On the existence and uniqueness of estimates in robust and heteroscedastic regression models

Thesis submitted to the University of London for the degree
of Doctor of Philosophy in the Faculty of Science

by

Adam Crisp
University College London

October 1994

Abstract

The thesis studies redescending M-estimators for the ordinary linear regression model, and maximum likelihood estimators for heteroscedastic regression models. In general, redescending M-estimators do not yield unique estimates of the model parameters, and the thesis shows that the difficulties associated with this have not always been fully appreciated in the literature. This motivates the development of an approach whereby unique redescending M-estimates can be reliably obtained. This is achieved by embedding the linear model within a multivariate t location-scatter framework, which is known in the literature for its desirable uniqueness properties. M-estimates derived from the conditional t distribution are also considered, but it is shown that the resulting objective function is intrinsically multimodal, with modes of infinity.

The nonregularity result for the conditional t model is found to have implications for heteroscedastic regression models. Two classes of commonly proposed models are considered. The first is found to yield an unbounded likelihood at points corresponding to nonreplicated observations, whilst for the second a much stronger linear independence condition is obtained for the likelihood to be unbounded.

The thesis concludes with a discussion of efficient methods for testing analytical conditions arising from the preceding studies.

Keywords: conditional distribution, heteroscedastic regression, mean-variance relationship, multimodality, multivariate t distribution, redescending M-estimator, robust regression, singularity, uniqueness.

Acknowledgements

I would like to thank my supervisor, Jim Burridge, for the ideas, encouragement and friendship that he has offered over the last few years. Thanks are also due to the many people who have given me support and advice. I would like to mention in particular Maria Efstathiou, Costas Goutis, Kostas Skouras, Steve Gallivan and an Associate Editor of *Biometrika.* I also thank Charles Hayward, who got me started.

Contents

Introduction

The work reported here relates to the study of redescending M-estimators for the ordinary linear regression model and maximum likelihood estimators for heteroscedastic regression models. The connection between these topics arises from the development of an approach that overcomes the well-known uniqueness problems associated with redescending M-estimators.

An M-estimator is one of many estimators that define statistics which are in some sense *robust* (Hampel, Ronchetti, Rousseeuw, and Stahel, 1986). Such estimators arise from the need to construct procedures which are not too sensitive to deviations from assumptions, such as normality and independence, that form the foundations upon which many classical procedures are built. For example, a common type of deviation is the presence of outliers, namely data which are away from the pattern set by the majority. These may be due to a genuinely long-tailed distribution or perhaps mistakes made in data collection and can, for example, have an arbitrarily large effect on estimates obtained from methods of least-squares.

Robust methods probably date back to the prehistory of statistics (Hampel et al., p. 34), although a general theory regarding robust estimation has really only evolved over the last thirty years or so. A key development in this area was the introduction of the M-estimator (Huber, 1964) for robust estimation of a location parameter, and generalisations to more complicated estimation problems have followed. Of particular interest are the so-called redescending M-estimators, which are attractive in terms of their robustness properties but do not usually guarantee unique estimates of the model parameters. This has led to research into obtaining sufficient conditions under which a given sample will yield a unique redescending estimate, but these are not always useful. It is therefore of interest to consider an approach whereby unique redescending M-estimates can be reliably obtained, and this provides the principal motivation for the work presented here, which is structured as follows.

Chapter 1 presents an introduction to M-estimators and reviews the case for using the sub-class known as redescenders. This discussion is continued in Chapter 2 where attention is brought to an important, but generally unrecognised difficulty that can arise within the redescending framework. This is found to affect strongly the uniqueness results for the linear model obtained by Rivest (1989), and provides further motivation for the development of a new redescending approach.

Such an approach is considered in Chapter 3, where it is proposed that unique redescending estimates may be obtained if the linear model is embedded within a multivariate location-scatter model. Sufficient conditions for the existence of unique estimates in the location-scatter case have been developed by Kent and Tyler (1991), and these are of particular use when applied to the multivariate t distribution. Unique estimates for the linear model parameters are obtained from unique redescending estimates of location and scatter, and examples based on real data sets presented. They suggest that the multivariate t approach is a useful addition to the methods for robust regression already available. M-estimates derived from the conditional t distribution are also considered, but it is shown that the resulting objective function is intrinsically multimodal, with modes of infinity.

In Chapter 4 the nonregularity result for the conditional t model is found to have wider implications. It is observed that the essence of the problem lies in the heteroscedastic form of the conditional t model, and this motivates a more general investigation of heteroscedastic regression models. Two classes of commonly proposed models are considered. The first is found to yield an unbounded likelihood at points corresponding to nonreplicated observations, whilst for the second a much stronger linear independence condition is obtained for the likelihood to be unbounded. Numerical examples explore the practical difficulties that can arise in these cases.

Chapter 5 discusses efficient methods for testing two existence conditions arising from the multivariate t and heteroscedastic regression studies, and concluding remarks are presented in Chapter 6. Some computational details are included in Appendix A.

Chapter 1

M-estimators

This chapter presents background material in order to 'set the scene' for following chapters. It is divided into two sections. The first provides an introduction to M-estimators and how they are defined for the linear model. The second section deals with the sub-class known as redescending M-estimators.

1.1 Introduction

1.1.1 Simple Location Model

The definition of an M-estimator may be motivated by considering the method of maximum likelihood estimation for a one-dimensional location parameter. Suppose $X_1, \ldots, X_n$ are independent and identically distributed (i.i.d.) one-dimensional random variables. A parametric model may be defined as a family of probability distributions F_θ on the sample space, indexed by an unknown location parameter θ belonging to some parameter space Θ. Denoting the densities as f_θ the well-known maximum likelihood estimator is defined as the value $T_n = T_n(X_1, \ldots, X_n)$ which maximises $\prod_i f_{T_n}(X_i)$, or equivalently by the value T_n which minimises $-\sum_i \log f_{T_n}(X_i)$. Huber (1964) proposed that minimizing some other function may achieve more robustness, and so considered estimators that can be defined by a more general minimization principle of the form:

$$T_n \quad \text{minimises} \quad \sum_{i=1}^{n} \rho\,(X_i, T_n) \tag{1.1}$$

where ρ is a non-constant function. If ρ has a derivative $\psi(x,\theta) = (\partial/\partial\theta)\rho(x,\theta)$, then for useful choices of ρ the estimator T_n satisfies the implicit equation

$$\sum_{i=1}^{n} \psi(X_i, T_n) = 0. \tag{1.2}$$

Definition 1.1. Any estimator defined by (1.1) or (1.2) is an M-estimator.

Though (1.1) and (1.2) are not always equivalent, for the sake of brevity M-estimators are often defined through a given ψ-function. For example, $\psi = -f'/f$ is equivalent to maximum likelihood estimation. Typically ψ is odd and, if strictly monotonically increasing, will yield a unique solution to (1.2) due to the convexity of ρ. Otherwise a solution to (1.2) may not be unique and difficulties may arise. These are discussed in Section 1.2 and in Chapter 2. There follow some examples of monotone ψ-functions.

Example 1.1. The Huber M-estimator (Huber, 1964) with cut-off point c is given by

$$\psi_c(x) = \min\{c, \max\{x, -c\}\}$$

for $0 < c < \infty$. It has minimax variance over the distributions contained in a particular neighbourhood of the normal distribution (Hampel et al., 1986, p. 172).

Example 1.2. A similar, and perhaps aesthetically more pleasing example arises from the distribution function of the scaled logistic distribution $F_\sigma(x) = \{1 + \exp(-x/\sigma)\}^{-1}$, for which we obtain the strictly monotone

$$\psi_\sigma(x) = \{2F_\sigma(x) - 1\}/\sigma,$$

where $0 < \sigma < \infty$. The scale parameter σ is equivalent to the Huber cut-off point c in that the larger c and σ are, the more similar the solution will be to least-squares, whilst smaller values tend to give more robustness.

Robustness properties of an estimator may be obtained via the ***influence function*** (Hampel et al., 1986, p. 84). It describes, at some underlying model distribution F, the effect of an infinitesimal contamination at a point on an estimator, standardised by the mass of the contamination. For M-estimators it reduces to

$$\mathrm{IF}(x; \psi, F) = \frac{\psi(x)}{\int \psi' \, dF}$$

(Hampel et al., 1986, p. 103), under the assumption that the denominator is nonzero. Among the measures that can be derived from the IF are the following (Hampel et al., 1986, pp. 85–88): the asymptotic variance of the estimator, which is simply the expected square of the IF; the *gross-error sensitivity* of ψ at F, given by

$$\gamma^* = \sup_x |\,\mathrm{IF}(x; \psi, F)|\,;$$

and the *rejection point*

$$\rho^* = \inf\{r > 0;\ \mathrm{IF}(x; \psi, F) = 0 \text{ when } |x| > r\}. \qquad (1.3)$$

(If there exists no such r, then $\rho^* = \infty$.) Thus all observations further away than ρ^* are rejected completely. The Huber and logistic M-estimators both have $\rho^* = \infty$. M-estimators for which ρ^* is finite will be discussed in Section 1.2.

The gross-error sensitivity measures the worst influence which a small amount of contamination of fixed size can have on the value of the estimator, and so it is extremely desirable that γ^* be finite. In such cases ψ is described as ***B-robust*** at F. It can be shown that, if F has a unimodal and symmetric density f that satisfies certain differentiability conditions (Hampel et al., 1986, p. 125), then ψ is B-robust if and only if $|\psi|$ is bounded (Hampel et al., 1986, p. 132). It follows that the ψ-functions considered in Examples 1.1 and 1.2 are B-robust for such F.

Furthermore, for a given upper bound on γ^*, the Huber ψ is optimal (in terms of asymptotic variance) B-robust at $F = \Phi$. Clearly an estimator cannot be B-robust if ψ is unbounded: for example, the sample mean is not B-robust, since it corresponds to $\psi(x) = x$.

1.1.2 Linear Regression Model

We now move on to consider the linear regression model, which contains the location model as a special case:

$$y_i = \mathbf{x}_i^T \boldsymbol{\beta} + \epsilon_i, \tag{1.4}$$

where $\{(y_i, \mathbf{x}_i) \colon i = 1, \ldots, n\}$ is a multivariate sample of size n for $y_i \in \mathbb{R}$ and $\mathbf{x}_i \in \mathbb{R}^p$, $\boldsymbol{\beta}$ is a vector of unknown parameters belonging to $\mathbb{R}^p$ and the $\{\epsilon_i\}$ are i.i.d. with distribution function $F(\epsilon_i/\sigma)$, where $\sigma > 0$ is an unknown scale parameter. The form of (1.2) can be easily extended to cope with this more complicated estimation problem. A more general form of (1.1) defines the M-estimator for $\boldsymbol{\beta}$ (Huber, 1973; Hampel et al., 1986, p. 311) as the value $\hat{\boldsymbol{\beta}}$ that minimises

$$\sum_{i=1}^{n} \rho\left(\frac{y_i - \mathbf{x}_i^T \boldsymbol{\beta}}{\sigma}\right), \tag{1.5}$$

or, on taking derivatives, solves the vector equation

$$\sum_{i=1}^{n} \mathbf{x}_i \psi\left(\frac{y_i - \mathbf{x}_i^T \boldsymbol{\beta}}{\sigma}\right) = 0, \tag{1.6}$$

where again, $\psi = \rho'$. Thus $\rho(t) = t^2$ defines the least-squares estimator. There remains the question of estimating σ. A natural approach is to consider an M-estimator for σ, which may be defined as the solution of

$$\sum_{i=1}^{n} \chi\left(\frac{y_i - \mathbf{x}_i^T \hat{\boldsymbol{\beta}}}{\sigma}\right) = 0, \tag{1.7}$$

where χ is some given function. For example, if we take $\chi(t) = \operatorname{sgn}(|t| - 0.6745)$, (Hampel et al., 1986, p. 107), we obtain

$$\hat{\sigma} = \frac{\text{MED}\{|y_i - \mathbf{x}_i^T\hat{\boldsymbol{\beta}}|\}}{0.6745}, \tag{1.8}$$

which is a robust estimator for scale. The division by $\Phi^{-1}\left(\frac{3}{4}\right) \approx 0.6745$ ensures Fisher-consistency (Hampel et al., 1986, p. 83) when the data are normally distributed. Estimates $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}$ of $\boldsymbol{\beta}$ and σ may then be calculated by solving the system of equations (1.6) and (1.7). Algorithms for this task are considered by Huber (1981, Chapter 7).

Unfortunately, the M-estimator defined by (1.6) can only have bounded influence with respect to outliers in the response variable (Hampel et al., 1986, p. 313). Robustness against outliers in the x space may be obtained via a generalised M-estimator for $\boldsymbol{\beta}$, defined as a solution to

$$\sum_{i=1}^{n} \eta\{\mathbf{x}_i, (y_i - \mathbf{x}_i^T\boldsymbol{\beta})/\sigma\}\mathbf{x}_i = 0 \tag{1.9}$$

(Hampel et al., 1986, p. 315), where the function η may be written in the form

$$\eta(\mathbf{x}, t) = w(\mathbf{x}) \cdot \psi(t \cdot v(\mathbf{x})).$$

The Huber M-estimator thus corresponds to $w(\mathbf{x}) = v(\mathbf{x}) = 1$. Robustness against outliers in the x space is achieved through the weight function w, which involves a robust covariance matrix in the x space, to be determined by the solution of further implicit equations. For details, see Hampel et al. (1986, pp. 315–328).

1.2 Redescending M-estimators

1.2.1 General Remarks

We return now to the Huber M-estimator (1.6). In the preceding section it was mentioned that, if ψ is not monotonically increasing, then difficulties may arise from non-unique solutions. This is important since non-monotone ψ-functions are of great interest. The main examples of such functions are the so-called redescenders, and these are considered in the present section.

The motivation for redescending M-estimators arises from the rejection point ρ^* (1.3). Hampel et al. (1986, p. 88) say that it is desirable for ρ^* to be finite, that is to say for M-estimators that there exists a fixed constant $0 < r < \infty$ such that $\psi(t) = 0$ for all $|\,t\,| \geq r$. M-estimators with such ψ-functions are described by Hampel as being *redescending.* However, this definition has not always been used in the literature: for example, Maronna and Yohai (1981) and Novovičová (1990) use the term to describe functions such that $\psi(t) \to 0$ as $|\,t\,| \to \infty$, and Holland and Welsch (1977) use the terminology "soft redescender" to describe ψ-functions that merely tend to zero, and "hard redescender" when Hampel's condition is satisfied. Indeed, "soft" redescenders, such as that used in the maximum likelihood estimator for the Cauchy distribution, give very little influence to extreme observations and behave almost like estimators with low rejection point, even though their rejection point is infinite (Hampel, 1974). Following Kent and Tyler (1991) who use the term "strong redescender" for ψ-functions that vanish outside some central region, we shall observe the following:

Definition 1.2. If $\psi(t) \to 0$ as $|\,t\,| \to \infty$, ψ is *weakly* redescending. If $\psi(t) = 0$ for $|\,t\,| > r$, where $0 < r < \infty$, ψ is *strongly* redescending.

It should be noted that the distinction between the two is important, as some conditions related to the existence and uniqueness of redescending M-estimators do not apply in both cases (Kent and Tyler, 1991). Here then, are some examples of redescending ψ-functions, illustrated graphically in Figures 1.1, 1.2 and 1.3:

Example 1.3. A strong redescender is obtained from a single cycle of the *sine* *function*, advocated by Andrews (Andrews et al., 1972):

$$\psi_{\sin(a)}(t) = \begin{cases} \sin(t/a) & \text{if } |\,t\,| < a\pi, \\ 0 & \text{otherwise.} \end{cases}$$

Example 1.4. In Andrews et al. (1972) Hampel proposed a *three-part* (strongly) redescending M-estimator:

$$\psi_{a,b,r}(t) = \begin{cases} t & \text{if } 0 \le |\,t\,| \le a, \\ a\,\mathrm{sgn}(t) & \text{if } a \le |\,t\,| \le b, \\ a\left(\dfrac{r - |\,t\,|}{r - b}\right)\mathrm{sgn}(t) & \text{if } b \le |\,t\,| \le r, \\ 0 & \text{otherwise,} \end{cases}$$

where $0 < a < b < r < \infty$.

Example 1.5. A weakly redescending M-estimator is obtained from the p-dimensional t distribution on $\nu > 0$ degrees of freedom:

$$\psi_{\nu,p}(t) = \frac{(\nu + p)\,t}{\nu + t^2}, \qquad \forall\, t \in \mathbb{R}.$$

All redescending M-estimators are B-robust since $|\psi|$ is bounded by definition (Hampel et al., 1986, p. 153). However, it is their ability to reject extreme outliers which makes them preferable to monotone ψ-functions. Hampel et al. (1986, pp. 166–167) compare the asymptotic variances for some strongly redescending location M-estimators with those of the Huber and scaled logistic M-estimators. They find that the Huber and logistic are the most efficient at the normal model and that they do well at relatively short-tailed distributions. However, their variance goes up considerably at distributions which produce large outliers, where the redescenders are up to 20% more efficient. At such distributions, the monotone ψ suffer in that they can never reject an outlier, no matter how far away.

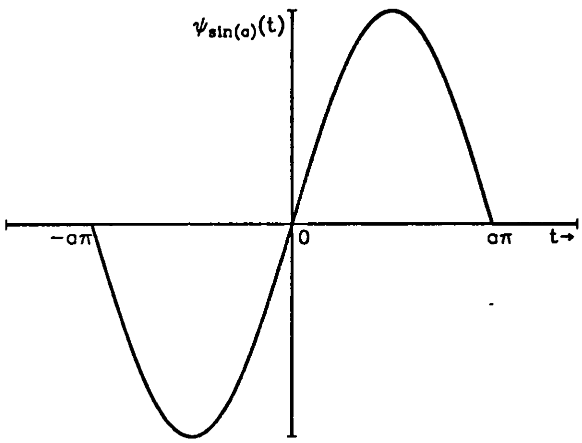

Figure 1.1: Andrews' sine function.

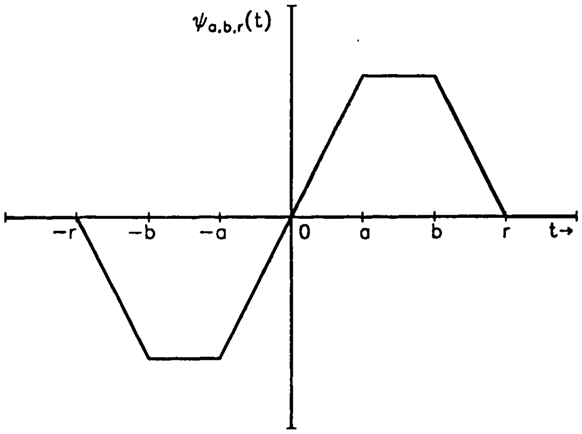

Figure 1.2: Shape of Hampel's three-part redescending ψ-function.

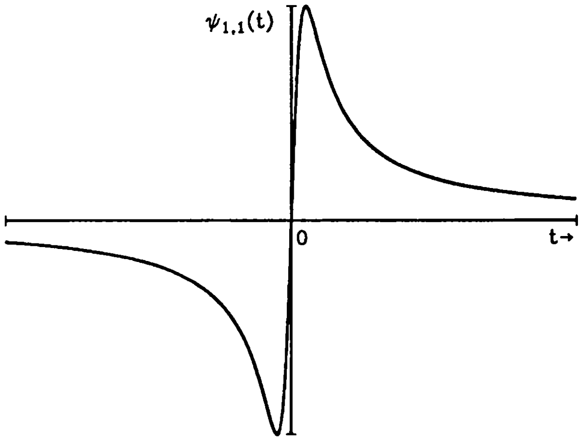

Figure 1.3: Shape of the weakly redescending $t_{\nu,p}$ M-estimator.

1.2.2 The Uniqueness Problem

Though the motivation for using redescending M-estimators is clear, their lack of monotonicity ensures that if a solution to (1.6) exists, it may not be unique. To counter this difficulty, several approaches have been proposed. Klein and Yohai (1981), in a study of asymptotic behaviour, overcome the difficulty by defining the estimate as the limit (if it exists) of a given iterative sequence, and Huber (1981, p. 192) suggests that one might 'start with a monotone ψ, iterate to death, and then append a few (1 or 2) iterations with the non-monotone ψ'. However, Huber does not discuss the choice of monotone ψ or the adequacy of the solution thus obtained. This is perhaps not too surprising, given his earlier comments (p. 103) to the effect that, in his opinion, the difficulties associated with redescenders more than offset the improvements in asymptotic variance obtained at heavy-tailed distributions.

A similar approach is to compute only a *one-step* M-estimator (Bickel, 1975), where the estimate is computed from one iteration of a Newton-Raphson type algorithm, using a robust estimate for the starting value. For non-monotone ψ the estimator is safe in that the problem of uniqueness is avoided, but the approach is somewhat circular in that one requires a robust estimate in order to obtain a robust estimate.

For the regression problem, a very robust starting value for $\boldsymbol{\beta}$ is the *least median of squares* estimator (Rousseeuw, 1984), defined as the value that minimises

$$m_{\boldsymbol{\beta}} = \mathrm{MED}\{(y_i - \mathbf{x}_i^T\boldsymbol{\beta})^2\}$$

over $\boldsymbol{\beta}$, but this proposal presents great computational difficulties, though algorithms are available (Souvaine and Steele, 1987).

Rather than develop alternative computational strategies, one might consider obtaining sufficient conditions for a system of estimating equations to yield a unique solution. This can be done even if one specifies only that the equations are real-valued and continuous (Zeidler, 1986), but the price paid for such generality is that the conditions obtained are extremely strong, and almost impossible to test for all but the simplest of problems. Conditions have been developed for the M-estimation framework by, for example, Maronna and Yohai (1981) and Rivest (1989). However, these do not seem to provide a straightforward means of testing the uniqueness (or indeed existence) of a solution to (1.6). Indeed, the results in Rivest (1989) warrant detailed consideration, for in the event that (1.6) does not admit a unique solution, Rivest defines $\hat{\boldsymbol{\beta}}$ as the value of $\boldsymbol{\beta}$ that corresponds to the global minimum of (1.5). The tacit assumption is that the global minimum is attained at a unique value of $\boldsymbol{\beta}$. However, this need not be the case, as demonstrated in the next chapter.

Chapter 2

On the uniqueness of *M*-estimates

2.1 Introduction

Rivest (1989) considers the uniqueness of M-estimators for the linear model (1.4). The M-estimator of $\boldsymbol{\beta}$ is defined as the solution of

$$\sum_{i=1}^{n} \mathbf{x}_i \psi \left(\frac{y_i - \mathbf{x}_i^T \boldsymbol{\beta}}{c\sigma} \right) = 0 \tag{2.1}$$

where c is a 'positive constant of robustness', and σ is to be estimated by (1.8). It is noted that (1.8) is a special case of a generalised estimator for σ, given by the solution to

$$\sum_{i=1}^{n} \chi \left(\frac{y_i - \mathbf{x}_i^T \boldsymbol{\beta}}{\sigma} \right) = (n-p)\gamma, \tag{2.2}$$

where $\chi(t)$ is an even function, increasing in $[0, \infty)$, and $\gamma = \mathrm{E}[\chi(Z)]$, where Z is a $\mathcal{N}(0,1)$ random variable. One of Rivest's objectives is to obtain sufficient conditions concerning ψ and χ so that the non-linear system (2.1) and (2.2) has a unique solution; however, they do not seem to be complete. Defining $\psi_k(t) = \psi(t/k)$, the non-linear system

$$\sum_{i=1}^{n} \mathbf{x}_i \psi_k (y_i - \mathbf{x}_i^T \boldsymbol{\beta}) = 0 \tag{2.3}$$

is considered. For monotonic ψ, the system admits a unique solution $\boldsymbol{\beta}_k$; in the contrary case, Rivest defines $\boldsymbol{\beta}_k$ as the solution of (2.3) which minimises

$$\sum_{i=1}^{n} \rho\left(\frac{y_i - \mathbf{x}_i^T \boldsymbol{\beta}}{k}\right). \tag{2.4}$$

For a given $\boldsymbol{\beta}_k$, Rivest also defines

$$\sigma_k = \frac{\text{MED}\{|y_i - \mathbf{x}_i^T \boldsymbol{\beta}_k|\}}{0.6745}. \tag{2.5}$$

Rivest claims that for strongly redescending ψ the definition of $\boldsymbol{\beta}_k$ excludes from the study all the solutions of (2.1) and (2.2) which do not minimise (2.4). However, the possibility that (2.4) does not uniquely define $\boldsymbol{\beta}_k$ has not been acknowledged.

To examine the uniqueness of the solution to (2.1) and (2.2), Rivest states that an equivalent problem is to find a value of k for which $k = c\sigma_k$, and proceeds to consider k/σ_k as a function of k. If k/σ_k is increasing, he writes, the system has a unique solution for all c. By way of example, for various ψ and sets of data Rivest evaluates the function k/σ_k at 60 equidistant values of k, and the points $(k, k/\sigma_k)$ thus obtained are joined to form a continuous curve. By joining the points Rivest implicitly assumes continuity, but this need not be the case. For example, if at some k the global minimum of (2.4) is attained at more than one value of $\boldsymbol{\beta}$, then $\boldsymbol{\beta}_k$ will not be uniquely defined. This means that σ_k will not be uniquely defined either, and so the function k/σ_k need not be continuous at values of k corresponding to a switch from one solution to another. Examples of this behaviour are given in the next section, and these are followed by discussion of the implications for Rivest's analytical conditions.

2.2 Examples

Example 2.1. Consider the univariate case $\mu = \mathbf{x}_i^T \boldsymbol{\beta}$, where μ is the location parameter of a Cauchy density with scale parameter σ:

$$f(y; \mu, \sigma) = \frac{\sigma}{\pi\{\sigma^2 + (y - \mu)^2\}}.$$

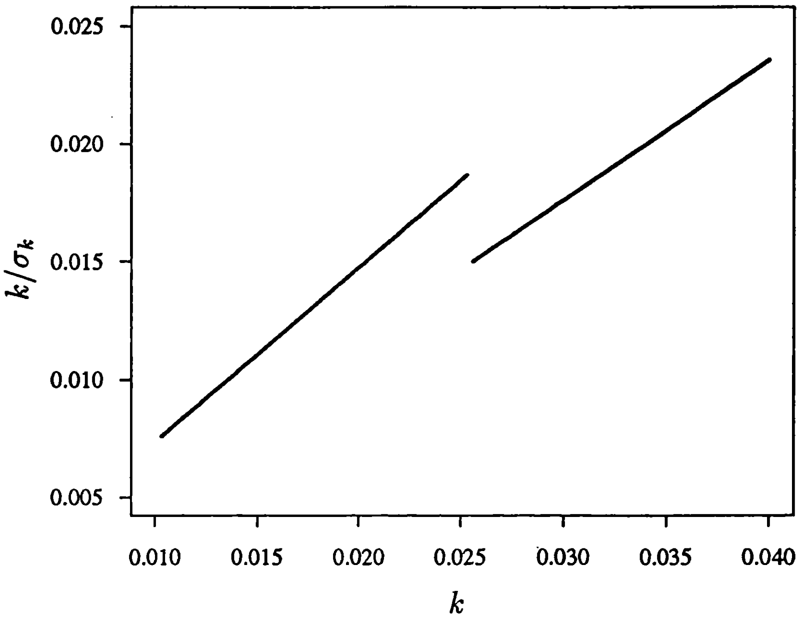

Figure 2.1: Discontinuity from use of Cauchy M-estimator on clustered data.

Take $\psi = -f'/f$ as a weakly redescending function, and without loss of generality take $c = 1$. The estimator μ_k is thus defined as the solution of

$$\sum_{i=1}^{n} \frac{y_i - \mu}{k^2 + (y_i - \mu)^2} = 0. \tag{2.6}$$

For the highly clustered data set

$$y = (-1.03, -1.02, -1.01, -1.00, -0.2, -0.1, 0.95, 1.0, 1.05, 1.10, 1.12),$$

solutions to (2.6) satisfying (2.4) were found for 100 equidistant values of k in the range (0.01,0.04), and the points $(k, k/\sigma_k)$ thus obtained are presented in Figure 2.1.

The discontinuity present at $k \approx 0.0255$ is caused by a switch in the estimate of the location parameter μ from the cluster around -1 to that around +1. This can be seen in the plots of $\sum \rho\{(y_i - \mu)/k\}$ against μ (i.e. minus the log-likelihood against μ) presented in Figure 2.2. For large k the function becomes unimodal, and non-uniqueness is no longer a problem, but for $k \approx 0.0255$ there exist two values of μ that minimise (2.4), and so σ_k is not uniquely defined. (It should be

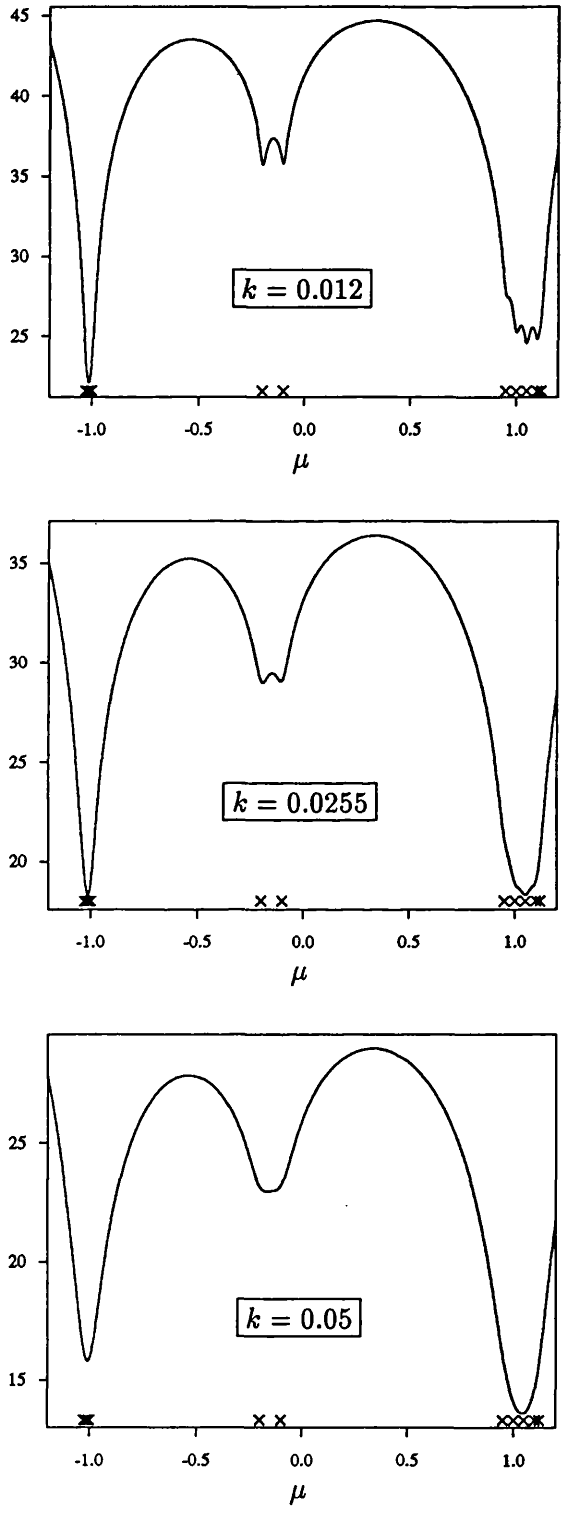

Figure 2.2: $\sum \rho\{(y_i - \mu)/k\}$ as a function of μ, with data points denoted by "×".

noted that in the narrow range of values for k considered here, the slow change in the estimates μ_k either side of the discontinuity gives rise to a near constant σ_k; hence the linearity present in Figure 2.1).

Example 2.2. Rivest gives an example based on the "Stack-Loss" data (Brownlee, 1965), comprising 21 observations on y = stack-loss and three covariates. Using the strong redescender $\psi = \psi_{\sin(a)}$, with $a = 1/\pi$, Rivest obtains a continuous graph of the function k/σ_k. However, it includes a suspiciously steep drop in the region of $k = 6.2$. When examined in detail, this is found to be a discontinuity caused by the same problem outlined in Example 2.1. For instance, at $k = 6.164$, both

$$\boldsymbol{\beta}_k^{(1)} = \begin{pmatrix} -37.395 \\ 0.808 \\ 0.548 \\ -0.070 \end{pmatrix} \quad \text{and} \quad \boldsymbol{\beta}_k^{(2)} = \begin{pmatrix} -39.098 \\ 0.890 \\ 10.537 \\ -0.094 \end{pmatrix}$$

are solutions to (2.3), with $\boldsymbol{\beta}_k^{(1)}$ minimising (2.4). This yields $\sigma_k = 1.4966$. However, at $k = 6.1645$, $\boldsymbol{\beta}_k^{(2)}$ is the solution that minimises (2.4), giving $\sigma_k = 1.6503$. There exists a $k^* \approx 6.16435$ for which the solutions give the same value of (2.4), and so σ_{k^*} is not uniquely defined. The function k/σ_k is therefore not continuous, as shown in the magnified section of Rivest's graph given in Figure 2.3.

The question arises as to which of the estimates for $\boldsymbol{\beta}_{k^*}$ one should take. An additional criterion may be to choose the one with minimal residual sum of squares. However, when other values of k are considered, this is not necessarily consistent with (2.4). For example, $\boldsymbol{\beta}_{6.164}^{(1)}$ above minimises (2.4) but $\boldsymbol{\beta}_{6.164}^{(2)}$ has a smaller residual sum of squares. In any case, Rivest's method for examining the function k/σ_k is clearly inadequate, and in the light of this result, other examples presented by him may also require correction.

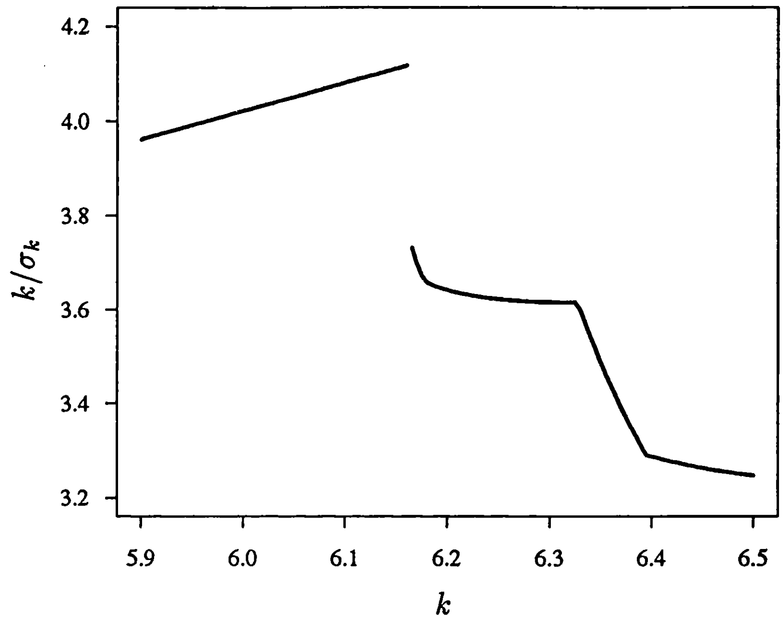

Figure 2.3: Discontinuity from use of sine M-estimator on Stack-Loss data.

2.3 Analytical Conditions

The implications of the previous discussion for Rivest's analytical conditions are now considered. Further to the generalised estimator for the scale parameter defined by (2.2), let σ_k be the solution to

$$\sum_{i=1}^{n} \chi\left(\frac{y_i - \mathbf{x}_i^T \boldsymbol{\beta}_k}{\sigma_k}\right) = (n-p)\gamma.$$

Rivest argues that the non-linear system (2.1) and (2.2) has a unique solution for a given value of c if

$$\sum_{i=1}^{n} \chi\left(c\frac{y_i - \mathbf{x}_i^T \boldsymbol{\beta}_k}{k}\right) = (n-p)\gamma \tag{2.7}$$

has a unique solution, and that therefore, in considering without loss of generality the case $c = 1$, this is equivalent to $\sum \chi_k(y_i - \mathbf{x}_i^T \boldsymbol{\beta}_k)$ being a decreasing function of k, where $\chi_k(t) = \chi(t/k)$. Following the results of Section 2.2, this argument is flawed, for values of k may exist where $\boldsymbol{\beta}_k$ minimising (2.4) is not unique. Therefore, $\sum \chi_k(y_i - \mathbf{x}_i^T \boldsymbol{\beta}_k)$ may be decreasing, but due to a discontinuity there may exist a value of c for which (2.7) has no solution, let alone a unique one. For

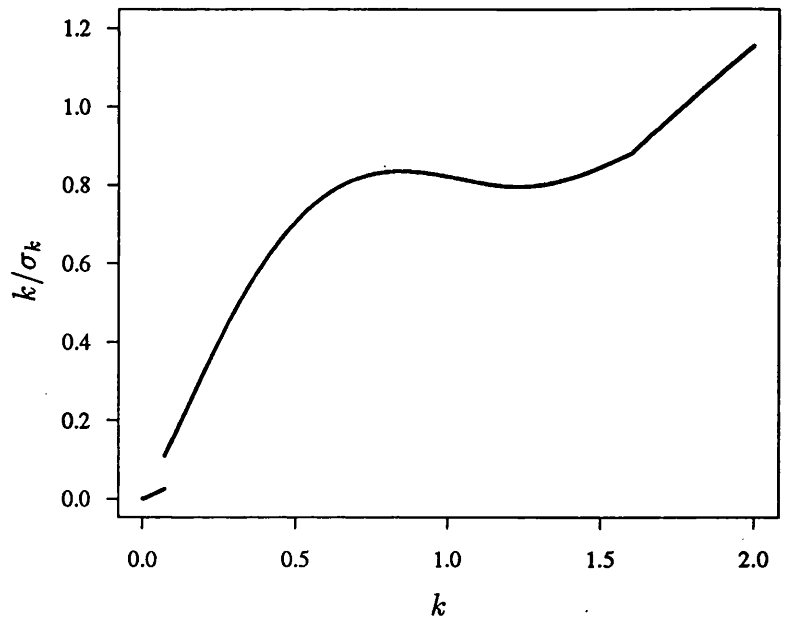

Figure 2.4: Second Cauchy M-estimator example. Note that k/σ_k increases monotonically for $k > 2$.

example, returning to the Cauchy M-estimator of Section 2.2, if the data set is changed to

$$y = (-1.1, -1.05, -1.03, -1.0, -0.97, -0.95, 0.97, 1.03, 1.05, 1.1, 1.2, 1.35, 1.5),$$

the k/σ_k function given in Figure 2.4 is obtained, where 20,000 equidistant values of k are taken in the range (0.0001, 2.0). Here, at $k \approx 0.0714$, the discontinuity jumps in the opposite direction to that shown in Figure 2.1, and since $k/\sigma_k \longrightarrow 0$ as $k \longrightarrow 0$, it seems that there exist values of c for which $k/\sigma_k = c$ has no solution. Therefore, Rivest's argument would be more accurately expressed by saying that if the function k/σ_k is increasing, then the system has *at most* a unique solution for all c.

Turning now to the analytical conditions, whilst assuming that ψ and χ are differentiable, with derivatives ψ' and χ' respectively, Rivest suggests that the function $\sum \chi_k(y_i - \mathrm{x}_i^T \beta_k)$ is decreasing for a given value of k if and only if

$$n^{-1} \sum_{i=1}^{n} \left\{ r_i - \mathrm{x}_i^T \left[\sum_{j=1}^{n} \mathrm{x}_j \mathrm{x}_j^T \psi_k'(r_j) \right]^{-1} \sum_{j=1}^{n} \mathrm{x}_j r_j \psi_k'(r_j) \right\} \chi_k'(r_i) > 0 \qquad (2.8)$$

where $r_i = y_i - \mathrm{x}_i^T \beta_k$, $\psi_k'(y) = \psi'(y/k)$ and $\chi_k'(y) = \chi'(y/k)$ for values of k for which $\sum \mathrm{x}_j \mathrm{x}_j^T \psi_k'(r_j)$ is an invertible matrix. Furthermore, in the case where the scale parameter is defined by (1.8), Rivest proposes that the system (2.1) and (1.8) has a unique solution if, for all k and for all $i = 1, \ldots, n$,

$$0.6745 > s_i \mathrm{x}_i^T \left[\sum_{j=1}^{n} \mathrm{x}_j \mathrm{x}_j^T \psi_k'(r_j) \right]^{-1} \sum_{j=1}^{n} \mathrm{x}_j r_j \psi_k'(r_j)/k, \qquad (2.9)$$

provided that $\sum \mathrm{x}_j \mathrm{x}_j^T \psi_k'(r_j)$ is invertible, where $s_i = \mathrm{sgn}(r_i)$.

Two points arise here. Firstly, it has already been shown that ψ and χ cannot be assumed to be continuously differentiable, and so the validity of (2.8) and (2.9) as conditions for uniqueness requires clarification. For example, (2.9) may be satisfied for all k where ψ_k' is defined, but the behaviour of the function k/σ_k at discontinuities and its effect on the existence of solutions to (2.1) and (1.8) remains unclear. Secondly, even if some proposed condition for the existence of ψ_k' for all k is satisfied, it is not immediately obvious how the conditions (2.8) and (2.9) may be put to profitable use, for to test (2.9) one is apparently faced with the task of finding β_k for all k — a non-trivial problem in all but the simplest cases.

2.4 Discussion

Attention has been drawn to a problem that all too easily may be overlooked. That the system of equations (2.3) may not have a unique solution has been appreciated by Rivest, but he has not considered the possibility that the solution corresponding to the global minimum of (2.4) may also be non-unique. In this he does not appear to be alone, for in comments on the non-uniqueness of strongly redescending M-estimates for the univariate case, both Huber (1981, p. 54) and Hampel et al. (1986, p. 152) state that one way of overcoming the problem is to take the global minimum. This is curious, since when discussing the linear model, Hampel et al. (1986, p. 339) note that there may be multiple global solutions to the minimization problem.

In considering solutions to (2.3), it has been demonstrated, in examples using both artificial and real data, that even the "optimal" estimate of the regression parameter $\boldsymbol{\beta}_k$, in the sense of minimising (2.4), is not necessarily unique. This suggests that, if redescending ψ-functions are to be used, then the estimator for $\boldsymbol{\beta}$ should be defined in a different way. An alternative definition is considered in the following chapter.

Chapter 3

Unique Redescending M-estimates for the linear model

3.1 Introduction

In the previous chapter we saw the complications that can arise when using redescending M-estimators for the linear model. In this chapter, a redescending approach is developed that does not suffer these drawbacks, since unique estimates will, in general, be obtained. This is achieved by application of recently developed results for the existence of unique weakly redescending M-estimates for the location-scatter model, and in particular, for location-scatter estimates corresponding to the multivariate t distribution.

A p-dimensional t distribution on $\nu > 0$ degrees of freedom gives rise to a weakly redescending ψ-function:

$$\psi_{\nu,p}(t) = \frac{(\nu + p)\, t}{\nu + t^2}. \tag{3.1}$$

This ψ is appealing in that the degrees of freedom parameter ν may be regarded as a parameter of robustness: as can be seen from (3.1), the degree of down-weighting on outliers increases with decreasing ν, and since in the limit as $\nu \to \infty$, $\psi_{\nu,p}(t) \to t$, (3.1) includes least-squares as a special case.

The t distribution has been widely considered as an alternative to the strict assumption of normality that is often made in classical statistical inference. Maronna (1976) presents an example using t M-estimates for multivariate location and scatter, and Pendergast and Broffitt (1985) mention the t distribution as a potential M-estimator for growth-curve models. Zellner (1976) considered inference under an assumed multivariate t distribution on the vector of errors in the linear regression model, and Sutradhar and Ali (1986) generalised this to multivariate regression. However, the models contained therein are of no use in the context of robust regression, since the resulting objective function is maximised at the least-squares estimate.

Lange, Little, and Taylor (1989) report a study of maximum likelihood estimation for regression models with assumed t errors, and note its equivalence to redescending M-estimation. However their approach to the uniqueness problem of redescending M-estimators is given thus: 'multiple maxima of the likelihood seem possible, particularly when ν is small; however, we did not find any for our problems'. Since Gabrielsen (1982) notes that one can show that for all ν and all linear models, there exist, with probability greater than zero, data such that the joint likelihood for the regression parameter and the scale parameter is multimodal, the acknowledgement of Lange et al. that 'widely distributed software should recognize and deal with the possibility of multiple maxima of the likelihood' is certainly a valid one.

Unique t M-estimates may be obtained if the linear model is embedded within a multivariate location-scatter framework. Some uniqueness results related to the location-scatter model are reviewed in Section 3.2, and particularly how they apply to the multivariate t distribution. Unique redescending t M-estimates for the linear model are then obtained from within the location-scatter framework by utilising a standard result for the form of a conditional location parameter of an elliptically symmetric distribution. Examples based on real data are presented in Section 3.3. For many data sets it is not reasonable to embed the covariates within a multivariate t framework. Therefore, in Section 3.4, we consider M-estimates derived from the conditional t distribution, so that, as with least-squares, estimates may be defined without regard to the joint marginal distribution of the covariates. However, it is shown that the resulting objective function is extremely nonregular, being, in general, unbounded at each of the data points. Attempts to obtain estimates from a local mode are generally not successful. Finally, the merits and limitations of the methodologies used are discussed in Section 3.5.

3.2 Multivariate t Methodology

3.2.1 Location-Scatter Model

The existence and uniqueness of redescending M-estimates for the multivariate location-scatter model have been considered by numerous authors, including Maronna (1976), Tyler (1988) and Kent and Tyler (1991). Specifically, let $\{\mathbf{z}_i : i = 1, \ldots, n\}$, be a data set in $\mathbb{R}^p$, and denote $\mathcal{P}_p$ as the set of symmetric $p \times p$ positive definite matrices. Kent and Tyler (1991) consider estimates $\hat{\mu} \in \mathbb{R}^p$ and $\hat{\Sigma} \in \mathcal{P}_p$ to maximise objective functions of the form

$$L(\mu, \Sigma) = -\tfrac{1}{2} n \log |\Sigma| - \sum_{i=1}^{n} \rho\left\{(\mathbf{z}_i - \mu)^T \Sigma^{-1} (\mathbf{z}_i - \mu)\right\}, \tag{3.2}$$

where ρ is continuous. If ρ is differentiable, then setting the derivative of (3.2) with respect to μ and Σ to 0 yields the estimating equations

$$\begin{aligned} \hat{\mu} &= \operatorname{ave}\{\omega_i \mathbf{z}_i\} / \operatorname{ave}\{\omega_i\}, & (3.3) \\ \hat{\Sigma} &= \operatorname{ave}\{\omega_i (\mathbf{z}_i - \hat{\mu})(\mathbf{z}_i - \hat{\mu})^T\}, & (3.4) \end{aligned}$$

where $\omega_i = u(s_i)$, $u(s) = 2\rho'(s)$ and $s_i = (\mathbf{z}_i - \hat{\mu})^T \hat{\Sigma}^{-1} (\mathbf{z}_i - \hat{\mu})$. Here "ave" stands for the arithmetic average over $i = 1, \ldots, n$. Kent and Tyler give a sufficient condition for the existence of a unique solution $\hat{\mu} \in \mathbb{R}^p$, $\hat{\Sigma} \in \mathcal{P}_p$ to (3.3) and (3.4) for the weakly redescending case (which in their notation implies $s^{1/2}u(s)$ is increasing near 0 and decreasing near ∞), and as an example consider multivariate t M-estimates. Such estimates correspond to the solutions of the likelihood equations for the location-scatter families of elliptically symmetric t distributions (Cornish, 1954) on $\nu > 0$ degrees of freedom. In this case the log-likelihood, up to an additive constant independent of μ and Σ, is given by (3.2) with $\rho(s)$ taken as

$$\rho_\nu(s) = \tfrac{1}{2}(\nu + p)\log\{(\nu + s)/\nu\},$$

and the likelihood equations correspond to (3.3) and (3.4) with $u(s)$ taken as

$$u_\nu(s) = (\nu + p)/(\nu + s).$$

The sufficient condition for the existence of a solution $(\hat{\mu}_\nu, \hat{\Sigma}_\nu)$ is then:

> Condition D_ν^*. For any hyperplane $H \in \mathbb{R}^p$ with $0 \leq \dim(H) \leq p - 1$,
> $P_n(H) < \{\nu + \dim(H)\}/(p + \nu)$.

Here, $P_n(\cdot)$ denotes the empirical distribution of the $\{\mathbf{z}_i\}$. This condition becomes increasingly strict as ν decreases, as the upper bound on the proportion of data points lying in lower dimensional hyperplanes also decreases.

Kent and Tyler also prove that, for $\nu \geq 1$, if a solution exists it will be unique, and they note that when sampling from continuous distributions in $\mathbb{R}^p$, Condition D_ν^* holds with probability 1 for samples of size $n \geq p + 1$. However, the effectively discrete nature of real data may negate this welcome property, and so the question of whether or not to test Condition D_ν^* must be addressed. Since the sufficient condition D_ν^* can be made a necessary condition for the existence of a solution $(\hat{\mu}_\nu, \hat{\Sigma}_\nu)$ by replacing the strict inequality with a simple inequality, in order to demonstrate that a solution does not exist we must find, for some q such that $0 \leq q \leq p - 1$, a subsample of size $n_q > n\{(\nu + q)/(p + \nu)\}$ that lies in a hyperplane of dimension q. An efficient method for verifying the existence of such a subsample is discussed in Chapter 5. However, in practice, if an estimate is obtained, it must be unique. The results on uniqueness do not apply for $0 < \nu < 1$.

3.2.2 Linear Regression Model

The linear model is embedded in the location-scatter framework as follows. Let us interpret a p-dimensional observation $(y, \mathbf{x}^T)$ from the linear model as an observation $\mathbf{z}^T = (z_1, \ldots, z_p)$ from a multivariate $t_p(\mu, \Sigma, \nu)$ distribution; i.e. for $\mathbf{z} = (z_1, \ldots, z_p)^T$,

$$f(\mathbf{z}) = c_{\nu,p}|\Sigma|^{-1/2}\left\{1 + (\mathbf{z} - \mu)^T\Sigma^{-1}(\mathbf{z} - \mu)/\nu\right\}^{-(\nu+p)/2},$$

where $c_{\nu,p}$ is a normalizing constant independent of μ and Σ (Mardia, Kent, and Bibby, 1979, p. 57). For samples of size $n \geq p + 1$, and values of $\nu \geq 1$, there is a unique estimate $(\hat{\mu}_\nu, \hat{\Sigma}_\nu)$ if Condition D_ν^* holds for the sample. Hence a unique estimate $\hat{\beta}$ of the parameter for the "regression" of Z_1 on $(Z_2, \ldots, Z_p)$ may be obtained from the component-wise location parameters of the conditional distribution of $Z_1 \mid Z_2, \ldots, Z_p$. More generally, consider the partition $\mathbf{z} = (\mathbf{z}_1^T\ \mathbf{z}_2^T)^T$, $\mu = (\mu_1^T\ \mu_2^T)^T$ and

$$\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{12}^T & \Sigma_{22} \end{bmatrix},$$

where $\mathbf{z}_i$, $\mu_i \in \mathbb{R}^{p_i}$ $(i = 1, 2)$ with $p_1 + p_2 = p$, and the submatrices Σ_{ij} are of order $p_i \times p_j$. Then, using elliptical symmetry, it can be shown (Fang, Kotz, and Ng, 1990) that the conditional location parameter, $\tilde{\mu}$, of $\mathbf{Z}_1 \mid \mathbf{Z}_2 = \mathbf{z}_2$ is given by

$$\tilde{\mu} = \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(\mathbf{z}_2 - \mu_2). \tag{3.5}$$

For the linear model (1.4), $p_1 = 1$, $p_2 = p - 1$, $\mathbf{z}_1 = y$ and $\mathbf{z}_2 = \mathbf{x}$. A unique estimate $\hat{\beta}$ of the regression of $\mathbf{Z}_1$ on $\mathbf{Z}_2$ may then be obtained from a unique estimate $(\hat{\mu}, \hat{\Sigma})$ of location and scatter. Specifically, we will have

$$\hat{\beta} = \begin{pmatrix} \hat{\mu}_1 - \hat{\Sigma}_{12}\hat{\Sigma}_{22}^{-1}\hat{\mu}_2 \\ \hat{\Sigma}_{22}^{-1}\hat{\Sigma}_{12}^T \end{pmatrix}. \tag{3.6}$$

A major advantage of this approach over the M-estimator defined by the solution to (1.6) is that we may obtain robustness against outliers in the response variable y *and* in the covariates $\mathbf{x}$, as with generalised M-estimators. This can be seen by noting that

$$u_\nu \left\{ (\mathbf{z}_i - \hat{\boldsymbol{\mu}})^T \hat{\Sigma}^{-1} (\mathbf{z}_i - \hat{\boldsymbol{\mu}}) \right\} = u_\nu \left\{ (\mathbf{x}_i - \hat{\boldsymbol{\mu}}_2)^T \hat{\Sigma}_{22}^{-1} (\mathbf{x}_i - \hat{\boldsymbol{\mu}}_2) + r_i^2 / \hat{\sigma}^2 \right\},$$

where $r_i = y_i - (1 \ \mathbf{x}_i^T)\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}^2 = \hat{\Sigma}_{11} - \hat{\Sigma}_{12} \hat{\Sigma}_{22}^{-1} \hat{\Sigma}_{12}^T$. This can be proved by employing a standard identity for a partitioned quadratic form (DeGroot, 1970, p. 54). Hence both outlying covariates and outlying responses receive less weight than non-outlying observations, since $u_\nu(s)$ is decreasing in s.

The idea of embedding the linear model in a location-scatter framework has already been proposed by Maronna and Morgenthaler (1986), who consider the implicit equations (3.3) and (3.4) with arbitrary u and no underlying objective function. They note that the influence function for the scatter estimator (Huber, 1981, pp. 223–226), and hence for the regression estimator, is bounded only if $su(s)$ is bounded, but they do not discuss how the choice of u affects the existence and uniqueness of estimates. This is an important omission, because if $su(s)$ is bounded, then $s^{1/2}u(s)$ must redescend to 0 as $s \to \infty$ (Kent and Tyler, 1991). In general, one cannot assume the existence and uniqueness of redescending location-scatter estimates, and so the function u should be chosen with care. The multivariate t choice $u = u_\nu$ is extremely favourable, as $su_\nu(s)$ is bounded and the estimator enjoys the existence and uniqueness properties already discussed.

Unfortunately, the regression models available for consideration are limited in that estimates cannot be obtained if Condition D_ν^* is not satisfied. For smaller values of ν this may prevent the inclusion of factors and interaction terms in the model, due to the large number of indicator variables required to represent the various levels. Similar problems occur if the covariates arise from a designed experiment, but then the use of a multivariate t distribution jointly for y and $\mathbf{x}^T$ would be hard to justify, as it also is in the case of factors and interactions.

When it exists, $(\hat{\mu}_\nu, \hat{\Sigma}_\nu)$ may be obtained via a guaranteed-convergent algorithm given by Kent and Tyler. The estimate $(\hat{\mu}, \hat{\Sigma})$ can then be taken as a value $(\hat{\mu}_\nu, \hat{\Sigma}_\nu)$ corresponding to a global maximum of the likelihood function (3.2) over ν. In the event that such an estimate is not unique, the 'uniqueness problem' is generally reduced to consideration of the 1-dimensional parameter ν, so it will be readily apparent if competing estimates exist. This does not apply, however, if the data suggest an estimate corresponding to $0 < \nu < 1$, where Kent and Tyler's uniqueness results do not hold. In such cases one would have to settle for the best local-maximum estimate that could be found.

There remains the problem of assessing the adequacy of the multivariate t model. This may be achieved by considering probability plots, since if $\mathbf{Z}$ is a p-dimensional $t(\mu, \Sigma, \nu)$ random variable, then $S = (\mathbf{Z} - \mu)^T \Sigma^{-1} (\mathbf{Z} - \mu)/p \sim F_{p,\nu}$. Hence 'observed' values s_i of S may be calculated from an estimate $(\hat{\mu}, \hat{\Sigma})$, and the ordered values of $P(S \leq s_i)$ may be plotted against $i/(n+1)$, for $i = 1, \ldots, n$.

3.3 Examples

Example 3.1. *Stack-Loss data.* The first example is the stack-loss data set presented by Brownlee (1965). This data set has been examined by numerous authors, including Andrews (1974), Lange et al. (1989) and, of course, Rivest (1989). Assuming that the 21 observations on y = stack-loss, x_1 = air flow, x_2 = temperature and x_3 = acid concentration may be regarded as a multivariate sample from a 4-dimensional t distribution, estimates $(\hat{\mu}_\nu, \hat{\Sigma}_\nu)$ were calculated for a broad range of ν values. A selection of estimates of the regression parameter (3.6) thus obtained is given in Table 3.1. Maximised log-likelihoods are given in the second column of the table; they describe the profile log-likelihood as a function of ν. It can be seen that the multivariate t approach actually points to the least-squares estimator, given by $\nu = \infty$. The probability plots given in Figure 3.1 also suggest that the data are moderately consistent with a random sample from a multivariate normal distribution. This conclusion is quite different from that reached by Lange et al. (1989), who modelled the data under the assumption of *univariate* homoscedastic t_ν errors, and obtained $\hat{\nu} = 1.1$. However, whether or not this represents a significant improvement in fit over $\nu = \infty$ they leave open to question. The estimate given by Lange et al. is similar to that of Andrews, and they seem to give the best fit for the majority of the observations in this data set.

Example 3.2. *Scottish Hill Races data.* Staudte and Sheather (1990, pp. 265-268) analyse a data set comprising 35 observations on a dependent variable $y =$ recorded time in minutes, and independent variables $x_1 =$ distance in miles and $x_2 =$ climb in feet. They give generalised M-estimates, defined as the solution to (1.9), with $\omega(\mathbf{x}_i) = v(\mathbf{x}_i) = (1 - h_i)/\sqrt{h_i}$ and $\psi(t_i \cdot v(\mathbf{x}_i)) = \psi_c(t_i/v(\mathbf{x}_i))$, where h_i is the ith diagonal element of the matrix $\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$ and ψ_c is the Huber ψ-function defined in Example 1.1 with $c = k\sqrt{(p+1)/n}$. A selection of t M-estimates for the data are given in Table 3.2, together with the estimate obtained by Staudte and Sheather when $k = 1$. The maximised log-likelihoods given in column 2 suggest taking $\hat{\nu} \approx 1.6$, and this is confirmed by the probability plots given in Figure 3.2. The largest residuals calculated from the resulting estimate $\hat{\boldsymbol{\beta}}$ of the regression parameter correspond to the cases 7, 18, and 33, and are respectively 52.9, 64.8 and 24.7 — consistent with the equivalent residuals from the Staudte and Sheather estimate, which are 47.8, 64.9 and 18.6.

Staudte and Sheather also consider the median of the absolute deviations from the median (MAD) and interquartile range (IQR) of the residuals as criteria for judging how well the fitted model agrees with the bulk of the data — small values indicating a good fit. Their estimate has an MAD of 3.66 and an IQR of 7.46, whereas the multivariate t M-estimate for $\nu = 1.6$ has an MAD of 3.42 and an IQR of 6.90, indicating a slight improvement in fit.

Finally, Staudte and Sheather note that the robust approach they used is perhaps too inefficient at the normal model to be recommended for general usage.

Method	Log-likelihood	Intercept	Air-Flow	Temp.	Acid
$t,\ \nu = 1$	-243.01	-35.61	0.81	0.65	-0.12
$t,\ \nu = 3$	-236.79	-38.19	0.82	0.81	-0.13
$t,\ \nu = 10$	-234.26	-39.81	0.78	1.05	-0.14
$t,\ \nu = \infty$	-233.15	-39.92	0.72	1.30	-0.15
Andrews		-37.20	0.82	0.52	-0.07
Lange		-38.50	0.85	0.49	-0.07

Table 3.1: Stack-Loss results.

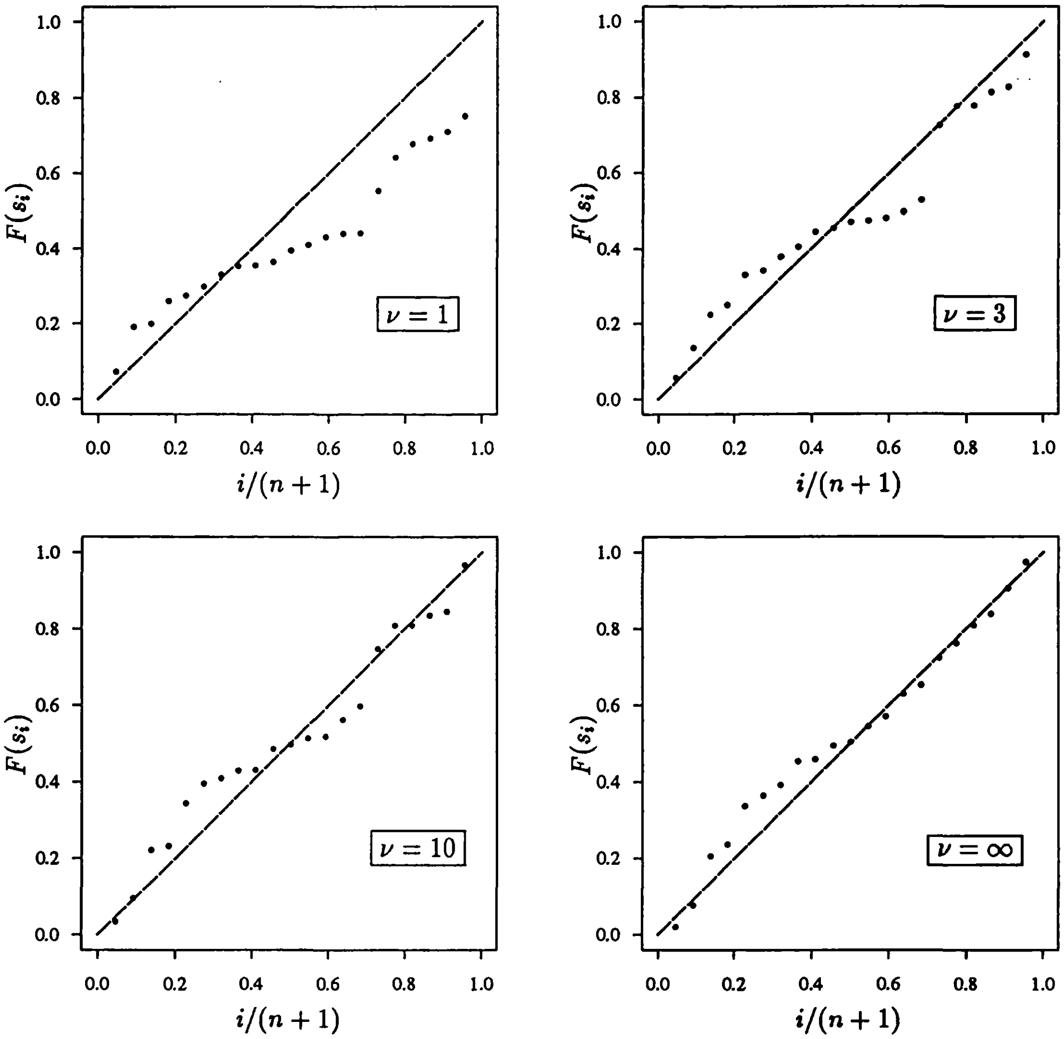

Figure 3.1: Probability plots for Stack-Loss data.

Method	Log-likelihood	Intercept	Distance	Climb
t, $\nu = 1.0$	-512.80	-7.11	6.07	0.008
t, $\nu = 1.6$	-511.48	-7.42	6.16	0.008
t, $\nu = 5.0$	-518.90	-8.83	6.42	0.008
t, $\nu = \infty$	-549.18	-8.99	6.22	0.011
Staudte		-8.92	6.61	0.008

Table 3.2: Scottish Hill Race results.

This disadvantage is not shared by the multivariate t approach, as the degrees of freedom parameter ν is estimated from the data, so that for multivariate normal data the least-squares estimate can be obtained.

Example 3.3. *Water Salinity data.* The water salinity data is a widely studied example in the robustness literature. See for example, Ruppert and Carroll (1980), Staudte and Sheather (1990, pp. 264-265) and references therein. The data consist of 28 observations on y = water salinity, x_1 = salinity lagged two weeks (sallag), x_2 = trend, which is one of the six biweekly periods in March-May, and x_3 = H_2O Flow — the river discharge. Various t M-estimates for the data are given in Table 3.3, along with a trimmed least-squares estimate obtained by Ruppert and Carroll and the Staudte and Sheather estimate with $k = 2$. The maximised log-likelihoods suggest taking $\hat{\nu} \approx 5$, and the outlying cases are then 15, 16 and 17 — with residuals of -2.38, 5.68 and -2.20, broadly agreeing with the results of Ruppert and Carroll and Staudte and Sheather. Note that the t M-estimate of $\boldsymbol{\beta}$ is very similar to the estimate given by Staudte and Sheather, and both are quite different from least squares.

In both of the studies cited above the MAD and IQR of the residuals are considered as means of comparing the fit of competing estimates. The MAD and IQR of the t M-estimate given by $\nu = 5$ are respectively 0.45 and 1.06, figures which are almost the same as those given by Staudte and Sheather and

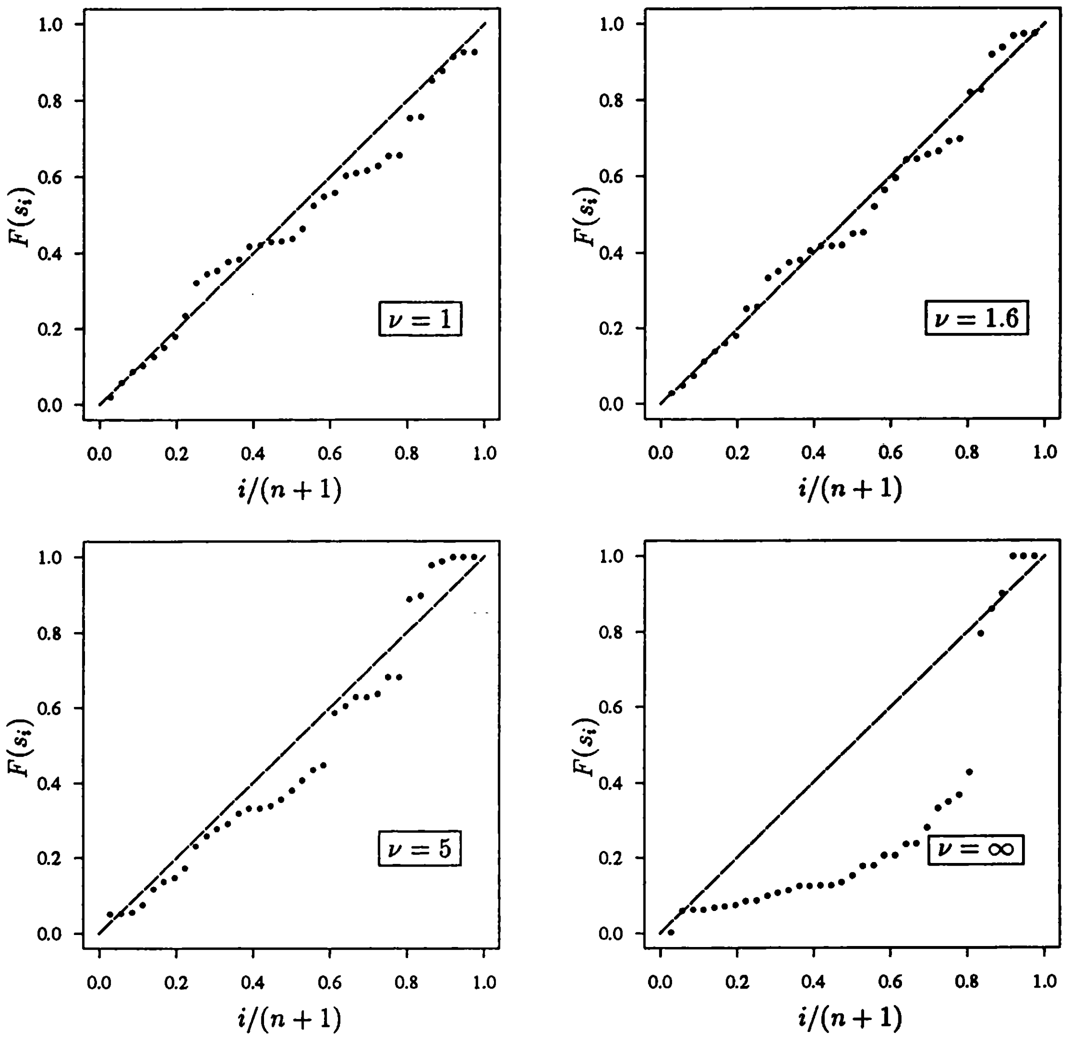

Figure 3.2: Probability plots for Scottish Hill Race data.

Method	Log-likelihood	Intercept	Sallag	Trend	Flow
$t,\ \nu = 1$	-239.40	20.15	0.707	-0.173	-0.704
$t,\ \nu = 3$	-233.06	18.87	0.715	-0.166	-0.653
$t,\ \nu = 5$	-232.50	17.56	0.723	-0.148	-0.601
$t,\ \nu = 7$	-232.68	16.46	0.729	-0.131	-0.558
$t,\ \nu = \infty$	-235.75	9.59	0.777	-0.026	-0.295
Ruppert		14.49	0.774	-0.160	-0.488
Staudte		16.89	0.715	-0.142	-0.570

Table 3.3: Water Salinity results.

Ruppert and Carroll for their estimates. Therefore in terms of these criteria the t M-estimate compares very favourably with those obtained from other robust methods.

Example 3.4. *Twickenham Run-Times.* This is a much larger data set, consisting of the finishing times of 556 runners (472 male, 84 female) who completed an 8 mile race in Twickenham, England, in 1983. For each individual there are measurements on the following:

$$\begin{aligned}
y &= \log(\text{finishing time}),\ x_1 = \log(\text{miles run per week} + 1),\ x_2 = \log(\text{age}),\\
x_3 &= \log(\text{weight}), \qquad\quad x_4 = \log(\text{height}),\\
a &= \text{number of other active sports per week (0, 1, or 2), and}\\
s &= \text{sex } (0 = \text{male},\ 1 = \text{female}).
\end{aligned}$$

A "body mass index" variable $x_5 = \log(\text{weight}/\text{height}^2)$ is also defined. For simplicity, the runners who gave only partial information have been omitted.

We seek a simple linear model relating finishing time to the various explanatory variables, and a conventional multiple regression model selection procedure suggested the following:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_2^2 + \beta_4 x_5 + \beta_5 s + e.$$

The data cannot be regarded as a random sample from a multivariate t distribution, particularly as both x_2 and x_2^2 are included in the model, but nevertheless, multivariate t M-estimates of the regression parameter are given in Table 3.4. For comparison, we also include the estimate obtained by using the Staudte and Sheather method with $k = 1$. The results suggest that the log-likelihood is maximised with $\hat{\nu} \approx 4$. The difference in log-likelihood between the best fitting t and the normal model yields a likelihood ratio chi-squared statistic of 356.3 on 1 df, an apparently significant improvement in fit. However, there is a negligible difference in terms of the MAD and IQR: for the t, $\nu = 4$ model we obtain an MAD of 0.066 and an IQR of 0.132, whereas for the multivariate normal estimate the figures are

Method	Log-likelihood	Int.	log(mr)	log(age)	log(age)^2	log(bmi)	sex
t, $\nu = 1$	646.90	6.942	-0.101	-1.335	0.204	0.437	0.219
t, $\nu = 2$	710.84	6.997	-0.101	-1.375	0.210	0.428	0.217
t, $\nu = 3$	723.77	7.032	-0.100	-1.401	0.214	0.421	0.215
t, $\nu = 4$	723.93	7.056	-0.099	-1.419	0.216	0.416	0.215
t, $\nu = 5$	719.75	7.074	-0.099	-1.433	0.218	0.412	0.215
t, $\nu = 10$	689.36	7.127	-0.096	-1.474	0.224	0.403	0.214
t, $\nu = 20$	647.24	7.170	-0.094	-1.507	0.229	0.397	0.215
t, $\nu = \infty$	545.78	7.283	-0.085	-1.590	0.241	0.393	0.219
Staudte		7.569	-0.097	-1.722	0.260	0.410	0.214

Table 3.4: Twickenham Run-Time results.

0.067 and 0.133. The Staudte and Sheather estimate yields the same values for the MAD and IQR as the t, $\nu = 4$ estimate, and so it appears that the least-squares estimate is adequate for this data set.

To examine whether or not the indicator variable for the sex factor has had an undue influence on the estimate $\hat{\nu}$, separate estimates for males and females are also presented in Tables 3.5 and 3.6. For males only, $\hat{\nu} \approx 5$, and for females, $\hat{\nu} \approx 4$. It seems therefore that the estimate of ν obtained for the combined data has not been largely influenced by the inclusion of an indicator variable. The MAD and IQR for the t, $\nu = 5$ male estimate are 0.065 and 0.130; for the t, $\nu = 4$ female estimate they are 0.075 and 0.153; the corresponding figures from the Staudte and Sheather estimates are 0.063, 0.127 and 0.074, 0.154.

Probability plots are given in Figure 3.3. The plot for $\nu = 4$ is quite good, save for the noticeable discrepancy at $i/(n+1) \approx 0.8$, which may be due to the univariate probability plot obscuring the fact that the data are not multivariate t. The $\nu = 4$ plot is not, however, as good as the univariate normal probability plot for the least-squares estimate (not shown), which indicates no departure from normality whatsoever.

Method	Log-likelihood	Int.	log(mr)	log(age)	log(age)^2	log(bmi)
t, $\nu = 1$	637.86	6.925	-0.101	-1.326	0.203	0.436
t, $\nu = 3$	764.90	6.992	-0.100	-1.381	0.211	0.418
t, $\nu = 4$	774.99	7.016	-0.099	-1.400	0.214	0.412
t, $\nu = 5$	777.38	7.034	-0.098	-1.415	0.216	0.407
t, $\nu = 6$	776.34	7.048	-0.098	-1.423	0.218	0.404
t, $\nu = \infty$	659.13	7.083	-0.083	-1.502	0.229	0.359
Staudte		7.527	-0.096	-1.704	0.257	0.405

Table 3.5: Twickenham Run-Time results — Males only.

Method	Log-likelihood	Int.	log(mr)	log(age)	log(age)^2	log(bmi)
t, $\nu = 1$	135.47	5.644	-0.098	-0.534	0.088	0.323
t, $\nu = 3$	155.06	6.345	-0.103	-0.904	0.140	0.360
t, $\nu = 4$	155.55	6.543	-0.104	-1.004	0.155	0.376
t, $\nu = 5$	154.82	6.683	-0.105	-1.072	0.164	0.391
t, $\nu = 6$	153.64	6.784	-0.106	-1.119	0.171	0.403
t, $\nu = \infty$	120.34	7.698	-0.107	-1.540	0.231	0.555
Staudte		7.020	-0.109	-1.233	0.187	0.427

Table 3.6: Twickenham Run-Time results — Females only.

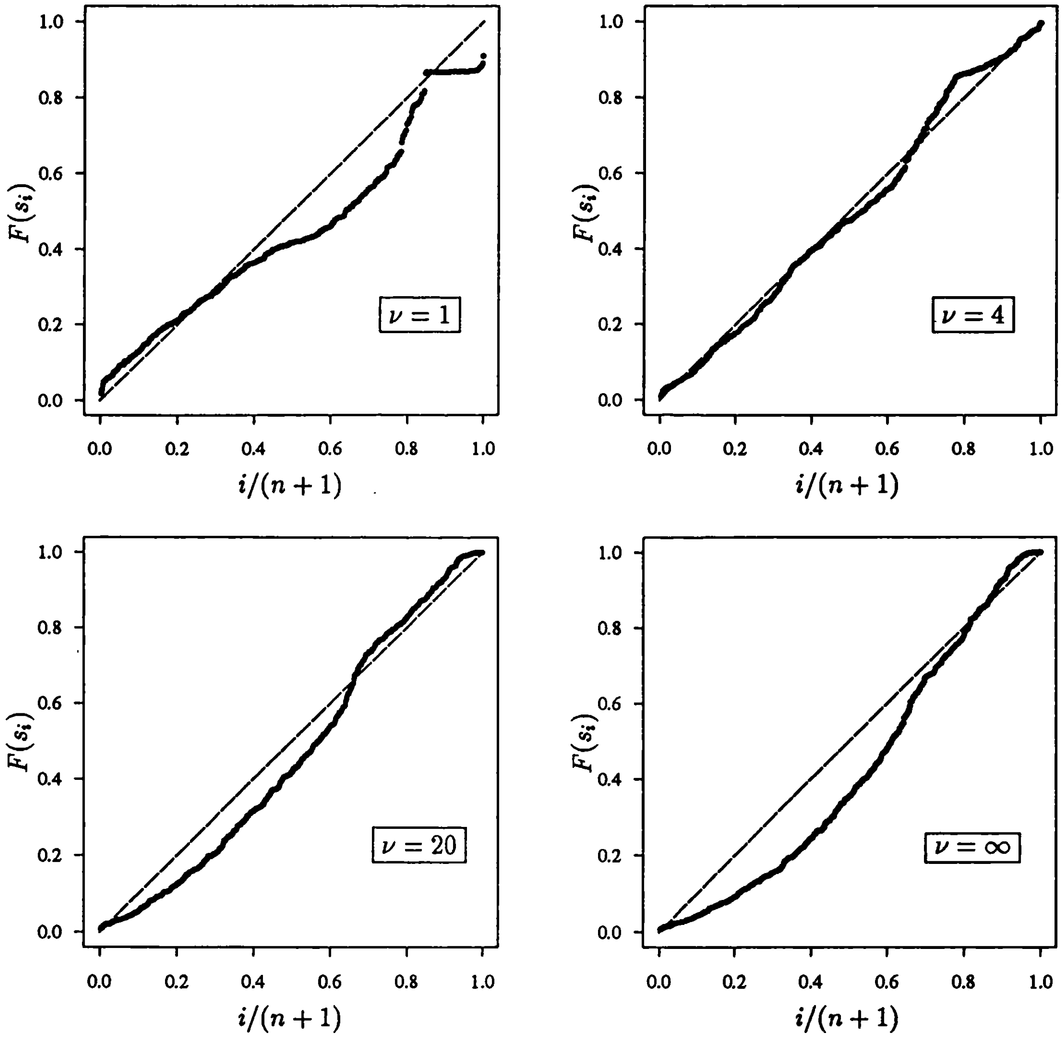

Figure 3.3: Probability plots for Twickenham Run-Time data.

3.4 Conditional t M-estimates

3.4.1 Main Result

For many data sets, such as those arising from designed experiments and those involving curvilinear terms such as Example 3.4 discussed earlier, it is not reasonable to impose a multivariate t framework on the covariates. In such cases it would be preferable to estimate directly the conditional distribution of a response y given a set $\mathbf{x}$ of explanatory variables, i.e., without regard to the joint marginal distribution of $\mathbf{x}$. This is effectively what occurs in conventional least-squares regression, which can be viewed as a method for estimating a conditional normal distribution.

Unfortunately, this approach breaks down when we consider the conditional t distribution, for as we prove in this section, it yields a highly nonregular likelihood function, being, in general, singular at each of the data points. The likelihood is thus intrinsically multimodal, with modes of infinity, so that unrestricted maximum likelihood estimation breaks down. We proceed as follows.

The *conditional t M*-estimate of $\boldsymbol{\beta}$ is the estimate defined by (3.6), but with its component terms obtained by maximising the log-likelihood function corresponding to the conditional distribution of $\mathbf{Z}_1|\mathbf{Z}_2 = \mathbf{z}_2$, rather than the full joint distribution of $\mathbf{Z}_1$ and $\mathbf{Z}_2$. DeGroot (1970) notes that the conditional distribution is a p_1-dimensional t distribution on $\nu_{1\cdot 2} = \nu + p_2$ degrees of freedom, with location parameter $\tilde{\boldsymbol{\mu}}$ given by (3.5) and scatter matrix

$$\tilde{\Sigma} = g_\nu(\mathbf{z}_2, \boldsymbol{\mu}_2, \Sigma_{22})\,\Sigma_{1\cdot 2},$$

where

$$g_\nu(\mathbf{z}_2, \boldsymbol{\mu}_2, \Sigma_{22}) = \nu_{1\cdot 2}^{-1}\left\{\nu + (\mathbf{z}_2 - \boldsymbol{\mu}_2)^T \Sigma_{22}^{-1}(\mathbf{z}_2 - \boldsymbol{\mu}_2)\right\}$$

and $\Sigma_{1\cdot 2} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{12}^T$. The conditional log-likelihood, up to an additive constant involving only $\nu_{1\cdot2}$ and p_1, is given by

$$L_{\mathbf{Z}_1|\mathbf{Z}_2}(\mu, \Sigma, \nu) = -\tfrac{1}{2}\sum_{i=1}^{n} \log|\tilde{\Sigma}_i| - \tfrac{1}{2}\sum_{i=1}^{n} \rho_{\nu_{1\cdot2}}\left\{(\mathbf{z}_{1i} - \tilde{\mu}_i)^T \tilde{\Sigma}_i^{-1}(\mathbf{z}_{1i} - \tilde{\mu}_i)\right\}, \quad (3.7)$$

where $\tilde{\mu}_i$ and $\tilde{\Sigma}_i$ are respectively $\tilde{\mu}$ and $\tilde{\Sigma}$ evaluated at $\mathbf{z}_2 = \mathbf{z}_{2i}$. For the linear model, we have $\mathbf{z}_{2i} = \mathbf{x}_i$, $\tilde{\mu}_i = (1 \;\; \mathbf{x}_i^T)\beta$ and we may denote $\Sigma_{1\cdot2}$ and $\tilde{\Sigma}_i$ as, respectively, σ^2 and σ_i^2, where $\sigma_i^2 = g_\nu(\mathbf{x}_i, \mu_2, \Sigma_{22})\,\sigma^2$.

Ideally, maximisation of the full log-likelihood $L_{\mathbf{Z}}(\mu, \Sigma, \nu)$ as carried out in Section 3.2 would be equivalent to maximisation of the conditional log-likelihood $L_{\mathbf{Z}_1|\mathbf{Z}_2}(\mu, \Sigma, \nu)$, so that the results on existence and uniqueness may still apply. This is true for the multivariate normal ($\nu = \infty$) case, as can be shown by writing $L_{\mathbf{Z}}(\mu, \Sigma, \nu)$ as

$$L_{\mathbf{Z}}(\mu, \Sigma, \nu) = L_{\mathbf{Z}_1|\mathbf{Z}_2}(\mu, \Sigma, \nu) + L_{\mathbf{Z}_2}(\mu_2, \Sigma_{22}, \nu),$$

where $L_{\mathbf{Z}_2}(\mu_2, \Sigma_{22}, \nu)$ is the log-likelihood corresponding to the marginal distribution of $\mathbf{Z}_2$. In the normal case, it is easy to show that the marginal components (corresponding to the marginal distribution of $\mathbf{Z}_2$) of the maximum likelihood estimate $(\hat{\mu}, \hat{\Sigma})$ for the full log-likelihood are the maximum likelihood estimates $(\hat{\mu}_2, \hat{\Sigma}_{22})$ for the marginal log-likelihood. It follows that the derivative of $L_{\mathbf{Z}_1|\mathbf{Z}_2}(\mu, \Sigma)$ with respect to μ and Σ must equal 0 at the value $(\mu, \Sigma) = (\hat{\mu}, \hat{\Sigma})$, and so for infinite ν, maximising the full likelihood is equivalent to maximising the conditional likelihood. For finite ν, this is not the case, since the conditional likelihood contains information about μ_2 and Σ_{22}. This difficulty is, however, minor in comparison to the more fundamental existence problem given in the following theorem. Before presenting the theorem, however, it is convenient to make a minor change to the notation, by denoting the number of occurrences of an observation $\mathbf{z}_{2i}$ in a given data set as r_i.

Theorem 3.1. *For a data set $\{\mathbf{z}_i = (\mathbf{z}_{1i}^T \; \mathbf{z}_{2i}^T)^T \in \mathbb{R}^p \colon i = 1, \dots, n\}$, the conditional likelihood* (3.7) *is unbounded at the points $\mu = \mathbf{z}_i$ whenever $r_i = 1$.*

Proof. Assume that the scatter matrix $\mathbf{\Sigma}$ of the joint distribution of $\mathbf{Z}_1$ and $\mathbf{Z}_2$ is of the form

$$\mathbf{\Sigma} = \frac{1}{\theta}\begin{bmatrix}\mathbf{\Gamma}_{11} & \mathbf{\Gamma}_{12}\\ \mathbf{\Gamma}_{12}^T & \mathbf{\Gamma}_{22}\end{bmatrix},$$

where $\theta > 0$ is a 1-dimensional dummy parameter, and the $\mathbf{\Gamma}_{ij}$ are any constant matrices such that $\mathbf{\Sigma} \in \mathcal{P}_p$. For some observation $\mathbf{z}_k = (\mathbf{z}_{1k}^T\ \mathbf{z}_{2k}^T)^T$ for which $r_k = 1$, put $\boldsymbol{\mu}_1 = \mathbf{z}_{1k}$ and $\boldsymbol{\mu}_2 = \mathbf{z}_{2k}$. Then $\tilde{\boldsymbol{\mu}}_k = \mathbf{z}_{1k}$, $\tilde{\mathbf{\Sigma}}_k = \nu(\mathbf{\Gamma}_{11} - \mathbf{\Gamma}_{12}\mathbf{\Gamma}_{22}^{-1}\mathbf{\Gamma}_{12}^T)/\theta\nu_{1\cdot 2}$ and the contribution to the conditional log-likelihood from the observation $\mathbf{z}_k$ reduces to $-\frac{1}{2}\log|\tilde{\mathbf{\Sigma}}_k|$, which $\to \infty$ as $\theta \to \infty$.

To complete the proof it is now shown that as $\theta \to \infty$, the contributions to the conditional log-likelihood from all observations $\mathbf{z}_j$ $(j \neq k)$ remain finite, a sufficient condition for which is that, in the limit, $\tilde{\mathbf{\Sigma}}_j$ be strictly positive definite. Thus

$$\begin{aligned}\tilde{\mathbf{\Sigma}}_j &= \frac{1}{\theta\nu_{1\cdot 2}}\left\{\nu + \theta(\mathbf{z}_{2j} - \mathbf{z}_{2k})^T\mathbf{\Gamma}_{22}^{-1}(\mathbf{z}_{2j} - \mathbf{z}_{2k})\right\}\left(\mathbf{\Gamma}_{11} - \mathbf{\Gamma}_{12}\mathbf{\Gamma}_{22}^{-1}\mathbf{\Gamma}_{12}^T\right)\\ &\to \frac{1}{\nu_{1\cdot 2}}(\mathbf{z}_{2j} - \mathbf{z}_{2k})^T\mathbf{\Gamma}_{22}^{-1}(\mathbf{z}_{2j} - \mathbf{z}_{2k})\left(\mathbf{\Gamma}_{11} - \mathbf{\Gamma}_{12}\mathbf{\Gamma}_{22}^{-1}\mathbf{\Gamma}_{12}^T\right) \quad \text{as} \quad \theta \to \infty.\end{aligned}$$

So as $\theta \to \infty$, $\tilde{\mathbf{\Sigma}}_j$ approaches a constant positive definite matrix. □

Thus the conditional likelihood for the multivariate t distribution is an example of a nonregular maximum likelihood estimation problem. In the event that $r_i > 1$ for all i, the conditional likelihood will not diverge to infinity under the conditions described in the proof; however, the existence of a unique maximum in this case seems unlikely.

3.4.2 Additional Remarks

The singularity problem seems to be very similar to the well-known difficulties associated with estimating finite normal-mixtures, where the resulting likelihood function can also be made singular at each of the data points. However, in a review of the normal-mixture literature, Titterington, Smith, and Makov (1985) note that despite its singularity problems, maximum likelihood is still viable as consistent estimates may be obtained from a well-defined local maximum of the likelihood function. Since the conditional t distribution can be regarded as an infinite normal-mixture, and as the singularity properties of its likelihood function are of a similar nature to those of the finite normal-mixture, it seems natural to suggest that conditional t M-estimates might still be obtained if (3.7) has a local maximum. Finding such a maximum is, however, another matter, as the singularities in the likelihood surface seem to present a substantial problem in the multivariate setting, although estimates have been found for the data set considered in the following example.

Example 3.5. Carroll and Ruppert (1988) and Mak (1992) consider a set of 85 paired-measurements from two hormone assay methods, y = 'test method' and x = 'reference method', and one aspect of the study is to see how the test measurement y is related to the reference measurement x. A linear model $y = \beta_0 + \beta_1 x + e$ is thought reasonable, but as can be seen in Figure 3.4 the data are obviously heteroscedastic. The conditional t model, with its 'built-in' heteroscedastic function $g_\nu(\cdot)$ seems highly appropriate for this data set, and the regression estimates given by homoscedastic normal (least-squares), full bivariate t and univariate conditional t models are shown in Table 3.7.

All three methods yield similar estimates of the regression parameter and give comparable values of the MAD and IQR. However, the estimated standard errors,

Method	$\hat\beta_0$	(s.e.)	$\hat\beta_1$	(s.e.)	MAD	IQR
Normal	0.0849	(0.5057)	0.9520	(0.0314)	0.894	1.836
Full t, $\hat\nu \approx 1.15$	-0.2532	(0.2012)	0.9551	(0.0390)	0.879	1.837
Conditional t, $\hat\nu \approx 3.48$	-0.2669	(0.1457)	0.9583	(0.0393)	0.887	1.837

Table 3.7: Hormone assay results.

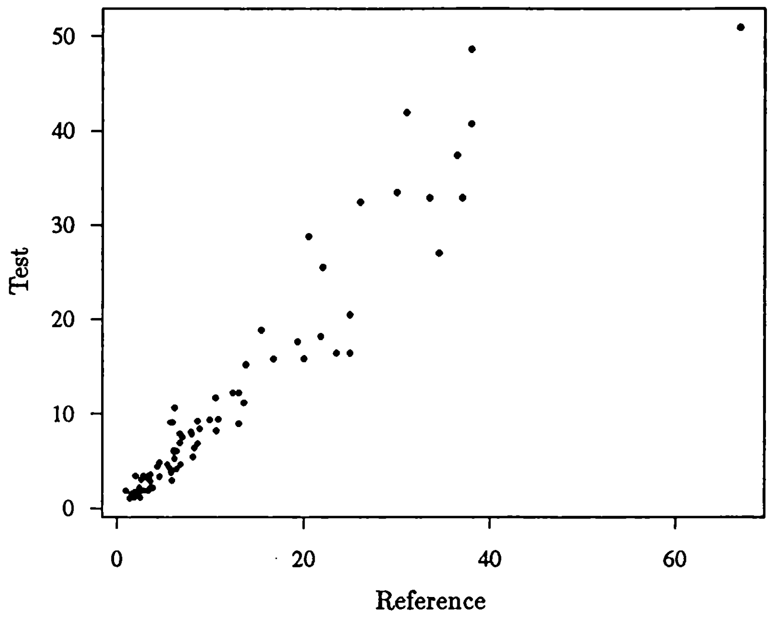

Figure 3.4: Scatter-plot of hormone assay data.

calculated from the observed information matrices, are more interesting. The t standard errors for $\hat{\beta}_0$ are much smaller than that for the normal model, whilst the estimated standard errors of $\hat{\beta}_1$ are slightly larger. This is not surprising, since the variability of the test measurement increases with the size of the reference measurement, and so the heteroscedastic model can estimate the intercept more precisely than can the homoscedastic normal model. Probability plots for each of the three models are given in Figure 3.5. It can be seen that the conditional t model fits the data very well and, as expected, the homoscedastic normal is extremely poor. It is interesting to note that the conditional t estimates and standard errors are very similar to the iterative weighted-least-squares estimates for these data, using a quadratic mean-variance function (Mak, 1992, Table 1). However, Mak excluded the last pair of observations from his study, and does not discuss how his estimates are affected by its inclusion.

The conditional t M-estimates for the previous example were obtained via an EM algorithm (see Appendix A.2). However, we were unable to compute estimates for the data sets considered in Section 3.3, where the number of covariates is greater

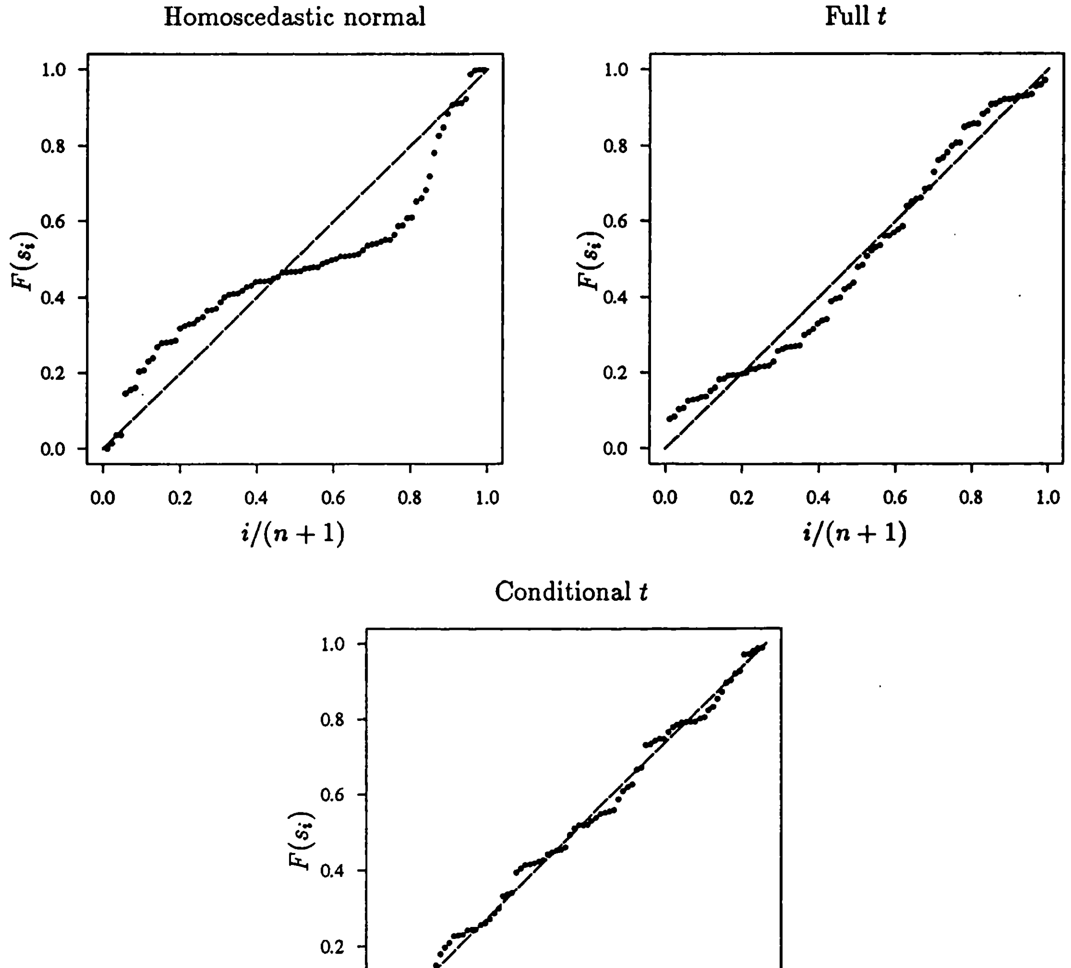

Figure 3.5: Probability plots for hormone assay data. For homoscedastic normal, $s_i = (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)/\hat{\sigma}$ and $F \sim N(0,1)$; for full t, $s_i = (\mathbf{z}_i - \hat{\boldsymbol{\mu}})^T \hat{\Sigma}^{-1} (\mathbf{z}_i - \hat{\boldsymbol{\mu}})/2$ and $F \sim F_{2,1.15}$; for conditional t, $s_i = (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)/\hat{\sigma}_i$ and $F \sim t_{3.48}$.

than one. This may be because a sensible local maximum of the conditional likelihood function does not exist for those data sets, or because the starting point used for the algorithm was not sufficiently close to a well-defined solution in order to prevent the estimates being captured by one of the singularities.

Further evidence for the computational fragility of the conditional t model has been obtained from a simulation study. Conditional t data was obtained by simulating samples of covariates from a bivariate t_4 distribution with location $\mu_2 = (1 \ \ 2)^T$ and scatter matrix $\Sigma_{22} = (1 \ \ 0.59, \ \ 0.59 \ \ 4)$, and responses were then simulated from a univariate t_6 distribution with location $\tilde{\mu}_i = 1 + x_{1i} + x_{2i}$ and dispersion $\sigma_i^2 = 2\,g_4(\mathrm{x}_i, \mu_2, \Sigma_{22})$. For each sample local maximum likelihood estimates were sought using the EM algorithm described in Appendix A.2, with the true parameter values used as starting points. Samples yielding well-defined estimates were found, but n had to be quite large for convergence to occur reliably. For example, at $n = 100$ only 4 samples from 10 generated yielded local maximum likelihood estimates, but at $n = 400$ all but one of 50 samples generated led to satisfactory convergence. At $n = 2000$, estimates were obtained for all 50 samples generated. When convergence occurred, estimated standard errors were calculated from the observed information matrix, which in turn was obtained by numerically differentiating analytical derivatives of the conditional t likelihood (see Appendix A.3). Summary statistics of the results obtained from 49 samples of 400 observations and from 50 samples of 2000 observations are given in Table 3.8. It can be seen that the sample estimates of β and σ^2 are much less variable than those of μ_2 and Σ_{22}. This is not too surprising, given that we are effectively trying to estimate the location and scatter matrices of the covariates without making any distributional assumptions. Of greater interest is the indication that the estimates may share the properties of maximum likelihood estimates obtained from regular likelihoods. For example, the observed standard deviations are consistent with the averages of the standard errors calculated by using regular asymptotic theory; as n increases from 400 to 2000 there is an approximate $1/\sqrt{n}$ reduction in the standard deviations of the β and σ^2 estimates; and at $n = 2000$, probability plots (not shown) indicate that the sample estimates for all parameters are approximately normally distributed. On the other hand, there is a greater-than-expected reduction in the standard deviations for the μ_2 and Σ_{22} estimates, possibly due to the presence of several outliers in the estimates observed at $n = 400$. It is not clear if these outliers are a result of a small sample problem (at $n = 400$!) or a failure to find the best local maximum. However, the magnitude of the average standard errors does suggest that the data may be consistent with a simpler model, even though we know a more complex model is true. It would appear, therefore, that even at $n = 400$, the likelihood contains little information with respect to the μ_2 and Σ_{22} parameters.

To summarize, when there is more than one covariate, large samples of conditional t data are required in order to obtain local maximum likelihood estimates reliably. We therefore cannot expect to obtain such estimates from much smaller samples of real data, such as those considered earlier in this chapter. Regrettably, we seem obliged to reject the conditional t likelihood as a mechanism for obtaining robust regression estimates. However, modifications to the standard likelihood approach can be employed for nonregular cases, and these are discussed in the next section.

n		$\hat{\beta}$			$\hat{\sigma}^2$	$\hat{\mu}_2$		$\hat{\Sigma}_{22}$		
	Mean	1.009	0.993	0.996	1.941	1.006	1.897	1.800	0.733	5.606
400	S.D.	0.130	0.080	0.036	0.309	0.515	0.810	1.852	2.037	4.483
	A.S.E.	0.127	0.082	0.041	0.335	0.648	0.986	2.141	2.621	5.559
	Mean	0.999	0.993	1.001	1.975	1.015	2.073	1.057	0.608	4.167
2000	S.D.	0.063	0.042	0.018	0.145	0.131	0.288	0.246	0.317	0.973
	A.S.E.	0.057	0.037	0.018	0.139	0.141	0.276	0.233	0.346	0.914

Table 3.8: Summary statistics for estimates obtained from simulated conditional t data. A.S.E. denotes 'Average Standard Error'.

3.4.3 Alternative Approaches

As mentioned in the previous section, the unbounded likelihood problem is not restricted to the conditional t model. As well as the finite normal-mixture model, it is often studied in connection with continuous univariate distributions with an unknown threshold parameter, such as the three-parameter Weibull distribution. These studies have led to the proposal of modifications to the likelihood function so that singularities may be avoided. We now consider two that might be applied to the conditional t likelihood.

Alternative 1

Cheng and Amin (1983) introduced maximum product of spacings (MPS) estimation for data arising from continuous univariate distributions. Consider an ordered sample $x_1 < x_2 < \cdots < x_n$ drawn from a distribution with density $f(x, \theta_0)$, assumed strictly positive on some interval (α_1, α_2) and to be zero outside this interval. The MPS method is then simply to choose θ to maximise

$$H = (n+1)^{-1} \sum_{i=1}^{n+1} \log D_i,$$

where D_i is a spacing defined by

$$D_i = \int_{x_{i-1}}^{x_i} f(x, \theta)\, dx \quad (i = 1, 2, \ldots, n+1),$$

such that $x_0 \equiv \alpha_1$ and $x_{n+1} \equiv \alpha_2$. This approach avoids the singularity problem, as H is bounded above, due to the constraint that $\sum D_i = 1$. This implies that the maximum value of H is attained when the D_i's are equal.

In the regression context, the spacings would seem to have to be defined with respect to the residuals. For example, suppose we have the ordinary regression model (1.4). Then an ordered, univariate sample may be obtained by calculating, for given values of $\boldsymbol{\beta}$ and σ, scaled residuals $r_i = (y_i - \mathbf{x}_i^T \boldsymbol{\beta})/\sigma$, to yield an ordered sample $r_{(1)} < r_{(2)} < \cdots < r_{(n)}$. Values of $\boldsymbol{\beta}$ and σ would then be chosen to maximise

$$H^* = (n+1)^{-1} \sum_{i=1}^{n+1} \log \int_{r_{(i-1)}}^{r_{(i)}} f(u)\, du,$$

where $r_{(0)} = -\infty$ and $r_{(n+1)} = +\infty$. However, one need not consider how to compute such estimates, since the MPS approach actually breaks down when applied to regression models in this way. This is because H^* attains its upper bound when

$$\int_{r_{(i-1)}}^{r_{(i)}} f(u)\, du = \frac{1}{n+1} \quad \text{for all } i, \tag{3.8}$$

that is to say when $r_{(i)} = F^{-1}\{i/(n+1)\}$ for all i, where F denotes the distribution function of f. It becomes apparent that there exist data sets for which totally inappropriate values of $\boldsymbol{\beta}$ yield the upper bound of H^*. For example, suppose, in the simple case $y_i = \beta_0 + \beta_1 x_i$, we have data $x_i = F^{-1}\{i/(n+1)\}$ and we observe $y_i = x_i$ for all i. Then $\sigma = 1$, $\beta_0 = 0$, $\beta_1 = 1$ gives a perfect fit to the data, but on taking $\beta_1 = 0$ we obtain residuals that satisfy (3.8). In this case, a plot of the residuals against fitted values would indicate an extremely poor fit, whilst a probability plot would indicate a perfect fit! This simple example highlights the point that probability plots should never be used in isolation to measure the adequacy of a regression model. The MPS approach does not therefore appear to extend to the regression context, though there remains the possibility that some other implementation of the MPS approach could yield useful results.

Alternative 2

Titterington (1985) remarked that MPS estimation may be regarded as a maximum likelihood approach based on grouped data, that is to say the likelihood obtained from a sample $y_1, \ldots, y_{n+1}$ of which all that is known is that $x_{i-1} \leq y_i < x_i$. Titterington points out that usually the grouping would be achieved by allocating the x_i's to histogram bins (i.e. grouped) and that, for example, this may be used to eliminate the likelihood singularities that arise from a mixture of two univariate normals. A similar procedure might, therefore, be applied to linear model data in order to remove singularities from the conditional t likelihood.

Grouping is introduced to the linear model if a response observation y_i is not known exactly but is known only to lie between known constants a_i and b_i (Burridge, 1981). In practice, of course, when dealing with 'continuous' data, it is *always* the case that y_i is not known exactly since, as Heitjan (1989) notes, 'in a fundamental sense, all continuous variables are eventually rounded or coarsened, i.e., grouped'.

In the conditional t framework, the $\{y_i\}$ may be regarded as observations from a heteroscedastic location-scale density of the form $\sigma_i^{-1} f_\nu\{(y_i - \mu_i)/\sigma_i\}$, where f_ν is the standard t_ν density. If it is known only that, for $i = 1, \ldots, n$, $a_i < y_i < b_i$, then the conditional log-likelihood (3.7) is replaced by a grouped data likelihood of the form

$$L_G = \sum_{i=1}^{n} \log F_\nu(u_i, v_i),$$

where $u_i = (a_i - \mu_i)/\sigma_i$, $v_i = (b_i - \mu_i)/\sigma_i$ and

$$F_\nu(u, v) = \int_u^v f_\nu(\epsilon)\, d\epsilon.$$

Maximum grouped-likelihood estimates would be obtained by maximising L_G with respect to $\boldsymbol{\beta}$, σ^2, $\boldsymbol{\mu}_2$, $\boldsymbol{\Sigma}_{22}$ and ν. Unfortunately, whilst the grouped approach does prevent the singularity problem, it also reduces the amount of information contained in the likelihood. This is a severe drawback in the conditional t case, because the computational results of the previous section indicate that generally there is too little information in the *un*grouped likelihood. Attempts to maximise the grouped conditional likelihood arising from the various data sets considered in this chapter were only successful for the hormone assay data, which gave results similar to those presented in Table 3.7. We therefore seem to be forced to abandon the conditional t framework as a means of providing robust estimates for regression models.

3.5 Discussion

The method of least-squares for obtaining a regression estimate is equivalent to treating the data as a random sample from a multivariate normal distribution, with location vector $\boldsymbol{\mu}$ and scatter matrix $\boldsymbol{\Sigma}$. The least-squares estimate $\hat{\boldsymbol{\beta}}$ may then be obtained from maximum likelihood estimates $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\Sigma}}$, through (3.6). With this motivation for least-squares, a robust estimate of $\boldsymbol{\beta}$ might then be obtained by considering a generalisation of the multivariate normal distribution, namely the multivariate t distribution, as an M-estimator. The methods proposed in Section 3.2 may then be employed to obtain a unique and robust bounded influence estimate of $\boldsymbol{\beta}$. This is achieved by using the (unique) conditional distribution derived from the full estimated joint distribution for the data. The general approach outlined in Section 3.2 also allows for multivariate regression models, where the required estimate is a matrix of regression coefficients.

For genuinely multivariate data, this form of modelling the data is plausible, although the same parameter ν is used to measure the degree of non-normality across all of the variables. The examples presented in Section 3.3 demonstrate that the method can also be successfully applied even when the multivariate t assumption cannot be justified, and estimates could be obtained for data sets arising in designed experiments; but here the usefulness of the t approach would be substantially reduced, as robust estimates may not always exist for models appropriate to the data.

One can obtain robust t M-estimates by modelling the error distribution in (1.4) as a univariate homoscedastic t distribution (Lange et al., 1989), but the question of uniqueness remains unresolved (Gabrielsen, 1982), and robustness is only achieved against outliers in the response variable. The conditional t model seems to provide a balance between the multivariate and univariate approaches, as it provides robustness against outliers in all of the variables, but without regard to the joint marginal distribution of the covariates. The resulting likelihood function is, however, markedly nonregular, although the examples based on the hormone assay and simulated data sets indicate that the conditional t likelihood can have a well-defined local maximum. In general though the conditional t framework seems to have too many parameters for reliable estimation from "small" data sets. Furthermore, conditions for the existence of a local maximum remain unclear, and so the reliability of the EM algorithm described in Appendix A.2 cannot as yet be ascertained.

In conclusion, when a regression model is desired for genuinely multivariate data, the multivariate t M-estimator should be strongly considered as a means of providing an estimate of the regression parameter that will be unique and robust for all of the variables contained in the linear model. For other types of data the conditional t model is more appropriate, but singularities in the likelihood surface make the approach too unstable to be of general use. In the next chapter, this nonregularity is found to have implications for likelihoods derived from other heteroscedastic models and sufficient conditions for two well-known models to yield unbounded likelihoods are presented.

Chapter 4

Nonregular Likelihoods in Heteroscedastic Regression

4.1 Introduction

4.1.1 General Remarks

In the previous chapter it was established that the likelihood function for the conditional t distribution is unbounded at points corresponding to nonreplicated observations. The problem arises from the form of the conditional dispersion matrix, which allows the fitted dispersion to be made singular for one observation, yet nonsingular and bounded for all other observations. Thus the essence of the problem lies in the heteroscedastic nature of the model, rather than some special property associated only with t distributions. This result forms the motivation for the present chapter, for although unbounded likelihood problems are well-known for finite normal-mixture models (Titterington, Smith, and Makov, 1985, Chapter 4), and also for certain models arising in extreme value theory (Smith, 1985), they have not yet been noted in the context of heteroscedastic regression.

This chapter considers two well-known classes of heteroscedastic regression models. For the first class, it is shown that, under a weak condition similar to that noted in Theorem 3.1, the likelihood is singular at points corresponding to nonreplicated observations, causing unrestricted maximum likelihood estimation to break down. For the second, a much stronger linear independence condition is obtained for the likelihood to be unbounded, and it is suggested that in this case, the singularity difficulty will, in general, be avoided.

4.1.2 Heteroscedastic Regression Models

Techniques for analyzing data with nonconstant variability have been examined by numerous authors, including McCullagh and Nelder (1989), Davidian and Carroll (1987), Carroll and Ruppert (1988) and Mak (1992). In particular, the maximum likelihood approach has been criticised for its sensitivity to misspecification of the error density and the assumed model for the dispersion. However, as shown in this chapter, there is also a problem of a more practical nature, as the likelihood surface may contain singularities, causing unrestricted maximum likelihood estimation to break down. As in the conditional t case, the problem arises if the dispersion model allows one fitted dispersion to be singular, whilst the remainder are kept nonsingular and finite.

The models under consideration are of the following form. Let $\{y_i,\ i = 1, \ldots, n\}$ be a set of independent observations in $\mathbb{R}$, which have a location-scale error density of the form $\sigma_i^{-1} f\{(y_i - \mu_i)/\sigma_i\}$, for some fixed function f such that $0 < f(u) < \infty$ for all $u \in \mathbb{R}$, where $\mu_i = \mu_i(\boldsymbol{\beta})$ is a real-valued regression location function of known form indexed by a set of unknown parameters $\boldsymbol{\beta}$, and $\sigma_i = \sigma_i(\boldsymbol{\alpha})$ is a dispersion function of known form indexed by a set of unknown parameters $\boldsymbol{\alpha}$. The log-likelihood function, $L(\boldsymbol{\beta}, \boldsymbol{\alpha})$, is then given by

$$L(\boldsymbol{\beta}, \boldsymbol{\alpha}) = -\sum_i \log \sigma_i(\boldsymbol{\alpha}) + \sum_i \log f\{(y_i - \mu_i(\boldsymbol{\beta}))/\sigma_i(\boldsymbol{\alpha})\}.$$

The functions μ and σ may involve known fixed covariates $\mathbf{x}_i$, recorded for each unit i. The terminology 'location-dispersion', rather than 'mean-variance', is used, since the framework allows for densities, such as the Cauchy, which do not have finite moments. The key observation is that if there exist directions in the $\boldsymbol{\alpha}$-space along which, for some j,

$$\sigma_j \to 0 \quad \text{and} \quad \sigma_i \to s_i \quad \text{where} \quad 0 < s_i < \infty \quad \text{for all} \quad i \neq j, \tag{4.1}$$

then at any value $\boldsymbol{\beta}_0$ of $\boldsymbol{\beta}$ where $\mu_j(\boldsymbol{\beta}_0) = y_j$, the likelihood will diverge to infinity.

The singularity problem does not seem to be restricted to the location-scale model outlined above. Problems might also occur if the objective function is a quasi-likelihood derived from a mean-variance relationship, or if the $\{\mu_i, \sigma_i\}$ are estimated nonparametrically. However, for clarity of exposition, we shall concentrate on the existence of singularities in the likelihood framework. This is considered in the next section, where sufficient conditions for the existence of directions satisfying (4.1) are developed for two well-known dispersion models.

4.2 Main Results

4.2.1 Model 1

The existence of directions in the $\boldsymbol{\alpha}$-space that satisfy (4.1) depends on the data and the form of the dispersion model. First, consider

$$\sigma_i(\boldsymbol{\alpha}) = \mathbf{z}_i^T \boldsymbol{\alpha}, \tag{4.2}$$

where the $\{\mathbf{z}_i\}$ are either fixed covariates or functions of the $\{\mu_i\}$. For example, one might have a quadratic model in the location, that is

$$\sigma_i = \alpha_0 + \alpha_1 \mu_i + \alpha_2 \mu_i^2. \tag{4.3}$$

Models similar to (4.3) have been considered by Davidian and Carroll (1987), Carroll and Ruppert (1988), and Mak (1992) gives maximum likelihood estimates for a real data set, using a quadratic function for the dispersion. However, as is shown by the following theorem, maximum likelihood estimates need not always exist:

Theorem 4.1. *The log-likelihood, $L(\boldsymbol{\beta}, \boldsymbol{\alpha})$, is unbounded at the points for which $\mu_j(\boldsymbol{\beta}_0) = y_j$ if there exists an $\boldsymbol{\alpha}_0$ such that*

$$\mathbf{z}_j^T \boldsymbol{\alpha}_0 = 0 \quad \textit{and} \quad 0 < \mathbf{z}_i^T \boldsymbol{\alpha}_0 < \infty \quad \textit{for all} \quad i \neq j. \tag{4.4}$$

Proof. Suppose there exists an $\boldsymbol{\alpha}_0$ for which (4.4) is true. Then, apart from the trivial case $\mathbf{z}_j = 0$, it is clear that there exists an $\boldsymbol{\alpha}_1$ such that $0 < \mathbf{z}_i^T \boldsymbol{\alpha}_1 < \infty$ for all i. Consider the value $\boldsymbol{\alpha}_\lambda = \lambda \boldsymbol{\alpha}_1 + (1 - \lambda)\boldsymbol{\alpha}_0$, where $0 < \lambda < 1$. Then $0 < \mathbf{z}_i^T \boldsymbol{\alpha}_\lambda < \infty$ for all i, and

$$L(\boldsymbol{\beta}_0, \boldsymbol{\alpha}_\lambda) = -\log \lambda - \log \mathbf{z}_j^T \boldsymbol{\alpha}_1 + \log f(0) - \sum_{i \neq j} \log \mathbf{z}_i^T \boldsymbol{\alpha}_\lambda + \sum_{i \neq j} \log f\{(y_i - \mu_i)/\mathbf{z}_i^T \boldsymbol{\alpha}_\lambda\}.$$

As $\lambda \to 0$, all terms except the first tend to a finite limit and hence $L(\boldsymbol{\beta}_0, \boldsymbol{\alpha}_\lambda) \to \infty$ as $\boldsymbol{\alpha}_\lambda \to \boldsymbol{\alpha}_0$ as $\lambda \to 0$. □

Therefore previously reported maximum likelihood estimates may at best correspond to *local* maxima (or indeed minima).

The condition that there exists an $\boldsymbol{\alpha}_0$ that satisfies (4.4) has a useful geometrical interpretation. It simply says there is a half-space with boundary through the origin that contains $\mathbf{z}_j$ in its boundary, whilst all other zs are contained in its interior. So for example, if $\mathbf{z}_j^T = (1 \;\; \mu_j \;\; \mu_j^2)$, such a half-space exists provided $\mathbf{z}_j$ is nonreplicated.

4.2.2 Model 2

Let us now consider dispersion models of form such as

$$\sigma_i(\boldsymbol{\alpha}) = \exp(\mathbf{z}_i^T \boldsymbol{\alpha}), \tag{4.5}$$

which are commonly proposed in the literature. Since (4.5) is strictly positive for finite $\mathbf{z}_i^T\boldsymbol{\alpha}$, the difficulty seems unlikely to occur, except in quite special cases, for models of this form. This would appear to be true, since it is now shown that there exists a direction in the $\boldsymbol{\alpha}$-space that satisfies (4.1) if and only if $\mathbf{z}_j$ is linearly independent of the other $\mathbf{z}$ values. Note, however, that (4.1) may not be absolutely necessary for singularities to occur.

First, let $\mathbf{Z} = (\mathbf{z}_1, \ldots, \mathbf{z}_n)^T$, $\mathbf{Z}^{(j)}$ be the matrix obtained by deleting the jth row of $\mathbf{Z}$, $\mathbf{c} = \mathbf{Z}\boldsymbol{\alpha}_0$ and $\mathbf{c}^{(j)} = \mathbf{Z}^{(j)}\boldsymbol{\alpha}_0$ for some arbitrary finite starting point $\boldsymbol{\alpha}_0$. A nonzero direction $\mathbf{u}$ is required such that

$$\mathbf{z}_j^T\mathbf{u} < 0 \quad \text{and} \quad \mathbf{Z}^{(j)}\mathbf{u} = 0, \tag{4.6}$$

which by elementary linear algebra (Towers, 1988, p. 147) exists if and only if $\mathbf{z}_j^T$ is linearly independent of the rows of $\mathbf{Z}^{(j)}$, i.e. if and only if $\text{rank}(\mathbf{Z}) = \text{rank}(\mathbf{Z}^{(j)})+1$. Now let $\boldsymbol{\alpha}_\lambda = \boldsymbol{\alpha}_0 + \lambda\mathbf{u}$, so that $\mathbf{z}_j^T\boldsymbol{\alpha}_\lambda \to -\infty$ as $\lambda \to \infty$, and $\mathbf{Z}^{(j)}\boldsymbol{\alpha}_\lambda = \mathbf{c}^{(j)}$ for all λ. Any such direction $\mathbf{u}$ will satisfy (4.1). If $\mathbf{z}_j^T$ is linearly dependent on the rows of $\mathbf{Z}^{(j)}$ there cannot exist a direction satisfying (4.1), since whenever $\mathbf{Z}^{(j)}\boldsymbol{\alpha}_\lambda$ tends to a finite limit, so must $\mathbf{z}_j^T\boldsymbol{\alpha}_\lambda$.

The linear independence condition can be tested immediately when the $\{\mathbf{z}_i\}$ are known, but this is not possible when they depend on unknown parameters, as in (4.3). For such cases it would be desirable to test whether or not there exists a point in the parameter space where the linear independence condition is true. In general this will not be straightforward, but for some models the linear independence condition for $\mathbf{Z}$ has a simple geometrical interpretation in terms of the model for the location. For example, suppose $\mu_i = \mathbf{x}_i^T\boldsymbol{\beta}$ and $\mathbf{z}_i^T = (1 \;\; \mu_i \;\; \mu_i^2)$. For some j and some $\boldsymbol{\beta}$, it is required that $\text{rank}(\mathbf{Z}) = \text{rank}(\mathbf{Z}^{(j)}) + 1$ and $\mu_j = y_j$. Since $\mathbf{Z}$ is in Vandermonde form (Isaacson and Keller, 1966, p. 188), $\text{rank}(\mathbf{Z}) = \text{rank}(\mathbf{Z}^{(j)})$ whenever $\mathbf{Z}^{(j)}$ contains three or more distinct rows, and so $\text{rank}(\mathbf{Z}) = \text{rank}(\mathbf{Z}^{(j)}) + 1$ if and only if for all $i \neq j$, $\mu_i \neq \mu_j$ and $\mathbf{Z}^{(j)}$ has no more than two distinct rows. Hence a $\boldsymbol{\beta}$ is required such that

$$\mathbf{x}_i^T\boldsymbol{\beta} = \begin{cases} y_j & \text{if } i = j, \\ k_1 \text{ or } k_2 & \text{if } i \neq j, \end{cases}$$

where $k_1 \neq k_2$ are arbitrary constants such that $k_1 \neq y_j$ and $k_2 \neq y_j$. It is easily seen that this has a solution in $\boldsymbol{\beta}$ if and only if the $\{\mathbf{x}_i\}$ satisfy the very strong requirement that they consist of two or three parallel subsets of co-planar points, such that one subset is a singleton. This contrasts with the weaker result for model (4.2), where any singleton yields an unbounded likelihood. In simple regression, for example, this means that the regressor variable can take at most three distinct values, otherwise there is no direction in the $\boldsymbol{\alpha}$-space for which (4.1) will be true. The argument extends easily to the case where $\mathbf{z}_i^T\boldsymbol{\alpha}$ is a polynomial in μ_i.

4.3 Examples

We now illustrate some of the points raised in the preceding discussion with two numerical examples. Both examples indicate that local maximum likelihood estimation can be a viable approach when the data satisfy conditions for the likelihood to be unbounded.

Example 4.1. We first consider a dispersion model of the form (4.2). Let $\mu_i(\boldsymbol{\beta}) = \beta_0 + \beta_1 x_i$, $\sigma_i^2(\boldsymbol{\alpha}) = \alpha_0 + \alpha_1\mu_i + \alpha_2\mu_i^2$, where α_0, $\alpha_2 > 0$, $\alpha_1^2 - 4\alpha_0\alpha_2 < 0$, and let f be the standard normal density. It was proved in Section 4.2.1 that if a data set includes a nonreplicated observation, then this model yields an unbounded likelihood. To demonstrate this, it is convenient to make a minor change in parameterization, and write $\sigma_i^2(\boldsymbol{\alpha})$ as $\sigma_i^2(\boldsymbol{\alpha}^*) = \alpha_0^* + \alpha_1^*(\mu_i - \alpha_2^*)^2$, for α_0^*, $\alpha_1^* > 0$. The profile likelihood for α_0^*, calculated by maximising

$$L = -\tfrac{1}{2}\sum_{i=1}^{n}\log\sigma_i^2 - \tfrac{1}{2}\sum_{i=1}^{n}\frac{(y_i - \mu_i)^2}{\sigma_i^2} \tag{4.7}$$

with respect to $(\boldsymbol{\beta}, \alpha_1^*, \alpha_2^*)$ as α_0^* varies, should then tend to infinity as $\alpha_0^* \to 0$.

Taking $\boldsymbol{\beta} = (0,1)^T$, $\alpha_0^* = 3$, $\alpha_1^* = 2$ and $\alpha_2^* = 5$, a data set was generated by simulating values of y from f for $x = \{1, 1.5, \ldots, 10.5\}$. A scatter plot of the data is given in Figure 4.1. For $\alpha_0^* \in [0.001, 5.0]$ the estimated profile shown in Figure 4.2 has been obtained, and the divergence as $\alpha_0^* \rightarrow 0$ is clear. However, it should be noted that the numerical methods employed may only have converged to local maxima, since we know that the likelihood function becomes extremely nonregular as α_0^* becomes small. The true profile likelihood for α_0^* may therefore diverge faster than the curve shown in Figure 4.2.

It is interesting to note the existence of a well-defined local maximum in the estimated profile, which suggests that, despite the presence of singularities, the likelihood approach may still yield useful results. This is in common with previously reported applications of maximum likelihood to the normal-mixture model.

Example 4.2. The second example considers a dispersion model of form (4.5). We now let $\sigma_i(\boldsymbol{\alpha}) = \exp(\alpha_0 + \alpha_1 \mu_i + \alpha_2 \mu_i^2)$, and take $\mu_i(\boldsymbol{\beta}) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i}$. It was shown in Section 4.2.2 that this model will yield an unbounded likelihood if the $\{\mathbf{x}_i\}$ consist of two or three parallel subsets of co-planar points, such that one subset is a singleton. For example, consider the data in Table 4.1, where the $\{\mathbf{x}_i\}$ have been chosen to satisfy this condition, and the $\{y_i\}$ have been simulated under the assumption that $\boldsymbol{\beta} = (\,2 \;\; -1 \;\; 0.5\,)^T$, $\boldsymbol{\alpha} = (\,-2 \;\; -2 \;\; 0.5\,)^T$ and f is the standard normal density. Local maximum likelihood estimates were sought for these data using two methods: the first used the numerical routines of the matrix programming language GAUSS 3.0 to maximise the log-likelihood (4.7), using a routine to evaluate the analytical derivatives

$$\frac{\partial L}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n} \mathbf{x}_i \left[\frac{(y_i - \mu_i)}{\sigma_i^2} \{1 + (y_i - \mu_i)(\alpha_1 + 2\alpha_2 \mu_i)\} - (\alpha_1 + 2\alpha_2 \mu_i) \right]$$

and

$$\frac{\partial L}{\partial \boldsymbol{\alpha}} = \sum_{i=1}^{n} \mathbf{z}_i \left\{ \frac{(y_i - \mu_i)^2}{\sigma_i^2} - 1 \right\},$$

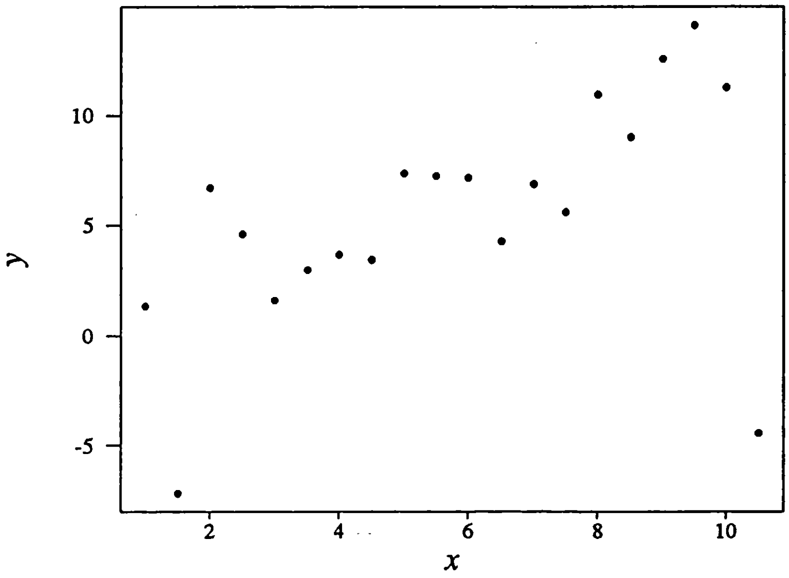

Figure 4.1: Scatter-plot of Example 4.1 data.

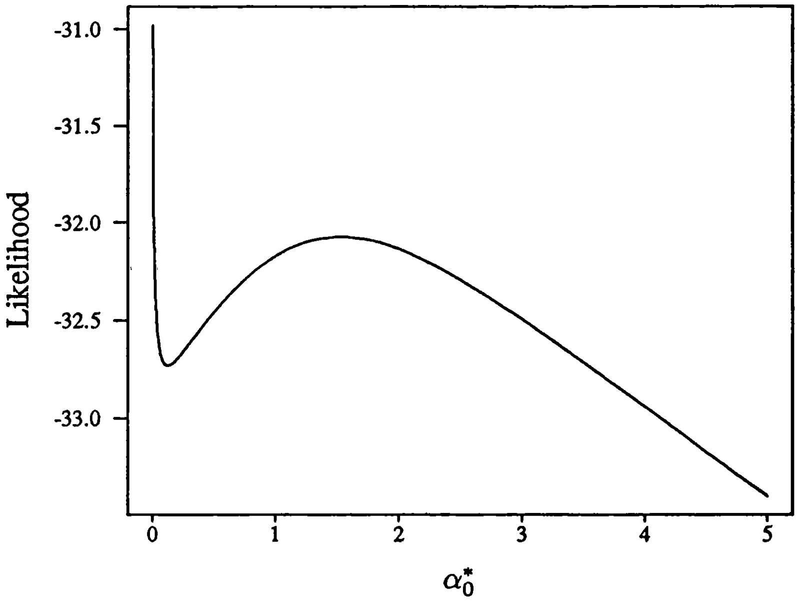

Figure 4.2: Estimated profile likelihood for α_0^*.

where $\mathbf{z}_i = (\ 1 \quad \mu_i \quad \mu_i^2\ )^T$; the second used the GLIM 3.77 macros given by Aitkin (1987), which maximise the likelihood by alternating between weighted-normal and gamma/log-link regression models. The results obtained are presented in Table 4.2. It can be seen that the estimates obtained by the two methods only agree to two significant figures, and the consequent discrepancy between the log-likelihoods is even greater. Also, the absolute values of the derivatives $\partial L/\partial\beta_1$ and $\partial L/\partial\beta_2$ are very large when evaluated at the GLIM estimate. The GLIM results were not improved by increasing the number of iterations and accuracy used at the normal and gamma regression steps, and only a minor improvement was obtained by using the higher-precision GLIM 4. Further checks revealed that when GLIM evaluated the log-likelihood at the GAUSS estimate, the value returned (0.759 to 3 significant figures) was much closer to that returned by GAUSS, and that both sets of parameter estimates yield negative definite matrices of second derivatives. However, the second derivatives do indicate extreme curvature with respect to the $\boldsymbol{\beta}$ parameters. A possible explanation for the discrepancies may therefore be that GAUSS, working at the unusually high level of 80-bit precision, is less susceptible than GLIM to the numerical condition of the problem. This explanation is consistent with the fact that similar results were obtained for a larger sample ($n = 65$) that satisfied the parallel hyperplanes condition, but when uniform 'noise' was added to the $\{x_{2i}\}$ in Table 4.1, the level of agreement between GAUSS and GLIM estimates increased as the sample size increased.

To summarise, the results indicate that, as in the previous example, local maximum likelihood estimates can be obtained when the data satisfy the condition for the likelihood to be unbounded. However, there is evidence to suggest that the accuracy of estimates obtained by GLIM macros may be less than that achieved for data sets that do not satisfy the unboundedness condition. It is therefore of practical interest to know if the condition is satisfied. An efficient method for checking this will be developed in the next chapter.

$\mathbf{x}_i$		y_i
2	6	2.9630
3	7	2.5250
4	8	1.9980
5	9	1.5140
6	10	1.0250
7	11	0.4390
8	12	0.1230
9	13	-0.1980
2	1	0.5360
3	2	-0.0630
4	3	0.3620
5	4	-4.4150
6	5	13.1530
7	6	22.3940
8	7	1003.6790
9	8	-1983.7790
5	6	-0.0040

Table 4.1: Data for Example 4.2.

Method	$\hat{\beta}$			$\hat{\alpha}$			L
GAUSS	2.002	-0.990	0.496	-1.916	-2.193	0.537	0.761
	$< 10^{-10}$	$< 10^{-10}$	$< 10^{-10}$	$< 10^{-10}$	$< 10^{-10}$	$< 10^{-10}$	
GLIM	2.012	-0.989	0.494	-1.911	-2.202	0.540	0.734
	$< 10^{-3}$	21.87	43.81	$< 10^{-5}$	$< 10^{-6}$	$< 10^{-5}$	

Table 4.2: Estimates and absolute derivatives for Example 4.2.

4.4 Discussion

Unbounded likelihoods seem likely to occur widely for dispersion models of form (4.2), given the rather weak condition (4.4) for the existence of a direction that satisfies (4.1), and so the use of (4.2) cannot be recommended. As an alternative, if the dispersion model (4.5) is considered inappropriate or computationally too demanding, one might consider the commonly proposed power functions

$$\sigma_i(\alpha) = (\alpha_0 + |\mu_i|)^{\alpha_1} \quad \text{and} \quad \sigma_i(\alpha) = |\mu_i|^{\alpha},$$

where $\alpha_0 > 0$, since it is easily seen that these only yield unbounded likelihoods when $y_i = 0$ for some i.

When the dispersion is modelled by (4.5), a direction satisfying (4.1) exists if and only if the rank of the dispersion model can be decreased by deleting one observation. Whilst it has not been verified that this condition is necessary for the likelihood to be unbounded, it does seem that by using models of form (4.5) the difficulty will largely be avoided. Exceptions are perhaps most likely to occur with data arising from factorial experiments, in which case a specific structure is built-in to the covariates. Heteroscedastic models based on (4.5) have been considered for replicated factorials by Aitkin (1987) and Nair and Pregibon (1988), although in the full replication case it is highly unlikely that there will exist a direction in the parameter space for which (4.1) will be true. However, if the model is applied in the nonreplicated case we would strongly recommend that the linear independence condition be checked, if it is possible to do so. As well as indicating the presence of singularities in the likelihood surface, it may also warn of potential reductions in accuracy if estimates are to be computed using GLIM (as they often are), as demonstrated by Example 4.2. If numerical problems are identified one might then consider an alternative dispersion model that yields a bounded likelihood.

As noted in the introduction, the scope of the problem is not restricted to the framework discussed here. An apparent example of this can be found in Rigby and

Stasinopoulos (1994), who estimate semi-parametric heteroscedastic models using quasi-likelihood. They consider counts of AIDS cases over time and present graphs of estimated location and dispersion functions. However, the final six observations are fitted perfectly and have dispersion values of zero. This is not commented on by the authors, despite the fact that the observations in question are contained in the *most* variable part of the time series.

The results in this chapter provide a salutary warning for those who work with heteroscedastic models. If conditions for the likelihood to be unbounded hold, then numerical routines may fail to converge, or converge to spurious estimates. In the absence of a guarantee that the likelihood function will have one finite mode, great care should be taken to ensure that a satisfactory local maximum has indeed been found.

In the next chapter we propose a method for testing the parallel hyperplanes condition that was developed in Section 4.2.2. The condition for the existence of t M-estimates that arose in Chapter 3 is also considered, since it too is of a co-planar points form.

Chapter 5

Testing the existence of maximum likelihood estimates

5.1 Introduction

In this chapter efficient methods are proposed for testing two conditions for the existence of maximum likelihood estimates. The first condition is the 'parallel hyperplanes' condition for the heteroscedastic regression model of Section 4.2.2; the second is 'Condition D_ν^*' (Kent and Tyler, 1991) for the location-scatter model of Chapter 3. In each case it has been shown that maximum likelihood estimates will not exist if the data satisfy a given geometrical property, but practical methods for testing the data have not been discussed. In this chapter we provide such methods and illustrate their use with examples. We begin by reviewing the conditions listed above.

Chapter 4 considered heteroscedastic regression models for data $\{(y_i, \mathbf{x}_i) : i = 1, \ldots, n\}$, where the $\{y_i\}$ are univariate responses and the $\{\mathbf{x}_i\}$ are p-dimensional covariates. It was shown in Section 4.2.2 that the location-scale relationship $\mu_i =$

$\mathbf{x}_i^T\boldsymbol{\beta}$ and $\sigma_i = \exp(\alpha_0+\alpha_1\mu_i+\alpha_2\mu_i^2)$ gives rise to an unbounded likelihood function if the $\{\mathbf{x}_i\}$ consist of two or three parallel subsets of co-planar points, such that one subset is a singleton. A simple method for testing the $\{\mathbf{x}_i\}$ would be to enumerate all possible combinations of two or three subsets such that one is a singleton and test each one in turn. A calculation reveals the total number of such combinations to be

$$n + \binom{n}{2} + n\left\{\sum_{r=2}^{n_2-1}\binom{n-1}{r} + (n/2-n_2)\binom{n-1}{n_2}\right\},$$

where n_2 denotes the integer part of $(n-1)/2$. Clearly this quantity explodes in size as n increases: for example, at $n = 20$ it is around 5 million, but at $n = 30$ it increases to 8053 million and so an approach based on complete enumeration will only be possible for very small data sets.

A similar problem arises for the condition derived by Kent and Tyler (1991), which is also of a co-planar points form. For a set of p-dimensional observations $\{\mathbf{y}_i\}$, they show that, for $\nu \geq 1$, estimates of the location and scatter parameters, defined implicitly by the equations (3.3) and (3.4), exist only if Condition D_ν^* is satisfied. Although Kent and Tyler point out that the condition is satisfied with probability one for random samples of size $n \geq p+1$ from any continuous distribution, the effectively discrete nature of real data may negate this welcome property. One might therefore wish to check if, for some q such that $0 \leq q \leq p-1$, there exists a subsample from the data of size $n_q = [n(\nu+q)/(\nu+p)]+1$ that lies in a hyperplane of dimension q, where '$[\cdot]$' denotes integer part. In practice it is useful to test the condition for $\nu = 1$, since if the condition is not satisfied for $\nu = 1$, it is not satisfied for all $\nu > 1$. Unfortunately the total number of subsamples available for testing grows extremely quickly with n and p. For example, with $n = 30$, $p = 4$ and $\nu = 1$ there are

$$\binom{30}{25} + \binom{30}{19} + \binom{30}{13} + \binom{30}{7} \approx 176 \text{ million}$$

subsamples for consideration. So once again it will only be possible to enumerate and test all possible subsamples when n is small. Therefore, in order to test either of these conditions large numbers of combinations must be excluded from consideration *before* they are enumerated. In the following section algorithms are described that achieve such a reduction in computation.

5.2 Description of algorithms

5.2.1 Parallel Hyperplanes

We consider first the parallel hyperplanes condition. The objective is to determine if the minimum number r, say, of parallel hyperplanes that contain all the data, subject to a singleton constraint, is less than or equal to 3. If we determine that $r \leq 3$ then maximum likelihood estimates will not exist; otherwise, they almost certainly will.

Matters can be simplified slightly by first performing two operations on the matrix $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^T$. Firstly, if the matrix has a constant column it may be deleted in order to reduce the dimension of the problem and hence the computational burden. Secondly, we may delete any replications, so that the $\{\mathbf{x}_i\}$ we actually test are all distinct. This is justified since if the condition is not satisfied for a set of distinct points it cannot be satisfied when we add replications. Furthermore, if it is satisfied for a set of distinct points we need only then check if the point in the singleton hyperplane, $\mathbf{x}_s$ say, has a replication. If there is no replication the condition is satisfied. Otherwise the likelihood will be unbounded only if all replications of $\mathbf{x}_s$ yield responses equal to y_s. This is of course unlikely, but it may occur if the procedure for measuring the response involves gross rounding. Therefore in the following it may be assumed that the $\{\mathbf{x}_i\}$ are distinct and that $\mathbf{X}$ does not contain a constant column. To avoid trivial cases we shall also assume the following: that the dimension, p, of the $\{\mathbf{x}_i\}$ is greater than one, since to test the condition for $p = 1$ one need only check if $n > 3$; that $\mathbf{X}$ has full rank; and finally, that $n > p + 1$ and there exists a set of $p + 1$ points in general position.

In order to develop an efficient method we exploit the geometry of the condition. Let us suppose the condition is true and we have available a set of $p+1$ points in general position that does not include the singleton. Since $p+1$ points in general position determine the vertices of a simplex, we can obtain from its facets $p+1$ candidate hyperplanes such that one corresponds to a 'solution' hyperplane, that is, one of the two or three parallel hyperplanes that between them contain all of the data. For example, consider the data shown in Figure 5.1. In case (a) the simplex yields a solution hyperplane through points $\mathbf{x}_1$ and $\mathbf{x}_2$. If, however, the set of $p+1$ points does include the singleton, as in case (b), then the facets of the simplex need not define a solution hyperplane. In practice, of course, we do not actually know if a given set of $p+1$ points includes the singleton, so to be sure of obtaining a solution hyperplane we must consider $p+2$ non co-planar points and find the, not necessarily distinct, candidate hyperplanes for each of the $\frac{1}{2}(p+1)(p+2)$ selections of p points, as in (c). If the condition is true, one of the candidate hyperplanes must be a solution. This leads to a simple algorithm for testing the condition.

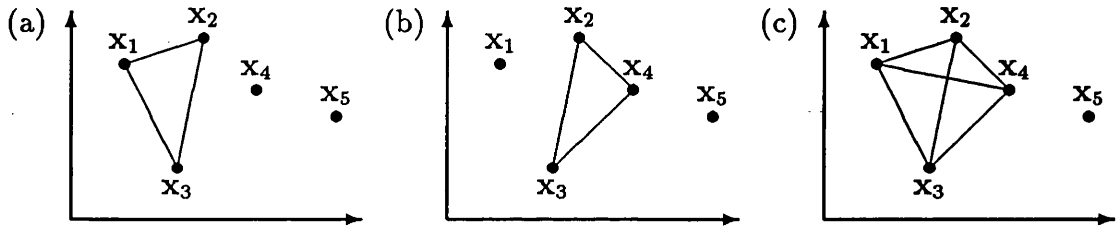

Figure 5.1: Simple example for $p = 2$.

Step 1: Find a sample of $p+2$ non co-planar points.

Step 2: Calculate the hyperplanes defined by each selection of p points.

Step 3: Check if each hyperplane could be one of two or three parallel hyperplanes that contain all the data. If no such hyperplanes are found, the condition is not satisfied.

There follow some comments regarding the practical implementation of these steps. Step 1 is easily achieved, so let us assume without loss of generality that $\mathbf{x}_1, \ldots, \mathbf{x}_{p+1}$ are non co-planar, so that we can take $S = \{\mathbf{x}_1, \ldots, \mathbf{x}_{p+2}\}$ as the sample. For step 2, we first define, for $i = 1, \ldots, p+1$, H_i to be the hyperplane defined by the p points obtained by deleting $\mathbf{x}_i$ and $\mathbf{x}_{p+2}$ from S. We now have $p+1$ distinct hyperplanes. In order to calculate the remaining $\frac{1}{2}p(p+1)$ hyperplanes we proceed as follows: for all $i = 2, \ldots, p+1$ and $j = 1, \ldots, i-1$, delete $\mathbf{x}_i$ and $\mathbf{x}_j$ from S and define a new hyperplane from the remaining p points. Note, however, that if for some i, $\mathbf{x}_{p+2} \in H_i$, we need not consider the deletions for any j since we will just duplicate H_i. Similarly if, for some j, $\mathbf{x}_{p+2} \in H_j$, we need not consider deleting $\mathbf{x}_i$ and $\mathbf{x}_j$ since we will duplicate H_j. Incorporating these checks into step 2 will ensure that we only define (and hence test) distinct candidate hyperplanes.

We now consider step 3. For a given candidate, the approach is to check each observation in turn, defining a new hyperplane parallel to the candidate if the current observation is not contained in any of the available hyperplanes. This continues until we are forced to define a fourth hyperplane, in which case the candidate cannot be one of two or three parallel hyperplanes that contain all of the data. This procedure is described exactly in the following algorithm, which we start with $i = 1$ and d equal to the number of hyperplanes defined in step 2.

(i) If $i > d$ go to (vi). Otherwise, let $P_1 = H_i$, *fail* $= 0$, $r = 1$, $n_1 = 0$, $n_2 = 0$, $n_3 = 0$ and $j = 1$. Go to (ii).

(ii) If *fail* $= 1$ or if $j > n$ go to (v). Otherwise, let $in = 0$, $k = 1$ and go to (iii).

(iii) If $k > r$ go to (iv). If $\mathbf{x}_j \in P_k$, let $n_k = n_k + 1$, $in = 1$ and go to (iv). Otherwise let $k = k + 1$ and go to (iii).

(iv) If $in = 1$ let $j = j + 1$ and go to (ii). Otherwise let $r = r + 1$. If $r > 3$ let *fail* $= 1$ and go to (ii). Otherwise let P_r be the hyperplane through $\mathbf{x}_j$ parallel to P_1. Let $n_r = n_r + 1$, $j = j + 1$ and go to (ii).

(v) If *fail* $= 0$ and $n_1 = 1$ or $n_2 = 1$ or $n_3 = 1$, write "condition true" and stop. Otherwise let $i = i + 1$ and go to (i).

(vi) Write "condition false" and stop.

5.2.2 Condition D_ν^*

We now turn to Condition D_ν^*. For given values of n, p and ν such that $n > p$ and $\nu \geq 1$, we wish to verify if for some value q, such that $0 \leq q \leq p - 1$, there exists a sample of size $n_q = [n(\nu + q)/(\nu + p)] + 1$ that is contained in a hyperplane of dimension q.

In this case the condition does not require the entire data set to satisfy a given geometrical property, and so it has not been possible to develop as efficient an algorithm as that developed for the parallel hyperplanes condition. We are, however, still able to make significant gains over complete enumeration by adopting a 'branch and bound' approach. The idea is to start with an empty sample and employ a simple branching strategy that either includes or excludes each observation in turn. At each branching, the sample is checked to see if it is not contained in a hyperplane of dimension q or if the number of observations excluded from the sample is greater than $n - n_q$. If either of these conditions is true we must bound the sample and re-start the branching strategy from the last included observation. This process continues until a sample is found that satisfies the condition or until the first $n - n_q$ observations have been excluded, in which case the condition is not true. In practice, this process can be simplified slightly by noting that the hyperplane check need not be carried out if the sample contains $q + 1$ or fewer observations. Furthermore, it will not be necessary to compare the sample size with n_q in these cases, since it is easily shown that $n_q > q + 1$ whenever $n > p$.

A detailed algorithm to implement the strategy outlined above is now presented. It uses indicator variables, a_i, such that $a_i = 1$ if the ith observation is included in the sample, and $a_i = 0$ otherwise. Initially we let $a_i = 0$ for all i. We now start the algorithm by setting $q = p$ and going to step 1.

Step 1: Let $q = q - 1$. If $q < 0$ write "condition false" and stop. Otherwise, let $n_q = [n(\nu + q)/(\nu + p)] + 1$, set $a_i = 0$ for all i and set $i = 0$. Go to step 2.

Step 2: Let $i = i + 1$. If $a_i = 1$ let $a_i = 0$. Otherwise let $a_i = 1$. Go to step 3.

Step 3: Calculate $\bar{n} = i - \sum_j a_j$. If $\bar{n} > n - n_q$, go to step 7. Otherwise step 4. (*If more than $n - n_q$ observations have been excluded, we cannot form a sample of size n_q and must therefore attempt to exclude an observation from the sample.*)

Step 4: If $a_i = 0$ go to step 1. Otherwise, if $\sum_j a_j \leq q + 1$, go to step 2, else go to step 5. (*Since any set of $q + 1$ or fewer points is contained in a q-dimensional hyperplane, there is no need to check the actual observations*).

Step 5: If the observations for which $a_i = 1$ are contained in a q-dimensional hyperplane, go to step 6, else go to step 7. (*If the current sample does not consist of co-planar points, we must attempt to exclude an observation from the sample.*)

Step 6: If $\sum_j a_j = n_q$ write "condition true" and stop. Otherwise go to step 2.

Step 7: If $a_i = 0$ for all i go to step 1. Otherwise let k be the largest value of i for which $a_i = 1$, set $i = k - 1$ and go to step 2.

5.3 Examples

Example 5.1. (Parallel Hyperplanes). To demonstrate the computational saving of the algorithm over complete enumeration, we applied it to an $\mathbf{X}$ matrix of standard normal random deviates, with $n = 30$ and $p = 4$. As already noted, for $n = 30$ an approach based on complete enumeration would have to consider some 8053 million partitions in order to prove that $r > 3$. However, the parallel planes algorithm needs to consider just 15 candidate hyperplanes and, coded in the matrix programming language GAUSS, took 0.2 seconds (on a 486-DX2 50 PC) to prove that $r > 3$, whereas a program based on complete enumeration would certainly take hundreds of hours.

It is clear that the major factor in the running time is the value of p, as it determines the number of candidate hyperplanes to be tested. Increasing the value of p does have a large effect, but the running times can still be short. For example, increasing p to 8 increased the running time to 1.2 seconds and $p = 16$ gave 10.3 seconds. The running time is generally unaffected by an increase in n because to reject the condition we need only verify that the condition does not hold for a subset of the data.

Example 5.2. (Condition D_ν^*). Once again we consider a random normal data matrix, with $n = 30$ and $p = 4$. We take $\nu = 1$ and so an approach based on complete enumeration would need to test some 176 million samples in order to show the condition does not hold. The branch and bound algorithm visited step 2 just 7866 times, but performed the hyperplane test in step 5 only 3057 times, due to the bounding steps 3 and 4. The program took 15 seconds to run. Increases in n and p do have a large effect on the running time, due to the increased number of samples that must be considered. For example, at $n = 100$ and $p = 4$ the running time increased to 21 minutes; with $n = 100$ and $p = 5$ to 110 minutes. Nevertheless, a data set of this size could not possibly be tested by complete enumeration (the number of samples is of the order of 10^{29}) and so branch and bound still presents a huge gain in efficiency.

It should be noted that the run-times quoted in this example and in Example 5.1 should be regarded as lower bounds for testing data sets that do not satisfy the conditions. This is because random data sets, such as those considered in these examples, will be in general position and hence candidate solutions will be rejected at the earliest possible stage.

5.4 Discussion

This chapter has discussed approaches for testing conditions for the existence of maximum likelihood estimates in two models. The conditions are such that naive approaches based on complete enumeration of available combinations will generally be computationally too expensive to be viable.

The results discussed in Example 5.1 suggest that the parallel hyperplanes algorithm will be useful for virtually all data sets for which a heteroscedastic linear model is thought appropriate. Many data sets can now be tested for Condition D_{ν}^{*}, but certainly not all. This is because even the branch and bound approach will become too expensive for large enough values of n and p. Nevertheless, the branch and bound approach does enable the condition to be tested for many data sets of interest, where previously this may not have been the case.

Finally, it should be noted that the time taken for either algorithm to stop will depend on the order of the observations. However, investigation into a procedure for finding an optimal ordering has not been undertaken, as it does not seem useful.

Chapter 6

Concluding Remarks

The original work presented in the thesis commenced in Chapter 2 with an examination of the uniqueness problems that can arise when using redescending M-estimators for linear regression models. The results obtained by Rivest (1989) were shown to be flawed in that they neglect a continuity problem which arises when the solution corresponding to the global minimum of (2.4) is not unique. This motivated the development of a redescending approach from which unique estimates of the model parameters could be reliably obtained. Such an approach was considered in Chapter 3, where estimates were obtained by embedding the linear model in a multivariate t location-scatter framework. This enabled the location-scatter existence and uniqueness results of Kent and Tyler (1991) to be applied in the regression context. Whilst robust regression estimates have already been defined in terms of robust estimates of location and scatter by Maronna and Morgenthaler (1986), their work is limited in that the key question of uniqueness is not addressed.

In practice, with the exception of the results obtained for the Stack-Loss data set, the multivariate t M-estimates compared favourably with estimates obtained from other robust methods. They are easily computed, and for genuinely multivariate data the approach to modelling the data is plausible, although examples have demonstrated that useful results can also be obtained when the multivariate t assumption cannot be justified. For such cases it would be preferable to model the data via the conditional t approach, since the estimates will not be affected by an assumed joint marginal distribution for the covariates. However, as noted in Theorem 3.1, the resulting objective function is extremely nonregular, being unbounded at points in the parameter space corresponding to nonreplicated observations. Attempts to identify local modes were generally unsuccessful for real data sets, although modes were found for large samples of simulated conditional t data. It would be desirable to obtain a sufficient condition for the conditional t likelihood to have a well-defined local maximum, but this has not been attempted due to the theoretical difficulty of the task and the lack of practical success in identifying modes for real data sets. Modifications to the conditional t approach were considered, but not found to be useful.

Although the conditional t discussion proved to be of limited importance in the context of robust regression, it led to the discovery, in Chapter 4, of hitherto unrecognised likelihood problems for certain well-known heteroscedastic regression models. Sufficient conditions were obtained for a location-scale framework to yield an unbounded likelihood when the dispersion was modelled by either

$$\sigma_i(\alpha) = \mathbf{z}_i^T \alpha \quad \text{or} \quad \sigma_i(\alpha) = \exp(\mathbf{z}_i^T \alpha).$$

For the first model the likelihood was shown to be unbounded under a very weak condition noted in Theorem 4.1, whereas for the second model a much stonger linear independence condition was obtained for the likelihood to be unbounded. The choice of dispersion function can therefore have a great effect on the regularity of the likelihood. For example, when the $\{\mathbf{z}_i\}$ are quadratic in a linear location function, the first model yields an unbounded likelihood at any unreplicated observation, whereas the second yields an unbounded likelihood if the covariates satisfy the very strong parallel hyperplanes condition.

Examples based on simulated data suggested that, in the unbounded case, local maximum-likelihood estimation can be viable for both of the dispersion models considered, although questions have been raised over the accuracy of estimates obtained using GLIM when the covariates satisfy the parallel hyperplanes condition. However, this condition may be tested very efficiently by using the simplex-based algorithm discussed in Chapter 5.

Finally, there is scope for further work in this area since the problem may occur in more complicated heteroscedastic models than those considered here. In the meantime, one should proceed with caution if an objective function derived from a heteroscedastic model is not known to be bounded.

Appendix A

Likelihood Fitting

A.1 Full Multivariate t M-estimates

We have experience with two algorithms for maximising the multivariate t log-likelihood: a simple approach based on the estimating equations (3.3) and (3.4) and a guaranteed-convergent algorithm given by Kent and Tyler, which involves embedding the location-scatter estimation problem within a scatter-only problem of greater dimension. In both cases the degrees of freedom parameter ν is held fixed in order to calculate an estimate $(\hat{\mu}_\nu, \hat{\Sigma}_\nu)$. This process is then repeated over a grid of values of ν, in order to obtain a final estimate $(\hat{\mu}, \hat{\Sigma})$ that corresponds to a maximum of the profile log-likelihood as a function of ν. For comments on the uniqueness aspects of this approach, refer to the text of Section 3.2.2.

First we consider the less complicated of the two algorithms. Given initial estimates $\mu^{(0)} \in \mathbb{R}^p$ and $\Sigma^{(0)} \in \mathcal{P}_p$, define

$$\begin{aligned} \mu^{(m+1)} &= \text{ave}\{\omega_i \mathbf{z}_i\} / \text{ave}\{\omega_i\} \\ \Sigma^{(m+1)} &= \text{ave}\{\omega_i(\mathbf{z}_i - \mu^{(m)})(\mathbf{z}_i - \mu^{(m)})^T\} \end{aligned}$$

where $\omega_i = u_\nu(s_i)$, $u_\nu(s) = (\nu+p)/(\nu+s)$ and $s_i = (\mathbf{z}_i - \mu^{(m)})^T(\Sigma^{(m)})^{-1}(\mathbf{z}_i - \mu^{(m)})$. This algorithm is attractive in its simplicity, but is not guaranteed-convergent for all possible starting values. In practice however, no problems were encountered with this algorithm. The validity of estimates obtained from it was examined by using them as starting values for the Kent and Tyler algorithm. No discrepancies unattributable to the limitations of machine precision were encountered.

The algorithm presented by Kent and Tyler, when applied to the multivariate t distribution, is as follows: for a data set $\{\mathbf{z}_i,\ i = 1, \ldots, n\} \in \mathbb{R}^p$, define $\mathbf{d}_i = (\mathbf{z}_i^T, 1)^T \in \mathbb{R}^{p+1}$. Also, given $\Sigma \in \mathcal{P}_p$ and $\mu \in \mathbb{R}^p$, define $\mathbf{A} \in \mathcal{P}_{p+1}$ by

$$\mathbf{A} = \begin{bmatrix} \Sigma + \mu\mu^T & \mu \\ \mu^T & 1 \end{bmatrix}.$$

Given some initial estimate $\mathbf{A}^{(0)} \in \mathcal{P}_{p+1}$ define

$$\mathbf{A}^{(m+1)} = \text{ave}\left[u_\nu^* \left\{\mathbf{d}_i^T \left(\mathbf{A}^{(m)}\right)^{-1} \mathbf{d}_i\right\} \mathbf{d}_i\mathbf{d}_i^T\right],$$

where $u_\nu^*(s) = u_\nu(s-1)$ for $s \geq 1$ and $u_\nu^*(s) = u_\nu(0)$ for $s < 1$. The new estimate $\mathbf{A}^{(m+1)}$ is then scaled by the value of its $(p+1, p+1)$ element, so that $\mathbf{A}_{p+1,p+1}^{(m+1)} = 1$. For details, see Kent and Tyler, where it is shown that this algorithm will always converge to a unique estimate

$$\hat{\mathbf{A}} = \begin{bmatrix} \hat{\Sigma} + \hat{\mu}\hat{\mu}^T & \hat{\mu} \\ \hat{\mu}^T & 1 \end{bmatrix} \in \mathcal{P}_{p+1},$$

such that $L_{\mathbf{Z}}(\hat{\mu}, \hat{\Sigma}, \nu) \leq L_{\mathbf{Z}}(\mu, \Sigma, \nu)$ for all $\mu \in \mathbb{R}^p, \Sigma \in \mathcal{P}_p$.

In practice the Kent and Tyler algorithm proved to be the fastest, by a narrow margin over the simple algorithm, results typically being obtained in just a few seconds (using GAUSS 3.0 on a 486-DX2 50 PC). Alternatively, one could attempt to estimate μ, Σ and ν simultaneously using the EM algorithm (Dempster, Laird, and Rubin, 1977). However, this notoriously slow algorithm was not implemented due to the success of the profile approach.

A.2 Conditional t M-estimates

In the following discussion it is assumed that ν is fixed. The procedure may then be repeated over a grid of values of ν, in order to obtain a final estimate.

First, consider the model

$$y \mid u \sim \mathcal{N}(\mu, \sigma^2/u) \quad \text{and} \quad u \sim \chi^2_\nu/\nu, \tag{A.1}$$

where $\nu > 0$. As noted by Lange et al. (1989), this framework yields the following standard results:

(i) $y \sim t\,(\mu, \sigma^2, \nu)$;

(ii) $u \mid y \sim \chi^2_{\nu+1}/(\nu + \delta^2)$, where $\delta^2 = (y - \mu)^2/\sigma^2$. Hence

$$\mathrm{E}[u \mid y] = (\nu + 1)/(\nu + \delta^2).$$

Lange et al. use this framework in order to calculate maximum likelihood estimates via the EM algorithm. In order to maximise the conditional t log-likelihood (3.7), we extend (A.1) thus: let

$$y \mid (\mathbf{x}, u) \sim \mathcal{N}(\beta_0 + \mathbf{x}^T \boldsymbol{\beta}_1, \sigma^2/u) \quad \text{and} \quad u \mid \mathbf{x} \sim \chi^2_{\nu_{1\cdot 2}}/\nu_{1\cdot 2}\, g_\nu(\mathbf{x}, \boldsymbol{\mu}_2, \boldsymbol{\Sigma}_{22}). \tag{A.2}$$

A straightforward but tedious calculation shows that this formulation yields the required conditional t distribution for $y \mid \mathbf{x}$, i.e.

$$y \mid \mathbf{x} \sim t\,(\beta_0 + \mathbf{x}^T \boldsymbol{\beta}_1, g_\nu(\mathbf{x}, \boldsymbol{\mu}_2, \boldsymbol{\Sigma}_{22})\sigma^2, \nu_{1\cdot 2}).$$

Maximum likelihood estimation for the conditional t model may therefore be attempted by applying the EM algorithm, with missing data $\{u_i : i = 1, \ldots, n\}$. Note, however, that the normal model in (A.2) does not include the function g_ν and has no parameters in common with the chi-squared model, so the models may be estimated independently for given values of u. In the normal case this can be achieved simply using weighted-least-squares. Given estimates $\beta_0^{(m)}$, $\boldsymbol{\beta}_1^{(m)}$, $\sigma^{2(m)}$,

$\mu_2^{(m)}$ and $\Sigma_{22}^{(m)}$ at iteration m, a calculation reveals the weight computed at the E step to be

$$\begin{aligned} w_i^{(m)} &= \mathrm{E}\left[u_i \,|\, y_i, \mathrm{x}_i, \beta_0^{(m)}, \beta_1^{(m)}, \sigma^{2(m)}, \mu_2^{(m)}, \Sigma_{22}^{(m)}\right] \\ &= \frac{\nu + p}{\nu_{1\cdot 2}\, g_\nu(\mathrm{x}_i, \mu_2^{(m)}, \Sigma_{22}^{(m)}) + (y_i - \beta_0^{(m)} - \mathrm{x}_i^T \beta_1^{(m)})^2/\sigma^{2(m)}}. \end{aligned}$$

The M step involves maximising the complete data log-likelihood, which is the sum of a normal and a chi-squared log-likelihood. We first calculate estimates $(\beta_0^{(m+1)}, \beta_1^{(m+1)}, \sigma^{2(m+1)})$ to maximise the normal likelihood, with weights $w_i^{(m)}$, i.e. we find the value $(\beta_0^{(m+1)}, \beta_1^{(m+1)})$ that minimises

$$\sum_i w_i^{(m)}(y_i - \beta_0 - \mathrm{x}_i^T \beta_1)^2$$

and then calculate

$$\sigma^{2(m+1)} = \frac{1}{n} \sum_i w_i^{(m)} (y_i - \beta_0^{(m+1)} - \mathrm{x}_i^T \beta_1^{(m+1)})^2.$$

We then find $\mu_2^{(m+1)}$ and $\Sigma_{22}^{(m+1)}$ to maximise the chi-squared log-likelihood which, up to an additive constant, is given by

$$L(\mu_2, \Sigma_{22}) = \frac{\nu_{1\cdot 2}}{2} \left\{ \sum_{i=1}^{n} \log g_\nu(\mathrm{x}_i, \mu_2, \Sigma_{22}) - w_i^{(m)} g_\nu(\mathrm{x}_i, \mu_2, \Sigma_{22}) \right\}. \qquad (A.3)$$

Differentiating (A.3) with respect to μ_2 and Σ_{22} yields estimating equations which may be solved iteratively by, for example, successively redefining the estimates in terms of the estimating equations, as in the simple algorithm considered in Appendix A.1.

The question of what starting values to use for this EM algorithm does not seem to have a clear-cut answer. Obvious candidates are the parameter estimates obtained by using the full multivariate t approach, and when the data are multivariate t, these should be adequate. Otherwise, such starting points may not be sufficiently close to a well-defined local maximum to enable the EM algorithm to converge satisfactorily. Modifications to this algorithm should therefore be investigated, with the aim of ensuring convergence to a satisfactory solution for those data sets where such a solution exists. Such possibilities are not investigated here.

A.3 Calculation of Derivatives

Section 3.4.2 presented examples where standard errors were quoted for conditional t M-estimates. The standard errors were obtained from observed information matrices, which in turn were calculated by numerically differentiating the analytical first derivatives of the log-likelihood (3.7). Including the term for the degrees of freedom parameter ν, this may be written as

$$\begin{aligned}
&L(\beta_0, \boldsymbol{\beta}_1, \sigma^2, \boldsymbol{\mu}_2, \boldsymbol{\Sigma}_{22}, \nu) \\
&\quad = n \log\left\{\frac{\Gamma((\nu+p)/2)}{(\pi\nu_{1\cdot 2})^{1/2}\Gamma(\nu_{1\cdot 2}/2)}\right\} - \frac{1}{2}\sum_{i=1}^{n} \log \sigma_i^2 - \frac{(\nu+p)}{2}\sum_{i=1}^{n}\left\{1 + \frac{r_i^2}{\nu_{1\cdot 2}\,\sigma_i^2}\right\},
\end{aligned}$$

where $r_i = y_i - \beta_0 - \mathbf{x}_i^T\boldsymbol{\beta}_1$, $\sigma_i^2 = \{\nu + (\mathbf{x}_i - \boldsymbol{\mu}_2)^T\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{x}_i - \boldsymbol{\mu}_2)\}\sigma^2/\nu_{1\cdot 2}$ and $\nu_{1\cdot 2} = \nu + p - 1$. Straightforward calculations reveal that

$$\begin{aligned}
\frac{\partial L}{\partial \beta_0} &= (\nu+p)\sum_{i=1}^{n}\frac{r_i}{\nu_{1\cdot 2}\,\sigma_i^2 + r_i^2}, \\
\frac{\partial L}{\partial \boldsymbol{\beta}_1} &= (\nu+p)\sum_{i=1}^{n}\frac{\mathbf{x}_i r_i}{\nu_{1\cdot 2}\,\sigma_i^2 + r_i^2}, \\
\frac{\partial L}{\partial \sigma^2} &= \frac{1}{2\sigma^2}\sum_{i=1}^{n}\left\{\frac{(\nu+p)r_i^2}{\nu_{1\cdot 2}\,\sigma_i^2 + r_i^2} - 1\right\}, \\
\frac{\partial L}{\partial \boldsymbol{\mu}_2} &= \frac{\sigma^2}{\nu_{1\cdot 2}}\sum_{i=1}^{n}\frac{1}{\sigma_i^2}\left\{1 - \frac{(\nu+p)r_i^2}{\nu_{1\cdot 2}\,\sigma_i^2 + r_i^2}\right\}\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{x}_i - \boldsymbol{\mu}_2).
\end{aligned}$$

We now introduce the operators vec and vech, such that for a $p \times p$ matrix $\mathbf{A}$, vec($\mathbf{A}$) returns the $p^2 \times 1$ vector obtained by writing the columns of $\mathbf{A}$ one below the other starting with the first, and vech($\mathbf{A}$) returns the $p(p+1)/2 \times 1$ vector obtained by returning only the lower triangular portion of $\mathbf{A}$. We also let $\mathbf{D}_p$ denote the $p(p+1)/2 \times p^2$ matrix of indicators that satisfies, for symmetric $\mathbf{A}$,

$$\mathbf{D}_p \operatorname{vec}(\mathbf{A}) = \operatorname{vech}(2\mathbf{A} - \operatorname{diag}\mathbf{A}).$$

For example,

$$\mathbf{D}_2 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

We can now differentiate L with respect to the distinct components of $\boldsymbol{\Sigma}_{22}$ by introducing the parameterization $\boldsymbol{\xi} = \text{vech}(\boldsymbol{\Sigma}_{22})$ and employing the results summarised by Rao (1985, Table 4). We obtain

$$\frac{\partial L}{\partial \boldsymbol{\xi}} = \frac{\sigma^2}{2\nu_{1\cdot 2}} \sum_{i=1}^{n} \frac{1}{\sigma_i^2} \left\{ 1 - \frac{(\nu + p) r_i^2}{\nu_{1\cdot 2}\, \sigma_i^2 + r_i^2} \right\} \mathbf{D}_{p-1} \,\text{vec} \left\{ \boldsymbol{\Sigma}_{22}^{-1} (\mathbf{x}_i - \boldsymbol{\mu}_2)(\mathbf{x}_i - \boldsymbol{\mu}_2)^T \boldsymbol{\Sigma}_{22}^{-1} \right\}.$$

Finally,

$$\frac{\partial L}{\partial \nu} = \frac{n}{2} \left\{ \psi\left(\frac{\nu + p}{2}\right) - \psi\left(\frac{\nu_{1\cdot 2}}{2}\right) - \frac{1}{\nu_{1\cdot 2}} \right\} - \frac{1}{2\nu_{1\cdot 2}} \sum_{i=1}^{n} \left(\frac{\sigma^2 - \sigma_i^2}{\sigma_i^2} \right)$$
$$- \frac{1}{2} \sum_{i=1}^{n} \log \left\{ 1 + \frac{r_i^2}{\nu_{1\cdot 2} \sigma_i^2} \right\} + \frac{(\nu + p)\sigma^2}{2\nu_{1\cdot 2}} \sum_{i=1}^{n} \frac{r_i^2}{\sigma_i^2 (\nu_{1\cdot 2}\, \sigma_i^2 + r_i^2)},$$

where $\psi(\alpha) = \dfrac{\partial}{\partial \alpha} \log \Gamma(\alpha)$, the so-called digamma function (Abramowitz and Stegun, 1964). In practice, this was calculated using an algorithm due to Bernardo (1976).

The following algorithm was devised for the calculation of $\mathbf{D}_p$: start with a $p(p+1)/2 \times p^2$ matrix of zeros. Then, for all (i, j) such that $1 \leq j \leq i \leq p$, set for each $r = (j-1)(p - j/2) + i$, $c_1 = (j-1)p + i$ and $c_2 = (i-1)p + j$, the elements (r, c_1) and (r, c_2) to one.

Estimated second derivatives were compared with analytical second derivatives for all terms except $\partial^2 L / \partial \nu^2$. However, only negligible differences were observed. The evaluation of $\partial^2 L / \partial \nu^2$ was not attempted as it was felt that additional 'precision' gained analytically would be offset by lack of precision in the evaluation of $\psi'(\alpha)$, the trigamma function.

References

Abramowitz, M. and Stegun, I. A. (Eds.). (1964). *Handbook of Mathematical Functions*, pp. 258–259. U.S. Department of Commerce.

Aitkin, M. (1987). Modelling variance heterogeneity in normal regression using GLIM. *Appl. Statist.*, **36**, 332–339.

Andrews, D. F., Bickel, P. J., Hampel, F. R., Huber, P. J., Rogers, W. H., and Tukey, J. W. (1972). *Robust Estimates of Location: Survey and Advances.* Princeton Univ. Press.

Andrews, D. F. (1974). A robust method for multiple linear regression. *Technometrics*, **16**, 523–531.

Bernardo, J. M. (1976). Psi (digamma) function. *Appl. Statist.*, **25**, 315–317.

Bickel, P. J. (1975). One-step Huber estimates in the linear model. *J. Amer. Statist. Assoc.*, **70**, 428–434.

Brownlee, K. A. (1965). *Statistical Theory and Methodology in Science and Engineering* (2nd edition). Wiley, New York.

Burridge, J. (1981). A note on maximum likelihood estimation for regression models using grouped data. *J. Roy. Statist. Soc. Ser. B*, **43**, 41–45.

Carroll, R. J. and Ruppert, D. (1988). *Transformation and weighting in regression.* Chapman and Hall, London.

Cheng, R. C. H. and Amin, N. A. K. (1983). Estimating parameters in continuous univariate distributions with a shifted origin. *J. Roy. Statist. Soc. Ser. B*, **45**, 394–403.

Cornish, E. A. (1954). The multivariate t distribution associated with a set of normal sample deviates. *Aust. J. Physics*, **7**, 531–542.

Davidian, M. and Carroll, R. J. (1987). Variance function estimation. *J. Amer. Statist. Assoc.*, **82**, 1079–1091.

DeGroot, M. H. (1970). *Optimal Statistical Decisions.* McGraw-Hill, USA.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the *EM* algorithm. *J. Roy. Statist. Soc. Ser. B*, **39**, 1–38.

Fang, K. T., Kotz, S., and Ng, K. W. (1990). *Symmetric Multivariate and Related Distributions.* Chapman and Hall, London.

Gabrielsen, G. (1982). On the unimodality of the likelihood for the Cauchy distribution: some comments. *Biometrika*, **69**, 677–678.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., and Stahel, W. A. (1986). *Robust Statistics: The Approach Based on Influence Functions.* Wiley, New York.

Hampel, F. R. (1974). The influence curve and its role in robust estimation. *J. Amer. Statist. Assoc.*, 69, 383–393.

Heitjan, D. F. (1989). Inference from grouped continuous data: a review. *Statistical Science*, **4**, 164–183.

Holland, P. W. and Welsch, R. E. (1977). Robust regression using iteratively reweighted least-squares. *Comm. Statist.-Theor. Meth.*, **7**, 813–827.

Huber, P. J. (1964). Robust estimation of a location parameter. *Ann. Math. Statist.*, **35**, 73–101.

Huber, P. J. (1973). Robust regression: Asymptotics, conjectures and Monte Carlo. *Ann. Statist.*, **1**, 799–821.

Huber, P. J. (1981). *Robust Statistics.* Wiley, New York.

Isaacson, E. and Keller, H. B. (1966). *Analysis of Numerical Methods.* Wiley, New York.

Kent, J. T. and Tyler, D. E. (1991). Redescending *M*-estimates of multivariate location and scatter. *Ann. Statist.*, **19**, 2102–2119.

Klein, R. and Yohai, V. J. (1981). Asymptotic behavior of iterative *M*-estimators for the linear model. *Comm. Stat.-Theor. Meth.*, **10**, 2373–2388.

Lange, K. L., Little, R. J. A., and Taylor, J. M. G. (1989). Robust statistical modelling using the *t* distribution. *J. Amer. Statist. Assoc.*, **84**, 881–896.

Mak, T. K. (1992). Estimation of parameters in heteroscedastic linear models. *J. Roy. Statist. Soc. Ser. B*, **54**, 649–655.

Mardia, K. V., Kent, J. T., and Bibby, J. M. (1979). *Multivariate Analysis.* Academic Press.

Maronna, R. and Morgenthaler, S. (1986). Robust regression through robust covariances. *Comm. Statist.-Theor. Meth.*, **15**, 1347–1365.

Maronna, R. A. and Yohai, V. J. (1981). Asymptotic behavior of general *M*-estimates for regression and scale with random carriers. *Z. Wahrscheinlichkeitstheorie verw. Gebiete*, **58**, 7–20.

Maronna, R. A. (1976). Robust *M*-estimators of multivariate location and scatter. *Ann. Statist.*, **4**, 51–67.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models* (2nd edition). Chapman and Hall, London.

Nair, V. N. and Pregibon, D. (1988). Analyzing dispersion effects from replicated factorials. *Technometrics*, **30**, 247–257.

Novovičová, J. (1990). *M*-estimators and gnostical estimators for identification of a regression model. *Automatica*, **26**, 607–610.

Pendergast, J. F. and Broffitt, J. D. (1985). Robust estimation in growth curve models. *Comm. Statist.-Theor. Meth.*, **14**, 1919–1939.

Rao, C. R. (1985). Matrix derivatives: Applications in Statistics. In Kotz, S., Johnson, N. L., and Read, C. B. (Eds.), *Encyclopedia of Statistical Sciences*, Vol. 5, pp. 320–325. Wiley, New York.

Rigby, R. A. and Stasinopoulos, M. D. (1994). An additive model for the dispersion parameter of a generalised model. In *Proceedings of the 9th International Workshop on Statistical Modelling.*

Rivest, L. P. (1989). De l'unicité des estimateurs robustes en régression lorsque le paramètre d'échelle et le paramètre de la régression sont estimés simultanément. *Can. J. Statist.*, **17**, 141–153.

Rousseeuw, P. J. (1984). Least median of squares regression. *J. Amer. Statist. Assoc.*, **79**, 871–880.

Ruppert, D. and Carroll, R. J. (1980). Trimmed least squares estimation in the linear model. *J. Amer. Statist. Assoc.*, **75**, 828–838.

Smith, R. L. (1985). Maximum likelihood estimation in a class of nonregular cases. *Biometrika*, **72**, 67–90.

Souvaine, D. L. and Steele, J. M. (1987). Time- and space-efficient algorithms for least-median of squares regression. *J. Amer. Statist. Assoc.*, **82**, 794–801.

Staudte, R. G. and Sheather, S. J. (1990). *Robust Estimation and Testing.* Wiley, New York.

Sutradhar, B. C. and Ali, M. M. (1986). Estimation of parameters of a regression model with a multivariate t error variable. *Comm. Statist.–Theor. Meth.*, **15**, 429–450.

Titterington, D. M., Smith, A. F. M., and Makov, U. E. (1985). *Statistical Analysis of Finite Mixture Distributions.* Wiley, New York.

Titterington, D. M. (1985). Comment on "Estimating parameters in continuous univariate distributions". *J. Roy. Statist. Soc. Ser. B*, **47**, 115–116.

Towers, D. A. (1988). *Guide to Linear Algebra.* Macmillan Education, Basingstoke.

Tyler, D. E. (1988). Some results on the existence, uniqueness, and computation of the M-estimates of multivariate location and scatter. *SIAM J. Sci. Stat. Comput.*, **9**, 354–362.

Zeidler, E. (1986). *Nonlinear functional analysis and its applications - I: Fixed-Point Theorems, II/B: Nonlinear Monotone Operators.* Springer-Verlag, London.

Zellner, A. (1976). Bayesian and non-Bayesian analysis of the regression model with multivariate Student-t error terms. *J. Amer. Statist. Assoc.*, **71**, 400–405.