Katholieke Universiteit Leuven

Departement Elektrotechniek

ESAT-SISTA/TR 98-124

A FAMILY OF ADAPTIVE ROBUST ESTIMATES BASED ON THE SYMMETRICAL GENERALIZED LOGISTIC DISTRIBUTION

Baibing Li and Bart De Moor[2]

December 1998

[1]This paper is also available by anonymous ftp from ftp.esat.kuleuven.ac.be as file rep98124.ps.Z in the directory pub/SISTA/baibing/reports.

[2]ESAT/SISTA, K.U.Leuven, Kardinaal Mercierlaan 94, B-3001 Heverlee (Leuven), Belgium, tel.+32-16-32.18.55, fax +32-16-32.19.70
email: baibing.li@esat.kuleuven.ac.be.

A FAMILY OF ADAPTIVE ROBUST ESTIMATES BASED ON THE SYMMETRICAL GENERALIZED LOGISTIC DISTRIBUTION

Baibing Li and Bart De Moor

ESAT-SISTA, Dept. of Electrical Engineering, Katholieke Universiteit Leuven
Kardinaal Mercierlaan 94, 3001 Leuven, Belgium

Key Words: L_1/L_2 estimate; regression; robust estimate; symmetrical generalized logistic distribution.

ABSTRACT

A family of adaptive robust estimates which range from L_1 to L_2 estimates is suggested based on the symmetrical generalized logistic (*SGL*) distribution. Depending on whether existing outliers and how serious for the outliers in sampling data, this adaptive *SGL* estimate family can automatically choose an L_2 estimate, L_1 estimate, or smoothed Huber estimate to fit the data. It is shown that the asymptotic efficiencies of the adaptive *SGL* estimates relative to L_2/L_1 estimates are 1 at the normal/Laplace distribution situations respectively. Practical examples show that the *SGL* estimates have satisfactory flexibility to deal with different patterns of outliers in data. Hence, the adaptive *SGL* robust estimates are useful for automatic analysis in systems identification and are very convenient for practitioners.

1. INTRODUCTION

A lot of attention has been paid to develop robust estimates for regression analysis in recent decades due to the fact that L_2 estimates are sensitive to outliers. *M*-estimate is one of most of important classes of robust estimates.

In practice, M-estimate seems not so successful as it is in theory. One of the important reasons for this is not that practitioners are not convinced by its importance, but many of practitioners often do not know which particular method should be used among many different M-estimate procedures, how to tune technical parameters in M-estimates, and why (see also, Lange et al., 1989).

Another related problem in practice is automatic analysis in systems identification in engineering. It is often desired that methods used in systems identification have reasonable flexibility for wide range of data types, and/or have ability to automatically adapt to different types of practical data.

Hogg (1967) perhaps is the earliest pioneer noticing these problems and proposing a new concept, adaptive M-estimates, to deal with them. Yuh and Hogg (1988) proposed a composite M-estimate based on t distribution, AT-estimate, that is, using $t(11)$ (t distribution with degrees of freedom v=11) for error model if the selector statistics indicate the underlying distributions are symmetrical and light-tailed, and using $t(3)$ if symmetrical and heavy-tailed.

Another closely related approach, t estimate, is to estimate regression coefficients based on t distribution as error model with any of degrees of freedom v, where v is determined by samples. Lange et al (1989) gave a review for this approach and investigated many practical cases. To make it more convenient for practical use, Jamshidian (1997) gave SAS and BMDP programs of t estimates for two examples.

According to Lange et al. (1989), the reason that they chose t rather than other distribution families mainly lies on grounds of familiarity and computational simplicity. The purpose of this paper is then to consider a new family of estimates, as an alternative of t estimates, with the same computational simplicity and on the other hand, possesses some other attractive characteristics.

It is known that L_1 and L_2 estimates have simplicity in intuition (corresponding to median and mean for location parameter) and then are very attractive to practitioners. Hence, it is desired that a family of adaptive robust estimates can not only cover L_2 estimate and behave approximately like t estimates for lighter and mild tailed error distributions, but also can be extended to L_1 estimates for relatively serious outliers. We will show in this paper that the *SGL* estimate family with shape parameter $\lambda\in[0,+\infty)$ is a very ideal candidate for these requirements, that is, when not existing any outliers in data, the *SGL* estimates automatically choose L_2 estimates; when existing relatively serious outliers, choose L_1 estimates; when only mild outliers, choose smoothed Huber estimates. Moreover, the asymptotic efficiency 1 relative to L_2/L_1 estimates can be achieved by the *SGL* estimates at the normal/Laplace distribution situations.

2. ADAPTIVE *SGL* ESTIMATES

Consider the density of the standardized *SGL* distribution $SGL\{\mu,\sigma,\lambda\}$:

$$f_{SGL}(x;\mu,\sigma,\lambda)=[1/(\sqrt{2\pi}\,\sigma)]R(\lambda+1/2)[\mathrm{sech}\{(x-\mu)/(c(\lambda)\sigma)\}]^{2\lambda} \quad \sigma>0, \quad \lambda>0$$

where $R(\lambda)=\lambda^{1/2}\Gamma(\lambda)/\Gamma(\lambda+1/2)$, $c(\lambda)=2\lambda/(2\lambda+1)^{1/2}$, $\Gamma(z)$ represents the gamma function, and $\mathrm{sech}(x)=2/[\exp(x)+\exp(-x)]$.

Denote the density of the t distribution $t\{\mu,\sigma,v\}$ as

$$f_t(x;\mu,\sigma,v)=[\Gamma((v+1)/2)/\{\sigma(v\pi)^{1/2}\Gamma(v/2)\}][1+(x-\mu)^2/(\sigma^2 v)]^{-(v+1)/2} \quad \sigma>0,\ v>0$$

the density of the Laplace distribution $L\{\mu,\sigma\}$ as

$$f_L(x;\mu,\sigma)=[1/(2\sigma)]\exp[-|x-\mu|/\sigma] \qquad \sigma>0$$

and the density of the standard normal as $\varphi(x)$. Denote their cumulative distribution as $F_{SGL}(x;\mu,\sigma,\lambda)$, $F_t(x;\mu,\sigma,v)$, $F_L(x;\mu,\sigma)$ and $\Phi(x)$ respectively.

First we state three lemmas.

LEMMA 1. (Davidson, 1980, see Balakrishnan, 1992, pp212) $SGL\{0,1,\lambda\}$ approaches to normal $N(0,1)$ as $\lambda\rightarrow+\infty$.

LEMMA 2. (Li and De Moor, 1997) $t\{0,1,v\}$ can be approximated by $SGL\{0,[(v+3)/(v+1)]^{1/2},v/6\}$ with the maximum absolute distance

$$\sup_x |F_t(x;0,1,v)-F_{SGL}(x;0,[(v+3)/(v+1)]^{1/2},v/6)|=O(1/v^2) \quad \text{for large } v$$

LEMMA 3. $SGL\{0,1,\lambda\}$ approaches to Laplace distribution $L(0,1)$ as $\lambda\to 0$.

The proof of the Lemma 3 is given in **APPENDIX.**

For data $y_1,\ldots,y_n$ independently sampled from a symmetrical distribution, the idea of the *SGL* regression is to use the assumption of *SGL* error model

$$y_j \sim SGL\{\mu_j(\theta),\sigma,\lambda\}$$

to replace following three assumptions on error distributions:

(I)normal assumption: $y_j \sim N\{\mu_j(\theta),\sigma^2\}$ (underlying distribution for L_2 estimates)

(II)t assumption: $y_j \sim t\{\mu_j(\theta),\sigma,v\}$ (underlying distribution for t estimates)

(III)Laplace assumption: $y_j \sim L\{\mu_j(\theta),\sigma\}$ (underlying distribution for L_1 estimates)

The assumption of *SGL* error model, as we shall see below, will automatically produce an L_2/L_1 or smoothed Huber estimate. The rationale for this replacement is directly from Lemma 1-3. Since choosing an underlying error distribution in regression analysis reflects a belief and meanwhile means a risk of mis-belief, a belief in a distribution family covering a broad range of distributions is a good compromise for this problem to reduce the risk. Moreover, a related advantage of the *SGL* model is that it gives a clear statement of assumption on distributions.

For *SGL* regression, we assume that $\mu_j(\theta)$ is a general nonlinear function of θ that relates the response y_j to p covariates x_j ($j=1,...,n$). Particularly, for linear regression, we have

$$\mu_j(\theta)=x_j^T\theta=\sum_{k=1}^{p} x_{jk}\theta_k \qquad j=1,...,n \qquad (1)$$

for given $n\times p$ design matrix $X=[x_1,...,x_n]^T$ with rank(X)=p. From now on, we concentrate on the linear regression problem (1).

From the assumption of *SGL* model $y_j \sim SGL\{x_j^T\theta,\sigma,\lambda\}$, the estimates of the parameters (θ,σ,λ) for *SGL* distribution can be obtained as the *ML* estimates by maximizing the following log-likelihood function

$$G(\theta,\sigma,\lambda)=\log[\prod_{j=1}^{n}\{\ f_{SGL}(y_j;x_j^T\theta,\sigma,\lambda)\}]$$

$$=n\log(R(\lambda+1/2))-n\log(\sqrt{2\pi}\,\sigma)+2\lambda\sum_{j=1}^{n}\log(\sec\mathrm{h}(r_i)) \qquad (2)$$

where $r_j=e_j/(c(\lambda)\sigma)$, $e_j=y_j-x_j^T\theta$ and $c(\lambda)=2\lambda/(2\lambda+1)^{1/2}$.

The first-order derivatives of the log-likelihood function are:

$$\partial G/\partial\theta=(2\lambda+1)^{1/2}\sigma^{-1}\sum_{j=1}^{n}\tanh(r_j)\,x_j \qquad (3)$$

$$\partial G/\partial\sigma=(n/\sigma^2)[(1/n)(2\lambda+1)^{1/2}\sum_{j=1}^{n}(y_j-x_j^T\theta)\tanh(r_j)-\sigma] \qquad (4)$$

$$\partial G/\partial\lambda=nR'(\lambda+1)/R(\lambda+1)+2\sum_{j=1}^{n}\log(\sec\mathrm{h}(r_j))$$

$$+2(\lambda+1)/(2\lambda+1)\sum_{j=1}^{n}r_j\tanh(r_j) \qquad (5)$$

Hence, from Eq.(3), the "normal equations" of θ are given by:

$$\sum_{j=1}^{n}\tanh(r_j)\,x_j=0 \qquad (6)$$

Observation for Eq.(6) gives:

(i) for large λ such that r_j is small enough, we have $\tanh(r_j)\sim r_j$. Then Eq. (6) approximately equals to

$$\sum_{j=1}^{n}r_j\,x_j=0 \qquad \text{or equivalently} \qquad \sum_{j=1}^{n}(y_j-x_j^T\theta)x_j=0$$

which is the normal equations of L_2 regression. Therefore, for large λ, the *SGL* estimates of regression coefficients are approximations of L_2 estimates.

(ii) for small λ such that r_j is large enough, we have $\tanh(r_j)\sim\mathrm{sign}(r_j)$. Then Eq. (6) is approximately equivalent to

$$\sum_{j=1}^{n}\mathrm{sign}(y_j-x_j{}^T\theta)x_j=0$$

which is the normal equations of L_1 regression. Therefore, for small λ, the *SGL* estimates of regression coefficients are approximations of L_1 estimates.

It is noticeable from this property that the *SGL* estimates provide an asymptotic approach to analytically investigate properties of L_1 estimates.

(iii) for large λ, from Eq.(4) and Taylor' theorem

$$\sigma^2=n^{-1}\sum_{j=1}^{n}e_j^2 +O(1/\lambda) \tag{7}$$

Hence, combining (6) and (7), we have

$$\sum_{j=1}^{n}x_j e_j[1-e_j^2 / (6\lambda\{n^{-1}\sum_{j=1}^{n}e_j^2\}) / +O(1/\lambda^2)]=0 \tag{8}$$

Similarly, for large v, directly from log-likelihood function of t distribution (see also, Lange et al, 1989), the solution θ of t-estimate satisfies

$$\sum_{j=1}^{n}x_j e_j[1-e_j^2 / (v\{n^{-1}\sum_{j=1}^{n}e_j^2\}) + O(1/v^2)]=0 \tag{9}$$

Thus, the solutions of (8) and (9) for θ are approximately the same when $v\approx 6\lambda$. This means that for large v and λ, *SGL* and t estimates of θ are almost the same.

A similar discussion can be given to the *ML* estimates of σ. Immediately from Eq.(4), it satisfies

$$\sigma=(2\lambda+1)^{1/2}(1/n)\sum_{j=1}^{n}[\{y_j - x_j^T\theta\}\tanh(r_j)\ ] \tag{10}$$

Therefore,

$$\sigma\to(1/n)\sum_{j=1}^{n}|y_j - x_j^T\theta| \qquad \text{as } \lambda\to 0,$$

which means that when $x_j^T\theta$ is reduced to a location parameter, the *SGL* estimates of σ are absolute deviation from the sample median. Moreover,

$$\sigma^2\to(1/n)\sum_{j=1}^{n}[y_j - x_j^T\theta]^2 \qquad \text{as } \lambda\to +\infty$$

which means that the *SGL* estimates of σ are standard deviation as $\lambda\to+\infty$.

Finally, we introduce an auxiliary index ranging in the interval [1, 2]:

$$q=(2\lambda+1)/(\lambda+1) \tag{11}$$

to describe the *SGL* estimate family. It is clear that for q=1, 1.5 and 2, the *SGL* estimate takes L_1 estimate, ordinary logistic robust estimate, and L_2 estimate respectively.

3. SOME PROPERTIES OF *SGL* ESTIMATES

3.1. Unbiasedness, variance and efficiency

For simplicity, consider estimates of one-dimensional location parameter θ.

First, from the Theorem 4 of Goodall (1983, p364), for any symmetrical underlying error distribution (not necessarily the *SGL*), the *SGL* estimates $\hat{\theta}_{SGL}$ of the location parameter θ, obtained by solving Eqs.(3)-(5), are unbiased.

Next, consider asymptotic variance of estimate $\hat{\theta}_{SGL}$. For the *SGL* distributions, from the symmetry, we can obtain following elements of the Fisher information matrix by direct computation from (3)-(5):

$$I_{\sigma\theta}= E\{(\partial G/\partial\sigma)(\partial G/\partial\theta)\}=0 \qquad I_{\lambda\theta}= E\{(\partial G/\partial\lambda)(\partial G/\partial\theta)\}=0$$

$$I_{\theta\theta}=E\{(\partial G/\partial\theta)^2\}=\sigma^{-2} \tag{12}$$

Hence, the asymptotic variance of $\hat{\theta}_{SGL}$ is σ^2. In addition, from (12) we conclude that the *ML* estimates of θ and (σ,λ) are asymptotically uncorrelated, and asymptotic standard errors of estimates of θ are unaffected by estimating the scale and shape parameters σ and λ.

For the location parameter θ, from the above asymptotic variance, it is easy to check that the asymptotic efficiency of the *SGL* estimate $\hat{\theta}_{SGL}$ relative L_2 estimate $\hat{\theta}(L_2)$ at the normal distribution $N(\theta,\sigma^2)$ is

$$\mathrm{eff}(\hat{\theta}_{SGL}\,|\,\hat{\theta}(L_2), N(\theta,\sigma^2))=\mathrm{var}(\hat{\theta}(L_2)|\, N(\theta,\sigma^2)) \,/\, \mathrm{var}(\hat{\theta}_{SGL}\,|\, N(\theta,\sigma^2))=1$$

Similarly, the asymptotic efficiency of the *SGL* estimate $\hat{\theta}_{SGL}$ relative L_1 estimate $\hat{\theta}(L_1)$ at the Laplace distribution $L(\theta,\sigma)$ is

$$\mathrm{eff}(\hat{\theta}\,|\,\hat{\theta}(L_1), L(\theta,\sigma))=\mathrm{var}(\hat{\theta}(L_1)|\, L(\theta,\sigma)) \,/\, \mathrm{var}(\hat{\theta}_{SGL}\,|\, L(\theta,\sigma))=1$$

Therefore, the *SGL* estimates have optimal performances at both the normal and Laplace distribution situations. Furthermore, the above conclusions can be extended to any member distribution of the *SGL* family, that is, at any particular generalized symmetrical logistic distribution as an underlying error distribution (for instance, the ordinary logistic, and the normal/Laplace distributions), the *SGL* estimates have asymptotic efficiency equal to 1.

Finally, for general linear regression problems (1), the covariance matrix of the *SGL* estimator can be derived similarly under some regular conditions:

$$\mathrm{cov}(\hat{\theta}_{SGL}) \approx \hat{\sigma}^2_{SGL}\,(X^T X)^{-1}$$

Other asymptotic properties of the *SGL* estimators are immediate from the counterpart results of *ML* estimators since the former belong to the latter.

3.2. The relationship with the Huber estimates

In this sub-section, we show that for any intermediate value of the shape parameter $\lambda \in (0,\infty)$, the *SGL* estimates are smoothed Huber estimates.

For normalized scale parameter estimate s, the Huber's ψ-function is

$$\psi_{Huber}(u;\, s,K) = \begin{cases} u/(sK) & \text{for } |u/s| \le K \\ \mathrm{sign}(u/s) & \text{for } |u/s| > K \end{cases}$$

From the relationship between *M*-estimator and *MLE* estimator, the *SGL* estimates can be regarded as an *M*-estimate with the ψ-function:

$$\psi_{SGL}(x-t;\, \sigma,\lambda) = \tanh((x-t)/(c(\lambda)\sigma)) \propto -d\log[f_{SGL}(x;t,\sigma,\lambda)]/dt$$

Noting that for $\lambda \in (0,\infty)$, we have

$$\lim_{u\to\infty} \psi_{Huber}(u;s,K)/\psi_{SGL}(u;\sigma,\lambda) = \text{constant}$$

and $$\lim_{u\to\infty} \psi_{Huber}(u;s,K)/\psi_{SGL}(u;\sigma,\lambda) = \text{constant}$$

we conclude that the ψ_{SGL} has the same tails and middle part as ψ_{Huber}, or equivalently, just like ψ_{Huber}, ψ_{SGL} has exponential (Laplace distribution) tails and a middle part of normal distribution. Hence, the *SGL* estimates, as the Huber estimates, are not sensitive to both gross errors and rounding/grouping errors for $\lambda \in (0,\infty)$. Then it can be expected that it possesses nice behaviors under such disturbances. Furthermore, it is clear by observation for ψ_{Huber} and

ψ_{SGL} that $\psi_{SGL}(u;\sigma,\lambda)$ can be regarded as a smoothed version of ψ_{Huber}. FIG. 1 gives a plot of Huber ψ-function with K=1.25 and s=1, and the function $y=\tanh(u/(c\sigma))$ with c=1.1 and σ=1. It can be seen that they are close each other except in the neighborhoods of the two unsmoothing points $u=\pm Ks$.

Similar to the target distribution of the Huber estimates (Huber, 1981, p85), the *SGL* distribution can be also represented as a contaminated normal model for a contaminating distribution $H(x)$ (see **Lemma 4** in **Appendix**):

$$F_{SGL}(x;0,1,\lambda)=(1-\varepsilon)\Phi(x)+\varepsilon H(x) \qquad \text{for } 0<\varepsilon<1 \text{ and } \lambda\in[0,\infty).$$

FIG. 2 gives the plot of densities of the *SGL* for several shape parameters.

To sum up, ψ_{SGL} has similar tails and central part as ψ_{Huber} and can be viewed as a smoothed version of ψ_{Huber}.

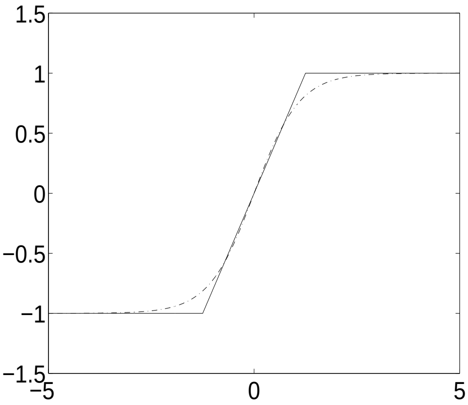

FIG. 1. Plot of $\psi_{Huber}(u;1,1.25)$ (—) and $\tanh(u/1.1)$ (-·-·)

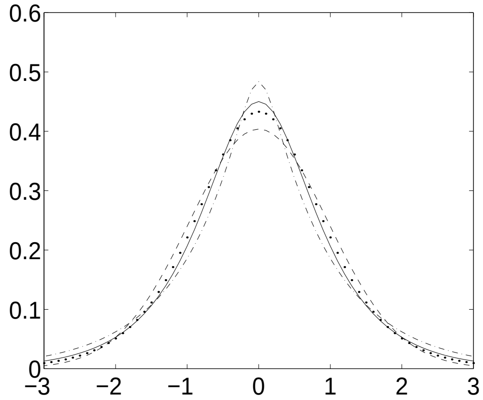

FIG. 2. Plot of $f_{SGL}(x;0,1,10)$ (---), $f_{SGL}(x;0,1,1)$ (...),$f_{SGL}(x;0,1,0.5)$(—), $f_{SGL}(x;0,1,0.1)$ (-·-·).

On the other hand, unlike the Huber estimates which have fixed parameter K in its ψ-function ψ_{Huber}, the *SGL* estimates determine the shape parameter based on sampling data. Hence, the *SGL* estimates have higher efficiencies than the Huber estimates at any of the member distribution of the *SGL* family, particularly, at the normal and Laplace distributions as mentioned before. TABLE I gives a comparison on numerical variances of the *SGL* and Huber estimates of location parameter for 1000 samples with size 40 at three important underlying distribution situations: the normal, Laplace and Cauchy distributions.

TABLE I. A comparison on numerical variances of the *SGL* and Huber estimates of the location parameter for 1000 samples with size 40

	Normal(0,1)	Laplace(0,1)	Cauchy
Huber estimates*	0.0278	0.0365	0.0968
SGL estimates	0.0266	0.0344	0.0677

* K=1.25 and s=MAD/0.6745, where MAD is the median absolute deviation.

4. PRACTICAL EXAMPLES

Four practical examples are examined in this section to investigate flexibility of the adaptive *SGL* estimates for different patterns of outliers.

***Example* 1.** (Pilot-Plant data). This set of data was from Daniel and Wood (1971), and also considered by Rousseeuw and Leroy (1987) and Yale and Forsythe (1976). The response variable (y) and explanatory variable (x) are the acid content determined by titration and organic acid content determined by extraction and weighing. The explanatory variable is re-scaled by multiplying 0.01 for its raw data in this study.

The scatter plot given by FIG. 3(a) suggests a linear statistical relation between response variable and explanatory variable. The *SGL* estimate gives:

$$\hat{y}=35.4583+32.1608x \qquad \text{with} \quad \operatorname{cov}(\hat{\theta}_{SGL})=\begin{bmatrix} 0.3630 & -0.2862 \\ -0.2862 & 0.2777 \end{bmatrix}$$

and $\hat{\lambda}=2.63\times10^{7}$, or equivalently by the index in Eq.(11), $\hat{q}=2$.

Obviously, $\hat{q}=2$ suggests that at this case, the adaptive *SGL* estimates automatically adopt normal distribution as the underlying error distribution to produce an L_2 estimate. Direct computations for L_2 estimate and its covariance matrix have verified that they are the same as their counterparts of the *SGL*.

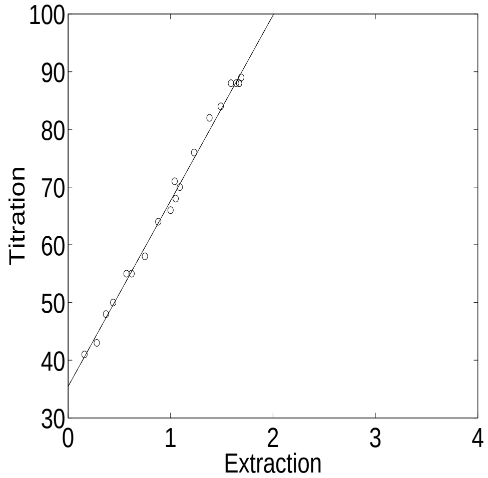

FIG. 3(a). Pilot-Plant data with L_2 fit (---) and *SGL* fit(—)

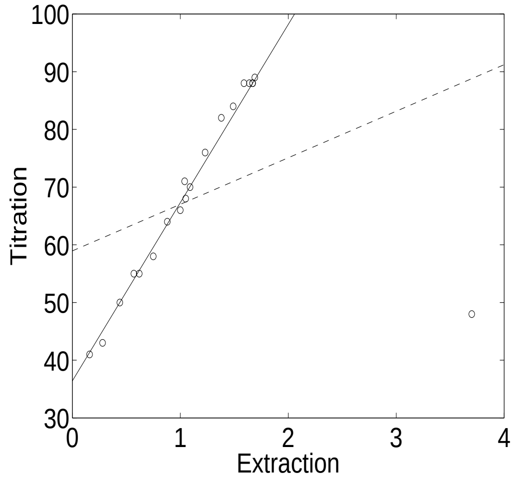

FIG. 3(b). Contaminated Pilot-Plant data with L_2 fit (---) and *SGL* fit(—)

To investigate the effects of outliers, Rousseeuw and Leroy (1987) considered the situation in which the *x*-value of the 6*th* observation was registered as 370 instead of 37 by mistake (re-scaled as 3.7 in this study). The scatter plot is given by FIG.3(b). With this contaminated data, the L_2 fit is

$$\hat{y}=58.9388+8.0712x \quad \text{with} \quad \text{cov}(\hat{\theta}_{LS})=\begin{bmatrix} 787.4569 & -474.9143 \\ -487.9143 & 396.7537 \end{bmatrix}$$

which, from FIG. 3(b), is strongly attracted by this single outlier.

On the contrast, the *SGL* gives its fit as

$$\hat{y}=36.4065+30.8943x \quad \text{with} \quad \text{cov}(\hat{\theta}_{SGL})=\begin{bmatrix} 6.8265 & -4.1170 \\ -4.1170 & 3.4395 \end{bmatrix}$$

and $\hat{\lambda}=6.1696\times10^{-13}$, or equivalently, $\hat{q}=1$.

It is clear that the *SGL* fit lies close to the L_2 fit applied to the uncontaminated data, and then is robust on the outlier, the 6*th* observation. $\hat{q}=1$ suggests that at this situation, the adaptive *SGL* estimates automatically produce an L_1 estimate. This has been verified by directly computing L_1 estimate and its covariance matrix.

***Example* 2.** (Brain and weight data) This set of data with size 28 was sampled by Rousseeuw and Leroy (1987) from larger data set in Weisgerg

(1980) to investigate the relation between brain and body weights. The response and explanatory variables are the logarithms (to the base 10) of these measurements for body and brain weights.

The scatter plot given by FIG. 4 suggests a linear relationship between response variable (y) and explanatory variable (x) for majority of data while there are three observations, number 6, 16 and 25 lying far from the others. The L_2 estimate gives its fit as

$$\hat{y}=1.1096+0.4960x \quad \text{with} \quad \mathrm{cov}(\hat{\theta}_{LS})=\begin{bmatrix} 0.8370 & -0.2602 \\ -0.2602 & 0.1589 \end{bmatrix}$$

which is pulled away by three observations 6, 16 and 25.

The *SGL* estimate gives

$$\hat{y}=0.815+0.741x \qquad \text{with} \qquad \mathrm{cov}(\hat{\theta}_{SGL})=\begin{bmatrix} 0.0132 & -0.0041 \\ -0.0041 & 0.0025 \end{bmatrix}$$

which fits the majority of data nicely. According to Rousseeuw and Leroy (1987), the three observations 6, 16 and 25 correspond to three dinosaurs, each of which possessed a small brain as compared with a heavy body. This suggests they do not obey the same trend as the one followed by the majority of the data.

***Example* 3.** (International phone call data) This set of data was from Belgian Statistical Survey and firstly investigated by Rousseeuw and Leroy (1987). The data contain the total number (response variable y) (in tens of millions) of international phone calls made each year (explanatory variable x).

The scatter plot is given by FIG. 5. Rousseeuw and Leroy (1987) noticed that the time series contain heavy contamination from 1964 to 1969 due to using another recording system (the years 1963 and 1970 are also partially affected).

The L_2 estimate gives its fit as

$$\hat{y}=-26.006+0.504x \quad \text{with} \quad \mathrm{cov}(\hat{\theta}_{LS})=\begin{bmatrix} 2.3162 & -0.0372 \\ -0.0372 & 0.0006 \end{bmatrix}\times 10^{3}$$

which is affected seriously by the contaminated data between 1964 and 1969.

On the contrast, the *SGL* estimate gives its fit as

$$\hat{y}=-7.466+0.151x \qquad \text{with} \quad \text{cov}(\hat{\theta}_{SGL})=\begin{bmatrix} 41.3801 & -0.6644 \\ -0.6644 & 0.0108 \end{bmatrix}$$

which is quite successful in avoiding the outliers.

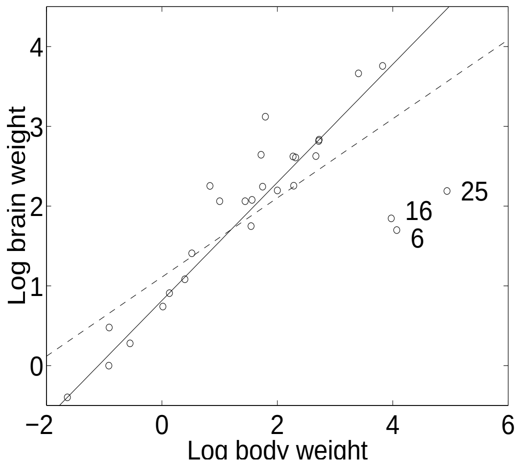

FIG. 4. Brain and weight data with L_2 fit (---) and *SGL* fit(—)

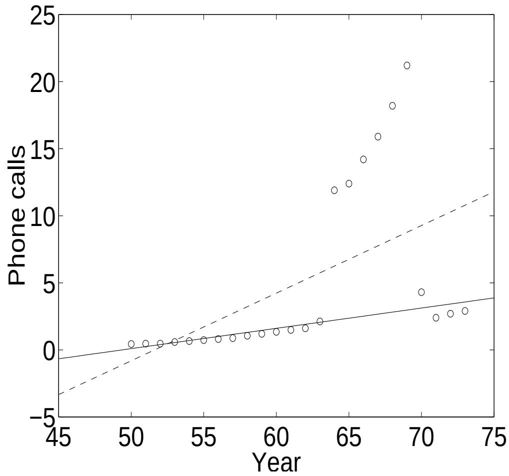

FIG. 5. International phone call data with L_2 fit (---) and *SGL* fit(—)

Finally, we consider an example with multiple explanatory variables. Normally, multiple regressions are much more complex than simple regressions because directly visual detection is no longer possible.

***Example* 4.** (Stack-Loss Data). The stack-loss data set was firstly given by Browlee and has become classical for testing robust estimation procedures. The response variable is stackloss (y). Three explanatory variables are rate of operation (x_1), the cooling water inlet temperature (x_2) and the acid concentration (x_3) (see, Daniel and Wood, 1971). Linear model is considered:

$$y_j=\theta_0+\theta_1 x_{j1}+\theta_2 x_{j2}+\theta_3 x_{j3}+\varepsilon_j \qquad j=1,...,21.$$

The L_2 estimate gives its fit as

$$\hat{y}=-39.92+0.72x_1+1.30x_2-0.15x_3$$

which is seriously affected by outliers according to many previous analyses. TABLE II lists several robust estimates of the regression coefficients, including the *SGL* estimate, Huber estimate, Andrews estimate and t-estimate. Since many previous researches suggested that observations 1, 3, 4 and 21 were outliers (see, for instance, Daniel and Wood, 1971), we also consider the L_2 estimate based on the data set deleting these four observations (denote as L_2-delete estimate).

TABLE II. Estimates of the regression coefficients for stack-loss data

	SGL	L_2-delete	Andrews†	Huber†	*t*†
$\hat{\theta}_0$	-39.69	-37.65	-37.20	-41.00	-38.50
$\hat{\theta}_1$	0.83	0.80	0.82	0.83	0.85
$\hat{\theta}_2$	0.57	0.58	0.52	0.91	0.49
$\hat{\theta}_3$	-0.06	-0.07	-0.07	-0.13	-0.07

† *Source*: Lange et al. (1989).

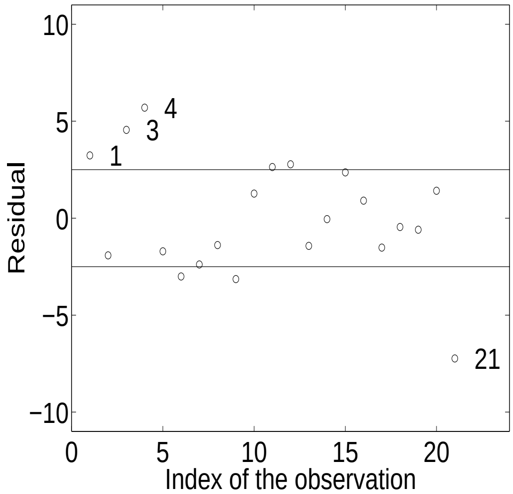

FIG. 6(a). Stack-Loss data: Residual plot associated with the L_2 fit.

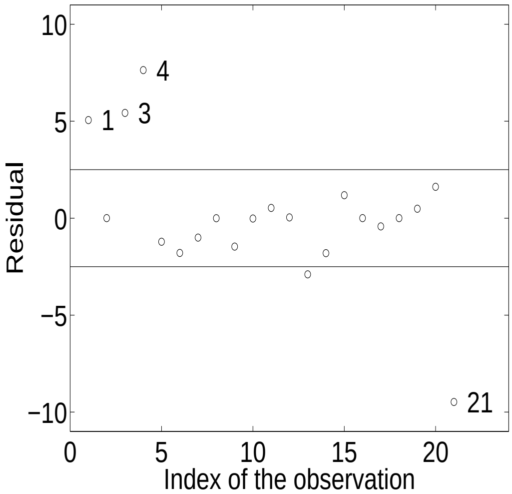

FIG. 6(b). Stack-Loss data: Residual plot associated with the *SGL* fit.

In general, most of robust estimates in TABLE II are quite similar. Comparatively, the *SGL* estimate and Andrews estimate are closer to the L_2-delete estimate.

The residual plots associated with the L_2 fit and with *SGL* fit are given by FIG. 6(a) and FIG. 6(b) respectively. It is clear that the four observations 1, 3, 4 and 21 are much more outlying in FIG. 6(b) than in FIG. 6(a).

5. CONCLUSIONS AND DISCUSSIONS

A family of adaptive robust estimates which range from L_1 to L_2 estimates is suggested. This *SGL* adaptive robust estimate family can automatically choose an adequate estimate based on sampling data, either L_1/L_2 estimates or smoothed Huber estimates. It provides a convenient approach for practitioners and also a flexible method for automatic analysis in systems identification.

It is clear that the *SGL* estimates are not panaceas for all kinds of robustness problems. It has its limitations. Since it has an exponential tails as Huber estimates, for data with extreme outliers such that its distribution has heavier tails than Laplace, the *SGL* estimates can not give better estimates than L_1 estimates. Even for error distributions with tails lighter than Laplace distribution, the *SGL* can only achieve its optimality in efficiency at the member distributions of the *SGL* family.

The *SGL* estimates are proposed as an alternative to the t estimates in this paper. It is interesting to make a comparison between them. Similar to the *SGL* estimates, it is easy to check that t estimates can reach its optimal efficiency 1 at t distribution family. Hence, one of the major differences between the *SGL* and t estimates is that they achieve their optimality at different underlying distributions. When underlying error distributions have lighter tails and are close to the normal, they have similar behaviors (note that their solutions are approximately same as both v and λ are large, see Eqs. (8) and (9)). When v and λ become small for relatively serious outliers, their underlying distributions significantly diverge: the t family approaches to the Cauchy as $v \rightarrow 1$; while the *SGL* family approaches to the Laplace as $\lambda \rightarrow 0$. Hence, as the asymptotic efficiencies indicate, each of the estimates has the superiority over another at its own underlying error distributions. For instance, the t estimates are better than the *SGL* at the Cauchy, while the *SGL* estimates are better than the t at the

Laplace. For a practical problem, which estimate is better depends on which assumption for the underlying distributions is closer to the specific problem.

APPENDIX. THE PROOF OF LEMMA 3 AND 4

Proof of Lemma 3. First, consider $h(x;\lambda)=[\text{sech}(x/c(\lambda))]^{2\lambda}$ for $c(\lambda)=2\lambda/(2\lambda+1)^{1/2}$. Letting $u=2\lambda$, we have

$$\lim_{\lambda\to 0+} h(x;\lambda)=\lim_{u\to 0+}\exp[-u\log\{\exp((u+1)^{1/2}x/u)+\exp(-(u+1)^{1/2}x/u)\}]$$

$$=\exp(-|x|)\exp[-\lim_{u\to 0+} u\log\{1+\exp(-2(u+1)^{1/2}|x|/u)\}]=\exp(-|x|)$$

Thus, the proof is completed by noticing $\lim_{\lambda\to 0+}[(1/(\sqrt{2\pi}))R(\lambda+1/2)]=1/2$.

Lemma 4. There exists a distribution $H(x)$ and ε such that

$$F_{SGL}(x;0,1,\lambda)=(1-\varepsilon)\Phi(x)+\varepsilon H(x) \qquad \text{for } 0<\varepsilon<1 \text{ and } \lambda\in[0,\infty).$$

Proof: Noting that the *SGL* has heavier tails than the normal, there exists $M>0$ such that $d(x)\hat{=}\varphi(x)/f_{SGL}(x;0,1,\lambda)<1$ for $|x|>M$. For $d(x)$ in the interval $[-M,M]$, suppose it achieves its maximum at $x=x_0\in[-M,M]$. Then we have $d(x_0)>1$ (otherwise, $\varphi(x)$ is not greater than $f_{SGL}(x;0,1,\lambda)$ everywhere and then the integral of the former integrated over $(-\infty, +\infty)$ is less than the one of the latter, which leads a contradiction). For $\varepsilon=1-1/d(x_0)$, we have $h(x)=(1/\varepsilon)[f_{SGL}(x;0,1,\lambda)-(1-\varepsilon)\varphi(x)]\geq 0$. Hence, the proof is completed by taking $h(x)$ as the density of $H(x)$ and noting $H(x)$ has total mass 1.

ACKNOWLEDGMENTS

The authors are grateful for the reviewers' valuable suggestions for the earlier version of this paper. The first author is a post-doctor of K.U.Leuven and the second author is a senior Research Associate with the FWO-Flanders. This work was supported by Sustainable Mobility Programme-Project MD/01/24 (1997-2000), Concerted Research Action GOA-MIPS, the FWO project G.0256.97: Numerical Algorithms for Subspace System Identification, the FWO

Research Communities: ICCOS and Advanced Numerical Methods for Mathematical Modeling, and the Belgian Government: Interuniversity Attraction Poles (IUAP P4-02 and P4-24).

BIBLIOGRAPHY

Balakrishnan, N. (1992). *Handbook of the Logistic Distribution*, New York: Marcel Dekker, Inc.

Daniel, C. and Wood, F.S. (1971). *Fitting Equations to Data*, New York: Wiley.

Goodall, C. (1983). "*M*-estimators of Location: an outline of the theory," in *Understanding Robust and Exploratory data analysis*, (Hoaglin, D. C., Mosteller, F. and Tukey, J. W. Eds.), New York: Wiley.

Hogg, R.V. (1967). "Some observations on robust estimation," *J. Amer. Statist. Assoc.*, **62**, 1179-86.

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

Jamshidian, M. (1997). "*t*-distribution modeling using the available statistical software," *Computational Statistics & Data Analysis*, **25,** 181-206.

Lange, K.L., Little, R.J.A. and Taylor, J.M.G. (1989). "Robust statistical modeling using the *t* distribution," *J. Amer. Statist. Assoc.*, **84,** 881-890.

Li, B. and De Moor, B. (1997). "Simple approximation relation between the symmetrical generalized logistic and the Student's *t* distributions," *Commun. Statist.-Simula.*, **26**, 1029-39.

Rousseuw, P. J. and Leroy, A. M. (1987). *Robust Regression & Outlier Detection*, New York: Wiley.

Weisberg, S. (1980). *Applied Linear Regression,* New York: Wiley.

Yale, C. and Forsythe, A.B. (1976). "Winsorized regression," *Technometrics*, **18**, 291-300.

Yuh, L. and Hogg, R.V. (1988). "On adaptive *M*-regression," *Biometrics*, **44,** 433-445.