Novel Approximation Estimators For Mixed Noise in Metrology

D. P. Jenkinson*, J. C. Mason*, A. Crampton*
& M. G. Cox**, A. B. Forbes**, R. Boudjemaa**

*School of Computing and Engineering,
University of Huddersfield, UK.

e-mail:
{d.p.jenkinson, j.c.mason, a.crampton}@hud.ac.uk.

**Centre for Mathematics and Scientific Computing,
National Physical Laboratory, Teddington, UK.

e-mail:
{maurice.cox, alistair.forbes,
redouane.boudjemaa}@npl.co.uk

Introduction

- A data contamination example concerns dust collections on the surface of a co-ordinate measuring machine (CMM). Data errors are a mixture of normally distributed errors and outliers.
- A least squares approximation may not be the most accurate form of approximation - the l_2 norm is susceptible to outliers. A "mixed" norm (eg $\ell_2 + \ell_1$ or $\ell_2 + \ell_0$) may be better.
- A robust estimator can be applied to solve the problem as a nonlinear "transferred least squares" (TLS) which can reduce the outlier effects.

Aims of Research

- Applies an estimator to fit polynomial and radial basis function (RBF) approximations to data with predominantly l_2 noise, but where some outliers are present in the data.

- Extends the estimator

$$G = \frac{\epsilon}{(1+\epsilon^2)^{\frac{1}{2}}},$$

suggested by Maurice Cox (NPL 1999 - private communication), where ϵ represents the error (residuals), to solve a "transferred least squares" (TLS) approximation problem.

The estimator treats small errors as themselves, but replaces large errors by constant values (eg 1 for G above).

Well Established Estimators: Huber

$$G(\epsilon) = \begin{cases} \epsilon^2, & \text{for} \quad |\epsilon| \leq c \\ 2c|\epsilon| - c^2, & \text{for} \quad |\epsilon| > c \end{cases}$$

1) is continuously differentiable

2) $G \approx \epsilon^2$ for small ϵ, (ℓ_2) (parabola)
$G \approx |\epsilon|$ for large ϵ, (ℓ_1) (straight line)

Further Work

Estimators currently under investigation are

1. $G = \tanh(c\epsilon)$
2. $G = \frac{\epsilon}{(1+c^2\epsilon^2)^{\frac{1}{2}}}$
3. $G = \frac{2}{\pi}\arctan\left(\frac{\pi c\epsilon}{2}\right)$
4. $G = 1 - \exp(-c|\epsilon|)$
5. $G = \sqrt{(1 - \exp(-c^2\epsilon^2))}$

and work is continuing at both institutions.

All satisfy
$G \approx \epsilon$ for small ϵ
$G \approx$ *constant* for large ϵ.

General Forms of Approximation

1. Polynomial

$$F = \sum_{j=1}^{n} b_j x^{j-1}$$

   or better

$$F = \sum_{j=1}^{n} b_j T_{j-1}(x)$$

   where $T_j(x)$ is a Chebyshev polynomial of degree j given by $T_j(x) = \cos(j\theta)$ for $x = \cos(\theta)$.

2. Radial Basis Function (RBF)

$$F = \sum_{j=1}^{n} b_j \phi(\| \mathbf{x} - \lambda_j \|)$$

   - b_j are the solution parameters.
   - ϕ is a univariate basis function.
   - $\{\lambda_j\}_{j=1}^{n}$ are a set of fixed centres.
   - $\mathbf{x}$ is the input abscissa vector.

Norms

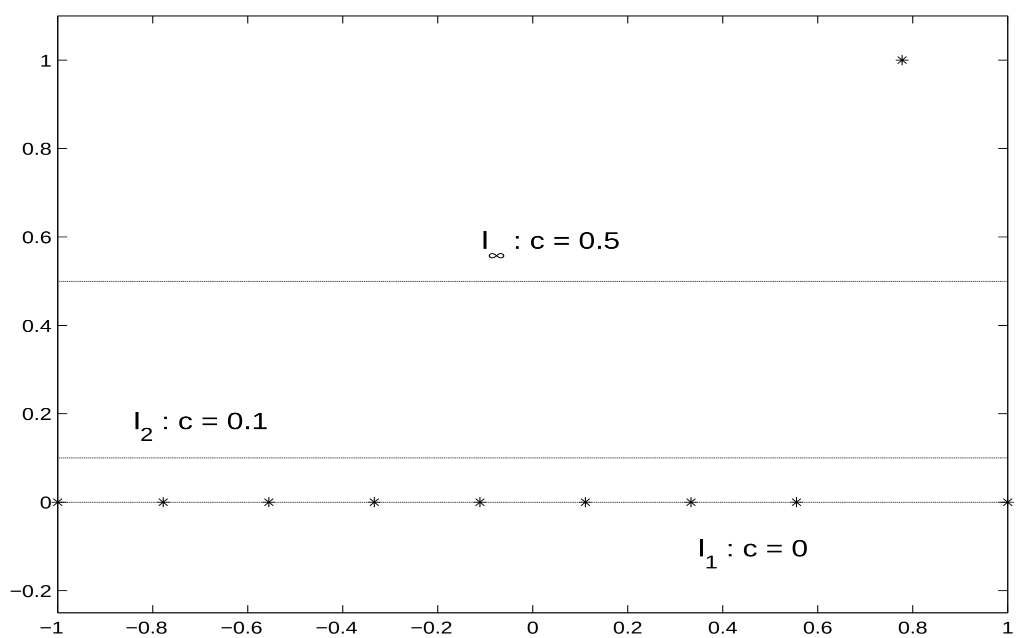
Approximation: $\quad F \approx f$ at $x = x_i$

$$l_1 : \quad \|f - F\|_1 = \sum |f(x_i) - F(x_i)| = \sum |\epsilon_i|$$

$$l_2 : \quad \|f - F\|_2 = \sqrt{\sum [f(x_i) - F(x_i)]^2} = (\sum \epsilon_i^2)^{\frac{1}{2}}$$

$$l_\infty : \quad \|f - F\|_\infty = \max |f(x_i) - F(x_i)| = \max |\epsilon_i|$$

Best (constant c) approximations to simple example data set:



Estimator $G(\epsilon) = \frac{\epsilon}{(1+\epsilon^2)^{\frac{1}{2}}}$

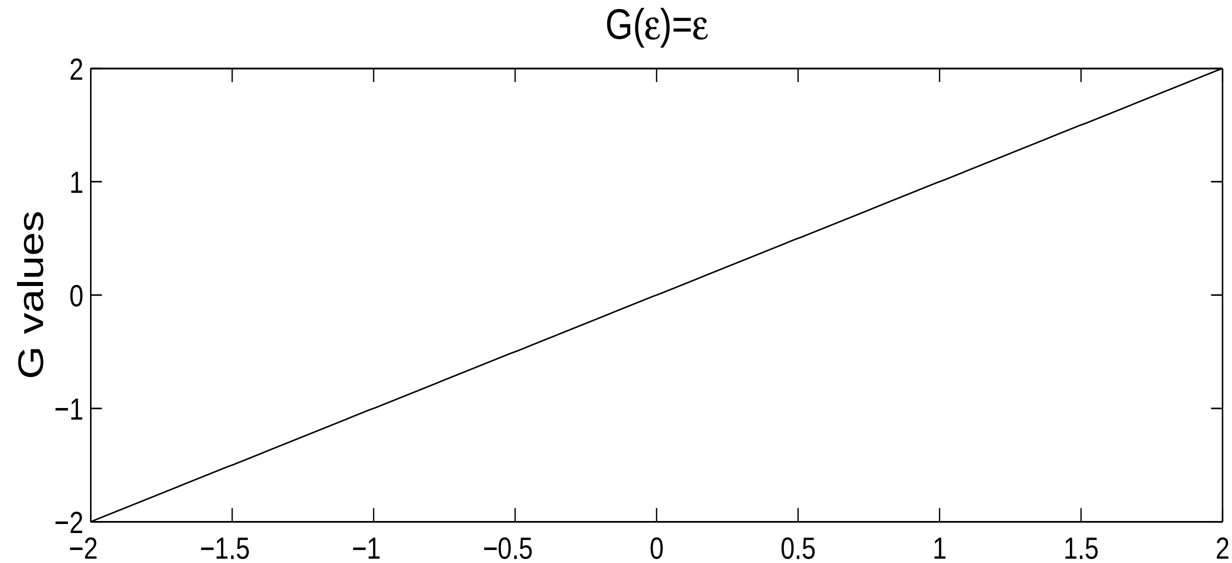
$$
\begin{aligned}
s &= \frac{(c-0)^2\,9}{1+(c-0)^2} + \frac{(c-1)^2\,1}{1+(c-1)^2} \\
&= 9\left(1-\frac{1}{1+c^2}\right) + \left(1-\frac{1}{1+(c-1)^2}\right)
\end{aligned}
$$

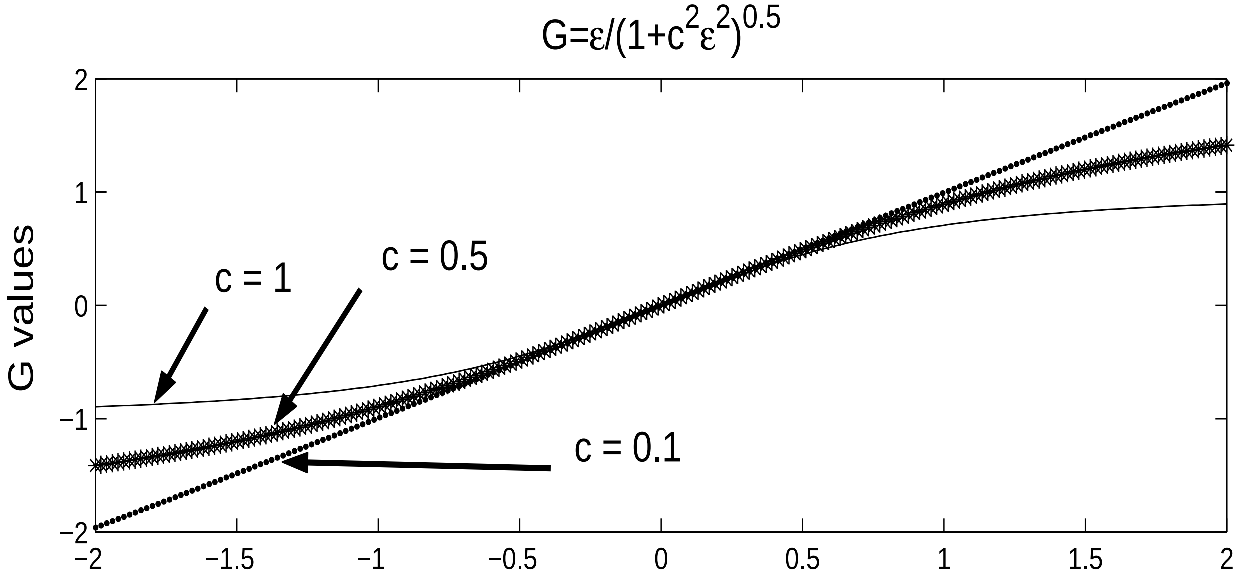
Taking the differential

$$
\begin{aligned}
M(c) &= \frac{ds}{dc} \\
&= \frac{9\,(2c)}{(1+c^2)^2} + \frac{2\,(c-1)}{\left(1+(c-1)^2\right)^2} \\
M(0) &= -\frac{1}{2}, \\
M(0.1) &= \frac{1.8}{1.01^2} + \frac{2\,(-0.9)}{1.81^2} = 1.22 > 0.
\end{aligned}
$$

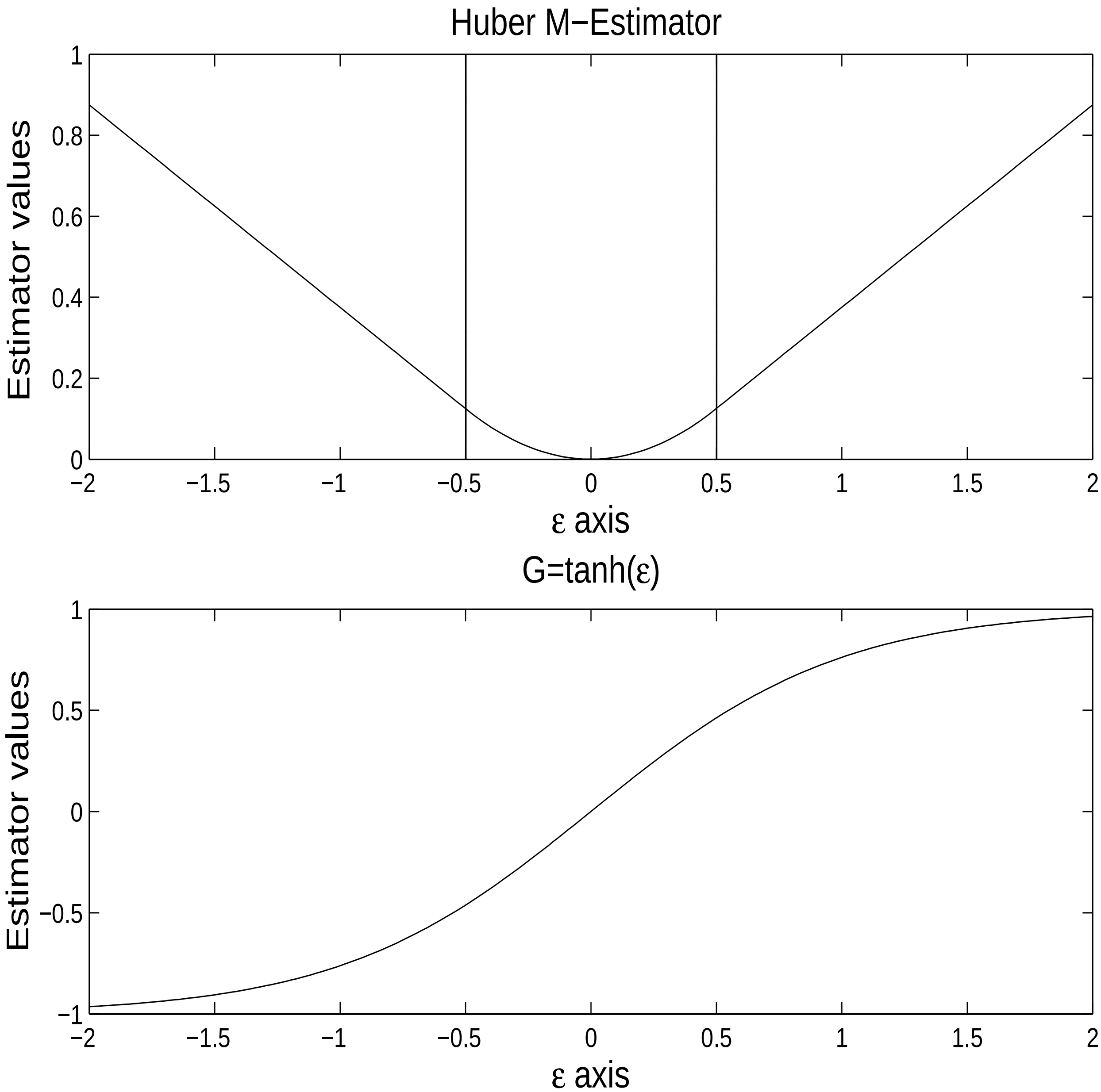
Maximum $(M=0)$ lies between $c=0$ and $c=0.1$ at approximately $c=0.03$.
So estimator is a compromise between l_1 and l_2.

Estimator Representations





Estimator Representations



Least squares

Least squares approximations take the form

$$\min_{\mathbf{b}} \sum_{i=1}^{m} \epsilon_i^2 (\mathbf{b})$$

where

- $\mathbf{b}$ is a vector of solution parameters $(b_1, b_2, \ldots, b_n)^T$
- ϵ is the approximation error (residual)

We extend least squares to include TLS by

$$\min_{\mathbf{b}} \sum_{i=1}^{m} [G(\epsilon)]^2 .$$

Iteratively Weighted Least Squares

We wish to minimise the l_2 norm

$$\sum_{i=1}^{m} \left[G\left(\epsilon_i\right)\right]^2$$

where

$$G\left(\epsilon\right) = \frac{\epsilon}{\left(1 + c^2\epsilon^2\right)^{\frac{1}{2}}}$$

and $\epsilon = \mathbf{f} - \mathbf{F}$.

Iterating over k, taking $F^{(k)}$ to be the k th approximation to $\mathbf{f}$, we minimise at step k

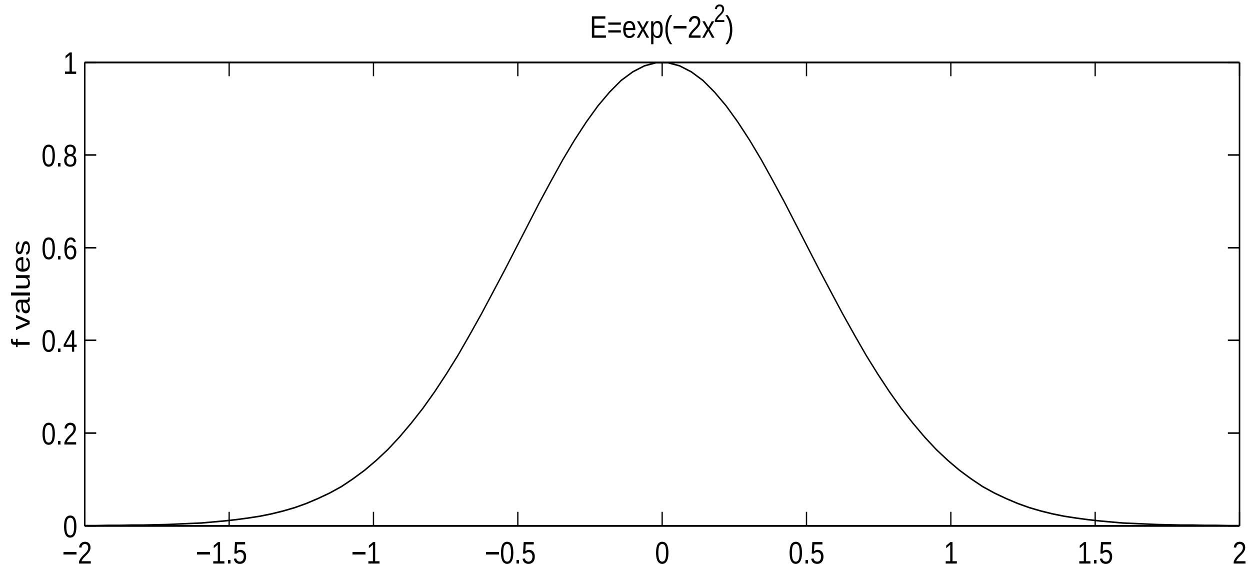
$$\sum \left[\epsilon_i^{(k+1)} \left(\frac{G\left(\epsilon_i\right)^{(k)}}{\epsilon_i^{(k)}}\right)\right]^2 . \quad (k = 0, 1, 2, \ldots)$$

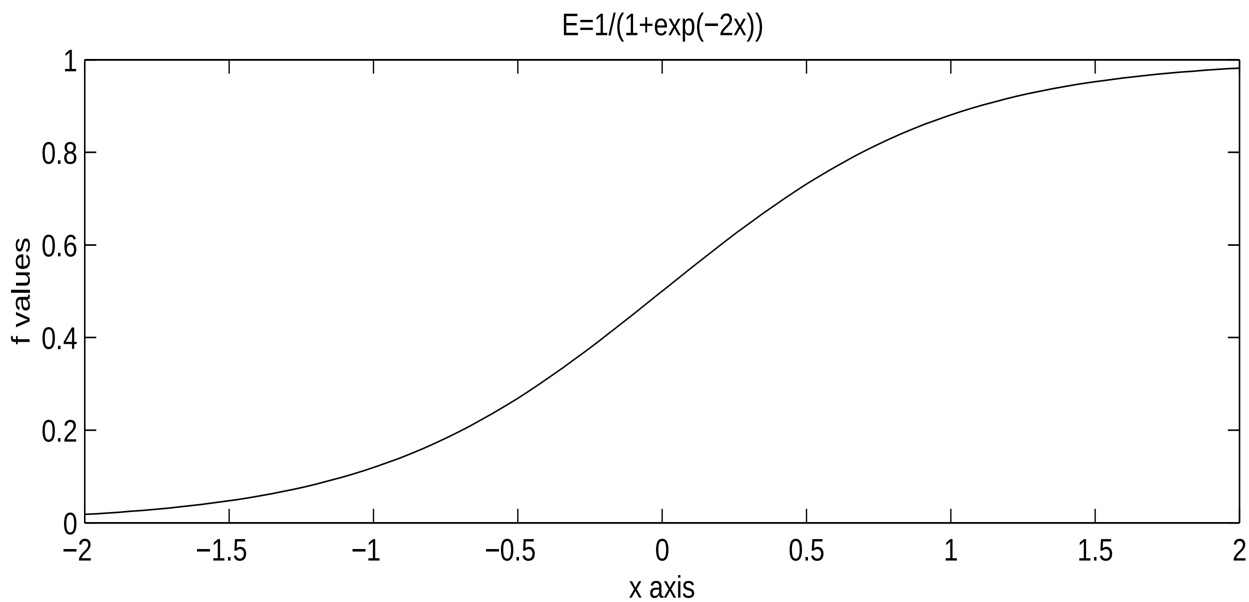
Here $G\left(\epsilon\right)/\epsilon$ is $\left(1 + c^2\epsilon^2\right)^{-\frac{1}{2}}$ and so the above is

$$\sum \left[\epsilon_i^{(k+1)} \left(1 + c^2 {\epsilon_i^{(k)}}^2\right)^{-\frac{1}{2}}\right]^2 .$$

- Algorithm usually converges to a near-best l_2 approximation.
- A linear least squares problem at each step.

Test Function Representations





Method of Function Approximation

- The test functions are sampled at 40 equally spaced points in the interval [-2, 2].
- Six data sets containing 2, 4, 6, 8, 10 and 12 outliers are constructed.
- 10 RBF centres, located at the Chebyshev zeros in the interval [-2, 2], are chosen for all approximations.
- The cubic radial basis function is used in the approximating form $F(x)$.
- The coefficients $b_j,\ j = 1, 2, \ldots, 10$, are calculated as a weighted least squares solution.
- The coefficients are then used to approximate at 100 points in [-2, 2] and the residual mean squares and the number of iterations taken to converge are compared.

Results: $f(x) = \exp(-2x^2)$

Alternative cubic	Number of Outliers		
estimator forms	8	10	12
cubic ϵ	0.50973 -	0.55179 -	0.63703 -
$\frac{\epsilon}{(1+\epsilon^2)^{\frac{1}{2}}}$	0.03389 6	0.03625 6	0.04932 10
$\tanh(\epsilon)$	0.04572 6	0.04934 6	0.07018 11
$1-\exp(-\epsilon)$	0.02058 8	0.02139 8	0.02799 13
$\sqrt{(1-\exp(-\epsilon^2))}$	0.02034 10	0.02136 10	0.03493 18

Results: $f(x) = \frac{1}{1+\exp(-2x)}$

Alternative cubic	Number of Outliers		
estimator forms	8	10	12
cubic ϵ	0.52620 -	0.62948 -	0.68764 -
$\frac{\epsilon}{(1+\epsilon^2)^{\frac{1}{2}}}$	0.04066 7	0.05584 9	0.06143 8
$\tanh(\epsilon)$	0.05614 8	0.07798 11	0.08556 10
$1-\exp(-\epsilon)$	0.02224 9	0.03199 12	0.03534 11
$\sqrt{(1-\exp(-\epsilon^2))}$	0.02599 12	0.04525 17	0.05012 16

Conclusions

- The estimators, solved as a weighted least squares problem, can all be shown to improve on a standard cubic approximation with outliers present in the data for the two test functions.

- The estimators $G = 1 - \exp(-|\epsilon|)$ and $G = \sqrt{(1 - \exp(-\epsilon^2))}$ are the most accurate forms of approximation for all levels of noise in the data sets.

- The estimator $G = \tanh(\epsilon)$ is the least accurate form of approximation for all levels of noise in the data sets.

- The estimator $G = \sqrt{(1 - \exp(-\epsilon^2))}$ takes the greatest number of iterations to converge in all cases.

- The estimator $G = \frac{\epsilon}{(1+\epsilon^2)^{\frac{1}{2}}}$ takes the least number of iterations to converge in all cases.

Analysis Method for $G(\epsilon) = \epsilon(1 + c^2\epsilon^2)^{-\frac{1}{2}}$

1. A Chebyshev degree 4 polynomial using 101 data on $[-1, 1]$.

2. A random curve is generated on this domain and the residual mean square (RMS) fit using TLS is calculated.

3. l_2 noise is added to the original data f by

   $$f = f + 0.001 * \text{randn}\,(m, 1)$$

   and outliers to every tenth f point as

   $$f = f + 0.01 * \text{randn}\,(10, 1)$$

   where randn are normally distributed.

4. Each calculation is repeated 300 times taking equally rising values of c.

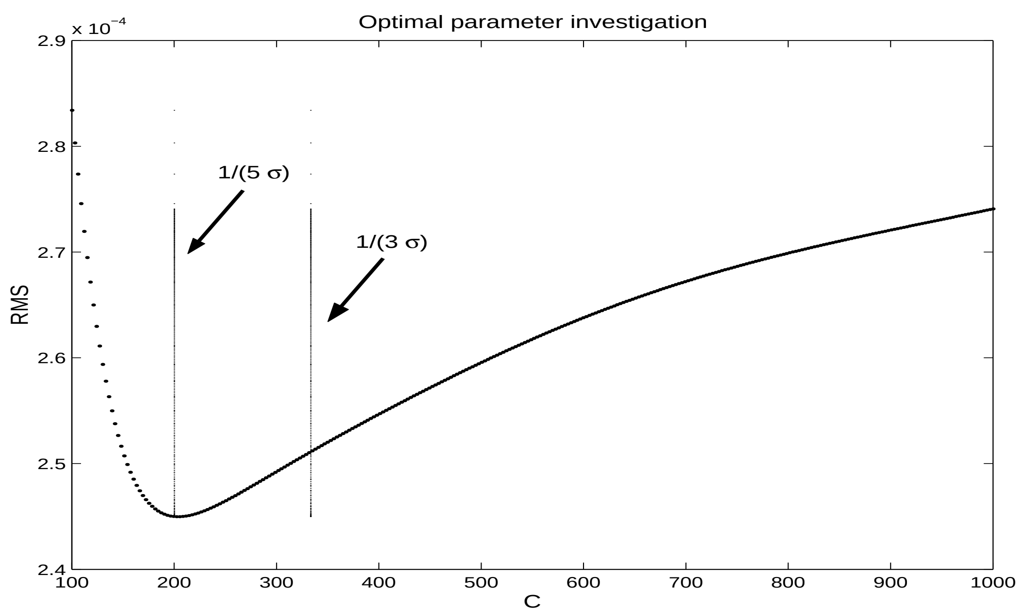
5. The alternative c values are compared with the predicted c value $c = 1/3\sigma$.

Optimising c in $G(\epsilon) = \epsilon(1 + c^2\epsilon^2)^{-\frac{1}{2}}$

To determine an optimum value for the parameter c, we again use repeated approximation.

300 equally spaced values ranging from $1/(10\sigma)$ to $1/\sigma$ are chosen for c. An approximation is constructed and the RMS evaluated for each c.

The graph below shows the RMS values plotted against the range of values for c when $\sigma = 0.001$.



A Robust Estimator? - Monte Carlo Simulation

Monte Carlo (Repeated Approximations) has been used to investigate the robustness of the estimator as follows.

- Construct initial uncorrupted data (x, y_0).
- Choose an approximating form (e.g polynomial).
- Choose number of simulations (say $k = 1000$).
- for each k construct

$$y_k = y_0 + l_2 \text{ noise} + \text{outliers}$$

and solve $C\mathbf{a}^{(k)} = \mathbf{y}^{(k)}$

If the estimator is robust, the variation in the fitted parameters (for each simulation) will be small.

For a degree 4 Chebyshev approximation the mean variation in the 5 fitted parameters using 1000 simulations is found to be

New estimator 0.00023440
L S estimator 0.00594347