loss function is initially high. The value of the parameter $\nu$ rather more rapidly, a local minimum energy state with higher energy may be found instead. Certain to temperature decreases they gradually lose mobility and form pure crystals which is in fact a state of global minimum energy. And as long as the temperature is decreased slowly, nature is almost certain to find this state of global minimum energy. If however, the temperature is decreased rather more rapidly, a local minimum energy state with higher energy may be found instead.

Similar to the temperature in the cooling process in thermodynamics, the simulated annealing algorithm we have used involves a parameter $T$ which is initially set at a high value in order that the probability of accepting a move that will increase the loss function is initially high. The value of $T$ is then gradually decreased as the search progresses. In the end, $T$ will be so small that the chance of accepting a move that will increase the loss function becomes vanishingly small.

Figure 1: $N$-point T-optimal designs for multiwavelet models with: (a) $N = 96$, $n = 48$, $\nu = 0.5$, $m = 3$; (b) $N = 96$, $n = 48$, $\nu = 5$, $m = 3$; (c) $N = 32$, $n = 64$, $\nu = 5$, $m = 2$; (d) $N = 32$, $n = 64$, $\nu = 10$, $m = 2$. $N$-point D-ratio optimal designs for multiwavelet models for fixed $\nu_\ast = \nu$ with: (e) $N = 96$, $n = 48$, $\nu = 0.5$, $m = 3$; (f) $N = 32$, $n = 64$, $\nu = 5$, $m = 2$. Path of loss function for accepted configuration during search for: (g) the T-optimal design in (a); (h) the D-ratio optimal design in (f).
This allows for occasional uphill excursions or jumps which makes it possible for the algorithm to hop out of local minima and finally find the global minimum. Generally, a simulated annealing algorithm consists of the following steps.

**Step 1.** Specify the starting number of observations at each of the \( N \) design points, \( \mathbf{n} = (n_1, \ldots, n_N) \) called the initial configuration.

**Step 2.** Specify a scheme by which subsequent configurations are randomly generated.

**Step 3.** Specify a criterion, which should be based on the loss function, for rejecting or accepting a new configuration and for the gradual reduction of the parameter \( T \).

The minimax design is found by iterating Steps 2 and 3 to convergence. Details of the simulated annealing algorithm we have used can be found in Fang and Wiens (2000). Clearly, the algorithm requires the design space to be finite. Thus, in our examples, we will construct a finite but dense and arbitrarily large design space consisting of the evenly spaced points \( x_i = (2i - 1)/2N, \ i = 1, 2, \ldots, N, \) from the \([0, 1]\) interval. It is necessary to make the design space dense and arbitrarily large in order to eliminate any practical restrictions that may arise as a result of finiteness.

**Designs For Multiwavelet Models**

For the purpose of illustration we have constructed four minimax T-optimal and two minimax D-ratio optimal \( N \)-point designs for discrimination when the model (2.1) is based on the multiwavelet system and the model parameters are to be estimated by (2.4). For both the T-optimal and D-ratio optimal designs, we considered the model (2.1) with (a) \( N = 96, \ n = 48, \ m = 3, \) and (b) \( N = 32, \ n = 64, \ m = 2, \) for fixed values of \( \nu \). In case (a), the designs were constructed for \( \nu = 0.5 \) and \( \nu = 5 \); whereas designs for case (b) were constructed for \( \nu = 5 \) and \( \nu = 10 \) as shown in Figure 1. Optimal frequencies \( n_i (i = 1, \ldots, N) \) in Figure 1 are indicated on the y-axis and points on the x-axis are the design points. One feature common to all the designs in Figure 1 is symmetry about the midpoint \( x = 0.5 \). For example, the T-optimal design shown in Figure 1(a) for \( N = 96, \ n = 48, \ m = 3, \) and \( \nu = 0.5 \) recommends taking 24 observations at 16 distinct points in the interval \([0, 0.5]\). The optimal design places 3 observations at the point 0.4427; 2 observations each at the points 0.057, 0.068, 0.182, 0.307, 0.318 and 0.432; and 1 each at 0.005, 0.12, 0.13, 0.193, 0.245, 0.255, 0.37, 0.38 and 0.495 (rounded to three decimal places). The design points in the interval \([0.5, 1]\) can then be generated by using the symmetric property of the designs.
Figure 2: $N$-point T-optimal designs for Daubechies wavelet models with: (a) $N = 96$, $n = 48$, $
u = 0.5$, $m = 3$; (b) $N = 32$, $n = 64$, $
u = 5$, $m = 2$. $N$-point D-ratio optimal designs for Daubechies wavelet models for fixed $\nu_r = \nu$ with: (c) $N = 96$, $n = 48$, $\nu = 0.5$, $m = 3$; (d) $N = 32$, $n = 64$, $\nu = 5$, $m = 2$.

In addition, we observe that the design points for the T-optimal designs in Figures 1(a) and 1(b) for the model with $N = 96$, $n = 48$ and $m = 3$ are exactly the same. They differ only in the number of observations $n_i$ assigned to each of the optimal design points. Similarly, we find that the design points in the optimal D-ratio designs shown in Figures 1(c) and 1(d) for the model with $N = 32$, $n = 64$ and $m = 2$ are the same except in the number of observations $n_i$ assigned to each of the optimal design points. This is an indication that the effect of the parameter $\nu$ on the designs is very minimal. The optimal design points for the two models are however completely different. This is also the case with the optimal D-ratio designs in Figures 1(e) and 1(f) since the designs are for models with $N = 96$, $n = 48$, $m = 3$ and $N = 32$, $n = 64$, $m = 2$ respectively.

We display in Figures 1(g) and 1(h) two typical paths of the loss functions (3.6) and (3.8), respectively, for accepted configurations as we gradually reduce the value of $T$ in the Splus code used in implementing the simulated annealing algorithm. Our experience show that if the initial value of $T$ in the algorithm is not sufficiently large, the result of implementing the algorithm will be a local minimum. Thus, for each model and for each fixed value of $\nu$ considered, we implemented the algorithm for various values of $T$, in order to be certain that a global optimum has been found.
Designs For Daubechies Wavelet Models

The optimal designs for discrimination when the model (2.1) is based on the Daubechies wavelet system is different from that of the multiwavelet system in the sense that the designs are not symmetric. We considered the Daubechies wavelet model with (a) $N = 96$, $n = 48$, $m = 3$ and $\nu = 0.5$ (b) $N = 32$, $n = 64$, $m = 2$ and $\nu = 5$. We constructed one T-optimal and one optimal D-ratio design for case (a) and one each for case (b) as shown in Figures 2(a) - 2(d). Just as we noted in the multiwavelet case, the design points in the T-optimal designs in Figure 2(a) and 2(b) are distinct because the models considered are not the same. The same situation holds for the design points in the optimal D-ratio designs in Figures 2(c) and 2(d). The T-optimal design in Figure 2(a) allocates the 48 observations to 44 distinct points, whereas the 64 observations are assigned to 16 of the 32 design points in the T-optimal design in Figure 2(b). For the optimal D-ratio designs in Figures 2(c) and 2(d), observations are assigned to 35 and 19 design points, respectively.

4. SEQUENTIAL DESIGNS

In this section we discuss an iterative procedure for constructing designs in which the optimal design is built up one trial at a time. Properties of such sequential designs for linear and nonlinear models have been studied extensively by several authors and their results properly documented in a large body of literature. Of particular interest to us are sequential designs for discrimination. Some authors that have used sequential procedures for constructing designs for discrimination are Atkinson and Cox (1974) and Tsitovich (1984).

We consider the situation where an analysis based on $n$ observations made on the basis of a non-sequential design leads to an inconclusive result as to which of the rival models is better. It then becomes necessary for the experimenter to perform further experimental runs in order to obtain additional data. If the experiments are run in sequence, the experimenter has the chance to evaluate each result before continuing or terminating further experimentation. In addition, computations prior to each new run can indicate where the next experiment ought to be conducted to provide maximum discrimination between competing models. Clearly, the procedures in §3 are not directly adaptable to be sequential. We therefore consider a new approach for constructing the designs in this section.

Suppose the wavelet based model of order $m - 1$ is better than the model of order $m$, one then expects that $q_m^T(x)M(\xi_N, f)q_m(x) \geq q_{m-1}^T(x)M_*(\xi_N, f_*)q_{m-1}(x)$. Thus, our designs will be choosen
to maximize the criterion \( d^{(n)}(x; \xi_N) = \max_{\nu} q_m^T(x) M(\xi_N, f) q_m(x) - \max_{\nu} q_{m-1}^T(x) M_*(\xi_N, f_*) q_{m-1}(x). \)

This criterion is similar to the G-optimality criterion used by Atkinson and Cox (1974, pp. 324) for iterative construction of designs. Now, by using the notations defined in §2.1 and the result (2.6), it can be shown that

\[
q_m^T(x) B^{-1} b b^T B^{-1} q_m(x) = \tau^2 N e^T \Theta^T a(x) a^T(x) \Theta e,
\]

where \( a(x) = M M_1^{-1} \Lambda^{-1} V^{-1} q_m(x). \) Define \( \kappa(x) = \|a(x)\|^2 - \|U^T a(x)\|^2. \) It then follows from \( \Theta \Theta^T = I - UU^T, \) that \( \max_{\nu} q_m^T(x) B^{-1} b b^T B^{-1} q_m(x) = \tau^2 N \lambda_{\max} \Theta^T a(x) a^T(x) \Theta = \tau^2 N \kappa(x). \)

If we define \( A = M_{\nu}^{1/2} M_1^1 \Lambda^{-1} V^{-1}, \) then \( \frac{\sigma^2}{n} q_m^T(x) H^{-1} q_m(x) = \frac{\sigma^2}{n} \|A q_m(x)\|^2. \) Let \( a_*(x), \kappa_*(x) \) and \( A_\ast \) correspond to \( a(x), \kappa(x) \) and \( A, \) respectively when \( M(\xi_N, f) \) is replaced by \( M_*(\xi_N, f_*) \) and \( q_m(x) \) is replaced by \( q_{m-1}(x). \) Using the results obtained, we now provide a simplified version of \( d^{(n)}(x; \xi_N) \) in Theorem 4.1.

**Theorem 4.1.** Let \( \kappa(x), \kappa_*(x), A \) and \( A_\ast \) be as defined above. Then, for fixed \( \nu = \frac{\sigma^2}{n \tau^2} \) and \( \nu_\ast = \frac{\sigma^2}{n \tau^2}, \)

\[
d^{(n)}(x; \xi_N) = \frac{\sigma^2}{n} \mathcal{Q} \nu^{-1} N \kappa(x) + \|A q_m(x)\|^2 \mathfrak{q} - \nu^{-1} N \kappa_*(x) + \|A_\ast q_{m-1}(x)\|^2 \mathfrak{q}. \quad (4.1)
\]

To construct sequential designs iteratively, we now proceed one trial at a time to add the \((n+1)\)-th observation at the design point at which (4.1) is maximized.

**Examples**

In our examples, we use the T-optimal and optimal D-ratio designs for Daubechies wavelet models with \( N = 32, n = 64 \) \( \nu = 5, m = 2 \) shown in Figure 2(b) and \( N = 96, n = 48 \) \( \nu = 0.5, m = 3 \) shown in Figure 2(c), respectively, as starting designs for constructing sequential designs for the Daubechies wavelet models. We seek to add 16 observations, one trial at a time, to the \( n \) observations in the starting design. In order to study the effect of \( \nu \) on the sequential designs, we used these starting designs with several values of \( \nu \) and found that the sequentially optimal designs remained unchanged in the cases we considered.

**Example 1:** In this example, the starting design (correct to three decimal places) with their corresponding number of observations in parenthesis is as follows: \( 0.047(7), 0.078(1), 0.172(3), 0.203(6), 0.297(7), 0.328(2), 0.422(4), 0.453(4), 0.547(7), 0.672(2), 0.703(5), 0.734(1), 0.797(7), 0.828(1), 0.922(6), 0.953(1) \). We found that 14 of the 16 additional observations were allocated to 9 of the 16 initial design points and the remaining 2 observations were assigned to the design point 0.984 at the 13th and 15th trial. Of the 14 observations, 3 additional observations were allocated to the point 0.047, 2 each to the points 0.203, 0.422 and 0.734 and 1 each to the points
Example 2: The starting design in this example consists of 48 observations allocated to 35 distinct design points. Following the sequential procedure, 6 of the 16 observations were allocated to design points in the starting design, whereas the remaining 10 observations were assigned to 7 points which are not in the starting design. The new design points and their assigned number of observations, in parenthesis, are 0.151(1), 0.172(1), 0.401(1), 0.411(1), 0.526(3), 0.599(1), and 0.786(2). The design points in the initial design to which 6 additional observations were allocated are 0.026(1), 0.089(1), 0.276(1), 0.339(2), and 0.474(1).

We note that the sequential procedure discussed in this section will lead to designs for discrimination which may not be optimal for model parameter estimation. If the objective is the construction of designs for simultaneous parameter estimation and model discrimination, one may adopt the multi-stage sequential procedure of Biswas and Chaudhuri (2002).

5. CONCLUDING REMARKS

In this paper, we have discussed sequential and nonsequential designs that will assist in determining the appropriate number of wavelet terms that should be used in approximating a nonparametric response curve. Several authors have shown that, in wavelet models the uniform design is optimal only when the Haar wavelet is used in defining the model. See for instance Herzberg and Traves (1994) and Oyet and Wiens (2000). Consequently, a comparison between the designs obtained in this paper and the uniform design was considered unnecessary. Our results show that symmetry appear to be an important property of designs for discriminating between multiwavelet models, whereas designs for Daubechies wavelet models are nonsymmetric. The mathematical developments in this paper is such that the methods can be applied to wavelet models other than those discussed in the examples.

APPENDIX: DERIVATIONS

Proof of Theorem 3.1: Clearly,

\[
\max_{\|e\|=1} |CM(\xi_N,e)C^T| = \frac{\mu}{\sigma^2} \|H\| 2^{(m+1)} \mu 1 + \frac{n}{\sigma^2} \max_{\|e\|=1} b^T \mathcal{D} b . \tag{A.1}
\]

Define \(G = \tilde{U}^T \tilde{U} M \tilde{U}^{-1} \mathcal{H}^{-1} \tilde{U}^T \tilde{U} \tilde{M} \tilde{U} \). Then, using (3.5) and results from quadratic forms we have that \(\max_{\|e\|=1} b^T \mathcal{D} b = \tau^2 N \max_{\|e\|=1} e^T Ge = \tau^2 N \lambda_{\text{max}}(G)\), where \(\lambda_{\text{max}}(G)\) is the
maximum eigenvalue of $\mathbf{G}$. By using matrix properties and the fact that $\tilde{\mathbf{U}}\tilde{\mathbf{U}}^T = \mathbf{I} - \mathbf{U}\mathbf{U}^T$ (see §2.1), we obtain

$$\tau^2 N \lambda_{\text{max}}(\mathbf{G}) = \tau^2 N \lambda_{\text{max}} \mathbf{1}^T \mathbf{M} \mathbf{1}^{-1} \mathbf{F}\mathbf{E} \mathbf{1} - \mathbf{F}\mathbf{E}\mathbf{1}^T,$$

(A.2)

where $\|\mathbf{A}\|^2 = \mathbf{A}^T \mathbf{A}$ and $\mathbf{E} = \mathcal{H}^{-1/2}$. The result (3.6) then follows by substituting (A.2) into (A.1) and also noting that $|\mathcal{H}| = \sum_{i=1}^{2^n} \lambda_i(\mathcal{H})$.

REFERENCES


