Robust recursive L_p estimation

D.W. McMichael, MA, DPhil

Indexing terms: Adaptive control, Optimisation, Algorithms, Power transmission and distribution

Abstract: Outlier-contaminated normal errors in regression problems are modelled by exponential power distributions and the resulting maximum likelihood estimators are shown to involve L_p minimisations ($1 < p \leqslant 2$). It is shown that L_{1+} estimation is minimax outlier-robust and minimax covariance-robust over the neighbourhood of exponential power distributions. Efficiency loss is negligible. Recursive gradient-type L_p estimators are derived and shown to be convergent and consistent. The major limitation on outlier robustness is seen to be the requirement for convergence of the recursive minimisation. The algorithm is validated with an application in adaptive control.

1 Introduction

Since Box coined the term, *robust* statistical procedures have received steadily increasing attention [6]. However, the design of recursive robust methods for multi-parameter regression has largely been side-tracked (with the notable exceptions of Poljak and Tsypkin [26, 27]). It has mainly been left to practising engineers to invent *ad hoc* methods for outlier rejection. These tend to involve ignoring cases with large prediction errors [32], and this has led to the proliferation of techniques which may not converge, are highly sensitive to the dead-banding thresholds, and may have poor tracking performance [14]. The aim of this paper is to present a family of recursive estimators, which are insensitive to the presence of outliers in the data, and yet are not significantly less efficient than conventional least squares. Underlying the discussion is the practical concern that the minimisation problems implied by a robust estimation criterion have to be solved by recursive gradient-type methods. This is a significant limitation.

1.1 Outliers

In data sets resulting from physical experiments it is not uncommon to find that some of the cases seem rather unlikely, and are inconsistent with the bulk of the remaining data. Detecting these outliers in multivariate regressions is not a simple problem, there is a significant body of literature concerned with methods for doing so [4, 8]. The alternative approach to outlier detection is to design estimation procedures that are outlier-robust: i.e. the estimates should not change much when outliers are introduced into the data. However, the key problem in designing robust statistical procedures is ensuring that they are reasonably efficient, even when outliers are not present.

We shall be concerned with the problem of estimating θ_0 in the linear-in-the-parameters model

$$y = x^T\theta_0 + \xi \tag{1}$$

where y is a physical measurement, x is a $n \times 1$ vector of regressors, θ_0 is a $n \times 1$ vector of parameters and ξ is a random error term. Conventional least-squares (LS) regression minimises the sum of squares of the fitting errors, and is the maximum likelihood estimator (MLE) when the measurement errors are normally distributed (Fig. 1). As can be seen, the probability of occurrence of normally distributed errors larger than $3 \times$ [standard deviation (σ)] is negligible. However, if outliers are present, the error distribution might look more like that in Fig. 2, where the normal errors have been 10% contaminated by Cauchy-distributed errors to form a

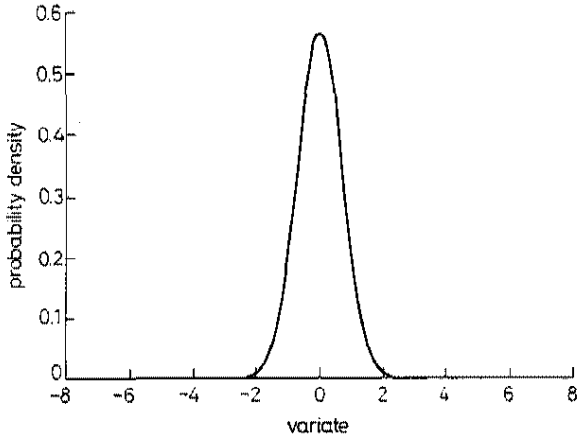

Fig. 1 *Normal probability density ($\sigma = 1/\sqrt{2}$)*

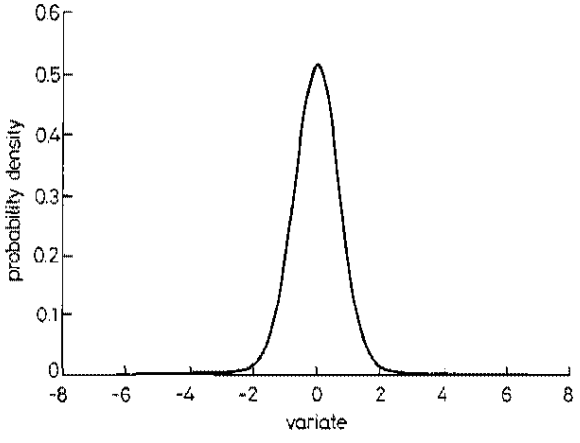

Fig. 2 *10% Cauchy-contaminated normal density*

Paper 7113D (C8), first received 28th March and in revised form 19th September 1989

The author is with the Control Systems Centre, University of Manchester Institute of Science and Technology, PO Box 88, Manchester M60 1QD, United Kingdom

mixture distribution with significantly longer and fatter tails. The outlier-contaminated error distribution models used here are symmetric, on the reasonable assumption that the sign of the error of the outliers is not known *a priori*. Barnet and Lewis [4] discuss the physical significance of outliers and the reasons why they occur.

1.2 Overview

The starting point for this design is finding a neighbourhood of error distribution models that can model variable levels of outlier contamination in normal data. A suitable class is the exponential power distributions, the maximum likelihood estimators for which are shown to involve L_p minimisations ($1 < p \leqslant 2$). Offline calculation of L_p estimates using the iteratively reweighted least-squares algorithm provides a direct analytical link to ordinary least-squares estimation, and is used to analyse the effect of individual outliers on L_p estimates. On maximum likelihood assumptions L_p estimation is shown to be consistent, asymptotically efficient, sufficient and normal. The robustness properties of L_p estimation are examined both in terms of sensitivity to single outliers, and asymptotic covariance of the estimates in the off-design case when the assumed distribution is wrong. Both single-case and asymptotic approaches to robustness indicate that p should be small (close to 1) and it is shown that $p = 1^+$ is the minimax choice according to both criteria. In this application, minimax robustification is not marked by a significant loss in efficiency if the errors are actually normal. Steepest decent and stochastic Newton recursive minimisation algorithms are shown to be consistent when the L_p criterion is slightly modified.

1.3 Previous work

After Box's comments on the vulnerability of classical tests on variance to non-normal samples [6], development of robust statistical methods was rather slow. Tukey [33] demonstrated the sensitivity of a range of location and scale estimators to failure of normality assumptions. Huber's paper [16] on the robust estimation of a location parameter was a classic, and marked the first successful use of minimax techniques in robust estimator design. Rejection rules for detecting and deleting outliers have long been available but the subject was put on firm theoretical foundations by Hampel [12] with the introduction of the influence curve as a measure of the effect on the estimates of a very general class of perturbations to the data. Large numbers of practical outlier detection methods have been proposed (for example, Cook [9], and Andrews and Pregibon [3]). Both residual analysis and the case influence curve are reviewed by Cook and Weisberg [8].

Since the early work of Huber and Hampel, a large number of MLE based (M-estimators) have been proposed: some based on common long-tailed distributions (Huber [15], and Poljak and Tsypkin [26]), and some on totally artificial distributions with redescending score functions; for example, Andrew's sine [2] and Tukey's biweight [1].

The robust recursive regression problem has received attention directly from Poljak and Tsypkin [26, 27] who discuss estimation critiera and give a preliminary treatment of the convergence of steepest descent algorithms. Interest has also been shown by those seeking to robustify Kalman filters: Masreliez and Martin [24] use minimax techniques to design synthetic score functions to reduce the effects of outliers both in the measurements and in the process noise. Their methods were criticised by Tsai and Kurz [31], mainly because its outlier robustness is strongly dependent on *a priori* knowledge of the level of contamination. But they themselves did not address the question of the efficiency-robustness tradeoff.

2 An outlier distribution model

Models for outlier-contaminated error distributions generally assume that the errors arise from two sources: most cases are affected by errors that are the sum of many small random disturbances, but some are affected by large errors which are sums of comparatively few large disturbances. Clearly, the former type are likely to be distributed in close approximation to normality, while the latter are not. This line of reasoning has frequently led to modelling with mixtures of normal and some long-tailed distribution (typically Cauchy, or a large variance normal distribution). The family of exponential power distributions $E_p(p, a)$, proposed here, are not of the mixture variety. They enable different degrees of outlier contamination to be modelled by choice of the parameter p ($1 < p \leqslant 2$) [25]. The net result is similar: they lie in a perturbation neighbourhood that includes the normal distribution, and with that exception are all leptokurtic. The exponential power distribution's probability density function (PDF) is

$$p_\xi(\xi) = \frac{p e^{-|\xi/a|^p}}{2a\Gamma(1/p)} \tag{2}$$

a is a scale (dispersion) parameter >0, and p is a dimensionless parameter affecting the kurtosis of the distribution. Fig. 3 shows the family of these distributions with

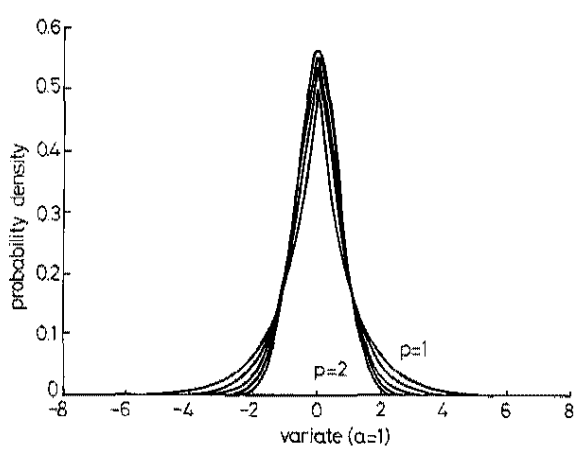

Fig. 3 *Exponential power PDFs for $p = 1, 1.25, 1.5, 1.75, 2$ ($a = 1$)*

$a = 1$ and p taking five equally spaced values between 1 and 2. When $p = 2$ the distribution is normal with variance $1/\sqrt{2}$, and when $p = 1$ the distribution is Laplacian. Its variance is

$$\text{var}\,(\xi) = a^2 \frac{\Gamma(3/p)}{\Gamma(1/p)} \tag{3}$$

As p decreases from 2 to 1, the distribution develops longer thicker tails, and, as a consequence, outlying errors are more probable. Underlying the concept of the outlier is rarity, they are 'surprising' and improbable data points: by extending the tails of the model distribution we are admitting that they are more likely.

3 MLEs based on exponential power distributions

The maximum likelihood estimator for a distribution parameter chooses estimates which maximise the probability (likelihood) of the data. The ith case of the linear model eqn. 1 is

$$y_i = \boldsymbol{x}_i^T \boldsymbol{\theta}_0 + \xi_i \tag{4}$$

The PDF of ξ_i is $p_i(\xi_i)$, and, because the error is additive,

$$p_{y_i}(y_i | \boldsymbol{\theta}_0) = p_{\xi_i}(y_i - \boldsymbol{x}_i^T \boldsymbol{\theta}_0) \tag{5}$$

If the ξ_i are mutually independent and identically distributed, the joint probability density of the observations y_i is

$$p(y_1, \ldots, y_N | \boldsymbol{\theta}_0) = \prod_{i=1}^{N} p_{\xi_i}(y_i - \boldsymbol{x}_i^T \boldsymbol{\theta}_0) \tag{6}$$

If the left-hand side of eqn. 6 is regarded as a function of $\boldsymbol{\theta}$, then it is interpreted as a likelihood function:

$$L(y_1, \ldots, y_N | \boldsymbol{\theta}) = \prod_{i=1}^{N} p_{\xi_i}(y_i - \boldsymbol{x}_i^T \boldsymbol{\theta}) \tag{7}$$

The MLE for $\boldsymbol{\theta}$ maximises the likelihood function over $\boldsymbol{\theta}$, giving

$$\hat{\boldsymbol{\theta}} = \arg \max_{\boldsymbol{\theta}} L(y_1, \ldots, y_N | \boldsymbol{\theta}) \tag{8}$$

Aside from the intuitive appeal of MLEs they possess a number of optimum properties. However, for exponential power distributions, when $p \neq 2$, $\hat{\boldsymbol{\theta}}$ is not a sufficient statistic for $\boldsymbol{\theta}_0$, but the asymptotic properties of the MLE criterion remain. As $N \rightarrow \infty$, the estimator is consistent ($\boldsymbol{\theta} \rightarrow \boldsymbol{\theta}_0$, with probability 1), sufficient, normal and attains the Cramer-Rao minimum-variance bound (MVB) [17].

3.1 Offline solutions of the MLE estimation problem

It is usually easier to maximise ln (L) rather than L itself, particularly if the error distribution is exponential. This approach results in the following minimisation problem:

$$\hat{\boldsymbol{\theta}} = \arg \min_{\boldsymbol{\theta}} \sum_{i=1}^{N} \phi(y_i - \boldsymbol{x}_i^T \boldsymbol{\theta}) \tag{9a}$$

$$\phi(x) \propto -\ln p_\xi(x) + \text{constant} \tag{9b}$$

$\phi(\cdot)$ is known as the case cost function (CCF). For exponential power distributions the MLE for $\boldsymbol{\theta}_0$ is

$$\hat{\boldsymbol{\theta}} = \arg \min_{\boldsymbol{\theta}} \sum_{i=1}^{N} | y_i - \boldsymbol{x}_i^T \boldsymbol{\theta} |^p \tag{10}$$

Notice that this is an L_p normed minimisation problem, and also that $\hat{\boldsymbol{\theta}}$ is independent of the scale parameter a. The case cost function and its derivative (the score function), for L_p estimators, are plotted in Figs. 4 and 5, for several values of p.

To prove asymptotic results using the methods of Kendall and Stuart [14] requires the existence of the first derivative of the CCF and the existence of the expectation of the second derivative. Unfortunately, when $p \neq 2$ $\phi(0)'$ does not formally exist, but this difficulty is overcome by augmenting $\phi(x)'$ by the point at zero, so that $\phi(0)' = 0$. The second derivative of $\phi(x)$ is infinite at $x = 0$, but its expectation does exist and the conventional proofs of the optimal asymptotic properties of MLEs apply.

When $p = 2$, eqn. 10 has the least-squares analytical solution (the normal equations):

$$\hat{\boldsymbol{\theta}} = [\boldsymbol{X}^T \boldsymbol{X}]^{-1} \boldsymbol{X}^T \boldsymbol{Y} \tag{11a}$$

$$\boldsymbol{X}^T = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N]^T \tag{11b}$$

$$\boldsymbol{Y}^T = [y_1, \ldots, y_N]^T \tag{11c}$$

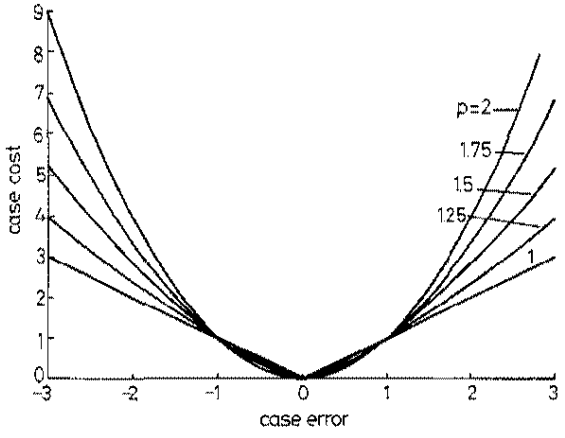

Fig. 4 *Case cost functions for L_p estimators*

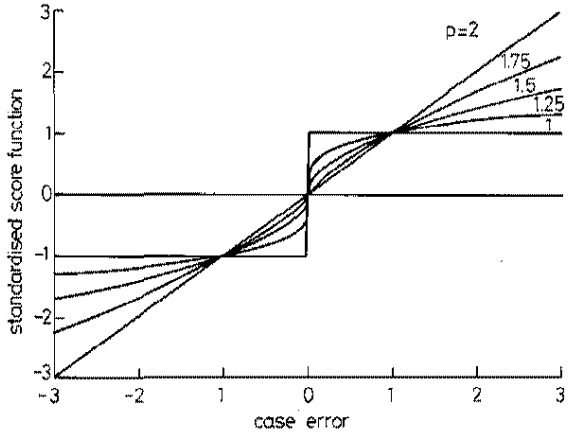

Fig. 5 *Standardised score functions for L_p estimators*

However, when $p \neq 2$, iterative methods are required (see Press *et al.* [29]). Insight into the process of robustification can be obtained from the numerically weak method of iteratively reweighted least squares (IRLS) mentioned by Box and Draper [7] and equivalent to the recursive algorithm described by Dutter [10]. (The technique is closely related to a method first proposed by Glaisher [11].) Dutter [10] has shown that, at each iteration, the method consistently reduces the estimator cost in eqn. 9a provided that $\phi(x)$ is convex and symmetric, $\phi'(x)/x$ is bounded and monotone decreasing for $x < 0$, and the minimum has not already been reached. The iterative scheme is as follows:

$$\hat{\boldsymbol{\theta}}_{(k)} = [\boldsymbol{X}^T \boldsymbol{W}_{(k-1)} \boldsymbol{X}]^{-1} \boldsymbol{X}^T \boldsymbol{W}_{(k-1)} \boldsymbol{Y} \tag{12a}$$

$$w_{(k)ij} = \begin{cases} \phi'(\varepsilon_{(k)i})/\varepsilon_{(k)i} & i = j \\ 0 & i \neq j \end{cases} \tag{12b}$$

$$\varepsilon_{(k)i} = y_i - \boldsymbol{x}_i^T \hat{\boldsymbol{\theta}}_{(k)} \tag{12c}$$

$\boldsymbol{W}_{(k)}$ is a diagonal weighting matrix which, when $\phi(x)$ is weaker than x^2, progressively penalises those cases with

large residuals ($\varepsilon_{(k)i}$). For this algorithm to be consistent, it may be necessary to modify $\phi(x')/x$ near the origin, to prevent unboundedness. It is revealing to compare eqn. 12*a* with the Markov estimator for uncorrelated data, in which the diagonal matrix W is the inverse of covariance matrix of the error vector:

$$[\xi_1, \ldots, \xi_N]^T$$

The Markov estimator weights cases in inverse proportion the corresponding error variance. Robust regression can be interpreted as Markov estimation with iterative estimation of var (ξ_i). We shall see later that this approach to solution of eqn. 10 is instrumental in estimating the influence of single outlying cases on the estimates.

3.2 An example

Before pursuing the analysis, we will examine, by example, the relative outlier sensitivity of L_1 and least-squares estimators. There are two sets of data, both consisting of ten cases, each perturbed by normally distributed errors, but, in the first set (Figs. 6*a* and 6*b*), the tenth case has an additional disturbance of -10. Positioning the outlier at this location maximises its leverage [8]. The second data set (Figs. 6*c* and 6*d*) contained no deliberate outliers and was generated by the equation

$$y_i = x_i + \xi_i \tag{13}$$

with ξ_i distributed as $N(0, 1)$, while the first set were generated by

$$y_i = x_i + \xi_i - 10\delta_{i,10} \tag{14}$$

where $\delta_{i,j}$ is the Kronecker delta. In each regression, the data were fitted to

$$y_i = \hat{c} + \hat{m}x_i \tag{15}$$

The plots in Fig. 6. show the data points (y_i), the deterministic relationship $y = x$ (solid line), and the fitted line $y = \hat{c} + \hat{m}x$ (broken line). By contrasting the fitted lined of Figs. 6*a* and 6*b*, it is clear that the L_1 regression is substantially less affected by the presence of the outlier than LS. However, of equal importance is the fact that, when no outliers are present, in Figs. 6*c* and 6*d*, the LS and L_1 regression lines are almost identical. This is good *prima facie* evidence that the efficiency loss involved in this form of robustisation may not be significant.

4 Robustness properties of L_p estimation

L_p estimation has been shown to be optimal in several senses, provided the error distribution is $E_p(p, a)$. We now go beyond the heuristic arguments of Section 1 and show that these distributions lead to MLEs that are less sensitive to outliers than least squares. Secondly, we shall examine the asymptotic behaviour of the $L_{\tilde{p}}$ estimator when the error distribution is $E_p(p, a)$ and p is not necessarily equal to $\tilde{p}$.

4.1 Influence of single outliers

Originated by Hampel [12] the influence curve can be used to measure of the differential effect of a single case

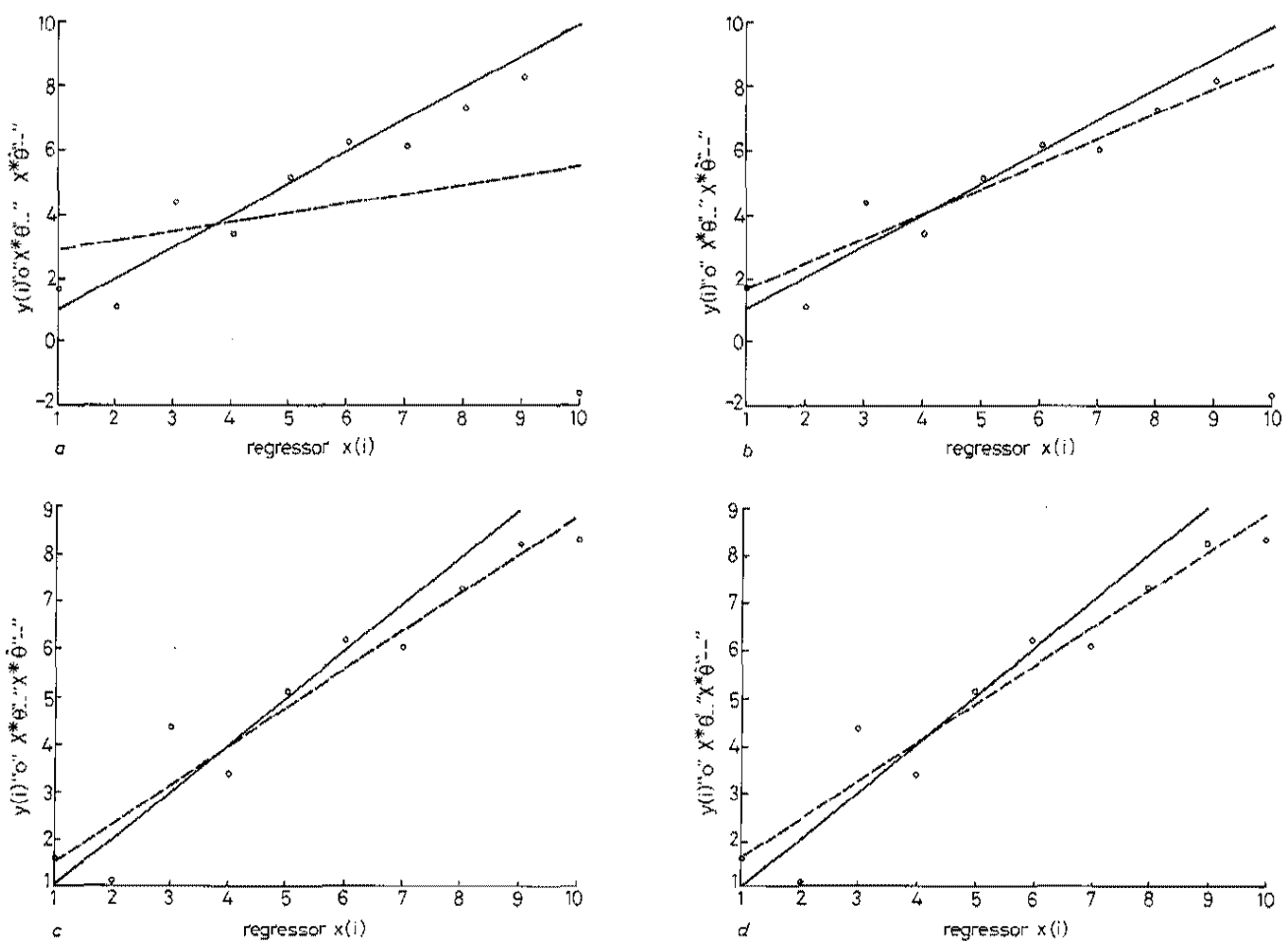

Fig. 6 *Example: Comparison of least squares and L_1 estimators*

a Least-squares linear regression (with an outlier) $p = 2$
b Robust linear regression (with an outlier) $p = 1$
c Least-squares linear regression (no outliers) $p = 2$
d Robust linear regression (no outliers) $p = 1$

on a general statistic. Although fundamental to the development of robust estimation, the mathematical diversion it introduces is unwarranted. In this case, more useful and clearly motivated results can be found by finding an expression for the change in the estimates when an outlier is added to the sample and given weight w_i, with all other cases weighted by 1. For weighted least squares, the new vector of estimates is

$$\hat{\theta}_i = \hat{\theta}_{(i)} + \frac{w_i[X_{(i)}^T X_{(i)}]^{-1} x_i \varepsilon_i}{1 + w_i x_i^T [X_{(i)}^T X_{(i)}]^{-1} x_i} \tag{16}$$

where the prediction error ε_i is

$$\varepsilon_i = y_i - x_i^T \hat{\theta}_{(i)} \tag{17}$$

and the suffix (i) indicates that the relevant quantity is assembled without the ith case. The proof of this is a direct consequence of the matrix inversion lemma. Thus, if $w_i x_i^T [X_{(i)}^T X_{(i)}]^{-1} x_i \ll 1$, then $\Delta\hat{\theta}_i := \hat{\theta}_i - \hat{\theta}_{(i)}$ (the sample influence curve (SIC) [23, 28]) is proportional to $w_i[X_{(i)}^T X_{(i)}]^{-1} x_i \varepsilon_i$. $\Delta\hat{\theta}_i$ can be measured by various quadratic norms and this leads directly to a class of outlier detection procedures [8]. The case is influential if x_i, ε_i, or w_i are large or $[X_{(i)}^T X_{(i)}]^{-1}$ is large in x_i direction (i.e. sample is small, or the previous x_j values have been small or weakly correlated with x_i). Thus, the relative influence of an outlier in weighted least squares is dependent on its leverage [13]:

$$\frac{w_i x_i^T [X_{(i)}^T W_{(i)} X_{(i)}]^{-1} x_i}{1 + w_i x_i^T [X_{(i)}^T W_{(i)} X_{(i)}]^{-1} x_i} \tag{18}$$

The same analysis can be used to find a simple approximate expression for the sample influence curve in non-LS estimation. The fully converged IRLS robust estimator for θ, based on the first $i-1$ cases, is

$$\hat{\theta}_{(i)} = [X_{(i)}^T W_{(i)} X_{(i)}]^{-1} X_{(i)}^T W_{(i)} Y_{(i)} \tag{19}$$

When the ith case is added and the estimator is outlier-robust, the first $i-1$ diagonal elements of the new fully iterated weighting matrix will tend not to be very different from the corresponding elements of $W_{(i)}$. We shall assume that they are the same. Furthermore, we shall only iterate w_{ii} once, giving the new weighting matrix:

$$W \simeq \begin{bmatrix} W_{(i)} & 0 \\ 0 & \dfrac{\phi'(\varepsilon_i)}{\varepsilon_i} \end{bmatrix} \tag{20}$$

For unmodified L_p estimation, this gives the sample influence curve (on dividing through all cases by p) as

$$\frac{[X_{(i)}^T W_{(i)} X_{(i)}]^{-1} x_i |\varepsilon_i|^{p-1} \operatorname{sign}(\varepsilon_i)}{1 + |\varepsilon_i|^{p-2} x_i^T [X_{(i)}^T W_{(i)} X_{(i)}]^{-1} x_i} \tag{21}$$

For large prediction errors, and $1 < p < 2$, the numerator of the L_p sample influence curve is smaller than that of least squares. Provided the case leverage is small (i.e. the denominator of eqn. 21 is not much greater than 1), the denominator is insensitive to large prediction errors ε_i. Consequently, L_p estimators are more outlier-robust than least squares. [A nominal upper bound has to be put on $\phi(x)/x(|\varepsilon_i|^{p-2})$ in the IRLS procedure, to satisfy the conditions for convergence.]

4.2 Asymptotic robustness properties

A second approach to analysing the robustness properties of estimators is to examine their asymptotic behaviour: principally, bias and covariance. Provided the score function is odd and the actual error distribution is symmetric, the estimator eqn. 9 will be unbiased if the expectation

$$E\left[\frac{\partial \ln \tilde{L}}{\partial \theta}\right]_{\theta=\theta_0} \tag{22}$$

exists. $\tilde{L}$ is the likelihood function corresponding to the assumed error distribution. To emphasise the fact that p can be viewed both as a parameter of the error distribution and as a design parameter, we shall denote the parameter of the error distribution $E_p(p, a)$ by 'p', and the design parameter in the minimisation problem eqn. 10 by '$\tilde{p}$'. In $L_{\tilde{p}}$ estimation, the condition 22 corresponds to a requirement for the existence of the $(\tilde{p}-1)$th moment of the error distribution ($1 < \tilde{p} \leqslant 2$): always satisfied by real data.

If, by chance, $\tilde{p} = p$, the asymptotic covariance reaches the minimum bound. A little integration shows that this is

$$\lim_{N\to\infty} \left\{ -\left[E\left(\frac{\partial^2 \ln L}{\partial \theta^2} \right) \right]^{-1} \right\} = \frac{a^2 \Gamma(1/p)}{p^2 \Gamma(2 - 1/p)} \lim_{N\to\infty} [X^T X]^{-1} \tag{23}$$

This is the lowest covariance obtainable when the errors are distributed as $E_p(p, a)$.

4.2.1 Asymptotic covariance of L_p estimates: We now examine the behaviour of $L_{\tilde{p}}$ estimation in the neighbourhood of exponential power distributions ($1 \leqslant p \leqslant 2$). It is thereby possible to analyse the ability of the $L_{\tilde{p}}$ estimator handle data with error distributions with varying proportional large errors. Expanding $\partial \ln \tilde{L}/\partial\theta$ in a zeroth-order Taylor series with error term, taking expectations of variances and repeated application of the strong law of large numbers leads to

$$\lim_{N\to\infty} \{\operatorname{cov}(\hat{\theta})\} = \frac{a^2 \Gamma(1/p)\Gamma((2\tilde{p}-1)/p)}{p^2 \Gamma((\tilde{p}-1)/p + 1)^2} \lim_{N\to\infty} [X^T X]^{-1} \tag{24}$$

The scalar factor of this expression (the asymptotic covariance factor) at $a = 1$ is plotted in Fig. 7, as a function of $\tilde{p}$ for various values of p, and in Fig. 8 as a function of p for various values of $\tilde{p}$. When plotted against $\tilde{p}$ notice that the minima of all the curves lie at $\tilde{p} = p$, that the highest variance occurs uniformly at $p = 1$ and the minimum at $p = 2$. When the noise is assumed nor-

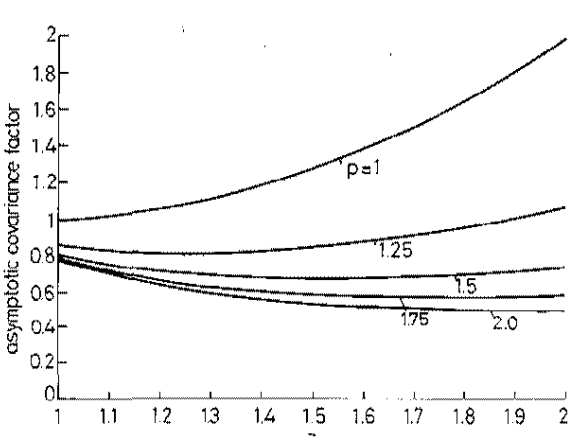

Fig. 7 *L_p asymptotic covariance factor against $\tilde{p}$, for various p*

mally distributed ($\tilde{p} = 2$), there is a substantial loss of efficiency if in fact $p = 1$: and outliers are present. However, if the estimator is chosen conservatively ($\tilde{p} = 1$), the

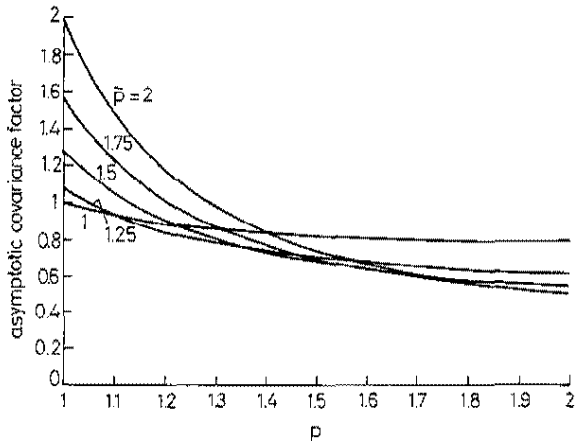

Fig. 8 *$L_{\tilde{p}}$ asymptotic-covariance factor against p, for various $\tilde{p}$*

increase in the variance over least squares is small, but the reduction in variance if $p = 1$ is large. These are just the properties required of a robust procedure: outlier insensitivity for a small loss in efficacy at the null case (normal errors, $p = 2$). The net result is that, as $\tilde{p}$ decreases, from 2 to 1 the sensitivity of the covariance of the estimates with respect to p decreases (see Fig. 7). This, in itself, is a useful property, because it suggests that,

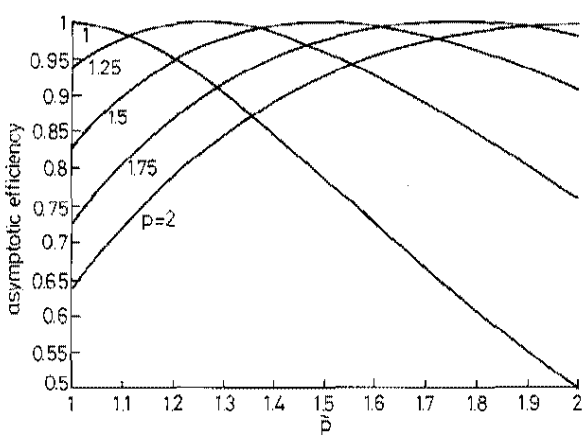

Fig. 9 *$L_{\tilde{p}}$ asymptotic efficiency against $\tilde{p}$, for various p*

when $\tilde{p}$ is close to 1, estimation of the covariance matrix may also be distribution-robust.

4.2.2 Asymptotic efficiency of $L_{\tilde{p}}$ estimates: In the established statistical literature [17], it is usual to measure the asymptotic variance of an estimator relative to the minimum-variance bound. For asymptotically normal estimators, the asymptotic efficiency is simply the ratio of the minimum-variance bound to the estimator's variance. As the scale and matrix factors of both are equal, the appropriate multivariate extension is to define the efficiency as the ratio of the scalar factor of eqn. 23 to that of eqn. 24

$$\eta = \frac{\Gamma((\tilde{p}-1)/p + 1)^2}{\Gamma(2 - 1/p)\Gamma((2\tilde{p}-1)/p)} \tag{25}$$

Plots of η against $\tilde{p}$ for various values of p are shown in Fig. 9. All the curves are concave and reach a maximum of 1 at $\tilde{p} = p$. The loss of efficiency, by assuming that the data are normally distributed, when in fact they are Laplacian, is greater than in the reverse situation. However, the benefits of underestimating p are masked by the fact that, as p decreases from 2 to 1, the minimum-variance bound increases. Asymptotic efficiency is therefore a less useful design tool than asymptotic covariance (Fig. 7).

4.3 Design: choosing $\tilde{p}$

In this Section, the chosen neighbourhood of error distributions, in which the robustness properties of $L_{\tilde{p}}$ estimation were analysed, is defended, and minimax design procedures are used to show that L_{1+} estimation is optimally robust, and realisable. In both IRLS and the recursive algorithms to be given in Section 5, it is required that the case cost function $\phi(x)$ be convex. Within the outlier-prone range of the exponential power distributions ($\tilde{p} < 2$), this condition requires that $\tilde{p} > 1$. As the motivation for $L_{\tilde{p}}$ estimation is the maximisation of the likelihood function, attention will be restricted to exponential power error distributions for which $1 \leqslant p \leqslant 2$.

In the absence of definite information about the error distribution, a sensible approach to design is to optimise the procedure for the 'worst case' within Ω, a nominated neighbourhood of possible perturbations. This argument is formalised by the minimax design procedure. If $C(\boldsymbol{Y}, \tilde{p}, \Omega)$ is a measure of estimator degradation due to perturbation of the data from the null model (normality), then the minimax optimal choice of $\tilde{p}$ is

$$\tilde{p} = \arg\min_{x} \left\{ \sup_{\Omega} C(\boldsymbol{Y}, x, \Omega) \right\} \tag{26}$$

The minimax criterion is conservative and may result in safeguarding the estimator from unlikely worst cases. To obtain sensible estimators, it may be necessary to choose Ω rather carefully. Minimax robust deisgns may themselves be unrobust to this choice. Happily, neither of these criticisms applies to the designs below:

(*a*) *Minimax SIC design:* The approximate sample influence curve expr. 21 is a measure of the effect of a single outlying case on the estimates in which the case error ξ_i is large. We seek to minimise this degradation. An appropriate norm $C(\boldsymbol{Y}, \tilde{p}, \xi_i)$ is the magnitude of the approximate SIC, the product of a scalar, and a vector independent of the case error ξ_i. Hence

$$C(\boldsymbol{Y}, \tilde{p}, \xi_i) \propto \frac{|\varepsilon_i|^{\tilde{p}-1}}{1 + |\varepsilon_i|^{\tilde{p}-2} \boldsymbol{x}_i^T [\boldsymbol{X}_{(i)}^T \boldsymbol{W}_{(i)} \boldsymbol{X}_{(i)}]^{-1} \boldsymbol{x}_i} \tag{27}$$

Where ξ_i is large, ε_i behaves as ξ_i, and $C(\boldsymbol{Y}, \tilde{p}, \xi_i)$ is always maximised in Ω by $\xi_i \to \infty$. It is minimised by setting $\tilde{p} = 1^+$. L_{1+} is therefore the minimax realisable SIC robust estimator.

(*b*) *Minimax asymptotic covariance design:* $L_{\tilde{p}}$ estimation is unbiased under the assumption that the error distribution is symmetric, so the asymptotic property of interest is covariance. For the reasons stated, we choose the perturbation neighbourhood Ω as $\{1 < p \leqslant 2\}$ and choose $C(\boldsymbol{Y}, \tilde{p}, \Omega)$ to be the asymptotic covariance factor:

$$C(\boldsymbol{Y}, \tilde{p}, \Omega) = a^2 \frac{\Gamma(1/p)\Gamma((2\tilde{p}-1)/p)}{p^2 \Gamma((\tilde{p}-1)/p + 1)^2} \tag{28}$$

Reference to Fig. 7 shows that, regardless of $\tilde{p}$, its highest value occurs uniformly when $p = 1$. Choosing $\tilde{p} = 1$ minimises this worst case variance. Thus, the minimax realisable covariance design also chooses $\tilde{p} = 1^+$.

As mentioned previously, there is little loss in efficacy incurred by using L_{1+} estimation, when the noise is

Gaussian, and the general criticism of minimax methods for being too conservative is not justified in this case. Furthermore, the lower boundary of Ω is not arbitrary.

5 Recursive L_p estimation

By recursive estimation, we mean numerical procedures that enable sequential cases to be included into the estimator one at a time, in such a way as the computational effort required is less than recalculating the estimates with offline techniques each time a new case arrives. The net result is usually that the cost minimisation is less accurate (recursive least squares is a notable exception). Recursive techniques are appropriate when the case data arrive serially and the estimates are needed in real time; for example, in self-tuning control. When there is a substantial cost associated with gathering each new case, sequential estimation procedures become attractive. Recursive procedures are particularly suited to parameter tracking problems where adaptivity is required [19].

Two sorts of gradient-type methods are given here: a stochastic Newton method [18] and a slower, but computationally less demanding, group of steepest descent stochastic approximation methods [30, 34]. Stochastic approximation finds the solution of the general equation:

$$\underset{y}{E}\,[g_i(\theta, y)] = 0 \tag{29}$$

with the iterative scheme:

$$\hat{\theta}(t) = \hat{\theta}(t-1) + K(t)g(\hat{\theta}(t-1), y(t)) \tag{30}$$

If the noise distribution and the case cost function $\phi(x)$ are symmetric, minimising

$$V(\theta) = E[\phi(y_i - x_i^T\theta)] \tag{31}$$

is asymptotically equivalent to the minimisation eqn. 9*a*. If $V(\theta)$ is convex, then we seek solutions to

$$\frac{\partial V(\theta)}{\partial \theta} = E[\phi'(\varepsilon_i(\theta))x_i] = 0 \tag{32}$$

For the stochastic Newton method, we also require an estimate of the Hessian of $V(\theta)$. Defining

$$\frac{\partial^2 V(\theta)}{\partial \theta^2} := \boldsymbol{R}(\theta) = E[\phi''(\varepsilon_i(\theta))x_i x_i^T] \tag{33}$$

we can construct a problem analogous to eqn. 29:

$$E[\phi''(\varepsilon_i(\theta))x_i x_i^T - \boldsymbol{R}(\theta)] = 0 \tag{34}$$

Eqns. 32 and 34 may be solved in parallel by the following recursive prediction error estimation algorithm:

$$\hat{\theta}(t) = \hat{\theta}(t-1) + \frac{\tilde{\boldsymbol{R}}(t)^{-1}}{t}x(t)\phi'(\varepsilon(t)) \tag{35}$$

$$\tilde{\boldsymbol{R}}(t) = \tilde{\boldsymbol{R}}(t-1) + \frac{1}{t}[\phi''(\varepsilon(t))x(t)x^T(t) - \tilde{\boldsymbol{R}}(t-1)] \tag{36}$$

$$\varepsilon(t) = y(t) - x(t)^T\hat{\theta}(t-1) \tag{37}$$

Note that $\tilde{\boldsymbol{R}}(t)$ is only an approximation for $\boldsymbol{R}(\hat{\theta}(t-1))$. It is of some practical importance that this scheme can be realised by a slight modification of the recursive least-squares algorithm:

$$\hat{\theta}(t) = \hat{\theta}(t-1) + \boldsymbol{P}(t)x(t)\phi'(\varepsilon(t-1)) \tag{38}$$

$$\boldsymbol{P}(t) = \frac{\boldsymbol{P}(t-1)}{\lambda(t)} \times \left[I - \frac{x(t)x(t)^T\boldsymbol{P}(t-1)}{\lambda(t)/\phi''(\varepsilon(t)) + x(t)^T\boldsymbol{P}(t-1)x(t)}\right] \tag{39}$$

In this realisation, $\boldsymbol{P}(t) = [\tilde{\boldsymbol{R}}(t)]^{-1}$ and $\lambda(t)$ is a forgetting factor ($0 < \lambda(t) \leqslant 1$). If $\lambda(t) < 1$, then the estimates will be able to track slowly varying parameters. The update eqn. 39 can be implemented accurately, efficiently and stably using Thornton and Bierman's $\boldsymbol{UDU}^T$ factorisation algorithm (see Bierman [5]). Notice that $\phi''(\varepsilon(t))$ must be bounded. This is not true for the L_{β} case cost function at $\varepsilon(t) = 0$, so the L_{β} case cost (or at least its second derivative) must be modified near the origin. A modification reminiscent of Huber's M-estimator (Huber [16]) is to introduce a quadratic section near the origin:

$$\phi_{pm}(x) = \begin{cases} \left(|x| - c\left(1 - \frac{p}{2}\right)\right)^p & |x| \geqslant c \\ x^2\left(\frac{p}{2}\right)^p c^{p-2} & |x| < c \end{cases} \tag{40}$$

which is continuous and its first derivative exists everywhere. c should be chosen to be the same order of magnitude as the noise variance. Its value is not critical, and becomes less so as p increases.

The time update requires $n(n+1)/2$ single precision storage locations and performs $(n-1)(n-2)$ divisions per update. If computational capacity is limiting, scalar gain steepest-decent methods are appropriate. Sensible choices for $K(t)$ in

$$\hat{\theta}(t) = \hat{\theta}(t-1) + K(t)x(t)\phi'(\varepsilon(t)) \tag{41}$$

are

$$K(t) = \frac{\alpha}{t}$$

$$\boldsymbol{K}(t) = \frac{\alpha}{t}[\beta + \phi''x(t)^Tx(t)]^{-1}$$

$$\boldsymbol{K}(t) = \alpha\left[\beta + \sum_{i=1}^{t}\phi''x(i)^Tx(i)\right]^{-1} \tag{42}$$

with $\alpha, \beta > 0$. These gains can be modified to include the forgetting factor $\lambda(t)$, by implementing a recursion for $\boldsymbol{K}_1(t)$ calculated from

$$\boldsymbol{K}_1(t) = \frac{\boldsymbol{K}_1(t-1)}{\lambda(t)} - [\boldsymbol{K}(t-1) - \boldsymbol{K}(t)] \tag{43}$$

5.1 Convergence and consistency

The proof of consistency is divided between showing that the recursive algorithm converges and showing that the convergence point corresponds to the actual parameter vector θ_0. Convergence of recursive prediction error algorithms has received much attention [19–22, 27]. Poljak and Tsypkin also provide concise conditions for consistency, but their proof requires that the case vectors $x(t)$ are independent and identically distributed and bounded in expectation. They only consider the case of

gains $K(t)$ with a constant positive definite matrix factor. Ljung and Soderstrom provide a rather general proof of convergence for prediction error algorithms. The proof is rather involved, but essentially consists of constructing a stochastic Lyapunov function related to distance of the estimates from the parameters, and, then, showing that under certain restrictions the estimates will converge to a region where the gradient of the Lyapunov function is zero. The main restrictions placed on estimator eqns. 35–37 are

(*a*) $\phi(x)$ is symmetric
(*b*) $\phi''(x) > 0$ (convexity)
(*c*) $|\phi'(x)| < C(1 + |x|), 0 < C < \infty$
(*d*) $\phi''(x) < C(1 + |x|^2), 0 < C < \infty$
(*e*) $\lambda(t) = 1, t > 0$
(*f*) $\hat{R}(0) > 0$
(*g*) $|x(t)| < \infty, t > 0$
(*h*) the minimum eigenvalue of

$$\sum_{t=1}^{\infty} x(t)x(t)^T$$

is ∞.

In the steepest descent algorithm, condition (*h*) corresponds to

$$\lim_{t\to\infty}\{tK(t)\} = \mu > 0 \tag{44}$$

and we require, in addition, the conventional constraints on the scalar gain in stochastic approximation:

$$\sum_{t=1}^{\infty} K(t) = \infty \tag{45}$$

and

$$\sum_{t=1}^{\infty} K(t)^2 < \infty \tag{46}$$

If the recursive algorithms are used to estimate the parameters of dynamic systems, there are further conditions, principally that the predictor is stable and that the (possibly closed-loop) system is exponentially stable [19].

The modified L_{pm} estimator, with case cost defined by eqn. 40, satisfies these conditions in both the stochastic Newton and steepest descent realisations. By the strong law, if the errors are uncorrelated, the point of convergence is θ_0, with probability 1.

6 Simulation study: adaptive control

In each of four simulations, an adaptive minimum-variance incremental controller [32] controlled the first-order system:

$$y(t) = -a_1 y(t-1) + b_0 u(t-1) + d(t) \tag{47}$$

where $y(t)$ is the system output, and $u(t)$ is the control input. $d(t)$ is a DC offset term that periodically underwent large jumps, generating large spike residuals in the incremental predictor,

$$\hat{y}(t) = y(t-1) - a_1\Delta y(t-1) + b_0\Delta u(t-1) \tag{48}$$

in which Δ is the backward-difference operator. The system was estimated by the regression

$$\Delta y(t) = -\hat{a}_1\Delta y(t-1) + \hat{b}_0\Delta u(t-1) \tag{49}$$

There was no random noise. Modified L_p (L_{pm}) estimation, using the case cost (5.12) with $c = 1$ was used throughout.

The first two simulation runs illustrated in Figs. 10 and 11 were designed to demonstrate the relative insensitivity to these spikes of the $L_{1.5m}$ (Fig. 11) estimator in comparison to the $L_{1.99m}$ estimator (Fig. 10). A forgetting

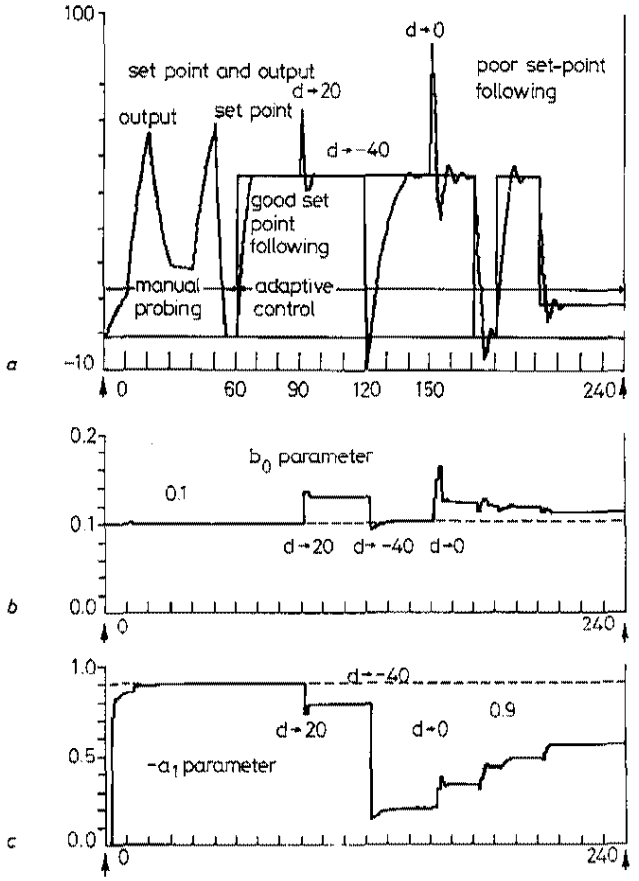

Fig. 10 *Minimum variance control — $L_{1.99m}$ estimation*
– – – – actual plant parameters, $-a_1$ and b_0

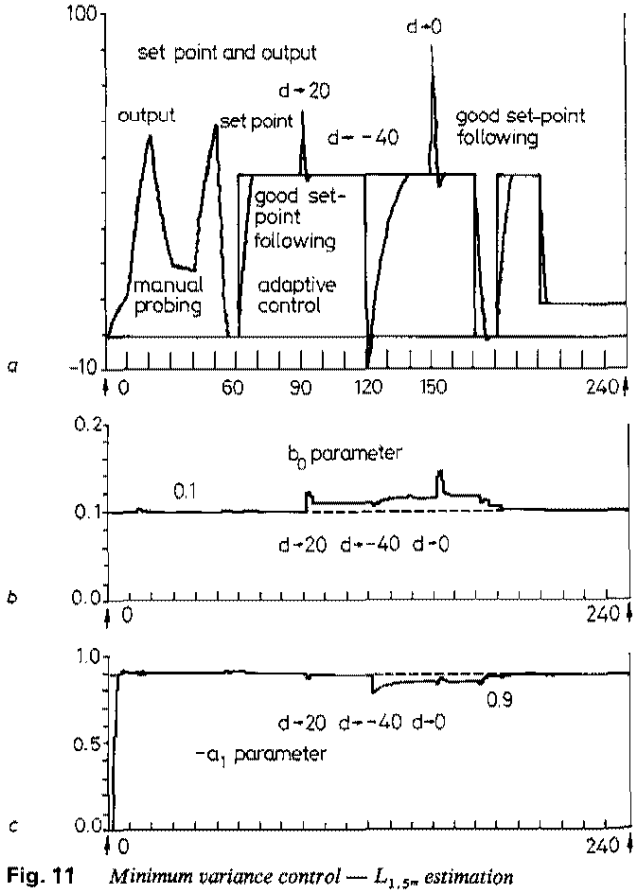

Fig. 11 *Minimum variance control — $L_{1.5m}$ estimation*
– – – – actual plant parameters, $-a_1$ and b_0

factor of 0.98 was used. For the first 60 samples the estimators were set running with the system under manual control. This period of manual probing allowed the estimators to obtain fairly accurate models of the simulated systems, as can be seen by comparing the full lines (estimates) with the broken lines (actual parameter values). At $t = 60$, manual probing stopped and the adaptive controller was switched on. The output climbed to the set point of 50. At $t = 90$ $d(t)$ increased from 0 to 20, at $t = 120$ it decreased to -40, and, finally, at $t = 150$ it was brought back to 0. Notice the large variations of the $L_{1.99m}$ estimates compared to those generated by the $L_{1.5m}$ estimator. This was reflected in the poor set-point following in Fig. 10 after the jumps in $d(t)$.

The second pair of simulations (Figs. 12 and 13) compare the adaptivity of $L_{1.99m}$ and $L_{1.5m}$ estimators, with the forgetting factor set at 0.97. After an initial period of manual probing, the adaptive controller was switched on. At $t = 70$, the system parameters were changed so as to increase the speed of the system and reduce its steady-state gain. At first, the $L_{1.99m}$ estimator adapted the faster, but, as the parameter error decreased, the $L_{1.5m}$ estimator caught up, and, by $t = 210$, its estimates were noticeably the better of the two sets. This pattern of adaptation is a consequence of the relative weighting of large and small prediction errors of the two estimators. $L_{1.99m}$ estimation weights smaller errors less than $L_{1.5m}$.

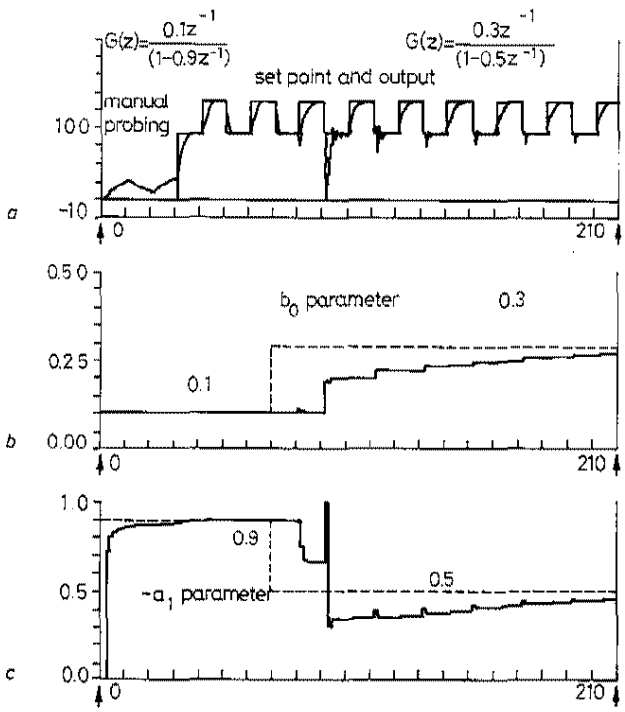

Fig. 12 *Adapting to a system change — $L_{1.99m}$ estimation*
- - - - actual plant parameters, $-a_1$ and b_0

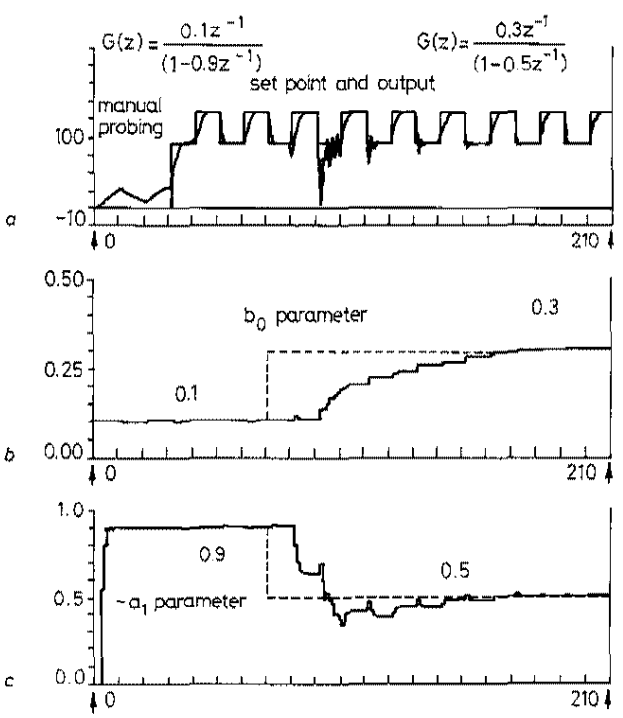

Fig. 13 *Adapting to a system change — $L_{1.5m}$ estimation*
- - - - actual plant parameters, $-a_1$ and b_0

These two pairs of simulations help to justify the dual claim that not only is recursive L_{pm} estimation more robust than recursive least squares, but adaptivity is not sacrificed in consequence.

7 Conclusion

This paper has presented a range of consistent recursive implementations of the L_p estimator, using algorithms that are closely related to recursive least squares and linear recursive steepest descent. The robustness and efficiency properties of this estimator are excellent as p becomes close to 1. Of all consistent gradient-type recursive estimators, L_{1+} estimation is both minimax SIC and asymptotic covariance robust. The estimator does not require any scale parameters to be known accurately, or to be estimated online. The recursive implementations can be made adaptive, so as to track time-varying parameters.

It is interesting to compare the two different approaches to handling outliers (detection and deletion, and robustification) in the light of the sample influence curve (Section 4). From eqn. 16 and expr. 21, it can be seen that robustification can be interpreted as a form of variably weighted outlier detection and removal. Robustification is, thus, the more general procedure. If the case cost $\phi(x)$ is smooth, robust procedures are not sensitive to detection thresholds as are detection and deletion procedures.

Finally, the foregoing discussion has concentrated on the sensitivity properties of M-estimators to measurement, given an experimental design X. Frequency, in real-time estimation problems, there is no choice and one has to take the data that comes. If not, careful experimental design can robustify estimation by reducing the variation of case leverage [13, 28].

8 References

1 ANDREWS, D.F., BICKEL, P., HAMPEL, F., HUBER, P., ROGERS, W.H., and TUKEY, J.W.: 'Robust estimates of location' (Princeton University Press, Princeton, NJ, 1972)
2 ANDREWS, D.F.: 'A robust method for multiple linear regression', *Technometrics*, 1974, **16**, pp. 523–531
3 ANDREWS, D.F., and PREGIBON, D.: 'Finding outliers that matter', *J. Roy. Statist. Soc., Ser. B*, 1978, **40**, pp. 85–93
4 BARNET, V., and LEWIS, T.: 'Outliers in statistical data' (Chichester-Wiley, 1978)
5 BIERMAN, G.J.: 'Factorization methods for discrete system estimation' (Academic Press, 1977)
6 BOX, G.E.P.: 'Non-normality and tests on variance', *Biometrica*, 1953, **40**, p. 318
7 BOX, G.E.P., and DRAPER, N.R.: 'Empirical model building and response surfaces' (John Wiley, New York, 1987)
8 COOK, R.D., and WEISBERG, S.: 'Residuals and influence in regression' (Chapman and Hall, 1982)
9 COOK, R.D.: 'Detection of influential observations in linear regression', *Technometrics*, 1977, **19**, pp. 15–19

10 DUTTER, R.: 'Robust regression: different approaches to numerical solutions and algorithms'. Report 6, Fachgruppe für Statistik, Eidgen. Techniche Hochschule, Zurich, 1975
11 GLAISHER, J.W.L.: 'On the rejection of discordant observations', *Monthly Notices Roy. Astr. Soc.*, 1874, **34**, p. 251
12 HAMPEL, F.: 'Contributions to theory of robust estimation'. PhD Thesis, University of California at Berkley, 1968
13 HOAGLIN, D.C., and WELSCH, R.: 'The hat matrix in regression and ANOVA', *Am. Stat.*, 1978, **32**, pp. 108–115
14 HODGSON, A.J.F.: 'Problems of integrity in the application of adaptive controllers'. DPhil Thesis, University of Oxford, OUEL report 1436/82, 1982
15 HUBER, P.: 'Robust statistics' (John Wiley, New York, 1981)
16 HUBER, P.: 'Robust estimation of a location parameter', *Ann. Math. Stat.*, 1964, **35**, pp. 73–101
17 KENDALL, M.G., and STUART, A.: 'The advanced theory of statistics — Vol. II' (Griffin, London, 1979, 4th edn.)
18 KUSHNER, H.J., and CLARK, D.S.: 'Stochastic approximation methods for constrained and unconstrained systems' (Springer-Verlag, New York, 1978)
19 LJUNG, L., and SÖDERSTROM, T.: 'Theory and practice of recursive identification' (MIT Press, 1983)
20 LJUNG, L.: 'Analysis of a general recursive prediction error algorithm', *Automatica*, 1981, **17**, pp. 89–100
21 LJUNG, L.: 'Analysis of recursive stochastic algorithms', *IEEE Trans.*, 1977, **AC-52**, pp. 551–575
22 LJUNG, L.: 'Convergence of recursive stochastic algorithms'. Report 7403, Department of Automatic Control, Lund Institute of Technology, Lund, Sweden, 1974
23 MALLOWS, C.L.: 'On some topics in robustness'. Bell Telephone Laboratories Report, Murray Hill, New Jersey, 1975
24 MASRELIEZ, C.J., and MARTIN, R.D.: 'Robust bayesian estimation for the linear model and robustifying the Kalman filter', *IEEE Trans.*, 1977, **AC-22**, pp. 361–371
25 McMICHAEL, D.W.: 'On-line fault detection: a system non-specific approach'. DPhil Thesis, University of Oxford, OUEL report 1729/88, 1987
26 POLJAK, B.T., and TSYPKIN, Y.Z.: 'Robust identification', *Automatica*, 1980, **18**, pp. 53–63
27 POLJAK, B.T., and TSYPKIN, Y.Z.: 'Adaptive estimation algorithms, convergence, optimality, robustness', *Autom. & Remote Control*, 1979, **3**, pp. 71–84
28 PREGIBON, D.: 'Data analytic methods for generalized linear models'. PhD Thesis, University of Toronto, 1981
29 PRESS, W.H., FLANNERY, B.P., TEUKOLSKY, S.A., and VETTERLING, W.T.: 'Numerical recipes' (Cambridge University Press, 1986)
30 ROBBINS, H., and MONRO, S.:: 'A stochastic approximation method', *Ann. Math. Stat.*, 1951, **22**, pp. 400–407
31 TSAI, C., and KURZ, L.: 'An adaptive approach to Kalman filtering', *Automatica*, 1983, **19**, (3), pp. 279–289
32 TUFFS, S.: 'Self-tuning control: algorithms and applications'. DPhil thesis, University of Oxford, OUEL report 1567/85, 1984
33 TUKEY, J.W.: 'A survey of sampling from contaminated distributions', *in* OLKIN, I. (Ed.): 'Contributions to probability and statistics' (Stanford University Press, 1960)
34 WASAN, M.T.: 'Stochastic approximation' (Cambridge University Press, 1969)