Katholieke Universiteit Leuven

Departement Elektrotechniek

ESAT-SISTA/TR 97-115

A FAMILY OF ADAPTIVE M-ESTIMATES BASED ON THE SYMMETRICAL GENERALIZED LOGISTIC REGRESSION

Baibing Li and Bart De Moor[2]

December 1997

[1]This paper is also available by anonymous ftp from ftp.esat.kuleuven.ac.be as file rep97115.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 M-ESTIMATES BASED ON THE SYMMETRICAL GENERALIZED LOGISTIC REGRESSION

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; linear/nonlinear regression; *M*-estimation; symmetrical generalized logistic distribution.

ABSTRACT

A family of adaptive *M*-estimates based on the symmetrical generalized logistic (*SGL*) distribution, adaptive *SGL*-estimate, is suggested. This family covers broad class of known *M*-estimates, varying from L_1 to L_2 estimates and including smoothed Huber *M*-estimate for its shape parameter $\lambda\in[0,+\infty)$. For large and moderate λ, the adaptive *SGL*-estimate can be used as an alternative of *t*-estimate since they are almost equivalent. For small λ, the adaptive *SGL*-estimate is very close to L_1 estimate due to its asymptotic property for extremely small λ. For unknown underline error distributions with different types of tails in practice, the adaptive *SGL*-estimate can automatically select appropriate distribution from its family for regression. All of its technical parameters needed in *SGL*-regression are produced simultaneously with the estimates of regression coefficients based on samples, rather than determined a prior. This is particularly useful for automatic analysis in systems identification and is also very convenient for practitioners since they do not have to firstly select a certain method among various *M*-regression procedures and then tune its technical parameters. Finally, several practical cases are examined.

1. INTRODUCTION

A lot of attention has been paid to develop robust estimates for regression analysis in recent decades due to the fact that least squares (L_2 norm) 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 most important reasons for this is not that practitioners are not convinced by its importance, but most of practitioners often do not know which particular method should be used among so many different *M*-estimate procedures, how to tune technical parameters needed in *M*-estimates, and why (see also, Lange et al., 1989).

Another related problem in practice is for automatic analysis in systems identification in engineering. It is often desired that methods used in systems identification have reasonable flexibility for wide range of types of data, i.e., they are desired to have ability to automatically adapt to different types of practical data. However, robust procedures designed for a certain kind of data types with pre-determined tuning parameters usually lack this kind of flexibility.

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 underline distributions are symmetrical and light-tailed, and using $t(3)$ if symmetrical and heavy-tailed. The simulation study given by them showed that, for seven symmetrical distributions with different tail-shapes and middle-shapes, the *AT*-estimate is the most efficient one among all of compared *M*-estimates including Huber's and Turkey's.

Another closely related approach, *t*-estimate, is to estimate regression coefficients based on *t* distribution as error model with any of degrees of freedom $v \geq 1$, where v is determined by samples. Lange et al (1989) gave an excellent 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*-estimate for two examples.

According to Lange et al. (1989), the reason that they chose *t* rather than other distribution family mainly lies on grounds of familiarity and computational simplicity. The purpose of this paper is then to consider a new family of estimates with the same computational simplicity, as an alternative of *t*-estimate, and meanwhile covering many important types of *M*-estimates. Specifically, it is desired that this new family of *M*-estimates can not only cover L_2 estimate and be close to *t*-estimate for lighter and mild tailed error distributions, but also can be extended to L_1 estimate for relatively serious outliers. We show that *SGL* distribution family with shape parameter $\lambda\in[0,+\infty)$ is a very ideal candidate for these requirements.

First, Li and De Moor (1997) showed that *t* distribution with degrees of freedom v is very close to *SGL* with $\lambda=v/6$ when both λ and v are large (the maximum absolute distance between them is less than 4.8×10^{-3} for $v\geq4$, 5.5×10^{-4} for $v\geq10$, and 6.4×10^{-5} for $v\geq30$). Both of them will approach to normal distribution as v and λ become to positive infinite. Hence, for lighter and mild outliers, *SGL* can behave approximately as *t*. Second, we shall show in this paper that when λ becomes to zero, *SGL* will approach to Laplace distribution. This means that when λ increases from zero to positive infinite, the adaptive *SGL*-estimates will vary from L_1 to L_2 estimates. Finally, for any intermediate value of λ, the *SGL*-estimate is equivalent to a smoothed version of Huber estimate. Therefore, the family of adaptive *SGL*-estimates has reasonable flexibility to deal with many types of practical problems and meanwhile provides a simple and unified framework in concept compared with other *M*-estimates. Practical examples show that the *SGL*-estimate indeed can automatically select adequate parameter λ and then adopt adequate distribution shape for different tailed error distributions.

In section 2 we introduce adaptive *SGL*-estimate. Section 3 discusses implementation of *SGL*-estimate. Practical cases are examined in section 4. Section 5 state conclusions and discussions.

2. ADAPTIVE *SGL*-ESTIMATE

Consider the density of the *SGL* distribution (denoted as $SGL\{\mu,\sigma,\lambda\}$)

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

the density of the *t* distribution (denoted as $t\{\mu,\sigma,v\}$)

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

and the density of the Laplace distribution (denoted as $L\{\mu,\sigma\}$)

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

Denote their cumulative distribution functions as $F_{SGL}(x;\mu,\sigma,\lambda)$, $F_t(x;\mu,\sigma,v)$ and $F_L(x;\mu,\sigma)$ respectively, where $\Gamma(z)$ represents the gamma function and $cosh(x)=[exp(x)+exp(-x)]/2$.

The basic idea of this paper is based on the following three lemmas.

LEMMA 1. (Davidson, 1980, see Balakrishnan, 1992, pp212) $SGL\{0, (2\lambda)^{-1/2},\lambda\}$ approaches to normal $N(0,1)$ as $\lambda\to+\infty$.

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

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

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

The proof of the Lemma 3 will be given in **APPENDIX.**

For data $y_1,\ldots,y_n$ independently sampled from a symmetrical distribution, consider *SGL* regression for which we use the assumption of *SGL* model $y_j \sim SGL\{\mu_j(\theta),\sigma,\lambda\}$ to replace following three assumptions about error distributions:

(I) the normal distribution assumption $y_j \sim N\{\mu_j(\theta),\sigma^2\}$

(II) the *t* distribution assumption $y_j \sim t\{\mu_j(\theta),\sigma,v\}$ (Lange et al, 1989).

(III) the Laplace distribution assumption $y_j \sim L\{\mu_j(\theta),\sigma\}$

The rationale for this is directly from Lemma 1-3. An advantage of the *SGL* regression model is that it gives a clear statement of assumptions on distribution.

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,

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

for given $n\times p$ design matrix $X=[x_1,...,x_n]^T$.

Hence, from the assumption of *SGL* model $y_j \sim SGL\{\mu_j(\theta),\sigma,\lambda\}$, the estimation problem for *SGL* regression is to obtain the *ML* estimates of (θ,σ,λ) by maximizing the log-likelihood function

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

$$=nR(\lambda)\text{-}nlog(\sigma)\text{-}nlog(2\pi^{1/2})\text{-}2\lambda\sum_{j=1}^{n}\log(\cosh(r_i)) \quad (2)$$

where $R(x)=log[\Gamma(x+1/2)/\Gamma(x+1)]$ and $r_j=[y_j\text{-}\mu_j(\theta)]/(2\sigma\lambda)$.

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

$$\partial G/\partial\theta=\sigma^{-1}\sum_{j=1}^{n}\tanh(r_j)\,\partial\mu_j(\theta)/\partial\theta \quad (3)$$

$$\partial G/\partial\sigma=(n/\sigma^2)[(1/n)\sum_{j=1}^{n}(y_j-\mu_j(\theta))\tanh(r_j)\text{-}\sigma] \quad (4)$$

$$\partial G/\partial\lambda=nR'(\lambda)\text{-}2\sum_{j=1}^{n}\log(\cosh(r_j))+2\sum_{j=1}^{n}r_j\tanh(r_j) \quad (5)$$

where $R'(\lambda)=\partial R(\lambda)/\partial\lambda=\psi(\lambda+1/2)\text{-}\psi(\lambda+1)$, $\psi(x)$ is digamma function, and $tanh(x)=[exp(x)\text{-}exp(\text{-}x)]/[exp(x)+exp(\text{-}x)]$.

Particularly, for linear regression problems (1), we can obtain “normal equations” from (3):

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

From Eq. (6), we obtain following properties for its solution θ:

(i) It is clear that the solution θ of Eq. (6) is an *M*-estimate of regression coefficients.

(ii) For large $(\lambda\sigma)$ such that r_j is small enough, we have $tanh(x/(2\lambda\sigma))\sim x/(2\lambda\sigma)$. Then Eq. (6) approximately equals to

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

which is the normal equations of L_2 regression. Therefore, for large $(\lambda\sigma)$, the *SGL*-estimate of regression coefficients is an approximation of L_2 estimate.

(iii) For small λ and fixed σ such that r_j is large enough, we have $\tanh(x)\sim\text{sign}(x)$. Then Eq. (6) is approximately equivalent to

$$\sum_{j=1}^{n}|y_j-x_j{}^T\theta|=0$$

which is the normal equations of L_1 regression. Therefore, for small λ, the *SGL*-estimate of regression coefficients is an approximation of L_1 estimate.

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

(iv) For large ($\lambda\sigma$), from (4) and Taylor' theorem

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

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

$$\sum_{j=1}^{n} x_j e_j [1-(6\lambda)^{-1}\{e_j^2/(n^{-1}\sum_{j=1}^{n} e_j^2)\} + O(1/(\lambda\sigma)^4]=0 \qquad (8)$$

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

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

where $e_j=y_j-x_j^T\theta$.

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

(iv) For any intermediate value of λ, fixed σ and $K>0$, $\phi(x)=tanh(x/(2\lambda\sigma))$ is a smoothed version of Huber function

$$\phi_{\mathrm{Huber}}(x,K)=\begin{cases}-1 & \text{for } x<-K \\ x/K & \text{for } |x|\le K \\ 1 & \text{for } x>K\end{cases}$$

Figure 1 gives plot of Huber function with K=1.25 and the function $y=tanh(x/(2\lambda\sigma))$ with λ=0.55 and σ=1 for *SGL*-estimate. It can seen that they are very close each other except in the neighborhoods of the two unsmoothing points $x=\pm K$.

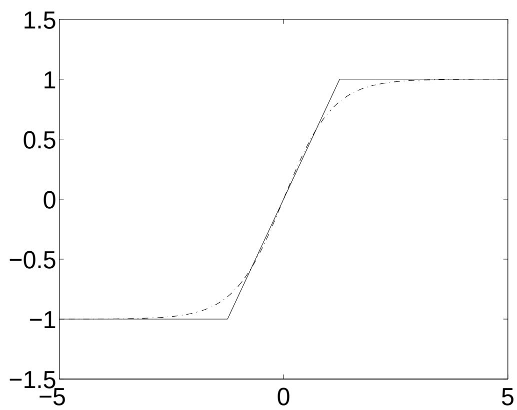

Figure 1. Plot of $\phi_{\mathrm{Huber}}(x;1.25)$ (—) and $tanh(x/1.1)$ (-·-·)

Hence, the solution of Eq. (6) is approximately equivalent to Huber *M*-estimates.

All of the above properties means that the *ML* estimate θ of *SGL* model covers broad class of well known *M*-estimates. More importantly, the technical parameters λ and σ are determined in an adaptive way based on samples, not needed to be set a prior. This makes it possible for automatic analysis in systems identification for different types of data and is obviously very convenient for practitioners.

Next, we briefly discuss the properties of *ML* estimate of σ. From (4), it satisfies

$$\sigma=(1/n)\sum_{j=1}^{n}[\{y_j-\mu_j(\theta)\}\tanh(r_j)\,] \tag{10}$$

Therefore, when $\lambda\to 0$, we have $\sigma\to(1/n)\sum_{j=1}^{n}|y_j-\mu_j(\theta)|$ which equals to the L_1 estimate for σ; while $\lambda\to+\infty$, we have $2\lambda\sigma^2\to(1/n)\sum_{j=1}^{n}[y_j-\mu_j(\theta)]^2$ which is the L_2 estimate for σ.

3. IMPLEMENTATION OF *SGL*-ESTIMATE

3.1. Algorithms

For general optimization problems with bound constraints like λ>0 and σ>0, one of frequently used methods is the constrained quasi-Newton algorithm which can guarantee superlinear convergence. This method is adopted by many software such as MATLAB (optimization toolbox).

As another approach, we suggest using following log-transforms to convert the constrained optimization problem into the unconstrained one:

$u=log(\lambda)$ and $w=log(\sigma)$

and then directly adopt unconstrained optimization methods.

For analytical methods (for instance, quasi-Newton) of this approach, Hesian matrix may become singular in iterations. For these cases, we recommend to use the simplex algorithm of Nelder and Mead (see, Seber and Wild, 1989).

Our experiences show that optimizations are not too sensitive for selections of initial values of (θ,σ,λ) in linear *SGL* regression. Hence, we can choose, for instance, $\lambda=\sigma=1.0$ and θ as L_2 estimate for initial values.

3.2. Bounds for parameter λ

One problem in the above approaches is that sometimes convergence is quite slow, especially in the situations in which λ becomes very small or large. In this subsection we consider some reasonable bounds for parameter λ since for relatively large (small) λ, the *SGL* will close to normal (Laplace) distribution satisfactorily.

First, we state some numerical results. Consider the distance between two distributions $F_1(x)$ and $F_2(x)$ as follows:

$$D\{F_1(x),F_2(x)\} = Max\{\underset{x\in S}{Max}\left|F_1(x)-F_2(x)\right|, F_1(4.3), F_2(4.3)\}$$

$$\text{for } x\in S=\{-4.3+0.005k \mid k=0,1,\cdots,860\}$$

which is a reasonable approximation of Kolmogorov-Smirnov distance (note that all of the density functions, *SGL*, normal, Laplace and Student's *t*, are symmetrical and decreasing for $x>0$). Denote $F_N(x)$ as the normal cumulative distribution.

Result (i) $D\{F_{SGL}(x;0,[3/(v+1)]^{1/2},v/6), F_t(x;0,1,v)\}<0.0048$ for $4\le v\le 40$

(ii) $D\{F_{SGL}(x;0,1,\lambda), F_L(x;0,1)\}<0.0068$ for $\lambda=0.12$

(iii) $D\{|F_{SGL}(x;0,(2\lambda)^{-1/2},\lambda)-F_N(x)\}<0.0062$ for $\lambda=7$

Since 0.005 is a reasonable accuracy for approximations between distributions, the above result suggests an acceptable lowest upper bound and highest lower bound for λ. Hence, if we use bounds

$$\lambda_{min}<\lambda<\lambda_{max} \tag{11}$$

and take $\lambda_{min}=0.12$, $\lambda_{max}=7$, the differences between *SGL* and normal, *SGL* and Laplace, and *SGL* and *t* distributions with $v\ge 4$, will be less than 0.0068.

In practice, the lower bound can be set as $\lambda_{min}=10^{-m}$ and the upper bound as $\lambda_{max}=10^{m}$ (m=1,2,3...) for different requirements for precision.

For the problem of maximizing log-likelihood function (2) with bound constraint (11), we can introduce following transform to eliminate bound constraints:

$$u=athan((\lambda-\lambda_a)/\lambda_b) \qquad \lambda_a=(\lambda_{\max}+\lambda_{\min})/2, \quad \lambda_b=(\lambda_{\max}-\lambda_{\min})/2 \tag{12}$$

and then use unconstrained optimization methods, where *atanh*(x) is the inverse hyperbolic tangent function.

For nonlinear *SGL* regression problems, to reduce possible complexity in optimizations for too small or large λ, it is recommended to set appropriate lower and upper bounds (11) and use transform (12) to search optimal solutions. Moreover, just as in the ordinary nonlinear LS regression, it is important for selections of initial value of (θ,σ,λ) in nonlinear *SGL* regressions. Sometimes trying several different initial values is necessary for optimization.

3.3. Using available statistical software

Jamshidian (1997) considered encoding of *t*-estimate using BMDP and SAS. This is very helpful for practitioners before it is finally absorbed into some well known software. Since the way for adaptive *SGL*-estimate is quite similar, the programs given by Jamshidian can be used for *SGL*-estimate by only minor modifications. By the way, besides the special square-root method for constituting objective function given by Jamshidian (1997), we can also adopt _LOSS_ to define objective functions for optimization problems which is provided by SAS NLIN for general non-linear optimizations (such as maximizing likelihood). See SAS reference books for details.

4. PRACTICAL EXAMPLES

Five practical examples are examined in this section to investigate flexibility of the adaptive *SGL*-estimate for different underline error distributions with light, mild and heavy tails, and 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 and explanatory variables are the acid content determined by titration and organic acid content determined by extraction and weighing.

The scatter plot given by Fig. 2(a) suggests a linear statistical relation between response variable (y) and explanatory variable (x). The *SGL*-estimate gives its fit as:

$\hat{y}$=35.4581+0.3216x

with $\hat{\lambda}$=6.6×10^6. Since from Lemma 1, the *SGL* distribution approaches to the normal distribution as λ becomes to positive infinite, the estimate $\hat{\lambda}$=6.6×10^6 actually suggests that at this case, the adaptive *SGL*-estimate automatically adopts normal distribution as the underline error distribution to produce an L_2 estimate. For comparison, we list the L_2 fit as follows:

$\hat{y}$=35.4583+0.3216x

It can be seen that both of them are almost completely same.

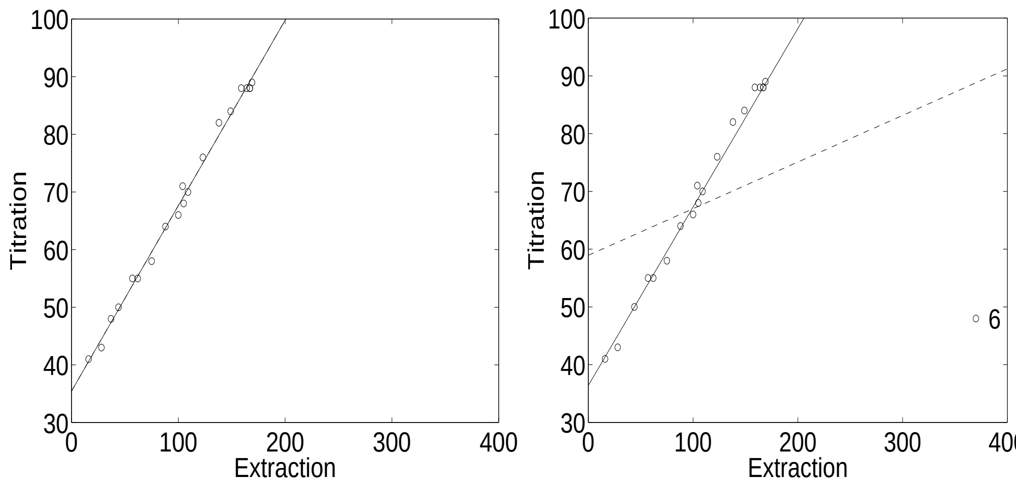

Figure 2(a). Pilot-Plant data with L_2 fit (---) and *SGL* fit(—)

Figure 2(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. The scatter plot is given by Fig.2(b). With this contaminated data, the L_2 fit is

$\hat{y}$=58.9388+0.0807x

which, from Fig. 2(b), is strongly attracted by this single outlier.

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

$$\hat{y}=36.4065+0.3089x \tag{13}$$

with $\hat{\lambda}$=6.5×10^{-12} which lies close to the L_2 fit applied to the uncontaminated data, and then is robust on the outlier, the 6*th* observation. Since from Lemma 3, the *SGL* approaches to the Laplace distribution as λ→0, the estimate $\hat{\lambda}$=6.5×10^{-12} actually suggests that at this situation, the adaptive *SGL*-estimate automatically produces an L_1 estimate. This can be verified by comparing (13) with following L_1 fit:

$$\hat{y}=36.4030+0.3090x.$$

In next example we consider a nonlinear regression problem which contains mild outliers among the 42 observations according to Lange and Sinshemer (1993).

***Example* 2.** (Lepidium data). This set of data was from Racine-Poon (1988) and also considered by Jamshidian (1997) and Lange and Sinshemer (1993). The response and explanatory variables are the growth of the plant lepidium measured as log weight (y) and soil concentration of an agrochemical (x). The response function is given by Racine-Poon (1988) as

$$\mu_j(\theta, x_j) = \begin{cases} \theta_1 / (1+\exp(\theta_2 + \theta_3 \log(x_j) + \theta_4 (\log(x_j))^2 & \text{for } x_j > 0 \\ \theta_1 & \text{for } x_j = 0 \end{cases}$$

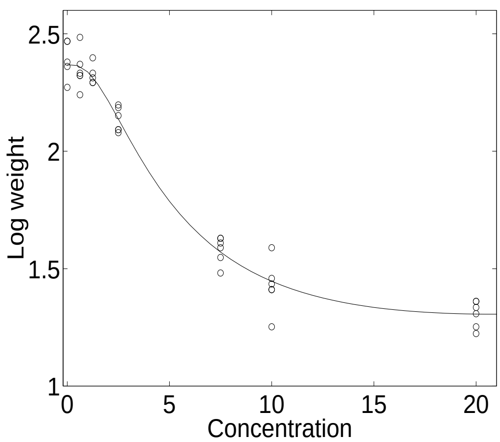

Figure 3. Lepidium data with *SGL* fit(—)

The *SGL*-estimate gives

$\hat{\theta}_1$=2.3607 $\hat{\theta}_2$=-4.3479 $\hat{\theta}_3$=2.7170 $\hat{\theta}_4$=-0.4473 with $\hat{\lambda}$=0.3583 which suggests there exist mild outliers. This can be verified by Fig . 3 for scatter plot and the *SGL* fit. Moreover, it can seen that the patterns of the outliers are symmetrical in some extents and then their effects are partly offset.

We give the nonlinear L_2 estimate for comparison:

$$\hat{\theta}_1\text{=2.3642} \quad \hat{\theta}_2\text{=-4.2332} \quad \hat{\theta}_3\text{=2.6280} \quad \hat{\theta}_4\text{=-0.4282}$$

As we expect, in this case the *SGL*-estimate is close to the L_2 estimate since the outliers almost do not matter.

Next, we consider two examples with more complicated outlier patterns.

***Example* 3.** (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 6,16,25 lying far from the others. The L_2 estimate gives its fit as

$$\hat{y}=1.110+0.496x$$

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

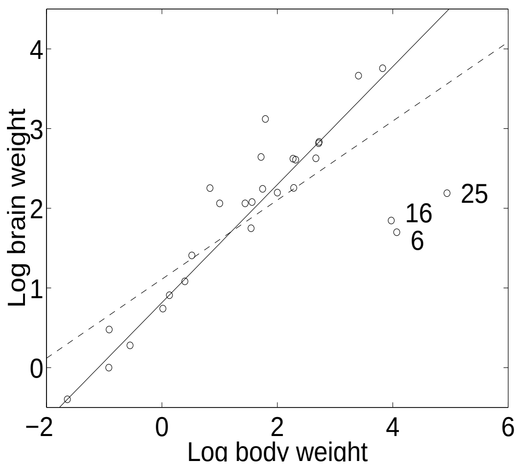

Figure 4. Brain and weight data with L_2 fit (---) and *SGL* fit(—)

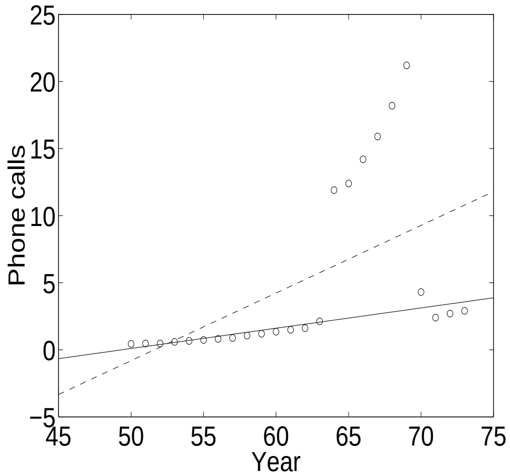

Figure 5. International phone call data with L_2 fit (---) and *SGL* fit(—)

The *SGL* estimate gives

$$\hat{y}=0.815+0.741x$$

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 that they do not obey the same trend as the one followed by the majority of the data.

***Example* 4.** (International phone call data) This set of data was from Belgian Statistical Survey and was firstly investigated by Rousseeuw and Leroy (1987). The data contain the total number (as response variable y) (in tens of millions) of international phone calls made each year (as explanatory variable x).

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

The L_2 estimate gives its fit as

$$\hat{y}=-26.006+0.504x$$

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

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

$$\hat{y}=-7.466+0.151x$$

which is quite successful in avoiding the outliers.

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

***Example* 5.** (Stack-Loss Data). The stack-loss data set was firstly given by Browlee and has become classical for testing robust estimation procedures. The response variables is stackloss (y), and three explanatory variables are rate of operation (x_1), the cooling water inlet temperature (x_2) and the acid concentration (x_3) (see, for instance, 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

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

Table 1 lists the *SGL*-estimate for regression coefficients and the ones by using other robust methods including Huber's estimate, Andrews' estimate, *t*-estimate (see, Lange et al., 1989) and LMS estimate (see Rousseeuw and Leroy, 1987). 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).

In general, most of robust estimates in Table 1 are quite similar. Comparatively, the *SGL*-estimate is closer to the L_2-delete estimate and Andrews' estimate.

Table 1. Estimates of regression coefficients for stack-loss data from six methods

	SGL	L_2-delete	Andrews†	Huber†	*t*†	LMS‡
$\hat{\theta}_0$	-39.69	-37.65	-37.20	-41.00	-38.50	-34.5
$\hat{\theta}_1$	0.83	0.80	0.82	0.83	0.85	0.71
$\hat{\theta}_2$	0.57	0.58	0.52	0.91	0.49	0.36
$\hat{\theta}_3$	-0.06	-0.07	-0.07	-0.13	-0.07	0.00

† *Source*: Lange et al. (1989).
‡ *Source*: Rousseeuw and Leroy (1987).

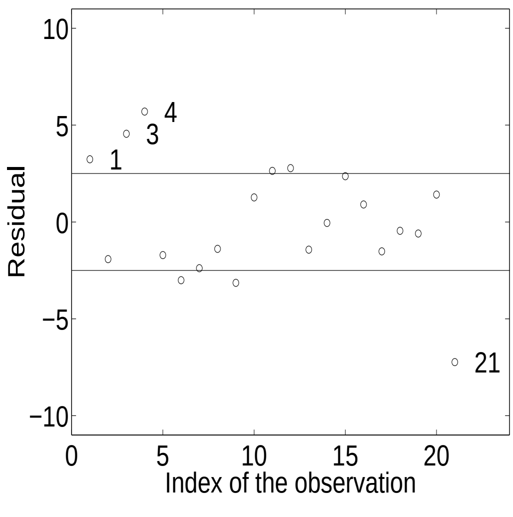

Figure 6(a). Stack-Loss data: Residual plot associated with the L_2 fit.

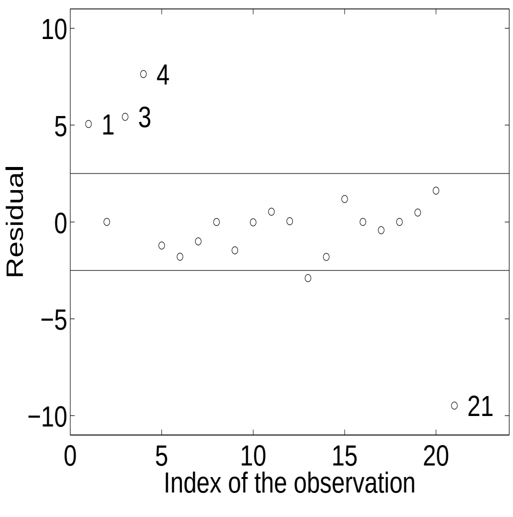

Figure 6(b). Stack-Loss data: Residual plot associated with the *SGL* fit.

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

In practice, facing "the bewildering choice of alternative (robust) procedures" (Lange et al.,1989), practitioners often do not know which one should be chosen and why. Therefore, it is desired to introduce a unified approach which can automatically determine (based on samples) an appropriate robust procedure. Moreover, automatic analysis in systems identification also demands the methods used in engineering have enough flexibility for different types of data. These are main motivations of this paper.

We believe that "adaptive *M*-estimate" approach, firstly proposed by Hogg (1967), is one of feasible and noticeable solutions for these kinds of problems. As we see in this paper, as an adaptive family, *SGL* estimate has satisfactory range to cover many important *M*-estimates—from L_1 to L_2 estimates, and approximately including Huber estimate and *t*-estimate. Meanwhile, practical examples show that, for underline error distributions with different types of tails, the *SGL* indeed has ability to automatically choose an appropriate distribution shape for data.

As a straightforward extension, the *SGL* family can be replaced by a more general one, the Type IV generalized logistic distribution family, which is skewed in general and includes *SGL* as its special case (see, Balakrishnan, 1992). This means that generalized logistic distribution family can provide a unified way to deal with both symmetrical and asymmetrical error distributions. However, when underline error distributions are known a prior as symmetrical ones, we suggest using *SGL*.

Finally, it should be noted that adaptive *SGL* estimate is not a panacea for all kinds of robustness problems. It has its limitations. It can only approximately cover some of distributions with tails longer than normal and shorter than Laplace distributions. Hence, for data with tails shorter than normal, it can not give better estimate than L_2 estimate; for data with extreme outliers such that its tails are longer than Laplace distribution, it can not give better estimate than L_1 estimate. Even for error distributions with tails longer than normal and shorter than Laplace distributions, the *SGL* can only give approximations of them in general. Therefore, this paper is just to provide a feasible and convenient alternative approach for practitioners to deal with robustness problems.

APPENDIX. THE PROOF OF LEMMA 3

First, consider $h(x;\lambda)=[cosh(x/(2\lambda))\,]^{-2\lambda}$. Letting $u=2\lambda$, we have

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

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

Thus, the proof is completed by noticing following result for normalization coefficient

$$\lim_{\lambda\to 0+} \Gamma(\lambda+1/2)/[2\pi^{1/2}\Gamma(\lambda+1)]=1/2.$$

ACKNOWLEDGEMENTS

The first author of this paper is a research fellow (post-doctor) of K.U.Leuven and the second author is a senior Research Associate with the FWO-Flanders. This work was supported by the Fellowship of K.U.Leuven (F97-72), 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.

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

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.

Lange, K.L. and Sinsheimer, J.S. (1993). “Normal/independent distributions and their applications in robust regression,” *J. Comput. Graphical Statist.*, **2**, 175-198.

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.

Racin-Poon, A. (1988). “A Bayesian approach to nonlinear calibration problems,” *J. Amer. Statist. Assoc*., **83,** 650-656.

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

Seber, G.A.F. and Wild, C.J. (1989). *Nonlinear Regression*. 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.