Acta Numerica (1998), pp. 337–377

Choice of norms for data fitting and function approximation

G. A. Watson
Department of Mathematics,
University of Dundee,
Dundee DD1 4HN, Scotland
E-mail: gawatson@mcs.dundee.ac.uk

For the approximation of functions and data, it is often appropriate to minimize a norm. Many norms have been considered, and a review is presented of methods for solving a range of problems using a wide variety of norms.

CONTENTS

1. Introduction

Two central problems in approximation theory are the approximation of data and the approximation of functions. Given a set of data or a function for which an approximation is required, two fundamental issues are the kind of approximation which should be chosen, and the measure of goodness of fit which should be used. These will of course depend very much on the precise nature of the underlying problem, and our main interest here is a consideration of methods for obtaining solutions to a wide range of such problems which are of practical importance.

As far as the approximation of functions is concerned, the goal is to replace the original function by one that is simpler or more manageable. The most useful measure is the Chebyshev norm: as Meinardus says in the preface to his book (Meinardus 1967), '... it has by far the greatest practical importance'. There has of course been much interest in the use of the L_2

norm, and one reason for this is that the analysis is relatively straightforward, with linear approximation leading directly to the solution of a linear system of equations. There are obvious advantages if the basis functions are orthogonal or orthonormal, so that this is in some ways a natural measure to use when approximating by orthogonal polynomials, Fourier series or orthogonal wavelets. The emphasis here is on methods for computing approximations, and so we will not pursue this further, but confine attention almost entirely to the Chebyshev norm: some relevant material is covered in Section 5. The L_1 norm has also attracted some interest, and we will deal briefly with it in Section 6.

For the approximation of data the situation is quite different, and there are many criteria that are of practical value. There are two main features of data approximation. Firstly, there is a wide range of characteristics which the data may possess. For example, the data may arise from the sampling of a known function, or they may be observed or measured data for which the underlying form is known: a simple example is data, generated experimentally, which might be expected to lie on a straight line. At the other extreme, the data may be irregularly positioned, with no discernible pattern. Secondly, observed data generally contain errors, and the nature of these is an important consideration in deciding a measure for goodness of fit.

The simplest and most direct way of choosing the unknown parameters in data-fitting problems is by interpolation. This may be appropriate, for example, if the model is linear, there is a large number of parameters, and it is known that a nonsingular system will occur. Other more sophisticated measures may indeed be unsuitable or impracticable because of the sheer size of the problem. This is the case in the approximation of scattered data using radial basis functions. Also, if there is significantly different behaviour of the data in different regions, so that considerable flexibility is required, then spline functions may be appropriate, and again interpolation may be the natural thing to use.

This article is, however, not concerned with interpolation, and thus, in the data-fitting context, it will be assumed that the data can be modelled by a function containing a number of free parameters, and minimizing a norm is appropriate. Perhaps the most commonly occurring criterion in such cases is the least squares norm. Its use has a long and distinguished history, it is relatively well understood, and there are good algorithms available. Yet there are often situations where it is not ideal. For example, a statistical justification for least squares requires certain assumptions about the error pattern in the data, and if these are not satisfied there may be bias in the estimate.

Therefore there are many other norms which are of interest in data fitting, and which have been studied from both a theoretical and a practical point of view. We will use this statement as an excuse for giving in Section 2 a very

general theoretical treatment of the conventional linear problem for arbitrary norms, and we will consider in particular the question of characterization of solutions of such problems. The analysis is of course contained in a treatment of approximation problems in completely general normed linear spaces, and this is well known to approximation theorists. However, whereas that requires some very sophisticated functional analysis, very powerful results can be obtained for the present problem in a comparatively straightforward manner, and this should be accessible to the general readership of this book. The main tool is the subdifferential of the norm, which extends the idea of the derivative to the non-differentiable case. Therefore we will give some attention to this, and then go on to use some of the results in special cases.

In Section 3, an important modification of the usual linear problem is addressed in some (though not complete) generality. In Section 4, nonlinear problems are briefly considered, again with some emphasis on a general treatment.

2. Approximation to data by linear models

Suppose that a relationship exists between variables so that one of the variables can be expressed as a linear combination of n functions of the others. Then, if a set of values of the variables is generated which is assumed to satisfy this relationship, the result is a system of linear equations, say

$$A\mathbf{x} = \mathbf{b}, \tag{2.1}$$

where $\mathbf{x} \in \mathbb{R}^n$ represents the unknown coefficients, $\mathbf{b} \in \mathbb{R}^m$ is a vector of values of the dependent variable, and $A \in \mathbb{R}^{m \times n}$ is formed from the values of the other variables. If the available data are subject to errors, and $m > n$, then (2.1) is an over-determined linear system which normally has no (exact) solution. If it is assumed that the errors are only present in the data values forming the vector $\mathbf{b}$, then we can introduce a vector $\mathbf{r}$ of perturbations of $\mathbf{b}$, representing these errors, so that

$$\mathbf{r} = A\mathbf{x} - \mathbf{b}, \tag{2.2}$$

and choose $\mathbf{x}$ to make the components of $\mathbf{r}$ small in some sense.

The problem of the solution of an overdetermined system of linear equations in this form has attracted enormous interest. Typically, this is done by solving the following problem:

$$\text{find } \mathbf{x} \in \mathbb{R}^n \text{ to minimize } \|\mathbf{r}\|, \tag{2.3}$$

where $\|\cdot\|$ is a given norm on $\mathbb{R}^m$. A solution always exists, and if the norm is differentiable, then we can easily characterize a minimum by zero-derivative conditions. Such conditions are readily extended to the general case through the use of the subdifferential. We will consider this next.

2.1. Characterization of solutions

Let $\|\cdot\|$ be a norm on $\mathbb{R}^m$. Then, for any $\mathbf{v} \in \mathbb{R}^m$, the *dual* norm is the norm on $\mathbb{R}^m$ defined by

$$\|\mathbf{v}\|^* = \max_{\|\mathbf{r}\| \leq 1} \mathbf{r}^T\mathbf{v}. \tag{2.4}$$

The relationship between a norm on $\mathbb{R}^m$ and its dual is symmetric, so that, for any $\mathbf{r} \in \mathbb{R}^m$,

$$\|\mathbf{r}\| = \max_{\|\mathbf{v}\|^* \leq 1} \mathbf{r}^T\mathbf{v}. \tag{2.5}$$

Important special cases are the l_p norms,

$$\begin{aligned} \|\mathbf{r}\|_p &= \left(\sum_{i=1}^m |r_i|^p\right)^{1/p}, \quad 1 \leq p < \infty, \\ \|\mathbf{r}\|_\infty &= \max_{1 \leq i \leq m} |r_i|. \end{aligned}$$

Then the dual norm is the l_q norm, where $1/p + 1/q = 1$.

Definition 1 The *subdifferential*, or set of subgradients of $\|\mathbf{r}\|$, is given by

$$\partial\|\mathbf{r}\| = \{\mathbf{v} \in \mathbb{R}^m : \|\mathbf{s}\| \geq \|\mathbf{r}\| + (\mathbf{s} - \mathbf{r})^T\mathbf{v}, \quad \text{for all } \mathbf{s} \in \mathbb{R}^m\}. \tag{2.6}$$

This set is closed, bounded and convex. It is also easily seen that it is just the set of vectors $\mathbf{v}$ such that equality holds in (2.5): in other words we have the following very useful result.

Lemma 1 Let $\mathbf{r} \in \mathbb{R}^m$. Then

$$\partial\|\mathbf{r}\| = \{\mathbf{v} \in \mathbb{R}^m : \|\mathbf{r}\| = \mathbf{r}^T\mathbf{v}, \ \|\mathbf{v}\|^* \leq 1\}. \tag{2.7}$$

Proof. Let $\mathbf{v}$ be in the set defined by (2.6). Then, setting $\mathbf{s} = 0$, and $\mathbf{s} = 2\mathbf{r}$, it follows that $\|\mathbf{r}\| = \mathbf{r}^T\mathbf{v}$, and further $\|\mathbf{v}\|^* \leq 1$, from the definition. Thus $\mathbf{v}$ lies in the set defined by (2.7). The reverse implication is immediate. □

If $\|\mathbf{r}\|$ is differentiable at $\mathbf{r}$, then the subdifferential is a singleton with

$$\partial\|\mathbf{r}\| = \left\{\mathbf{v} \in \mathbb{R}^m : v_i = \frac{\partial\|\mathbf{r}\|}{\partial r_i}, \quad i = 1, \ldots, m\right\}.$$

This follows from the inequality in (2.6), using convexity. If $\mathbf{r} = 0$, then obviously

$$\partial\|\mathbf{r}\| = \{\mathbf{v} \in \mathbb{R}^m : \|\mathbf{v}\|^* \leq 1\}.$$

If $\mathbf{r} \neq 0$, then it is a consequence of (2.5) that $\|\mathbf{v}\|^* = 1$. For the l_p norms, we have the sets

$$\partial\|\mathbf{r}\|_1 = \{\mathbf{v} \in \mathbb{R}^n : v_i = \operatorname{sign}(r_i), r_i \neq 0; \ |v_i| \leq 1, r_i = 0\}, \tag{2.8}$$

$$\partial\|\mathbf{r}\|_p = \{\mathbf{v} \in \mathbb{R}^n : v_i = \operatorname{sign}(r_i)|r_i|^{p-1}\|\mathbf{r}\|_p^{1-p}\}, \quad 1 < p < \infty, \tag{2.9}$$

and, defining

$$J = \{i : |r_i| = \|\mathbf{r}\|_\infty\},$$

$$\partial\|\mathbf{r}\|_\infty = \left\{\mathbf{v} \in \mathbb{R}^m : \operatorname{sign}(v_i) = \operatorname{sign}(r_i), \quad i \in J; \quad v_i = 0, \quad i \notin J; \right.$$

$$\left. \sum_{i \in J} |v_i| = 1 \right\}. \quad (2.10)$$

The concept of the subdifferential enables us to characterize readily the solutions of linear best approximation problems set in $\mathbb{R}^m$. Two preliminary lemmas are required. Lemma 2 can in fact be given in a much stronger form, but this version is sufficient for our purposes; Lemma 3 gives an expression for the directional derivative of the norm in terms of the subdifferential. With the assumption that the reader knows that the minimum of a continuous function over a closed and bounded set is attained, the rest of this section should be entirely self-contained.

Lemma 2 Let $\mathbf{r}$, $\mathbf{s} \in \mathbb{R}^m$, let $\gamma \in \mathbb{R}$, and let $\mathbf{v}(\gamma) \in \partial\|\mathbf{r} + \gamma\mathbf{s}\|$. Then the limit points of $\mathbf{v}(\gamma)$ as $\gamma \to 0$ lie in $\partial\|\mathbf{r}\|$.

Proof. Clearly $\{\mathbf{v}(\gamma)\}$ has limit points; let one of these be $\mathbf{v}$. Then $\|\mathbf{v}\|^* \leq 1$. Further,

$$\begin{aligned} \mathbf{v}(\gamma)^T\mathbf{r} &= \mathbf{v}(\gamma)^T(\mathbf{r} + \gamma\mathbf{s}) - \gamma\mathbf{v}(\gamma)^T\mathbf{s} \\ &= \|\mathbf{r} + \gamma\mathbf{s}\| - \gamma\mathbf{v}(\gamma)^T\mathbf{s}. \end{aligned}$$

Letting $\gamma \to 0$, the result follows. □

Lemma 3 Let $\mathbf{r}$, $\mathbf{s} \in \mathbb{R}^m$. Then

$$\lim_{\gamma \to 0+} \frac{\|\mathbf{r} + \gamma\mathbf{s}\| - \|\mathbf{r}\|}{\gamma}. \quad (2.11)$$

Proof. For all $\mathbf{v} \in \partial\|\mathbf{r}\|$, using Definition 1,

$$\|\mathbf{r} + \gamma\mathbf{s}\| \geq \|\mathbf{r}\| + \gamma\mathbf{v}^T\mathbf{s}.$$

Also, for all $\mathbf{v}(\gamma) \in \partial\|\mathbf{r} + \gamma\mathbf{s}\|$, again using Definition 1,

$$\|\mathbf{r}\| \geq \|\mathbf{r} + \gamma\mathbf{s}\| - \gamma\mathbf{v}(\gamma)^T\mathbf{s}.$$

Combining these inequalities shows that, for all $\mathbf{v} \in \partial\|\mathbf{r}\|$, $\mathbf{v}(\gamma) \in \partial\|\mathbf{r} + \gamma\mathbf{s}\|$, we have for $\gamma > 0$

$$\mathbf{v}^T\mathbf{s} \leq \frac{\|\mathbf{r} + \gamma\mathbf{s}\| - \|\mathbf{r}\|}{\gamma} \leq \mathbf{v}(\gamma)^T\mathbf{s}.$$

Letting $\gamma \to 0+$ and using Lemma 2, the result follows. □

These results enable a general, and simple, characterization result to be established.

Theorem 1 The problem (2.3) is solved by $\mathbf{x}$, with $\mathbf{r} = A\mathbf{x} - \mathbf{b}$, if and only if there exists $\mathbf{v} \in \partial\|\mathbf{r}\|$ with

$$A^T\mathbf{v} = 0. \tag{2.12}$$

Proof. Let $\mathbf{x}$ be a solution but assume that (2.12) is not satisfied. Consider the problem:

$$\text{find } \mathbf{v} \in \partial\|\mathbf{r}\| \text{ to minimize } \|A^T\mathbf{v}\|_2.$$

Then a solution exists, at $\mathbf{w}$, say. By convexity, $\lambda\mathbf{v} + (1-\lambda)\mathbf{w} \in \partial\|\mathbf{r}\|$, for $0 \leq \lambda \leq 1$, where $\mathbf{v} \in \partial\|\mathbf{r}\|$ is arbitrary. Thus

$$\begin{aligned} 0 &\leq \|A^T\Big(\lambda\mathbf{v} + (1-\lambda)\mathbf{w}\Big)\|_2^2 - \|A^T\mathbf{w}\|_2^2 \\ &= \lambda^2\|A^T(\mathbf{v}-\mathbf{w})\|_2^2 + 2\lambda(\mathbf{v}-\mathbf{w})^TAA^T\mathbf{w}. \end{aligned}$$

The last term on the right-hand side will actually be negative for small positive λ if the coefficient of λ is negative, which would lead to a contradiction, and so

$$\mathbf{v}^TAA^T\mathbf{w} \geq \mathbf{w}^TAA^T\mathbf{w} > 0.$$

Thus, setting $\mathbf{s} = -AA^T\mathbf{w}$ in Lemma 3, and using the fact that $\mathbf{v}$ is arbitrary, contradicts the assumption that $\mathbf{x}$ gives a minimum of the norm.

Now let the conditions hold for some $\mathbf{x} \in \mathbb{R}^n$, with $\mathbf{w} \in \partial\|\mathbf{r}\|$ satisfying (2.12). Then, if $\mathbf{y} \in \mathbb{R}^n$ is arbitrary,

$$\begin{aligned} \|A\mathbf{y} - \mathbf{b}\| &\geq \mathbf{w}^T(A\mathbf{y} - \mathbf{b}) \\ &= \mathbf{w}^T(A\mathbf{x} - \mathbf{b}) \\ &= \|A\mathbf{x} - \mathbf{b}\|. \end{aligned}$$

The proof is completed. □

This result, in conjunction with appropriate sets $\partial\|\mathbf{r}\|$, can be used to obtain specific characterization results.

2.2. l_1 approximation

There is particular interest in approximation using the l_1 norm because it has the property of de-emphasizing the effect of wild points or gross errors in $\mathbf{b}$. We will expand on that after giving a characterization result. For any $\mathbf{x} \in \mathbb{R}^n$, let I denote the set of indices where $r_i = 0$, and let I^c denote its complement. From Theorem 1 and (2.8), we have the following theorem.

Theorem 2 The vector $\mathbf{x}$ is a solution to the l_1 problem if and only if there exists $\boldsymbol{\lambda} \in \mathbb{R}^m$ with

$$\lambda_i = \text{sign}(r_i), \quad i \in I^c, \quad \text{and} \quad |\lambda_i| \leq 1, \quad i \in I,$$

such that

$$A^T \boldsymbol{\lambda} = 0. \tag{2.13}$$

The l_1 problem is said to be *primal nondegenerate* if, at any point $\mathbf{x}$, the rows of A corresponding to $i \in I$ are linearly independent; it is *dual nondegenerate* if, for any $\boldsymbol{\lambda}$ satisfying (2.13), at most $m - n$ components of $\boldsymbol{\lambda}$ are equal to 1 in modulus.

Suppose that, at a solution, one of the values of b_i, $i \in I^c$, is perturbed in such a way that $\text{sign}(r_i)$ is unchanged. Then clearly the characterization conditions are unaffected, so that the solution is unchanged. It is this property of robustness, or insensitivity to possibly large errors in the data, that makes the l_1 norm important.

Because $\|\mathbf{r}\|_1$ is a piecewise linear function, the most commonly used algorithms have been based on movement between intersections of points having sets of zero components of $\mathbf{r}$. In addition to methods based explicitly on linear programming (for example, Barrodale and Roberts 1973), variants based on reduced gradients (Osborne 1987, Shi and Lukas 1996), or projected gradients (Bartels, Conn and Sinclair 1978) are popular: for some relationships, see Osborne (1985). These methods are finite and, with efficient implementation, including line searches, have for a long time appeared to represent the right way to tackle the problem. Recently, however, there has been interest in iterative methods, which effectively smooth the problem. Partly, this has been due to the success of interior point methods for linear programming problems, and attempts have been made to use such ideas in the l_1 problem. We will examine first a method of this type, based on the idea of affine scaling.

Let A have rank n and let the QR decomposition of A be given by

$$A = YR,$$

where R is $n \times n$ upper triangular, and $[Y : Z]$ is an $m \times m$ orthogonal matrix. Then (2.13) is equivalent to

$$\boldsymbol{\lambda} = Z\mathbf{w}, \tag{2.14}$$

where $\mathbf{w} \in \mathbb{R}^{m-n}$ and (2.2) is equivalent to

$$Z^T(\mathbf{r} + \mathbf{b}) = 0. \tag{2.15}$$

Define D_r to be the diagonal matrix with (i,i) element r_i, and let $\mathbf{g} \in \mathbb{R}^m$ be the vector defined by

$$g_i = \begin{cases} \text{sign}(r_i), & i \in I^c, \\ 0, & i \in I. \end{cases}$$

It follows that $\mathbf{x}$ solves the l_1 problem if and only if there exists $\mathbf{r} \in \mathbb{R}^m$, $\mathbf{w} \in \mathbb{R}^{m-n}$, such that

$$D_r(\mathbf{g} - Z\mathbf{w}) = 0 \tag{2.16}$$

$$Z^T(\mathbf{r} + \mathbf{b}) = 0, \tag{2.17}$$

and, in addition,

$$-1 \leq (Z\mathbf{w})_i \leq 1, \quad i = 1, \ldots, m. \tag{2.18}$$

An affine scaling method attempts to solve (2.16), (2.17), (2.18) by an iterative method that computes a direction of progress from the current vector $\mathbf{r}$ by solving the following subproblem:

$$\begin{aligned} &\text{minimize}_{\mathbf{d} \in \mathbb{R}^m} \ \mathbf{g}^T\mathbf{d} \\ &\text{subject to} \ Z^T\mathbf{d} = 0, \\ &\|D^{-1}\mathbf{d}\|_2 \leq \tau, \end{aligned} \tag{2.19}$$

where D is a given positive definite diagonal matrix, and τ is a given positive number that restricts the size of $\mathbf{d}$: (2.19) can be thought of as scaling the solution. It is a straightforward exercise to show that $\mathbf{d}^* = \alpha\mathbf{d}$ is the solution, where α is a suitably chosen scalar and $\mathbf{d}$ is given by

$$\mathbf{d} = -A\left(A^T D^{-2} A\right)^{-1} A^T \mathbf{g}. \tag{2.20}$$

Assume that, at the current $\mathbf{r}$, (2.17) is satisfied (and so remains satisfied for subsequent $\mathbf{r}$), and also $I = \phi$. Then $\mathbf{g}$ is just the gradient of $\|\mathbf{r}\|_1$ at the current point, and it follows that the solution $\mathbf{d}$ is a descent direction for the l_1 norm (since $\mathbf{d} = 0$ implies that $A^T\mathbf{g} = 0$). A line search to minimize the piecewise linear function $\|\mathbf{r} + \alpha\mathbf{d}\|$ with respect to α can readily be incorporated, and, if we stop short of the optimal step length, then we can start again from a point with $I = \phi$. Note that no explicit value of τ is actually required.

Now consider the alternative of applying Newton's method to (2.16) and (2.17). The Newton step in $\mathbf{r}$ and $\mathbf{w}$ is given by solving the system

$$\begin{bmatrix} D_\beta & -D_r Z \\ Z^T & 0 \end{bmatrix} \begin{bmatrix} \boldsymbol{\delta}\mathbf{r} \\ \boldsymbol{\delta}\mathbf{w} \end{bmatrix} = \begin{bmatrix} -D_r(\mathbf{g} - Z\mathbf{w}) \\ 0 \end{bmatrix}, \tag{2.21}$$

where $D_\beta = \text{diag}\{\beta_i, \ldots, \beta_m\}$, with $\beta_i = g_i - (Z\mathbf{w})_i$, $i = 1, \ldots, m$. It follows from this system that

$$\boldsymbol{\delta}\mathbf{r} = -A\Big(A^T D_r^{-1} D_\beta A\Big)^{-1} A^T \mathbf{g}. \tag{2.22}$$

Note that (2.20) and (2.22) have the same general form. Let D be chosen by

$$D = \text{diag}\{|r_i|^{1/2}, \quad i = 1, \ldots, n\},$$

and write

$$\boldsymbol{\delta}_\theta \mathbf{r} = -A(A^T W_\theta A)^{-1} A^T \mathbf{g}, \tag{2.23}$$

where

$$W_\theta = \text{diag}\{|r_i^{-1}(g_i - (1-\theta)(Z\mathbf{w})_i)|, \quad i = 1, \ldots, m\}.$$

Then the choice $\theta = 1$ in (2.23) gives (2.20), and the choice $\theta = 0$ gives (2.22), provided that $D_r^{-1} D_\beta$ is positive definite. That this is true for a set of points $\mathbf{r}$ arbitrarily close to a solution but having no component of $\mathbf{r}$ zero is a key observation in the hybrid method of Coleman and Li (1992*a*). By providing a criterion for choosing θ so that $\theta \to 0$ as $(\mathbf{r}, \boldsymbol{\lambda})$ tend to optimal values, and working with (2.23), they develop a method which is globally convergent to a solution, with a quadratic convergence rate, if the l_1 problem is primal and dual nondegenerate. The full Newton step is not taken asymptotically because of the damping required to maintain differentiability; however, sufficiently accurate approximations to the full Newton step are achieved to permit quadratic convergence.

Note that (2.23) may be solved by first calculating the l_2 solution of the system

$$W_\theta^{1/2} A \, \mathbf{d}_1 = W_\theta^{-1/2} \mathbf{g},$$

followed by setting

$$\boldsymbol{\delta}_\theta \mathbf{r} = -A\mathbf{d}_1. \tag{2.24}$$

The new value of $\mathbf{r}$ is then obtained by a line search in the direction $\delta_\theta \mathbf{r}$. To obtain a new value of $\boldsymbol{\lambda}$, it helps to observe that use of the hybrid method corresponds to solving the system analogous to (2.21), namely

$$\begin{bmatrix} W_\theta & -Z \\ Z^T & 0 \end{bmatrix} \begin{bmatrix} \boldsymbol{\delta}_\theta \mathbf{r} \\ \boldsymbol{\delta}\mathbf{w} \end{bmatrix} = \begin{bmatrix} -(\mathbf{g} - Z\mathbf{w}) \\ 0 \end{bmatrix}. \tag{2.25}$$

It follows from this that the current value of $\boldsymbol{\lambda}$ can be updated to the value

$$Z(\mathbf{w} + \boldsymbol{\delta}\mathbf{w}) = \mathbf{g} + W_\theta \boldsymbol{\delta}_\theta \mathbf{r},$$

with $\delta_\theta \mathbf{r}$ given by (2.24). Of course, (2.13) remains satisfied. Although Z appears as an aid to the theoretical development of the method, note that its actual computation is unnecessary, and we need only work with $\mathbf{r}$ and $\boldsymbol{\lambda}$.

An initial approximation may be obtained by choosing the $\mathbf{r}$ satisfying (2.2) of minimum l_2 norm, and taking $\boldsymbol{\lambda}$ to be a multiple of that $\mathbf{r}$.

It may appear that difficulties are inevitable in practice as components of $\mathbf{r}$ tend to zero. However, Coleman and Li (1992*a*) show that theoretically this is not a problem, and neither is it in practice if the method is implemented carefully. They give numerical results which show improvement over an earlier attempt to provide a method based on an interior point approach. Although comparisons with other methods do not seem to be available, the affine scaling method seems promising for large problems, since it appears to be insensitive to problem size.

Another method that attempts to smooth the l_1 problem has been developed recently by Madsen and Nielsen (1993). It is based on the use of the Huber M-estimator, defined by

$$\psi_\gamma \equiv \psi_\gamma(\mathbf{r}) = \sum_{i=1}^{m} \rho(r_i), \tag{2.26}$$

where

$$\rho(t) = \begin{cases} t^2/2, & |t| \leq \gamma, \\ \gamma(|t| - \gamma/2), & |t| > \gamma, \end{cases} \tag{2.27}$$

$\mathbf{r}$ is given by (2.2) and γ is a scale factor or tuning constant. The function (2.26) is convex and once continuously differentiable, but has discontinuous second derivatives at points where $|r_i| = \gamma$. The mathematical structure of the Huber M-estimator is considered by Clark (1985). Clearly, if γ is chosen large enough, then ψ_γ is just the least squares function; in addition, if γ tends to zero, then limit points of the set of solutions minimize the l_1 norm (see Theorem 3). It is the latter property that concerns us here. It has been suggested by Madsen and Nielsen (1993) and also Li and Swetits (1998) that the preferred method for solving the l_1 problem is via a sequence of Huber problems for a sequence of scale values $\gamma \to 0$. This algorithmic development has led to increased interest in the relationship between the Huber M-estimator and the l_1 problem; see, for example, Madsen, Nielsen and Pinar (1994), Li and Swetits (1998).

Let a partition be defined by an index set σ and its complement σ^c as follows:

$$\sigma = \{i : |r_i| \leq \gamma\}, \qquad \sigma \cup \sigma^c = \{1, 2, \ldots, m\}. \tag{2.28}$$

Then the system of equations determined by the necessary conditions for $\mathbf{x}$ to be a solution is

$$A_\sigma^T A_\sigma \mathbf{x} = A_\sigma^T \mathbf{b}_\sigma - \gamma \sum_{i \in \sigma^c} g_i \mathbf{a}_i, \tag{2.29}$$

where $\mathbf{a}_i^T$ denotes the ith row of A, $g_i = \text{sign}(r_i)$ as before, A_σ is obtained from A by deleting rows corresponding to indices $i \in \sigma^c$, and $\mathbf{b}_\sigma$ is defined similarly.

For given γ, the Huber problem can be solved by Newton's method, or a variant, using a line search. There are other possibilities, and a comparison of eight algorithms for this problem (*inter alia*) is given by Ekblom and Nielsen (1996). Algorithms based on continuation are also given by Clark and Osborne (1986), Boncelet and Dickinson (1984). For given $\mathbf{x} \in \mathbb{R}^n$, define W as a diagonal matrix with elements 1 if $|r_i| \leq \gamma$ and 0 otherwise. Then, assuming that no value of $|r_i|$ is equal to γ, and letting $\mathbf{s} \in \mathbb{R}^m$ be defined by

$$s_i = \begin{cases} 0, & i \in \sigma, \\ \text{sign}(r_i), & i \in \sigma^c, \end{cases}$$

it is easily seen by differentiating the Huber function that ψ_γ is minimized if and only if

$$A^T \left[\frac{1}{\gamma} W\mathbf{r} + \mathbf{s} \right] = 0. \tag{2.30}$$

The formal Newton step $\mathbf{d}$ for solving this system of equations ignores the discontinuity in derivative. It satisfies

$$\frac{1}{\gamma} A^T W A\mathbf{d} = -A^T \left[\frac{1}{\gamma} W\mathbf{r} + \mathbf{s} \right], \tag{2.31}$$

or

$$A^T W A\mathbf{d} = -A^T (W\mathbf{r} + \gamma\mathbf{s}), \tag{2.32}$$

If A has full rank n, then the rank of W can always be taken to be at least n at the solution of the M-estimation problem: Osborne (1985). However, this does not ensure that W has rank at least n in a step of the Newton iteration. Problems with singularity of the linear system can be avoided by inserting additional 1s into the diagonal positions of W, or indeed the unit matrix can be used. The solution to (2.32) is most efficiently obtained through LU factorization of the matrix on the left-hand side; one step of iterative refinement is recommended in Madsen and Nielsen (1993). A line search may be needed to ensure descent.

As the iteration proceeds, the partition σ will change, until the partition valid at the solution is obtained. Then the iteration terminates in one further Newton step (because a quadratic is being minimized). Changes in the partition translate into changes in W (and $\mathbf{s}$), and corresponding changes to the LU factors of $A^T W A$. It is here that the efficiency of the algorithm is achieved, because changes of a single index in the partition mean a rank 1 change so that updating of the factors is all that is required, a typical iteration costing $O(n^2)$ operations. Changes of more than one index need

refactorization, at a cost of $O(n^3)$ operations: however, this is in practice needed only occasionally (see, for example, Table 2 of Madsen and Nielsen 1993). Once a solution has been obtained for a particular γ, the value of γ is reduced and the process repeated, using warm starts. Again, only a rank 1 change to A^TWA may be needed.

A key observation here is that it is not necessary to let γ go to zero, but the method can be terminated at a nonzero value. The relevant result is as follows.

Theorem 3 Let $\mathbf{x}_\delta$ minimize ψ_δ, and suppose that $\mathbf{s}$ and W remain constant for $0 < \gamma < \delta$. Then $\mathbf{x}_\delta + \delta\mathbf{v}$ solves the l_1 problem, where

$$\delta A^TWA\mathbf{v} = -A^TW\mathbf{r}(\mathbf{x}_\delta). \tag{2.33}$$

Proof. For $0 < \gamma < \delta$, define

$$\mathbf{x}_\gamma = \mathbf{x}_\delta + (\delta - \gamma)\mathbf{v}.$$

Then

$$\begin{aligned}
&A^TW\mathbf{r}(\mathbf{x}_\gamma) + \gamma A^T\mathbf{s} \\
&\quad = A^TW\Big(\mathbf{r}(\mathbf{x}_\delta) + (\delta - \gamma)A\mathbf{v}\Big) + \gamma A^T\mathbf{s} \\
&\quad = A^TW\mathbf{r}(\mathbf{x}_\delta) + (\delta - \gamma)A^TWA\mathbf{v} + \gamma A^T\mathbf{s} \\
&\quad = A^TW\mathbf{r}(\mathbf{x}_\delta) - \left(\tfrac{\delta-\gamma}{\delta}\right)A^TW\mathbf{r}(\mathbf{x}_\delta) + \gamma A^T\mathbf{s} \quad \text{using (2.33)} \\
&\quad = (\gamma/\delta)A^TW\mathbf{r}(\mathbf{x}_\delta) + \gamma A^T\mathbf{s} \\
&\quad = (\gamma/\delta)\Big(A^TW\mathbf{r}(\mathbf{x}_\delta) + \delta A^T\mathbf{s}\Big) \\
&\quad = 0,
\end{aligned}$$

using (2.30) and the definition of $\mathbf{x}_\delta$. It follows from (2.30) that $\mathbf{x}_\gamma$ minimizes the Huber function ψ_γ.

Now, for $0 < \gamma < \delta$,

$$A^TW\mathbf{r}(\mathbf{x}_\gamma) + \gamma A^T\mathbf{s} = 0$$

is equivalent to

$$\sum_{i\in\sigma} r_i(\mathbf{x}_\gamma)\mathbf{a}_i + \gamma \sum_{i\in\sigma^c} g_i(\mathbf{x}_\gamma)\mathbf{a}_i = 0,$$

or

$$\sum_{i\in\sigma} \frac{r_i(\mathbf{x}_\gamma)}{\gamma}\mathbf{a}_i + \sum_{i\in\sigma^c} g_i(\mathbf{x}_\gamma)\mathbf{a}_i = 0. \tag{2.34}$$

Using (2.28), continuity implies that

$$r_i(\mathbf{x}_0) = 0, \quad i \in \sigma,$$

where $\mathbf{x}_0 = \mathbf{x}_\delta + \delta\mathbf{v}$. Thus

$$\sigma \subset I = \{i : r_i(\mathbf{x}_0) = 0\}.$$

In addition,

$$g_i(\mathbf{x}_\gamma) = g_i(\mathbf{x}_0), \quad i \in \sigma^c.$$

Now,

$$\left|\frac{r_i(\mathbf{x}_\gamma)}{\gamma}\right| \leq 1, \quad i \in \sigma.$$

Thus there exist numbers v_i, $|v_i| \leq 1$, $i \in I$ such that

$$\sum_{i\in I} v_i\mathbf{a}_i + \sum_{i\in I^c} g_i(\mathbf{x}_0)\mathbf{a}_i = 0.$$

Thus $\mathbf{x}_0$ solves the l_1 problem and the proof is complete. □

Because the algorithm cannot return to the same sign vector, the condition $0 < \gamma < \delta$ will eventually be satisfied. Note that the matrix on the left-hand side of equation (2.33) is such that no new factorization is needed to compute $\mathbf{v}$.

Numerical results given by Madsen and Nielsen (1993) show that careful implementation of the method can make it superior to the algorithm of Barrodale and Roberts (1973). The larger scale calculations, however, are carried out only for randomly generated problems with $m/n = 2$, which does not seem appropriate for problems of practical interest. More efficient implementations of simplex-based methods, with careful attention to line search performance and scaling, are available for solving the l_1 problem: see Bloomfield and Steiger (1983), Osborne and Watson (1996). Thus, although the Huber-based approach seems to be a promising one, further investigation would be valuable.

2.3. l_p approximation, $1 < p < \infty$

For given $\mathbf{x} \in \mathbb{R}^n$, let $\mathbf{r} = \mathbf{r}(\mathbf{x})$, let D_r be as in the previous section, and let $D_{|r|}$ denote the matrix

$$D_{|r|} = \text{diag}\,\{|r_1|, \ldots, |r_m|\}.$$

Then, applying Theorem 1, or by direct differentiation, it follows that $\mathbf{x}$ minimizes $\|\mathbf{r}\|_p$ if and only if

$$A^T D_{|r|}^{p-1}\mathbf{g} = 0, \tag{2.35}$$

where, as before, $g_i = \text{sign}(r_i)$, $i = 1, \ldots, m$. When $p = 2$, this gives the usual normal equations; otherwise it is a nonlinear system of equations for a solution $\mathbf{x}$. If the value $p = 2$ is not satisfactory because of the error

pattern, there may be merit in moving p towards 1. The range $1 < p < 2$ is of particular interest computationally because there is reduced smoothness: problems with $p \geq 2$ are twice differentiable, problems with $1 < p < 2$ are once differentiable, and the case $p = 1$ is non-differentiable.

One way to proceed to find a solution is to recognize that (2.35) can be written as

$$A^T D_{|r|}^{p-2} \mathbf{r} = 0, \tag{2.36}$$

and this can be viewed as a weighted system of normal equations with weighting matrix $W = D_{|r|}^{p-2}$. This matrix is of course only defined for $1 < p < 2$ if no component of $\mathbf{r}$ is zero, and this will be assumed at present. Fixing this matrix at the current value of $\mathbf{x}$ and solving this weighted least squares problem for the new approximation gives the technique known as *iteratively reweighted least squares* (IRLS). This is attractive since least squares problems are easy to solve. If $\mathbf{x}$ is the current approximation, and $\mathbf{x} + \Delta\mathbf{x}$ is the new approximation, we have

$$A^T D_{|r|}^{p-2} A(\mathbf{x} + \Delta\mathbf{x}) = A^T D_{|r|}^{p-2} \mathbf{b}. \tag{2.37}$$

This simple iteration process will converge if started from close enough to a solution and also if p is close enough to 2. In fact, a stronger result is available for the special case when $1 < p < 2$. The following lemma, which is readily proved, is helpful.

Lemma 4 Let scalars a, b be given with $b \neq 0$. Then, if $1 < p < 2$,

$$|a|^p - |b|^p \leq \tfrac{p}{2} |b|^{p-2} \left(|a|^2 - |b|^2 \right), \tag{2.38}$$

with equality only if $|a| = |b|$.

Theorem 4 If $1 < p < 2$, the method of IRLS is convergent from any starting point to a point satisfying (2.36).

Proof. Let $\mathbf{r}^+$ be the vector $\mathbf{r}$ evaluated at $\mathbf{x}+\Delta\mathbf{x}$ defined by (2.37). Setting $a = \mathbf{r}_i^+$, $b = \mathbf{r}_i$ in (2.38) and summing from $i = 1$ to m gives

$$\begin{aligned} \|\mathbf{r}^+\|_p^p - \|\mathbf{r}\|_p^p &\leq \tfrac{p}{2} \left\{ \sum_{i=1}^m |r_i^+|^2 |r_i|^{p-2} - \sum_{i=1}^m |r_i|^p \right\} \\ &\leq 0, \end{aligned}$$

using the definition of $\mathbf{r}^+$. Thus there is strict reduction unless (2.36) is satisfied, and the result is proved. □

Unfortunately, the properties of guaranteed convergence provided by this theorem, coupled with the simplicity of the iteration process, are offset by the fact that the process is accompanied by a potentially unsatisfactory

convergence rate: this is linear, with convergence constant $|p-2|$ (see Osborne 1985). Thus, as p approaches 1, for example, convergence can be intolerably slow.

Consider now the alternative of using Newton's method. Let

$$\mathbf{f}(\mathbf{x}) = A^T D_{|r|}^{p-2}\mathbf{r}.$$

Then it is readily seen that

$$\nabla\mathbf{f}(\mathbf{x}) = (p-1)A^T D_{|r|}^{p-2}A,$$

so the Newton step $\mathbf{d}$ satisfies the linear system of equations

$$\begin{aligned}(p-1)A^T D_{|r|}^{p-2}A\mathbf{d} &= -A^T D_{|r|}^{p-2}\mathbf{r} \\ &= -A^T D_{|r|}^{p-2}(A\mathbf{x}-\mathbf{b})\end{aligned}$$

or

$$A^T D_{|r|}^{p-2}A(\mathbf{x}+(p-1)\mathbf{d}) = A^T D_{|r|}^{p-2}\mathbf{b}.$$

Comparing this with (2.37) shows that the IRLS step is just $(p-1)$ times the Newton step. It is easily seen that $\mathbf{d}^T\mathbf{f}(\mathbf{x}) < 0$, so that $\mathbf{d}$ is a descent direction for $\|\mathbf{r}\|_p^p$, and so it makes sense to incorporate a line search. If this is done in both methods, then essentially the same method is obtained.

As already indicated, one of the difficulties of the range $1 < p < 2$ is the fact that second derivatives do not always exist if any component of $\mathbf{r}$ becomes zero. Different strategies have been proposed to get round this difficulty. However, not just zero components but *nearly zero* components are potentially troublesome. There is some evidence, however, that these phenomena are not *by themselves* a major problem, but only if they are accompanied by p being close to 1. The main difficulty appears to be due to the fact that, as p approaches 1, we are coming closer to a discontinuous problem, effectively to a constrained problem. It seems necessary to recognize this in a satisfactory algorithm, and consider some of the elements of the l_1 problem in devising an approach that will deal with small values of p in a satisfactory way. This is the philosophy in the recent approach of Li (1993), which we will now describe.

Let Z be the matrix defined previously whose rows form a basis for the null space of A^T, so that

$$Z^T A = 0.$$

Define

$$\mathbf{g}_p = pD_{|r|}^{p-1}\mathbf{g},$$

the derivative of $\|\mathbf{r}\|_p^p$ with respect to $\mathbf{r}$. Then (2.35) is equivalent to

$$\mathbf{g}_p - Z\mathbf{w} = 0. \tag{2.39}$$

Of course, when $p = 1$, $\mathbf{g}_p$ is just $\mathbf{g}$. If D_r is nonsingular, then $\mathbf{x}$ solves the l_p problem if and only if there exists $\mathbf{r} \in \mathbb{R}^m$, $\mathbf{w} \in \mathbb{R}^{m-n}$ such that

$$D_r(\mathbf{g}_p - Z\mathbf{w}) = 0, \tag{2.40}$$

$$Z^T(\mathbf{r} + \mathbf{b}) = 0. \tag{2.41}$$

If D_r is singular, then an additional requirement is that if $r_i = 0$, then $(Z\mathbf{w})_i = 0$. Notice that (2.40) and (2.41) are just the system of equations (2.16) and (2.17) in the l_1 case, so that (2.40) incorporates the complementary slackness conditions which are an essential part of the l_1 characterization.

Applying Newton's method to (2.40) and (2.41) gives, as before, the step in $\mathbf{r}$ as

$$\boldsymbol{\delta}\mathbf{r} = -A\Big(A^T D_r^{-1} D_\beta A\Big)^{-1} A^T \mathbf{g}_p,$$

where in this case

$$\boldsymbol{\beta} = p\mathbf{g}_p - Z\mathbf{w}.$$

Clearly, when $p = 1$, this is just the matrix D_β defined before. In order to globalize the method, one can use a technique similar to that used for the l_1 method. In this case a suitable matrix W_θ is given by

$$W_\theta = \operatorname{diag}\{|r_i^{-1}(p(\mathbf{g}_p)_i - (1-\theta)(Z\mathbf{w})_i)|,\ i = 1, \ldots, m\},$$

and the step in the direction $\mathbf{r}$ is then

$$\boldsymbol{\delta}_\theta\mathbf{r} = -A\Big(A^T W_\theta A\Big)^{-1} A^T \mathbf{g}_p.$$

Clearly $\theta = 0$ gives

$$W_\theta = D_r^{-1} D_\beta,$$

and the Newton step. If $\theta = 1$, then

$$\boldsymbol{\delta}_1\mathbf{r} = p^{-1} A\Delta\mathbf{x},$$

or $1/p$ times the IRLS step. It is therefore possible to develop an algorithm for the l_p problem which is essentially equivalent to the method for $p = 1$ described above. A line search is of course required, and the details are given by Li (1993). For $1 < p < 2$, the method is globally convergent to $\mathbf{r}^*$ satisfying (2.35) if the rows of A corresponding to zero components of $\mathbf{r}^*$ are linearly independent; the convergence is superlinear if $\mathbf{r}^*$ has no zero component.

Numerical results given by Li (1993) show that the new method is clearly superior to IRLS (with the same line search) for values of p close to 1, with the gap between the two methods widening as p approaches 1. There is little difference for values of $p \geq 1.5$ or so. As with the l_1 case, the number of iterations appears to be independent of the problem size.

Finally we mention briefly the cases when $p > 2$. These are perhaps of less interest in practice, and also in theory, since second derivatives always exist. Newton's method (with line search) is perfectly satisfactory in most cases. Clearly, for very large p, scaling problems may well be a factor. We will not dwell on this, but will move on to consider the limiting case, the Chebyshev approximation problem.

2.4. Chebyshev approximation

For any $\mathbf{x} \in \mathbb{R}^n$, and

$$\mathbf{r} = A\mathbf{x} - \mathbf{b},$$

let J denote the set of indices

$$J = \{i : |r_i| = \|r\|_\infty\},$$

and let J^c denote its complement. Then, using Theorem 1 and (2.10) we have the following theorem.

Theorem 5 The vector $\mathbf{x}$ is a solution to (2.3) in the Chebyshev case if and only if there exists $\boldsymbol{\lambda} \in \mathbb{R}^m$ with

$$\lambda_i = 0, \quad i \in J^c, \quad \text{and} \quad \lambda_i \geq 0, \quad i \in J,$$

such that

$$A^T D_g \boldsymbol{\lambda} = 0, \tag{2.42}$$

where

$$D_g = \text{diag } \{g_1, \ldots, g_m\},$$

with $g_i = \text{sign}(r_i)$, $i = 1, \ldots, m$, as before.

A solution $\mathbf{x}$ is unique if A satisfies the Haar condition, that is, if every $n \times n$ submatrix is nonsingular. This condition is sufficient only if $m = n+1$ (see, for example, Watson 1980). The function $\|\mathbf{r}\|_\infty$ is piecewise linear, and the most popular methods have been based on the simplex method of linear programming or variants: see, for example, Barrodale and Phillips (1975), Bartels, Conn and Li (1989). These are finite, moving through sets of points with J containing $n + 1$ indices until optimality is achieved: the fact that there is always a solution at a point with $n + 1$ indices in J if A has rank n is just a restatement of the fact that the solution to a linear programming problem occurs at a basic feasible solution.

Recently, as in the l_1 case, there has been interest in the possibility of smoothing the problem. We describe here an approach that is the analogue of the methods previously described, due to Coleman and Li (1992*b*). Recall that their l_1 approach was able to cross lines of non-differentiability, avoided derivative discontinuities except in the limit, and combined descent steps

based on derivatives with Newton steps in an automatic way. In the current problem, $\|\mathbf{r}\|_\infty$ is differentiable provided that the norm is attained at just one component of $\mathbf{r}$. Let this be the jth component. The region of non-differentiability is therefore defined by the hyperplanes

$$|r_i| = |r_j|, \quad i \neq j.$$

Because j will change from iteration to iteration, it is not obvious how the l_1 method extends to this case; in particular, there is no global transformation that will result in new variables corresponding to the distances to these hyperplanes.

The approach used in Coleman and Li (1992*b*) is to define *local* transformations. At each step a transformation is defined which transforms $\mathbf{r}$ to a new variable $\mathbf{s}$ whose components measure the distance to the hyperplanes of non-differentiability. At the current point $\mathbf{x}$, with $\mathbf{r}$ defined as usual, let J be a singleton with

$$|r_j| = \|r\|_\infty.$$

Then we require $\mathbf{s} \in \mathbb{R}^m$ to be defined by

$$s_i = \begin{cases} g_j r_j - g_i r_i, & i \neq j, \\ g_j r_j, & i = j. \end{cases}$$

Alternatively, defining $T \in \mathbb{R}^{m \times m}$ by

$$T = (\mathbf{g} + g_j \mathbf{e}_j)\mathbf{e}_j^T - D_g,$$

then it is easily seen that

$$T^{-1} = g_j(\mathbf{e} + \mathbf{e}_j)\mathbf{e}_j^T - D_g,$$

and in addition

$$\mathbf{r} = T\mathbf{s}.$$

Of course, T depends on $\mathbf{r}$, and so the transformation is a local one.

Let Z be defined as before so that $A^T Z = 0$. Then the Chebyshev problem can be expressed in the form

$$\begin{aligned} &\text{minimize}_{\mathbf{s} \in \mathbb{R}^m} \quad \|T\mathbf{s}\|_\infty \\ &\text{subject to } Z^T(T\mathbf{s} + \mathbf{b}) = 0. \end{aligned} \tag{2.43}$$

Let $\mathbf{r}$ be such that $\|\mathbf{r}\|_\infty$ is differentiable, with $J = \{j\}$. Then the gradient vector is given by $D_g \mathbf{e}_j$. A descent direction $\mathbf{d}_s$ for the current $\mathbf{s}$, analogous to that defined for the l_1 case, can be obtained by solving the following subproblem:

$$\begin{aligned} &\text{minimize}_{\mathbf{d} \in \mathbb{R}^n} \quad \mathbf{e}_j^T D_g T \mathbf{d} \\ &\text{subject to } Z^T T \mathbf{d} = 0, \quad \text{and} \quad \|D^{-1}\mathbf{d}\|_2 \leq \tau. \end{aligned} \tag{2.44}$$

Again, D is a positive definite diagonal scaling matrix, and τ is a positive number that restricts the size of $\mathbf{d}_s$. As before, up to a scalar multiple, we can easily see that

$$\mathbf{d}_s = -T^{-1}A\Big(A^T T^{-T} D^{-2} T^{-1} A\Big)^{-1} A^T D_g \mathbf{e}_j,$$

so that the corresponding descent direction for $\mathbf{r}$ analogous to (2.20) is

$$\mathbf{d} = -A\Big(A^T T^{-T} D^{-2} T^{-1} A\Big)^{-1} A^T D_g \mathbf{e}_j. \tag{2.45}$$

This vector can be obtained via the solution of a least squares problem, and a line search enables $\mathbf{r}$ to be updated. Details of this and a suitable line search are given in Coleman and Li (1992*b*). This affine scaling algorithm may be slowly convergent, and we consider next the application of Newton's method to the problem. This requires suitable reinterpretation of the characterization conditions as a nonlinear system, analogous to (2.16) and (2.17). The following theorem is key to this.

Theorem 6 Necessary and sufficient conditions for $\mathbf{r}$ to solve the Chebyshev problem are that there exists $\mathbf{w} \in \mathbb{R}^m$ such that

$$\begin{aligned} D_s T^T (D_g \mathbf{e}_j - Z\mathbf{w}) &= 0, && (2.46)\\ Z\mathbf{r} + Z\mathbf{b} &= 0, && (2.47)\\ \lambda_i &\geq 0, \quad i \in I - \{j\},\\ 1 - \sum_{i \in J - \{j\}} \lambda_i &\geq 0, \end{aligned}$$

where

$$\lambda_i = g_i (Zw)_i, \quad i \in I - \{j\}.$$

Proof. Let $\mathbf{r}$ solve the Chebyshev problem. Then, obviously,

$$Z^T\mathbf{r} + Z^T\mathbf{b} = 0.$$

Further, from the characterization result, there exist numbers λ_i, $i \in I$, $\mathbf{w} \in \mathbb{R}^m$, such that

$$\sum_{i \in I} g_i \lambda_i \mathbf{e}_i = Z\mathbf{w},$$

where

$$\lambda_i \geq 0, \quad i \in J,$$

$$\sum_{i \in J} \lambda_i = 1.$$

It follows that

$$\sum_{i\in I} g_i\lambda_i T^T\mathbf{e}_i = T^T Z\mathbf{w},$$

and so

$$T^T D_g\mathbf{e}_j = \sum_{i\in I-\{j\}} \lambda_i e_i + T^T Z\mathbf{w}. \tag{2.48}$$

Thus

$$D_s T^T(D_g\mathbf{e}_j - Z\mathbf{w}) = 0,$$

and necessity is established.

Now let the stated conditions be satisfied. Clearly this implies that (2.48) is satisfied in the ith component, for $i \in J^c \cup \{j\}$. For $i \in J\backslash\{j\}$,

$$\Big(T^T D_g\mathbf{e}_j\Big)_i - \Big(T^T Z\mathbf{w}\Big)_i = g_i(Z\mathbf{w})_i,$$

and so, by the definition of λ_i, (2.48) also holds in these components. Thus (2.48) is satisfied, and reversing the argument of the proof of necessity leads to the required result. □

Consider now the application of Newton's method to the nonlinear system consisting of (2.46) and (2.46). Define D_β as before, where

$$\boldsymbol{\beta} = T^T(\mathbf{g} - Z\mathbf{w}).$$

Then the Newton step in $\mathbf{r}$ and $\mathbf{w}$ is given by the system of equations

$$\begin{bmatrix} D_\beta T^{-1} & -D_s T^T Z \\ Z^T & 0 \end{bmatrix} \begin{bmatrix} \boldsymbol{\delta}\mathbf{r} \\ \boldsymbol{\delta}\mathbf{w} \end{bmatrix} = \begin{bmatrix} -D_s T^T(D_g\mathbf{e}_j - Z\mathbf{w}) \\ 0 \end{bmatrix}. \tag{2.49}$$

It follows from this system that

$$\boldsymbol{\delta}\mathbf{r} = -A\Big(A^T T^{-T} D_s^{-1} D_\beta T^{-1} A\Big)^{-1} A^T D_g\mathbf{e}_j. \tag{2.50}$$

Coleman and Li (1992*b*) show that in a neighbourhood of a solution to (2.3), where $\|\mathbf{r}\|_\infty$ is differentiable, the matrix being inverted on the right-hand side of (2.50) is positive definite, so that the Newton direction becomes a descent direction.

Note that (2.45) and (2.50) have the same general form, and suitable definition of D enables a smooth transition to be made between the steps (2.45) and (2.50). Let D be chosen by

$$D = D_s^{1/2}.$$

Further, define

$$W_\theta = D_s^{-1}\Big((1-\theta)D_{|\beta|} + \theta\mathbf{e}\Big),$$

where $D_{|\beta|}$ denotes the diagonal matrix whose diagonal elements are the modulus of those of D_β. Define the search direction as

$$\mathbf{d} = -A\Big(A^T T^{-T} W_\theta T^{-1} A\Big)^{-1} A^T D_g \mathbf{e}_j. \tag{2.51}$$

Then, when $\theta = 1$, (2.51) just gives (2.45), and when $\theta = 0$, (2.51) gives (2.50), provided that D_β is positive definite: it is shown by Coleman and Li (1992*b*) that this holds in a neighbourhood of the solution excluding non-differentiable points.

Details of the way in which the parameter θ can be chosen, as well as other computational issues, are given by Coleman and Li (1992*b*). It is also shown that, under nondegeneracy conditions analogous to those for the l_1 problem, global convergence to a solution at a quadratic rate may be established. Numerical results show that the number of iterations is relatively insensitive to problem size.

The application of interior point methods for linear programming problems to l_1 and l_∞ problems has been considered by several other authors, for example Meketon (1987), Ruzinsky and Olsen (1989), Zhang (1993) and Duarte and Vanderbei (1994). An approach based on row (column) relaxation or proximal point methods has been used by Dax (1989, 1993), Dax and Berkowitz (1990): this may have potential for large sparse problems. It is interesting that all these smoothing methods have as a subproblem the solution of a weighted least squares problem.

The extent to which these new methods for both the l_1 and l_∞ problems will usurp methods of simplex type remains to be seen. The main methods of the latter type for which there is readily available software are not the most up to date, and in particular the issue of the provision of efficient line searches is a live one. It would seem that such issues are unlikely to be resolved until efficient codes for all these different types of method have been produced.

3. Total least norm problems

An assumption made in Section 2 was that errors were only present in $\mathbf{b}$, so that the model equations (2.2) were appropriate. However, in many situations, the independent variable values are also in error. The problem of dealing with errors in all variables is well known in the statistics literature, and solution methods go back to the end of the last century. The general situation is complicated by the need to make additional assumptions about the error structure in order to ensure identifiability (Moran 1959).

One way of taking errors also in the independent variables into account is to introduce a matrix E of perturbations of A. This leads to the relationship

$$\mathbf{r} = (A + E)\mathbf{x} - \mathbf{b}. \tag{3.1}$$

The analogue of the general problem considered in Section 2 is then the problem

$$\text{minimize } \|E : \mathbf{r}\| \text{ subject to (3.1)}, \tag{3.2}$$

where the norm is now a norm on $m \times (n+1)$ matrices. This problem, in which the norm is the Frobenius norm, was first analysed by Golub and Van Loan (1980), who referred to it as the *total least squares problem.* Since that time, problems of this kind have generated enormous interest, and there have been extensions in various directions, for example to structured problems. There are many applications, to systems identification, frequency estimation, superresolution, control theory, *etc.* In keeping with the underlying theme of this article, we will consider (3.2) for a general class of norms. This is not merely of theoretical interest, because (as in the previous section) norms other than the least squares norm are relevant in practice, and there has been algorithmic development.

It is not possible to deal with this problem in complete generality (as is the case when E is zero). There are some fundamental differences. As we show below, existence of solutions is not even guaranteed. It is not a convex problem, and general characterization results are not available. However, we can make progress in an analysis of the problem if we confine attention to a wide range of matrix norms known as separable matrix norms, a concept introduced in Osborne and Watson (1985). Before introducing these, we need the concept of the dual matrix norm: this is defined (analogous to (2.4)) by

$$\|M\|^* = \max_{\|N\| \leq 1} \text{trace}(N^T M). \tag{3.3}$$

Definition 2 A matrix norm on $m \times (n+1)$ matrices is said to be *separable* if, given vectors $\mathbf{u} \in \mathbb{R}^m$ and $\mathbf{v} \in \mathbb{R}^{n+1}$, there are vector norms $\|\cdot\|_A$ on $\mathbb{R}^m$ and $\|\cdot\|_B$ on $\mathbb{R}^{n+1}$ such that

$$\|\mathbf{u}\mathbf{v}^T\| = \|\mathbf{u}\|_A \|\mathbf{v}\|_B^*, \qquad \|\mathbf{u}\mathbf{v}^T\|^* = \|\mathbf{u}\|_A^* \|\mathbf{v}\|_B.$$

A useful property of separable norms is the following. Let $Z \in \mathbb{R}^{m \times (n+1)}$, $\|\mathbf{v}\|_B \leq 1$. Then, for separable norms,

$$\begin{aligned} \|Z\mathbf{v}\|_A &= \max_{\|\mathbf{u}\|_A^* \leq 1} \mathbf{u}^T Z \mathbf{v} \\ &\leq \max_{\|\mathbf{u}\|_A^* \leq 1, \|\mathbf{v}\|_B \leq 1} \mathbf{u}^T Z \mathbf{v} \\ &\leq \max_{\|\mathbf{u}\mathbf{v}^T\|^* \leq 1} \text{trace}(\mathbf{v}\mathbf{u}^T Z) \\ &\leq \max_{\|G\|^* \leq 1} \text{trace}(G^T Z) \\ &= \|Z\|, \end{aligned}$$

so that

$$\|Z\| \geq \max_{\|\mathbf{v}\|_B \leq 1} \|Z\mathbf{v}\|_A. \tag{3.4}$$

Examples of separable matrix norms are as follows.

Example 1 Operator norms defined by

$$\|M\| = \max_{\|\mathbf{x}\|_d \leq 1} \|M\mathbf{x}\|_c,$$

are separable with $\|\cdot\|_A = \|\cdot\|_c$, and $\|\cdot\|_B = \|\cdot\|_d$; in other words, equality holds in (3.4).

Example 2 The norms defined by

$$\|M\| = \left(\sum_{i,j} |m_{i,j}|^p \right)^{1/p}, \quad 1 \leq p < \infty,$$

are separable with $\|\cdot\|_A = \|\cdot\|_p$, and $\|\cdot\|_B = \|\cdot\|_q$, where $1/p + 1/q = 1$.

Example 3 Orthogonally invariant norms such that

$$\|M\| = \|UMV\|,$$

where U and V are orthogonal, are separable. Both norms are (unweighted) l_2 norms provided that $\|\mathbf{e}_1\mathbf{e}_1^T\| = 1$. This follows because if σ_1 is the largest singular value of M, then for any vectors $\mathbf{u}$ and $\mathbf{v}$ we will have

$$\|\mathbf{u}\mathbf{v}^T\| = \|\sigma_1 \mathbf{e}_1\mathbf{e}_1^T\| = \sigma_1,$$

by assumption, where σ_1 is the largest singular value of $\mathbf{u}\mathbf{v}^T$. But

$$\sigma_1 = \|\mathbf{u}\|_2\|\mathbf{v}\|_2,$$

and the result is established. For example, a class of such orthogonally invariant norms is the class of Schatten p norms, defined by

$$\|M\|_{C_p} = \left(\sum_{i=1}^{n} \sigma_i^p \right)^{1/p}, \quad 1 \leq p < \infty,$$

where $\sigma_1, \ldots, \sigma_n$ are the singular values of M.

To solve the total least norm problem it is necessary to minimize $\|E : \mathbf{r}\|$ subject to (3.1). This is greatly facilitated by replacing the problem by an equivalent but generally much more tractable problem:

$$\text{minimize } \|Z\mathbf{v}\|_A \text{ subject to } \|\mathbf{v}\|_B = 1, \tag{3.5}$$

where $Z = [A : -\mathbf{b}]$.

Theorem 7 Let $\mathbf{v}$ solve (3.5) with $v_{n+1} \neq 0$. Then (3.2) is solved by

$$[E : -\mathbf{r}] = -Z\mathbf{v}\mathbf{w}^T,$$

where $\mathbf{w} \in \partial\|\mathbf{v}\|_B$, and $\mathbf{x}$ is such that

$$\mathbf{v}^T = \alpha(\mathbf{x}^T, 1), \quad \alpha \in \mathbb{R}. \tag{3.6}$$

Proof. For E, $\mathbf{r}$ and $\mathbf{x}$ as defined,

$$\begin{aligned} \|[E : -\mathbf{r}]\| &= \|Z\mathbf{v}\|_A \|\mathbf{w}\|_B^* \\ &= \|Z\mathbf{v}\|_A. \end{aligned}$$

Now let E, $\mathbf{r}, \mathbf{x}$ be *any* feasible set for (3.1), with $\mathbf{v}$ defined by (3.6). Then

$$\begin{aligned} \|Z\mathbf{v}\|_A &= \|[E : -\mathbf{r}]\mathbf{v}\|_A \\ &\leq \|[E : -\mathbf{r}]\|, \end{aligned}$$

using (3.4). The result follows. □

Notice that the matrix $[E : -\mathbf{r}]$ which gives a solution is a rank 1 matrix. However, if there is no $\mathbf{v}$ solving (3.5) with last component nonzero, then there is no solution to the total least norm problem.

Example 4 Let

$$A = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 1 \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}.$$

Clearly $\mathbf{v} = (1, 1, 0)^T$ (suitably normalized) gives a zero value for the minimum in (3.5). Further, for any norm, $\|E : \mathbf{r}\|$ can be made arbitrarily small; however, it can never be zero because $\mathbf{b}$ is not in the range of A.

Consider the case when the matrix norm is the Frobenius norm. Then both vector norms occurring in the separable norm definition are least squares norms and the problem (3.5) becomes that of determining the smallest singular value of the matrix Z. The total least squares problem was first analysed by Golub and Van Loan (1980). It may happen that certain columns of A are known to be exact, so that the corresponding columns of E should be zero. This is easily dealt with by fixing the corresponding components of $\mathbf{v}$ in (3.5) to be zero: see Watson (1983), Osborne and Watson (1985). The solution of (3.5) for other norms is less straightforward, not least because the problem is not a convex one, and so local solutions are possible.

Necessary conditions for a solution of (3.5) can be given in the following form. (In the case when $\|\cdot\|_B$ is polyhedral, then these conditions are also sufficient for a local solution; see Watson 1983.)

Theorem 8 Let $\mathbf{v}$ solve (3.5). Then, for every $\mathbf{w} \in \partial\|\mathbf{v}\|_B$, there exists $\mathbf{g} \in \partial\|Z\mathbf{v}\|_A$ such that

$$Z^T\mathbf{g} = \|Z\mathbf{v}\|_A\mathbf{w}.$$

An algorithm for the minimization of the norm defined by

$$\|M\| = \sum_{i,j} |m_{i,j}|,$$

the total l_1 problem ($\|\cdot\|_A$ is the l_1 norm, and $\|\cdot\|_B$ is the l_∞ norm), is given by Osborne and Watson (1985). Variants of these problems, such as the total l_p problem and the orthogonal l_1 problem, have also been considered in Watson (1984), Späth and Watson (1987). The idea of introducing structure into the matrix E has also been investigated. The case of zero columns has already been mentioned, but there are important applications when perturbations of A should preserve other sparsity patterns, or Toeplitz, Vandermonde or Hankel forms. While (3.5) is useful for certain types of structure (see, for example, Watson 1988, 1991), for others it seems necessary to work with the original problem (3.2).

For example, Rosen, Park and Glick (1996, 1997) develop a method concerned with retaining given structure, and permitting the use of general norms. We will illustrate this in the important special case when A has Toeplitz structure that must be preserved, as occurs in system identification. Then

$$A_{ij} = \alpha_{n+i-j}, \quad i = 1, \ldots, m, \quad j = 1, \ldots, n,$$

so that we can define A entirely by its first row

$$\rho_1(A) = [\alpha_n, \alpha_{n-1}, \ldots, \alpha_1],$$

and its first column

$$\kappa_1(A) = [\alpha_n, \alpha_{n+1}, \ldots, \alpha_{n+m-1}]^T.$$

Thus, in (3.1) we can define E to have the same form, with unique unknown elements, say β_i, $i = 1, \ldots, n+m-1$. Assume that the matrix norm is an l_p norm defined for any matrix M by

$$\|M\| = \left(\sum |m_{ij}|^p\right)^{1/p}, \quad 1 \leq p \leq \infty.$$

Then

$$\|E : \mathbf{r}\| = \left\| \begin{array}{c} \mathbf{r}(\boldsymbol{\beta}, \mathbf{x}) \\ W\boldsymbol{\beta} \end{array} \right\| \tag{3.7}$$

where the vector norm is the l_p norm, where

$$\mathbf{r}(\boldsymbol{\beta}, \mathbf{x}) = (A + E)\mathbf{x} - \mathbf{b},$$

and where W is an $(m+n-1) \times (m+n-1)$ diagonal weighting matrix which accounts for repetitions of elements in E. The problem of minimizing (3.7) is of course nonlinear in $\boldsymbol{\beta}$ and $\mathbf{x}$, and methods for the l_1, l_2 and l_∞ norms based on linearization are given in Rosen et al. (1996). Extensions to problems where A depends nonlinearly on parameters which have to be estimated are given in Rosen et al. (1997). The techniques required are those for nonlinear problems, so we will not deal with this further but proceed to a more general study of this class.

4. Approximation to data by nonlinear models

When the model contains free parameters which occur nonlinearly, and it is assumed that errors are only present in the dependent variable, we obtain a problem which can be posed in the form

$$\text{minimize}_{\mathbf{x}\in\mathbb{R}^n} \quad \|\mathbf{f}(\mathbf{x})\|, \tag{4.1}$$

where $\mathbf{f} : \mathbb{R}^n \to \mathbb{R}^m$, $m > n$, the dependence of $\mathbf{f}$ on $\mathbf{x}$ is nonlinear, and the norm is a norm on $\mathbb{R}^m$. Provided that $\mathbf{f}$ is differentiable, we can write $A(\mathbf{x})$ for the $m \times n$ matrix of partial derivatives of $\mathbf{f}$ with respect to the components of $\mathbf{x}$. Allowing an arbitrary norm, we have the following result.

Theorem 9 Let $\mathbf{x}$ solve (4.1), with $\mathbf{f}$ such that

$$\mathbf{f}(\mathbf{z}) = \mathbf{f}(\mathbf{x}) + A(\mathbf{x})(\mathbf{z}-\mathbf{x}) + o(\|\mathbf{z}-\mathbf{x}\|_p)$$

for all $\mathbf{z}$ in a neighbourhood of $\mathbf{x}$, where $\|\cdot\|_p$ is a norm on $\mathbb{R}^n$. Then there exists $\mathbf{v} \in \partial\|\mathbf{f}(\mathbf{x})\|$ such that

$$A(\mathbf{x})^T\mathbf{v} = 0.$$

This result may be established along the lines of the proof of necessity in Theorem 1; the condition is not sufficient except in the special case when the norm is a convex function of $\mathbf{x}$. See Watson (1980), for example, for the details.

A general class of methods can be given for finding a point satisfying the conditions of this theorem (a stationary point). The basis is the solution of a sequence of linearized subproblems defined at the current approximation $\mathbf{x}$ to a stationary point, which enables an improved approximation to be obtained. A typical subproblem has the form

$$\begin{aligned} &\text{minimize}_{\mathbf{d}\in\mathbb{R}^n} \quad \|\mathbf{f}(\mathbf{x}) + A(\mathbf{x})\mathbf{d}\| \\ &\text{subject to } \|\mathbf{d}\|_A \le \tau, \end{aligned}$$

where τ is a suitably chosen positive scalar, and $\|\cdot\|_A$ is a suitably chosen norm. An analysis of the use of this subproblem is quite straightforward. Because $\mathbf{d} = 0$ is a candidate for a solution, we must have

$$\|\mathbf{f}(\mathbf{x}) + A(\mathbf{x})\mathbf{d}\| \le \|\mathbf{f}(\mathbf{x})\|.$$

But, for γ such that $0 < \gamma \leq 1$,

$$\begin{aligned} \mathbf{f}(\mathbf{x}+\gamma\mathbf{d}) &= \mathbf{f}(\mathbf{x}) + \gamma A(\mathbf{x})\mathbf{d} + o(\gamma) \\ &= \gamma(\mathbf{f}(\mathbf{x}) + A(\mathbf{x})\mathbf{d}) + (1-\gamma)\mathbf{f}(\mathbf{x}) + o(\gamma), \end{aligned}$$

and so

$$\begin{aligned} \|\mathbf{f}(\mathbf{x}+\gamma\mathbf{d})\| &\leq \gamma\|\mathbf{f}(\mathbf{x}) + A(\mathbf{x})\mathbf{d}\| + (1-\gamma)\|\mathbf{f}(\mathbf{x})\| + o(\gamma) \\ &= \|\mathbf{f}(\mathbf{x})\| + \gamma\Big(\|\mathbf{f}(\mathbf{x}) + A(\mathbf{x})\mathbf{d}\| - \|\mathbf{f}(\mathbf{x})\|\Big) + o(\gamma). \end{aligned}$$

Thus $\mathbf{d}$ is a descent direction for $\|\mathbf{f}\|$ at $\mathbf{x}$ unless

$$\|\mathbf{f}(\mathbf{x}) + A(\mathbf{x})\mathbf{d}\| = \|\mathbf{f}(\mathbf{x})\|.$$

Thus $\mathbf{d} = 0$ is a solution to the subproblem. Hence, using the characterization result given in Theorem 1 (noting that the bound constraint is inactive) and applying Theorem 9 shows that $\mathbf{x}$ is a stationary point.

In practice, a line search can be avoided: either $\mathbf{x}$ is replaced by $\mathbf{x} + \mathbf{d}$ to give a better approximation, or the subproblem can be re-solved with a reduced value of τ. If care is taken with the rules for choosing τ, this process can give convergence to a stationary point.

For the important cases of the l_2, l_1 and l_∞ norms, methods of this type are well known. A second-order rate of convergence is possible for polyhedral norm problems (which includes l_1 and l_∞). For example, consider the l_∞ norm, and suppose that $\mathbf{x}^*$ is a stationary point, with the current approximation $\mathbf{x}$ in a neighbourhood of $\mathbf{x}^*$. Suppose also that $\|\mathbf{f}(\mathbf{x}^*)\|$ is attained at exactly $n+1$ components of $\mathbf{f}(\mathbf{x}^*)$, say $j_1, \ldots, j_{n+1}$. Then, if the bound constraints on the solution of the linearized subproblems stay inactive, solutions can be interpreted as steps of Newton's method applied to the solution of

$$f_{j_i}(\mathbf{x}) - \sigma_i h = 0, \quad i = 1, \ldots, n+1, \tag{4.2}$$

where $\sigma_i = \pm 1$. A similar analogy is possible with the l_1 norm, with the corresponding requirement being that, at a stationary point $\mathbf{x}^*$, there are exactly n zeros of $\mathbf{f}(\mathbf{x}^*)$. In general, however, such conditions do not hold: there are too few nonlinear equations like (4.2) to determine the unknowns, and convergence can be slow. This has led to the incorporation of second derivative information into the subproblems. This can be done in different ways, for example by adding $\frac{1}{2}\mathbf{d}^T H\mathbf{d}$ to the objective function of the linearized subproblem, where H is the Hessian matrix (or some symmetric approximation to the Hessian) of the Lagrangian function at the current point for the problem posed as an optimization problem. The subproblems can usually be solved as quadratic programming problems. Good methods for the l_1 and l_∞ problems have been available for some time, and some references can be found in the review paper by Watson (1987).

Such methods are good for small problems. They are quite sophisticated, and relatively heavy computationally. In addition, sparsity in A, say, cannot easily be exploited. For the l_1 and l_∞ norms, with a suitable choice of $\|\cdot\|_A$, the linearized subproblems can usually be posed as linear programming problems, and efficient techniques are available for large sparse problems. Indeed, for very large problems, there may be little choice but to use the linearized subproblem as the basis for an algorithm, with, if possible, simple modifications introduced to help speed up convergence. Such ideas for l_∞ problems are proposed by Jonasson (1993), Jonasson and Madsen (1994). It remains to be seen how effective such methods will become, but, in any event, the basic subproblem will remain a very important and robust tool.

As already indicated, the problem (4.1) most often arises from the data-fitting problem analogous to (2.3) in the linear case, where we can write the ith component of $\mathbf{f}$ as

$$f_i(\mathbf{x}) = F(\mathbf{x}, \xi_i) - b_i, \quad i = 1, \ldots, m,$$

and where F depends nonlinearly on the parameters forming the vector $\mathbf{x}$. Here the data consists of sets of points (ξ_i, b_i), $i = 1, \ldots, m$, with only $\mathbf{b}$ containing errors. There is of course also the possibility of errors in the values of the variables ξ_i, and in this case the model equations can be written

$$f_i(\mathbf{x}) = F(\mathbf{x}, \xi_i + \delta_i) - b_i, \quad i = 1, \ldots, m. \tag{4.3}$$

Then it is appropriate to minimize some norm of the vector $\mathbf{v}$ in $\mathbb{R}^{2m}$ whose components are f_i, $i = 1, \ldots, m$, δ_i, $i = 1, \ldots, m$, where we assume for simplicity that ξ_i (and hence δ_i) are in $\mathbb{R}$ (although these can in practice have many components). In this form, the problem is referred to as an errors-in-variables problem. For example, we could minimize the (square of the) l_2 norm of all the errors,

$$\mathbf{v}^T\mathbf{v} = \sum_{i=1}^{n} (f_i^2 + \delta_i^2), \tag{4.4}$$

subject to (4.3). This problem is referred to as *orthogonal distance regression*, the reason for this name being that we are minimizing the sum of squares of the distances from the data points (ξ_i, b_i) to the model curve. An efficient method for minimizing (4.4) is given by Boggs, Byrd and Schnabel (1987), with software available in Boggs, Byrd, Donaldson and Schnabel (1989). The point here is that there is considerable structure which can be exploited.

The minimization of the l_1 norm of $\mathbf{v}$ has also been considered, again exploiting the structure; see, for example, Watson and Yiu (1991), Watson (1997).

5. Chebyshev approximation of functions

Let $f(x)$, $\phi_i(x)$, $i = 1, \ldots, n$, be continuous functions defined on a compact set X. Then the problem of approximating f in the Chebyshev sense by a linear combination of the functions ϕ_i can be stated as: find coefficients c_i, $i = 1, \ldots, n$, to minimize

$$\left\| f - \sum_{i=1}^{n} c_i \phi_i \right\| = \max_{x \in X} \left| f(x) - \sum_{i=1}^{n} c_i \phi_i(x) \right|.$$

The analogue of (2.42) is that there exist points x_j, $j = 1, \ldots, t$ in X where the norm is attained with sign g_i, $i = 1, \ldots, t$, and corresponding numbers μ_i, $i = 1, \ldots, t$ such that $\mu_i g_i > 0$, $i = 1, \ldots, t$ and

$$A^T \boldsymbol{\mu} = 0, \tag{5.1}$$

where A is the $t \times n$ matrix with (i, j) element $\phi_j(x_i)$. Suppose that $X = [a, b] \subset \mathbb{R}$. Then, if the system of functions ϕ_i, $i = 1, \ldots, n$ forms a Chebyshev set on $[a, b]$, so that no nontrivial linear combination of the functions has more than $(n-1)$ zeros in $[a, b]$, then it is readily shown from (5.1) that $t \geq n+1$, and the values of μ_i alternate in sign. This gives us the well-known classical alternation characterization property: there are $n+1$ points in $[a, b]$ where the norm is attained with the error alternating in sign as we go from left to right, or briefly

$$\mathcal{A}(f - \phi)_{[a,b]} \geq n + 1, \tag{5.2}$$

where ϕ denotes the approximation. The exchange algorithms of Remes (both single-point exchange and multi-point exchange) are effective ways of computing the (unique) best approximation.

Some general alternation theorems are also available for problems with constraints. For example, Brosowski and da Silva (1992) consider the problem of approximation on $[a, b]$ by a linear combination of functions forming a Chebyshev set on $[a, b]$, subject to certain side-conditions. Their results contain as a special case the classical theorem, and other known results for one-sided and restricted range approximation.

Once the Chebyshev set condition is dropped, then life becomes much more complicated. There may be nonuniqueness of solutions, but a more serious problem in practice is that, at a solution, there may be fewer than $n+1$ points where the norm is attained. If this possibility is ignored, then exchange methods can be applied, although convergence can be slow and ill-conditioning can occur (because of coalescing points in the set where the norm is attained). Multivariate problems are particularly susceptible to this difficulty. Methods of Newton type are available, with the Newton method used as a local method when information about the number of points where the norm is attained at a solution is known, along with associated

information. Methods of this type also apply to nonlinear problems (for example, Watson 1976).

A method for solving a wide class of continuous Chebyshev approximation problems, linear as well as nonlinear, is given by Jing and Fam (1987). The algorithm is shown to be convergent (possibly to a local minimum), and convergence is quadratic in nondegenerate cases. The method has some similarities to a method due to Jonasson and Watson (1982). Both approaches solve a sequence of linearized subproblems on the current set of points where the error function attains its local maxima, followed by a line search to obtain an improved approximation. Under appropriate conditions, both methods are equivalent locally to the Remes (multi-point) exchange method.

Most emphasis, however, has been on the use of particular approximating functions whose special properties can be exploited, and we consider some examples of these.

5.1. Chebyshev approximation by spline functions

Aside from interpolation, the use of spline functions for approximation has mainly been concerned with the use of the Chebyshev norm. Consider now the problem of approximating from the space of spline functions defined as follows.

Definition 3 Let integers m and k be given, and let $a = x_0 < x_1 < \cdots < x_{k+1} = b$. Then

$$S_m = \{s \in C^{m-1}[a,b] : s(x) \in \Pi_m \text{ on } [x_i, x_{i+1}], \quad i = 0, \ldots, k\},$$

is the space of polynomial splines of degree m with k fixed knots, where Π_m denotes the space of polynomials of degree m. S_m is a linear space with dimension $m + k + 1$.

The theory of approximation by Chebyshev sets does not apply to approximation from S_m. However, S_m is an example of a *weak Chebyshev space*: there exists at least one best approximation from S_m to any continuous function that has the classical alternation property (although there may be others that do not). The theory of Chebyshev approximation by splines with fixed knots is fully developed, and a characterization of best approximation goes back to Rice (1967) and Schumaker (1968). What is required is the existence of an interval $[x_p, x_{p+q}] \subset [a,b]$, with $q \geq 1$ such that there are at least $q + m + 1$ alternating extrema on $[x_p, x_{p+q}]$ or, in the notation introduced in (5.2),

$$\mathcal{A}(f - s)_{[x_p, x_{p+q}]} \geq q + m + 1,$$

where $s \in S_m$.

In addition to characterization of solutions, there has been interest in conditions for uniqueness and *strong uniqueness* of best approximations.

Definition 4 A function $s_f \in S_m$ is called a strongly unique best approximation to $f \in C[a,b]$ if there is a constant $K_f > 0$ such that, for all $s \in S_m$,

$$\|f - s\| \geq \|f - s_f\| + K_f \|s - s_f\|.$$

In general, best approximations are not unique. However, the uniqueness (and strong uniqueness) of best spline approximations is characterized by the fact that *all* knot intervals contain sufficiently many alternating extrema (see Nürnberger 1989).

An iterative algorithm for computing best Chebyshev approximations from spline spaces is due to Nürnberger and Sommer (1983). As in the classical Remes method, a substep at each iteration is the computation of a spline $s \in S_m$ such that

$$(-1)^i \Big(f(\xi_i) - s(\xi_i) \Big) = h, \quad i = 1, \ldots, m+k+2,$$

for some some real number h, and given points $\xi_1, \ldots, \xi_{m+k+2}$ in $[a,b]$. The number of equations reflects the fact that S_m has dimension $m+k+1$. Then one of the points ξ_i is replaced by a point where $\|f-s\|$ is attained in $[a,b]$ to get a new set of points $\{\xi_i\}$. The usual Remes exchange rule can result in a singular system of equations, so a modified exchange rule is needed. Such a rule is given by Nürnberger and Sommer (1983), which ensures that the new system has a unique solution. Because of possible nonuniqueness of best approximations, the proof of convergence is fairly complicated. However, a convergence result can be established. A multiple exchange procedure can also be implemented, and quadratic convergence is possible. The above results can be extended to more general spline spaces, where the polynomials are replaced by linear combinations of functions forming Chebyshev sets: see, for example, Nürnberger, Schumaker, Sommer and Strauss (1985).

To permit the full power of splines, one should allow the knots to vary, rather than be fixed in advance. The corresponding approximation problem is then a difficult nonlinear problem. To guarantee existence of best approximations, multiple knots have to be allowed. There may be local solutions; a characterization of best approximations is not known. For the case of k free knots, necessary and (different) sufficient conditions of the alternation kind given above may be proved. Let q' denote the sum of the knot multiplicities at the points $x_{p+1}, \ldots, x_{p+q-1}$. Then it is *necessary* for $s \in S_m$ to be a best Chebyshev approximation with k free knots to f in $[a,b]$ that there exists an interval $[x_p, x_{p+q}] \subset [a,b]$ with $q \geq 1$ such that

$$\mathcal{A}(f-s)_{[x_p, x_{p+q}]} \geq m + q + q' + 1$$

(Nürnberger, Schumaker, Sommer and Strauss 1989); it is *sufficient* for $s \in S_m$ to be a best Chebyshev approximation with k free knots to f in

$[a, b]$ that there exists an interval $[x_p, x_{p+q}] \subset [a, b]$ with $q \geq 1$ such that

$$\mathcal{A}(f - s)_{[x_p, x_{p+q}]} \geq m + k + q' + 2$$

(Braess 1971). The necessary condition is strengthened to a possibly longer alternant by Mulansky (1992). Other results on this topic are given by Kawasaki (1994), Nürnberger (1994*a*).

Although a characterization of best spline approximations with free knots is not known, a characterization of strongly unique best spline approximations with free simple knots is available: what is required is that *all* knot intervals contain sufficiently many alternating extrema (Nürnberger 1987; see also Nürnberger 1994*b*).

Some algorithms for computing best Chebyshev approximations by free knot splines are available. For example, Nürnberger, Sommer and Strauss (1986) (see also Meinardus, Nürnberger, Sommer and Strauss 1989) give an algorithm that converges through sequences of knot sets from an arbitrary set of knots. For each set of k knots, best Chebyshev degree m polynomial approximations to f are obtained on each subinterval using the classical Remes algorithm. The knots are then adjusted by a 'levelling' process, so that the maximum errors of the polynomial best approximations are equalized. Finally, the algorithm for fixed knots described above is applied on the levelled knot set.

Generalizations to multivariate splines have mainly been concerned with interpolation problems. But consider bivariate splines on $[a_1, b_1] \times [a_2, b_2]$. This region can be divided into rectangles by knot lines $x = x_i$, $y = y_i$, $i = 1, \ldots, s$, and a tensor product spline space can be defined. As in the univariate problem, partitions can be defined and improved systematically in such a way that best Chebyshev approximations are obtained in the limit. Some recent work on this problem is given by Meinardus, Nürnberger and Walz (1996), Nürnberger (1997). However, there are many unsolved problems: see Nürnberger (1996).

5.2. Chebyshev approximation by rational functions

Another important class of approximation problems is the best Chebyshev approximation of continuous functions by rational functions. The basic problem is as follows: define R_{nm} by

$$R_{nm} = \left\{ \frac{P(x)}{Q(x)} : P(x) = \sum_{j=0}^{n} a_j p_j(x), \quad Q(x) = \sum_{k=0}^{m} b_k q_k(x), \right.$$
$$\left. Q(x) > 0 \ \text{ on } [a, b] \right\}.$$

Then, given $f(x) \in C[a, b]$, we need to determine $R \in R_{nm}$ to minimize $\|f - R\|$, using the Chebyshev norm on $C[a, b]$. For the special case when

$P(x)$ and $Q(x)$ are polynomials of degree n and m, respectively, existence of a best approximation is guaranteed, is unique (up to a normalization), and is characterized by an alternating set of $n + m + 2 - d$ points,

$$\mathcal{A}(f - R)_{[a,b]} \geq n + m + 2 - d,$$

where d is the defect of the approximation, that is, the minimum difference between the *actual* degree of $P(x)$ and $Q(x)$ and n and m, respectively. If $d > 0$, the best approximation is said to be degenerate. For more general quotients, existence is no longer guaranteed, although characterization results are available (though not necessarily of alternation type), and uniqueness results may be extended.

For rational approximation by quotients of polynomials on an interval, the analogue of the Remes exchange method may be applied. It assumes nondegeneracy of the best approximation, and second-order convergence can be obtained. The system of linear equations that needs to be solved in the linear problem is replaced by a nonlinear system in the rational problem, equivalent to an eigenvalue problem, and various methods were proposed for this in the 1960s. Breuer (1987) has suggested a different direct approach to this subproblem, which uses continued fraction interpolation, and which, it is claimed, can lead to a considerable increase in efficiency, and also accuracy and robustness.

If attention is restricted to a discrete subset of $[a, b]$, with positivity of $Q(x)$ only required on the discrete set, then existence of best approximations is no longer guaranteed, even in the polynomial case, and characterization and uniqueness results are no longer valid. The Remes algorithm nevertheless may be applied, although a serious competitor is the differential correction algorithm, first proposed by Cheney and Loeb (1961), and further analysed by Barrodale, Powell and Roberts (1972), Cheney and Powell (1987). The method, which consists of a sequence of linear programming problems, has guaranteed convergence from any starting point, with quadratic convergence in the absence of degeneracy. It may in theory be applied to problems on intervals (Dua and Loeb 1973), but the solution of the subproblems is not straightforward.

Let the discrete subset on which a solution is required be x_i, $i = 1, \ldots, t$, and let R^D_{nm} be the set

$$R^D_{nm} = \left\{ \frac{P(x)}{Q(x)} : P(x) = \sum_{j=0}^{n} a_j p_j(x), \quad Q(x) = \sum_{k=0}^{m} b_k q_k(x), \right.$$
$$\left. Q(x_i) > 0, \quad i = 1, \ldots, t \right\}.$$

Let $P/Q \in R_{nm}^D$, and let Δ satisfy

$$|f(x_i) - P(x_i)/Q(x_i)| \leq \Delta, \quad i = 1, \ldots, t.$$

Then

$$|f(x_i)Q(x_i) - P(x_i)| \leq \Delta Q(x_i), \quad i = 1, \ldots, t. \tag{5.3}$$

Expanding the right-hand side, regarded as the function $g(\Delta, Q(x_i)) = \Delta Q(x_i)$, in a Taylor series about $(\Delta_k, Q_k(x_i))$ gives for each i

$$\Delta Q(x_i) = \Delta_k Q_k(x_i) + (\Delta - \Delta_k)Q_k(x_i) + (Q(x_i) - Q_k(x_i))\Delta_k + \ldots,$$

so to first-order terms the ith term of (5.3) can be written

$$\begin{aligned} &|f(x_i)Q(x_i) - P(x_i)| \\ &\quad \leq \Delta_k Q_k(x_i) + (\Delta - \Delta_k)Q_k(x_i) + (Q(x_i) - Q_k(x_i))\Delta_k \\ &\quad = (\Delta - \Delta_k)Q_k(x_i) + Q(x_i)\Delta_k. \end{aligned}$$

Thus, to first-order terms,

$$|f(x_i)Q(x_i) - P(x_i)| - \Delta_k Q(x_i) \leq (\Delta - \Delta_k)Q_k(x_i), \quad i = 1, \ldots, t,$$

or

$$\max_{1 \leq i \leq t} \left\{ \frac{|f(x_i Q(x_i) - P(x_i)| - \Delta_k Q(x_i)}{Q_k(x_i)} \right\} + \Delta_k \leq \Delta. \tag{5.4}$$

The differential correction algorithm is as follows.

(1) Choose an initial approximation $R_1 = P_1/Q_1 \in R_{nm}^D$; set $k = 1$.

(2) Determine $P(x)$ and $Q(x)$ to minimize the left-hand side of (5.4), where

$$\Delta_k = \max_{1 \leq i \leq t} |f(x_i) - P_k(x_i)/Q_k(x_i)|,$$

and Q_k is not identically zero.

(3) Set $P_{k+1} = P$, $Q_{k+1} = Q$, $k = k + 1$ and continue unless there is convergence.

This algorithm generates a monotonic sequence of numbers that decrease to the minimum error. Starting with $Q_1(x_i) > 0$, $i = 1, \ldots, t$, subsequent denominators retain this property. When the solution is unique, Cheney and Powell (1987) show that convergence is at least superlinear. Barrodale et al. (1972) show that quadratic convergence is obtained in the polynomial case in the absence of degeneracy.

A potentially unsatisfactory feature of approximation from R_{nm} (or R_{nm}^D) is that the denominator, although positive, can become arbitrarily close to zero at certain points. It is not sufficient simply to impose a lower bound on Q, because of the possibility of multiplying both numerator and denominator by an arbitrary constant. A modification of the differential correction algorithm that applies to problems with a lower bound on the denominator

and upper bounds on the absolute values of the coefficients b_i is given by Kaufman and Taylor (1981). It is more natural, however, to impose upper and lower bounds on the denominators themselves ('constrained denominators'). Some aspects of uniqueness are considered by Li and Watson (1997). A modified differential correction algorithm for this problem has been given by Gugat (1996*a*). This involves a constraint of the form

$$\mu(x) \leq Q(x) \leq \nu(x), \tag{5.5}$$

over the appropriate set, where μ and ν are continuous functions. (In fact the algorithm applies to much more general problems than the one considered here, possibly defined on intervals, even having nonlinear expressions P and Q.) The subproblem corresponding to (5.5) above differs in that the additional conditions

$$\mu(x_i) \leq Q(x_i) \leq \nu(x_i), \quad i = 1, \ldots, t \tag{5.6}$$

are imposed. However, it differs also in that, whereas the original method starts with an arbitrary approximation, with denominator Q_1 positive on $x_i, i = 1, \ldots, t$, and with error Δ_1, the method of Gugat (1996*a*) can start with an arbitrary number Δ_1 that is allowed to be smaller than the current error, and an arbitrary (feasible) denominator Q_1. This flexibility turns out to be an important advantage: for example, numerical results show that the choice $\Delta_1 = 0$ is a good one. Subsequent Δ_k are defined as in the original algorithm, but with the constraints (5.6) included in step (2). It is shown by Gugat (1996*a*) that convergence results for the original version carry over. The development outlined above shows that the differential correction algorithms have links with Newton's method. Other methods using variants of Newton's method are those of Hettich and Zenke (1990) and Gugat (1996*b*). However, in contrast to the methods considered here, these do not generate a monotonic sequence.

6. L_1 approximation of functions

While the theory of best Chebyshev approximation to functions has (perhaps quite naturally) received considerable attention, the same cannot be said for best L_1 approximation. Given the same setting as at the start of Section 5, the problem is

$$\text{minimize} \int_X |f(x) - \sum_{i=1}^{n} c_i \phi_i(x)| \, \mathrm{d}x. \tag{6.1}$$

For given $\mathbf{c}$, let $Z(\mathbf{c})$ denote the zeros of $f(x) - \sum_{i=1}^{n} c_i \phi_i(x)$ in X, and for points where this is nonzero, let $g(x, \mathbf{c})$ denote the sign. (We may define g to be zero at other points.) Define

$$V(\mathbf{c}) = \{v(x) : \|v\|_\infty \leq 1, \quad v(x) = g(x, \mathbf{c}), x \notin Z\}.$$

Then it may be shown (for example, Watson 1980) that $\mathbf{c}$ is a solution to (6.1) if and only if there exists $v \in V(\mathbf{c})$ such that

$$\int_X v(x)\phi_j(x)\,\mathrm{d}x = 0, \quad j = 1, \ldots, n. \tag{6.2}$$

Further, if the system of functions $\{\phi_i(x), \ldots, \phi_n(x)\}$ forms a Chebyshev set on $[a, b]$, the best approximation is unique.

If the measure of the set Z is zero (for example if Z just consists of a finite set of points) then clearly (6.2) can be written as

$$\int_X g(x, \mathbf{c})\phi_j(x) = 0, \quad j = 1, \ldots, n.$$

This corresponds to the case when the norm is differentiable, and the above equations are just zero-derivative conditions with respect to the components of $\mathbf{c}$. The likelihood of these being appropriate in practice means that usually the problem is a smooth one. It also means that great store is placed on the points where there are sign changes, or equivalently where the approximation interpolates f. If these points were known, and were exactly n in number, then we could compute the best approximation by interpolation, *provided that there were no other changes of sign in the error of the resulting approximation.*

Definition 5 The points $x_1 < \cdots < x_t \in (a, b) = (x_0, x_{t+1})$, where $1 \leq t \leq n$, are called *canonical points* if

$$\sum_{i=0}^{t} (-1)^i \int_{x_i}^{x_{i+1}} \phi_j(x)\,\mathrm{d}x = 0, \quad j = 1, \ldots, n.$$

In the Chebyshev set case, existence and uniqueness of such a set of points were established by Micchelli (1977). For approximation by polynomials of degree $n - 1$ in $[a, b]$, $t = n$ and the location of those canonical points is known – they lie at the zeros of the Chebyshev polynomial of the second kind of degree n (shifted if necessary). Thus their location is independent of f. Interpolation at these points can quite frequently result in the best polynomial approximation. (For further analysis of the L_1 problem, see Pinkus 1989.)

Example 5 Consider the approximation of $f(x) = 5 + 6e^{2x} + 2\sin(4x)$ by polynomials of degree $n - 1$ on $[-1, 1]$. Table 1 gives the outcome of determining a polynomial by interpolation at the zeros of the second kind Chebyshev polynomial $U_n(x)$. Shown are the number of zeros of the error in $[-1, 1]$, the value of the l_1 norm for the approximation given by the interpolant, and the minimum value of the norm. Clearly, when the number of zeros equals n, the best l_1 approximation is obtained and the norm is the minimum norm, otherwise it is not.

Table 1. *Interpolating polynomials*

n	no of zeros	norm	minimum norm
2	2	7.658748	7.658748
3	5	1.022593	0.816405
4	5	0.986141	0.812920
5	5	0.704263	0.704263
6	7	0.081135	0.063107
7	7	0.061799	0.061799
8	9	0.005072	0.003947
9	9	0.003937	0.003937
10	11	0.000192	0.000150
11	11	0.000150	0.000150

An algorithm for computing best L_1 approximations from general linear subspaces is given by Watson (1981). It is essentially of exchange type, based on the calculation of the zeros of the error at each iteration, and the construction of descent directions. It is also of Newton type, since it constructs the Hessian matrix of the error when it exists, and can have a second-order convergence rate. In a sense, it can be thought of as analogous to the second algorithm of Remes for Chebyshev problems, where a sequence of sets of zeros plays the role of a sequence of sets of extreme points; the connection with Newton's method under appropriate circumstances is also something the methods have in common. An algorithm for nonlinear problems is given by Watson (1982).

For best L_1 approximation by splines with fixed knots, it is known that every continuous function has a unique best approximation. Further, under certain assumptions, the best approximation can be determined by interpolation at canonical points. These results go back to Micchelli (1977). Little, if any, practical work has been done on this or more general problems.

Acknowledgement
I am grateful to Yuying Li, Kaj Madsen, Gunther Nürnberger, Mike Osborne and Mike Powell, who were kind enough to read parts of this paper and make constructive comments.

REFERENCES

I. Barrodale and C. Phillips (1975), An improved algorithm for discrete Chebyshev linear approximation, in *Proc. 4th Manitoba Conf. on Numer. Math.* (B. L. Hartnell and H. C. Williams, eds), University of Manitoba (Winnipeg, Canada), pp. 177–190.

I. Barrodale and F. D. K. Roberts (1973), 'An improved algorithm for discrete l_1 linear approximation', *SIAM J. Numer. Anal.* **10**, 839–848.

I. Barrodale, M. J. D. Powell and F. D. K. Roberts (1972), 'The differential correction algorithm for rational l_∞ approximation', *SIAM J. Numer. Anal.* **9**, 493–504.

R. Bartels, A. R. Conn and Y. Li (1989), 'Primal methods are better than dual methods for solving overdetermined linear systems in the l_∞ sense?', *SIAM J. Numer. Anal.* **26**, 693–726.

R. Bartels, A. R. Conn and J. W. Sinclair (1978), 'Minimisation techniques for piecewise differentiable functions: the l_1 solution to an overdetermined linear system', *SIAM J. Numer. Anal.* **15**, 224–241.

P. Bloomfield and W. L. Steiger (1983), *Least Absolute Deviations*, Birkhäuser, Boston.

P. T. Boggs, R. H. Byrd and R. B. Schnabel (1987), 'A stable and efficient algorithm for nonlinear orthogonal distance regression', *SIAM J. Sci. Statist. Comput.* **8**, 1052–1078.

P. T. Boggs, R. H. Byrd, J. R. Donaldson and R. B. Schnabel (1989), 'ODRPACK, Software for weighted orthogonal distance regression', *ACM Trans. Math. Software* **15**, 348–364.

C. G. Boncelet, Jr. and B. W. Dickinson (1984), 'A variant of Huber robust regression', *SIAM J. Sci. Statist. Comput.* **5**, 720–734.

D. Braess (1971), 'Chebyshev approximation by spline functions with free knots', *Numer. Math.* **17**, 357–366.

P. T. Breuer (1987), A new method for real rational uniform approximation, in *Algorithms for Approximation* (J. C. Mason and M. G. Cox, eds), Clarendon Press, Oxford, pp. 265–283.

B. Brosowski and A. R. da Silva (1992), A general alternation theorem, in *Approximation Theory* (G. A. Anastassiou, ed.), Marcel Dekker, Inc., New York, pp. 137–150.

E. W. Cheney and H. L. Loeb (1961), 'Two new algorithms for rational approximation', *Numer. Math.* **3**, 72–75.

E. W. Cheney and M. J. D. Powell (1987), 'The differential correction algorithm for generalized rational functions', *Constr. Approx.* **3**, 249–256.

D. I. Clark (1985), 'The mathematical structure of Huber's M-estimator', *SIAM J. Sci. Statist. Comput.* **6**, 209–219.

D. I. Clark and M. R. Osborne (1986), 'Finite algorithms for Huber's M-estimator', *SIAM J. Sci. Statist. Comput.* **7**, 72–85.

T. Coleman and Y. Li (1992*a*), 'A globally and quadratically convergent affine scaling algorithm for l_1 problems', *Math. Prog.* **56**, 189–222.

T. Coleman and Y. Li (1992*b*), 'A globally and quadratically convergent method for linear l_∞ problems', *SIAM J. Numer. Anal.* **29**, 1166–1186.

A. Dax (1989), 'The minimax solution of linear equations subject to linear constraints', *IMA J Numer. Anal.* **9**, 95–109.

A. Dax (1993), 'A row relaxation method for large minimax problems', *BIT* **33**, 262–276.

A. Dax and B. Berkowitz (1990), 'Column relaxation methods for least norm problems', *SIAM J. Sci. Statist. Comput.* **11**, 975–989.

S. N. Dua and H. L. Loeb (1973), 'Further remarks on the differential correction algorithm', *SIAM J. Numer. Anal.* **10**, 123–126.

A. M. Duarte and R. J. Vanderbei (1994), Interior point algorithms for lsad and lmad estimation, Technical Report SOR-94-07, Programs in Operations Research and Statistics, Princeton University.

H. Ekblom and H. B. Nielsen (1996), A comparison of eight algorithms for computing M-estimates, Technical Report 1996-15, Technical University of Denmark.

G. H. Golub and C. F. Van Loan (1980), 'An analysis of the total least squares problem', *SIAM J. Numer. Anal.* **17**, 883–893.

M. Gugat (1996*a*), 'An algorithm for Chebyshev approximation by rationals with constrained denominators', *Constr. Approx.* **12**, 197–221.

M. Gugat (1996*b*), 'The Newton differential correction algorithm for uniform rational approximation with constrained denominators', *Numer. Algorithms* **13**, 107–122.

R. Hettich and P. Zenke (1990), 'An algorithm for general restricted rational Chebyshev approximation', *SIAM J. Numer. Anal.* **27**, 1024–1033.

Z. Jing and A. T. Fam (1987), 'An algorithm for computing continuous Chebyshev approximations', *Math. Comp.* **48**, 691–710.

K. Jonasson (1993), 'A projected conjugate gradient method for sparse minimax problems', *Numer. Algorithms* **5**, 309–323.

K. Jonasson and K. Madsen (1994), 'Corrected sequential linear programming for sparse minimax optimization', *BIT* **34**, 372–387.

K. Jonasson and G. A. Watson (1982), A Lagrangian method for multivariate continuous Chebyshev approximation problems, in *Multivariate Approximation Theory 2* (W. Schempp and K. Zeller, eds), Birkhäuser, Basel, pp. 211–221.

E. H. Kaufman and G. D. Taylor (1981), 'Uniform approximation by rational functions having restricted denominators', *J. Approx. Theory* **32**, 9–26.

H. Kawasaki (1994), 'A second-order property of spline functions with one free knot', *J. Approx. Theory* **78**, 293–297.

C. Li and G. A. Watson (1997), 'Strong uniqueness in restricted rational approximation', *J. Approx. Theory* **89**, 96–113.

W. Li and J. J. Swetits (1998), 'Linear l_1 estimator and Huber M-estimator', *SIAM J. Optim.* To appear.

Y. Li (1993), 'A globally convergent method for L_p problems', *SIAM J. Optim.* **3**, 609–629.

K. Madsen and H. B. Nielsen (1993), 'A finite smoothing algorithm for linear l_1 estimation', *SIAM J. Optim.* **3**, 223–235.

K. Madsen, H. B. Nielsen and M. C. Pinar (1994), 'New characterizations of l_1 solutions of overdetermined linear systems', *Operations Research Lett.* **16**, 159–166.

G. Meinardus (1967), *Approximation of Functions: Theory and Numerical Methods*, Springer, Berlin.

G. Meinardus, G. Nürnberger and G. Walz (1996), 'Bivariate segment approximation and splines', *Adv. Comput. Math.* **6**, 25–45.

G. Meinardus, G. Nürnberger, M. Sommer and H. Strauss (1989), 'Algorithms for piecewise polynomials and splines with free knots', *Math. Comp.* **53**, 235–247.

M. S. Meketon (1987), Least absolute value regression, Technical report, AT&T Bell Laboratories, Murray Hill, New Jersey.

C. A. Micchelli (1977), 'Best L^1 approximation by weak Chebyshev systems and the uniqueness of interpolating perfect splines', *J. Approx. Theory* **19**, 1–14.

P. A. P. Moran (1959), 'Random processes in economic theory and analysis', *Sankhya* **21**, 99–126.

B. Mulansky (1992), 'Chebyshev approximation by spline functions with free knots', *IMA J. Numer. Anal.* **12**, 95–105.

G. Nürnberger (1987), 'Strongly unique spline approximation with free knots', *Constr. Approx.* **3**, 31–42.

G. Nürnberger (1989), *Approximation by Spline Functions*, Springer, Berlin.

G. Nürnberger (1994*a*), Approximation by univariate and bivariate splines, in *Second International Colloquium on Numerical Analysis* (D. Bainov and V. Covachev, eds), VSP, Utrecht, pp. 143–153.

G. Nürnberger (1994*b*), 'Strong unicity in nonlinear approximation and free knot splines', *Constr. Approx.* **10**, 285–299.

G. Nürnberger (1996), 'Bivariate segment approximation and free knot splines: research problems 96-4', *Constr. Approx.* **12**, 555–558.

G. Nürnberger (1997), Optimal partitions in bivariate segment approximation, in *Surface Fitting and Multiresolution Methods* (A. Le Méhauté, C. Rabut and L. L. Schumaker, eds), Vanderbilt University Press, Nashville, pp. 271–278.

G. Nürnberger and M. Sommer (1983), 'A Remez type algorithm for spline functions', *Numer. Math.* **41**, 117–146.

G. Nürnberger, L. L. Schumaker, M. Sommer and H. Strauss (1985), 'Approximation by generalized splines', *J. Math. Anal. Appl.* **108**, 466–494.

G. Nürnberger, L. L. Schumaker, M. Sommer and H. Strauss (1989), 'Uniform approximation by generalized splines with free knots', *J. Approx. Theory* **59**, 150–169.

G. Nürnberger, M. Sommer and H. Strauss (1986), 'An algorithm for segment approximation', *Numer. Math.* **48**, 463–477.

M. R. Osborne (1985), *Finite Algorithms in Optimisation and Data Analysis*, Wiley, Chichester.

M. R. Osborne (1987), The reduced gradient algorithm, in *Statistical Data Analysis Based on the L_1 Norm and Related Methods* (Y. Dodge, ed.), North Holland, Amsterdam, pp. 95–107.

M. R. Osborne and G. A. Watson (1985), 'An analysis of the total approximation problem in separable norms, and an algorithm for the total l_1 problem', *SIAM J. Sci. Statist. Comput.* **6**, 410–424.

M. R. Osborne and G. A. Watson (1996), Aspects of M-estimation and l_1 fitting problems, in *Numerical Analysis: A R Mitchell 75th Birthday Volume* (D. F. Griffiths and G. A. Watson, eds), World Scientific, Singapore, pp. 247–261.

A. Pinkus (1989), *On L^1 Approximation*, Cambridge University Press, Cambridge.

J. R. Rice (1967), 'Characterization of Chebyshev approximation by splines', *SIAM J. Math. Anal.* **4**, 557–567.

J. B. Rosen, H. Park and J. Glick (1996), 'Total least norm formulation and solution for structured problems', *SIAM J. Matrix Anal. Appl.* **17**, 110–126.

J. B. Rosen, H. Park and J. Glick (1997), Total least norm for linear and nonlinear structured problems, in *Recent Advances in Total Least Squares Techniques and Errors-in-Variables Modeling* (S. Van Huffel, ed.), SIAM, Philadelphia, pp. 203–214.

S. A. Ruzinsky and E. T. Olsen (1989), 'l_1 and l_∞ minimization via a variant of Karmarkar's algorithm', *IEEE Trans. Acoustics, Speech and Signal Processing* **37**, 245–253.

L. L. Schumaker (1968), 'Uniform approximation by Tchebycheffian spline functions', *J. Math. Mech.* **18**, 369–378.

M. Shi and M. A. Lukas (1996), 'On the reduced gradient algorithm for L_1 norm minimization with linear constraints'. Research Report, Department of Mathematics and Statistics, Murdoch University, Australia.

H. Späth and G. A. Watson (1987), 'On orthogonal linear l_1 regression', *Numer. Math.* **51**, 531–543.

G. A. Watson (1976), 'A method for calculating best nonlinear Chebyshev approximations', *J. Inst. Math. Appl.* **18**, 351–360.

G. A. Watson (1980), *Approximation Theory and Numerical Methods*, Wiley, Chichester.

G. A. Watson (1981), 'An algorithm for linear L_1 approximation of continuous functions', *IMA J. Numer. Anal.* **1**, 157–167.

G. A. Watson (1982), A globally convergent method for (constrained) nonlinear continuous L_1 approximation problems, in *Numerical Methods of Approximation Theory 1981* (L. Collatz, G. Meinardus and H. Werner, eds), Birkhäuser, Berlin, pp. 233–243. ISNM 59.

G. A. Watson (1983), The total approximation problem, in *Approximation Theory IV* (C. K. Chui, L. L. Schumaker and J. D. Ward, eds), Academic Press, New York, pp. 723–728.

G. A. Watson (1984), The numerical solution of total l_p approximation problems, in *Numerical Analysis Dundee 1983* (D. F. Griffiths, ed.), Springer, Berlin, pp. 221–238.

G. A. Watson (1987), Methods for best approximation and regression problems, in *The State of the Art in Numerical Analysis* (A. Iserles and M. J. D. Powell, eds), Clarendon Press, Oxford, pp. 139–164.

G. A. Watson (1988), 'The smallest perturbation of a submatrix which lowers the rank of the matrix', *IMA J. Numer. Anal.* **8**, 295–303.

G. A. Watson (1991), 'On a general class of matrix nearness problems', *Constr. Approx.* **7**, 299–314.

G. A. Watson (1997), The use of the L_1 norm in nonlinear errors-in-variables problems, in *Recent Advances in Total Least Squares Techniques and Errors-in-Variables Modeling* (S. Van Huffel, ed.), SIAM, Philadelphia, pp. 183–192.

G. A. Watson and K. F. C. Yiu (1991), 'On the solution of the errors in variables problem using the l_1 norm', *BIT* **31**, 697–710.

Y. Zhang (1993), 'A primal-dual interior point approach for computing the l_1 and l_∞ solutions of overdetermined linear systems', *J. Optim. Theory Appl.* **77**, 323–341.