



CMIS Technical Report 06/42

# Statistical Issues in Mapping and Monitoring Using Remote Sensing Data

**You-Gan Wang**

*CSIRO Mathematical and Information Sciences,  
65 Brockway Road, Floreat, Western Australia 6014, Australia.*

Tel +61 (08) 9333 6316

You-Gan.Wang@csiro.au

April 2006

## Preface

This report is provided in line with requirement in my APA objectives. It aims to provide some statistical challenges in the Stream of Mathematics for Mapping and Monitoring, Division of Mathematical & Information Sciences, CSIRO. It is only intended for CMIS internal use. Please do not quote or cite this report. Some materials have been submitted for publication and much more work are under further investigation for publication. Some materials are presented here to cast the ideas to consider (Some of them, to me, just seem to be wild flowers on the edge of the Great Wall). I wish to acknowledge Min Zhu for providing the tables and plots and Jared O'Connell for Figure 3. I also wish to thank Jeremy Wallace and Peter Caccetta for a few background discussions. I will keep dreaming.

# Contents

<b>1</b>	<b>Introduction</b>	<b>5</b>
<b>2</b>	<b><i>M</i>-estimation</b>	<b>6</b>
2.1	Standard errors . . . . .	8
2.2	Choosing the Optimal Resistance Level . . . . .	10
2.3	Application to Image Calibration . . . . .	11
<b>3</b>	<b><i>S</i>-estimation</b>	<b>14</b>
3.1	Standard errors of estimators . . . . .	17
3.2	Effects of tuning constant $c$ in $S$ -estimation . . . . .	18
3.3	Application to Land Monitor Project 2005 . . . . .	22
<b>4</b>	<b>Rank Regression</b>	<b>26</b>
4.1	Independence Model . . . . .	27
4.2	Weighted Ranks . . . . .	27
<b>5</b>	<b>Quantile Regression</b>	<b>28</b>
5.1	Introduction . . . . .	28
5.2	The Model . . . . .	29
5.3	More Details? . . . . .	30
<b>6</b>	<b>Least Trimmed-squares for Image Calibration</b>	<b>31</b>
6.1	Introduction . . . . .	31
6.2	Potential Targets . . . . .	33
6.3	The Model and the Algorithm . . . . .	34
6.4	Generalization to Time Series Images . . . . .	40

<b>7</b>	<b>Model Selection</b>	<b>41</b>
7.1	Introduction . . . . .	42
7.2	Working Covariance Models . . . . .	42
7.3	Selection Criteria . . . . .	47
7.4	Absolute Predictive error . . . . .	50
<b>8</b>	<b>Monitoring Land Changes</b>	<b>51</b>
8.1	Classification . . . . .	51
8.2	Some Statistics for Monitoring . . . . .	54
<b>9</b>	<b>Rank Estimation For Linear Models</b>	<b>55</b>
9.1	Introduction . . . . .	55
9.2	The Gehan-Wilcoxon Scheme . . . . .	59
9.3	Induced Smoothing . . . . .	60
9.4	Smoothed Rank Estimation for the AFT Model . . . . .	62
9.5	A Simulation Study . . . . .	64
9.6	Estimation of Perennial Vegetation Coverage . . . . .	65
<b>10</b>	<b>Other Potential Directions</b>	<b>67</b>
10.1	Wavelets . . . . .	67
10.2	Incorporating Spatial-Temporal Variations . . . . .	68
10.3	Internal Collaboration and Outreaching . . . . .	69

# 1 Introduction

Remote sensing data are widely accessible nowadays and has been of increasing importance in mapping and monitoring of land cover.

The Mathematics for Mapping and Monitoring (MMM) stream within CSIRO Mathematical and Information Sciences uses statistical and computational methods to integrate and analyse remotely sensed data and other spatial data to determine trends in land condition and to predict areas at risk from degradation. The major task of the MMM Stream is to develop and implement quantitative land-cover monitoring technologies for environmental management.

RAPG-MMM2: Development of new statistical models and methods for extracting (from remotely sensed data) and integrating continuous biophysical parameter estimates through time.

RAPG-MMM3: Statistical modeling and inference for multiple resolution (spectral, spatial, temporal) satellite imagery.

In line with these two research goals, I provide a few sections based on different statistical approaches on the statistical issues when analysing remote sensed data.

Remote sensing data are prone to contamination due to cloud and other atmospheric effects. One of the robust approach,  $S$ -estimation has been used in calibrating images from different dates (Furby and Campbell, 2001). However, there is a need to better understand (i) how  $S$ -estimation works including how to select the tuning constant, and (ii) impacts on efficiency and (iii) calculation of standard errors. Regarding (i), Wang *et al.* (2006) has provided details in the context of robust estimation using the Huber function and illustrated using the calibration data. I now provide more technical details in the following subsections.

There are a variety of approaches to robust inferences in statistics including  $M$ -estimation,  $S$ -estimation, rank regression, quantile estimation and least trimmed squares. These methods are also related to each other. For example,  $S$ -estimation is very similar to  $M$ -estimation, and is in fact a modified version of  $M$ -estimation.

Robustness against outliers is gained at the price of efficiency loss when the resistance is unnecessarily high. Therefore, efficient estimation is possible only if a dispersion function with appropriate resistance level is chosen. Little work has been done on how to choose such a dispersion function for a given data set. From a likelihood perspective, the ‘best’ loss function would be the negative loglikelihood function (Schrader and Hettmansperger, 1980).

For easy presentation, we shall consider the ordinary robust regression model,

$$y_i = X_i^T \beta + \sigma \epsilon_i, \quad i = 1, 2, \dots, n,$$

where  $X_i = (1, x_{i1}, \dots, x_{i,p-1})'$ , the covariate vector of dimension  $p$ ,  $\sigma^2 = \text{var}(y_i)/\text{var}(\epsilon_i)$  is a scale parameter, and  $\beta = (\beta_1, \beta_2, \dots, \beta_p)'$  is the parameter vector of interest. The errors ( $\epsilon_i$ ) are independent and identically distributed. The least squares approach leads to  $\hat{\beta} = (\sum_i X_i X_i^T)^{-1} (\sum_i X_i y_i)$ , which is not robust against outliers although consistent under the assumption of  $E\epsilon_i = 0$ .

## 2 $M$ -estimation

The idea of  $M$ -estimation is to replace the sum of the squared residuals by the sum of a loss function that is slowly varying,

$$l(\beta, \sigma, c) = \sum_{i=1}^n \rho\{(y_i - X_i^T \beta)/\sigma\}.$$

The sum of loss functions as the overall objective function is a reasonable choice here because the observations are independent. The most widely used dispersion function is

the Huber's function

$$\rho(u) = \begin{cases} \frac{1}{2}u^2 & (|u| \leq c) \\ c|u| - \frac{1}{2}c^2 & (|u| > c) \end{cases} .$$

The constant  $c$  must be pre-specified. This dispersion function can be regarded as the negative loglikelihood of a modified standard normal distribution with its tail (when  $|u| > c$ ) being replaced by an exponential distribution. As pointed out by Huber (1981, p. 18), 'The constant  $c$  regulates the amount of robustness; good choices are in the range between 1 and 2, say,  $c = 1.5$ '. Other values are also used in the literature, for example,  $c = 1.2$  (Canttoni and Ronchetti, 2001),  $c = 1.25$  (Chi, 1994; Street, Carroll and Ruppert, 1988). The default value of  $c$  in R package (*rlm* function) is 1.345. However, the choice of  $c$  can have a great impact on the estimation efficiency. If the errors are normally distributed and there is no contamination, the best choice of  $c$  is  $\infty$ . However, if the errors follow Laplace distribution, a heavily tailed distribution,  $c$  should be chosen to be 0 or a small positive value. Roughly speaking, the choice of  $c$  should reflect the possible proportion of outliers in the data. It is therefore sensible to adjust the  $c$  value accordingly based on the distribution of the data.

Another commonly used function is the so called Tukey biweight function

$$\rho(u) = \begin{cases} \frac{u^2}{2} - \frac{u^4}{2c^2} + \frac{u^6}{6c^4} & \text{if } |u| \leq c, \\ \frac{c^2}{6} & \text{elsewise,} \end{cases}$$

Note that unlike the Huber function, the Tukey function is not a convex function. Therefore extra computational care may be needed to eliminate local minimums.

For a given estimate of  $\hat{\sigma}$  of the scale parameter, we minimise

$$\sum_{i=1}^n \rho\{(y_i - X_i^T \beta) / \hat{\sigma}\},$$

to obtain estimates of  $\beta$ .

In general, if  $\rho$  has a sub-gradient  $\psi$ , we may define the  $M$ -estimator by the solution to

$$\sum_{i=1}^n X_i \psi\{(y_i - X_i^T \beta) / \hat{\sigma}\} \approx 0.$$

For example, the Huber- $\psi$  function is

$$\psi(u) = \begin{cases} u & (|u| \leq c) \\ \text{sign}(u) * c & (|u| > c) \end{cases}$$

A generalized version of the estimating functions is the Mallows's type,

$$U_w(\beta) = \sum_{i=1}^n w_i \psi\{(y_i - X_i^T \beta) / \hat{\sigma}\}, \quad (1)$$

where  $w_i$  are suitably chosen  $p$ -vectors to down-weight the points with large leverage values. The usual linear models assume the errors have expectation 0 and finite variance. But the  $M$ -estimation requires  $E\psi(\epsilon_i) = 0$  and  $E\psi^2(\epsilon_i) = \sigma_\psi^2$ . The fundamental assumption,  $E\psi(\epsilon_i) = 0$ , is required to ensure Fisher consistency of the estimating functions (1). It is often assumed that the distribution of  $\epsilon_i$  is symmetric. The Huber's approach is to gain robustness against outliers in  $y$  (or contaminated  $\epsilon_i$ ). The Mallows's approach will gain robustness in  $x$  by choosing appropriate  $w_i$  (as a trimmed function of  $X_i$ ).

## 2.1 Standard errors

Suppose

$$\frac{\partial E\psi(\epsilon_i + \delta)}{\partial \delta} \Big|_{\delta=0} = b$$

and  $E\psi^2(\epsilon_i) = \sigma_\psi^2$ . Under the finite variance assumption and  $E|\psi(\epsilon_i + \delta) - \psi(\epsilon_i)|^2 = o(1)$ , the estimator  $\hat{\beta}$  obtained by solving  $U_w(\beta) = 0$  has the following property,

$$\sqrt{n}(\hat{\beta} - \beta) \sim N(0, \tau^{-1} \sigma^2 V_n)$$

where

$$V_n = \left( n^{-1} \sum_{i=1}^n w_i X_i' \right)^{-1} \left( n^{-1} \sum_{i=1}^n w_i w_i' \right) \left( n^{-1} \sum_{i=1}^n X_i w_i' \right)^{-1},$$

and  $\tau = b^2/\sigma_\psi^2$  is a scalar. The  $\psi$  function with a larger  $\tau$  value will result in estimators with smaller variances for all components of  $\beta$  and any linear combinations of them. It is therefore sensible to treat  $\tau$  as an efficiency factor (Scharder and Hettmansperger, 1980) and we may choose the  $\rho(\cdot)$  function with the largest  $\tau$  value.

For the least squares estimator ( $c = \infty$ ), let  $\text{var}(\epsilon_i) = \sigma_\epsilon^2$ , we have  $\tau = \sigma_\epsilon^{-2}$ , and for the median estimator ( $c = 0$ ),  $\tau = 4f^2(0)$ , where  $f$  is the density of  $\epsilon_i$ . In the case when  $\sigma_\epsilon^2$  is  $\infty$ , such as Cauchy and  $t(2)$  distribution, the least squares method has efficiency 0 ( $\tau = 0$ ). For the Huber function with tuning constant  $c$ ,

$$\begin{aligned} \sigma_\psi^2 &= \int_{-c}^c \psi^2(u) dF(u) + c^2 \{1 - F(c) + F(-c)\} \\ b &= \int_{-c}^c dF(u). \end{aligned}$$

The following issues are of interest.

(i). How to obtain the best value for  $c$ ? Because the choice of  $c$  is case by case, it seems necessary to carry out extensive simulation studies to quantify efficiency gain/loss in the context of mapping and monitoring.

(ii). It appears that ‘contamination’ occurs not only in  $y$  but also in  $x$ , which requires us to study Mallows’ type estimator. Perhaps the impact of the choice of the down weight function  $w$  also needs to be investigated.

(iii). Comparison with other robust methods. So far, we have not carried out any studies in evaluating different robust approaches to better understand our needs in analysing the landsat data.

(vi). Remote sensing data usually consist of millions of observations. This may create a computational problem when the ranking inference involves the complexity

of  $O(N^2)$ . For example, the covariance matrix of the ranks is of this order. Special methods for computing large matrices can be developed, analytically or numerically.

In fact,  $\tau$  value is an efficiency measure relative to the least squares method. We suggest to choose the most appropriate  $\psi(\cdot)$  or more specifically, the tuning constant  $c$  by maximising  $\tau$ .

## 2.2 Choosing the Optimal Resistance Level

Suppose  $(\hat{\beta}, \hat{\sigma})$  are the current estimates of  $(\beta, \sigma)$ , denote  $\hat{\epsilon}_i = (y_i - X_i^T \hat{\beta})/\hat{\sigma}$ . Because  $b = \Pr(|\epsilon_i| \leq c)$ , we can use the approach of kernel density estimation to obtain an estimator for  $b$ ,  $1 - 2/n \sum_{i=1}^n \Phi\{-(y_i - X_i^T \hat{\beta})/(nh\hat{\sigma}) - c\}$ , in which  $h$  is the bandwidth. To avoid bandwidth selection, we may use the following fully nonparametric estimator and estimate  $\tau$  by

$$\hat{\tau}(c) = \frac{(\sum_{i=1}^n I(|\hat{\epsilon}_i| \leq c))^2}{n(\sum_{i=1}^n I(|\hat{\epsilon}_i| \leq c)\psi^2(\hat{\epsilon}_i) + c^2 I(|\hat{\epsilon}_i| > c))}$$

where  $I$  is the indicator function. Note that  $\hat{\tau}(c)$  is not a continuous function of  $c$ .

In practice, we suggest to evaluate  $\hat{\tau}(c)$  for a finite set of values of  $c$  in the range of 0 to 3, say. Simple smoothing method such as the nearest neighbors may be applied to  $\hat{\tau}(c)$  before selecting the optimal  $c$ .

We now consider the case when  $\sigma^2$  is unknown. An iteratively reweighted least-squares method used in robust regression model is discussed by Street et al. (1988). There are two commonly used robust estimators of  $\sigma$  (Street et al., 1988). The median absolute deviation (MAD) estimator is

$$\hat{\sigma} = \text{Median}\{|y_i - X_i^T \hat{\beta}|\}/0.6745. \quad (2)$$

The Huber's robust estimation can also be regarded as using the least informative distribution  $f(u) = \alpha^{-1} \exp\{-\rho(u)\}$  as a 'working likelihood' for parameter estimation.

I have investigated the performance of the Huber function using different  $c$  values for a variety of error distribution types. In particular, I suggest choosing a value for  $c$  depending on the data (data-driven approach), i.e. to make the value of  $c$  data-dependent. Simulation studies indicate that the efficiency can be improved as much as 40% compared to the traditional approach in which  $c$  is fixed at 1.345 (see Table 1 and 2).

### 2.3 Application to Image Calibration

Our dataset is generated by the satellite sensors (Landsat 5) for land monitoring and mapping in Australia. Image calibration is an important step for detecting environmental changes through time and assessment of vegetation productivity (e.g. Zhang, Guindon, and Cihlar, 2002). There are 148 target areas (pixels) carefully selected to represent different ground objects. The reference year is 1994 and the data were collected when the atmospheric conditions are ideal (i.e., minimum haze and clouds etc.). The analysis here is for illustrating the proposed methodology only. Detailed analysis will be reported elsewhere in line with other project objectives. Figure 1 shows the scatter plot of data from spectral band six for these 148 pixels (Year 2005 versus Year 1994). The estimated calibration lines by the least squares method and robust estimation using the Huber function with  $c = 1.345$  are quite different. The estimated regression coefficients are (14.399, 0.575) with standard errors (4.583, 0.042) for the LS approach, and (6.702, 0.673) with standard errors (3.189, 0.029) for the robust approach with  $c = 1.345$ . Figure 2(a) shows the efficiency plot at different  $c$  values. The vertical

Table 1: Relative efficiency of three  $\beta$  estimators, least squares (LS), using  $c = 1.5$  and data-dependent  $c$  (DD) when errors are contamination rate is  $\lambda$ . Panel (e) is for  $\lambda = 0$ . A value larger than one indicates more efficient than using  $c = 1.345$ .  $\bar{c}$  is the average  $\hat{c}$  values used in the DD method based on 1000 simulations, and  $\bar{\sigma}_c^2$  is the corresponding variance of the  $\hat{c}$  values.

$\lambda$	1%	5%	10%	20%	30%	40%	$\lambda$	1%	5%	10%	20%	30%	40%
	$\hat{\beta}_1$							$\hat{\beta}_2$					
(a). $N(0, 1)$ with contamination $N(0, 9)$													
LS	0.99	0.81	0.68	0.63	0.58	0.63		0.96	0.81	0.69	0.63	0.60	0.63
$c = 1.5$	0.99	1.00	1.00	1.02	0.97	0.96		0.99	0.99	1.00	1.02	0.97	0.96
DD	1.00	0.98	0.96	1.00	0.98	1.087		1.00	0.98	0.96	1.00	0.99	1.10
$\bar{c}$	2.05	1.47	1.16	0.85	0.66	0.53	$\bar{\sigma}_c^2$	0.33	0.22	0.19	0.14	0.10	0.08
(b). $N(0, 1)$ with contamination $\chi^2(4) - 4$													
LS	1.01	0.92	0.89	1.11	0.26	0.29		1.02	0.87	0.74	0.64	0.28	0.28
$c = 1.5$	0.98	0.99	1.01	1.02	0.99	1.00		0.98	0.99	1.00	1.02	1.00	0.99
DD	1.02	0.95	0.98	1.03	0.97	0.97		1.02	0.95	0.97	0.99	0.97	0.98
$\bar{c}$	2.14	1.57	1.25	0.92	1.10	0.96	$\bar{\sigma}_c^2$	0.34	0.24	0.20	0.17	0.20	0.17
(c). $N(0, 1)$ with contamination $t(2)$													
LS	0.98	0.58	0.52	0.40	0.27	0.22		0.98	0.71	0.56	0.36	0.09	0.13
$c = 1.5$	0.99	0.99	0.99	0.99	0.98	0.99		0.98	0.99	0.99	1.00	0.98	0.98
DD	1.01	1.00	0.99	0.98	1.08	1.10		1.01	1.01	0.98	0.97	0.98	1.12
$\bar{c}$	2.29	1.93	1.65	1.34	1.11	0.94	$\bar{\sigma}_c^2$	0.36	0.29	0.25	0.22	0.20	0.19
(d). Two-point contamination													
LS	0.90	0.58	0.42	0.43	0.56	1.01		0.91	0.56	0.43	0.43	0.57	1.02
$c = 1.5$	1.02	0.99	0.97	0.93	0.88	0.83		1.01	0.99	0.97	0.93	0.88	0.83
DD	1.02	0.97	1.00	1.14	1.49	2.63		1.02	0.97	1.00	1.13	1.45	2.63
$\bar{c}$	1.73	1.15	0.81	0.50	0.34	0.24	$\bar{\sigma}_c^2$	0.23	0.16	0.12	0.07	0.04	0.02
(e). $N(0, 1)$ , $\lambda = 0$ , no contamination													
	$\hat{\beta}_1$						$\hat{\beta}_2$						
LS	1.03						1.03						
$c = 1.5$	0.99						0.99						
DD	1.01						1.01						
$\bar{c}$	2.40			$\bar{\sigma}_c^2$			0.37						

Table 2: Relative efficiency of three  $\beta$  estimators, least squares (LS), using  $c = 1.5$  and data-dependent  $c$  (DD) when errors have a heavily tailed distribution. A value larger than one indicates more efficient than using  $c = 1.345$ .  $\bar{c}$  is the average  $\hat{c}$  values used in the DD method based on 1000 simulations, and  $\bar{\sigma}_c^2$  is the corresponding variance of the  $\hat{c}$  values.

$\hat{\beta}_1$				$\hat{\beta}_2$			
(a). $t$ distribution $f(u) \propto (1 + u^2/\nu)^{-(\nu+1)/2}$							
	$\nu = 1.5$	$\nu = 2$	$\nu = 2.5$		$\nu = 1.5$	$\nu = 2$	$\nu = 2.5$
LS	0.02	0.06	0.42	LS	0.02	0.03	0.40
$c = 1.5$	1.05	1.04	1.03	$c = 1.5$	1.05	1.04	1.03
DD	1.27	1.16	1.04	DD	1.25	1.14	1.04
$\bar{c}$	0.34	0.44	0.55	$\bar{\sigma}_c^2$	0.04	0.07	0.09
(b). Laplace distribution, $f(u) \propto \exp(- u /d)$							
	$d = 1$	$d = 2$	$d = 5$		$d = 1$	$d = 2$	$d = 5$
LS	0.77	0.73	0.72	LS	0.75	0.75	0.73
$c = 1.5$	1.03	1.04	1.03	$c = 1.5$	1.03	1.03	1.03
DD	1.37	1.38	1.34	DD	1.35	1.38	1.41
$\bar{c}$	0.20	0.21	0.21	$\bar{\sigma}_c^2$	0.02	0.02	0.02
(c). Cauchy distribution, $f(u) \propto \{1 + (u/d)^2\}^{-1}$							
	$d = 1$	$d = 3$	$d = 5$		$d = 1$	$d = 3$	$d = 5$
LS	-	-	-	LS	-	-	-
$c = 1.5$	1.07	1.08	1.07	$c = 1.5$	1.07	1.08	1.08
DD	1.27	1.04	1.22	DD	1.33	1.13	1.17
$\bar{c}$	0.26	0.17	0.17	$\bar{\sigma}_c^2$	0.02	0.01	0.01

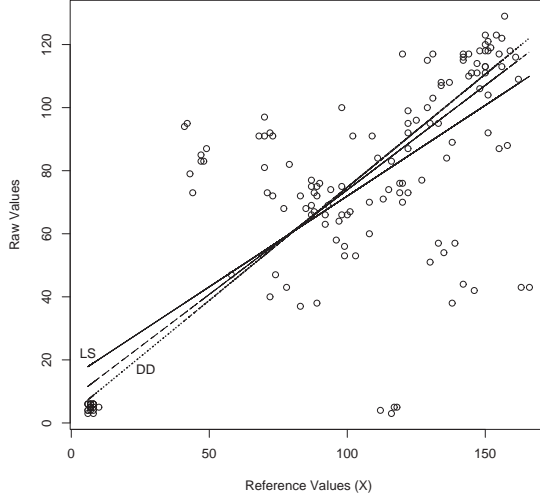


Figure 1: Scatter plot and estimated calibration lines by the least squares (LS) method and the Huber's  $M$ -estimation with data-dependent tuning constant  $\hat{c} = 0.7$  (DD). The middle line is based on  $c = 1.345$ .

line indicates that the DD approach gives  $\hat{c} = 0.7$  as the optimal value resulting the coefficient estimates as  $(3.015, 0.717)$  with standard errors  $(2.678, 0.025)$ . The choice of  $c$  values appear to impact the estimated coefficients as well as their standard errors. The kernel density plot (Fig. 2(b)) shows that the error distribution is quite long-tailed, justifying the need of a robust approach for parameter estimation.

### 3 S-estimation

I now introduce briefly the concept of  $S$ -estimation in the context of image calibration.

By treating digital counts from reference image as responses  $y_i$  and digital counts from raw image as covariates  $X_i$ , the ordinary robust regression model is

$$y_i = X_i^T \beta + \sigma \epsilon_i, i = 1, 2, \dots, n,$$

where  $X_i = (1, x_i)^T$  is the design matrix,  $\beta = (\beta_1, \beta_2)^T$  stands for gain and offset for a

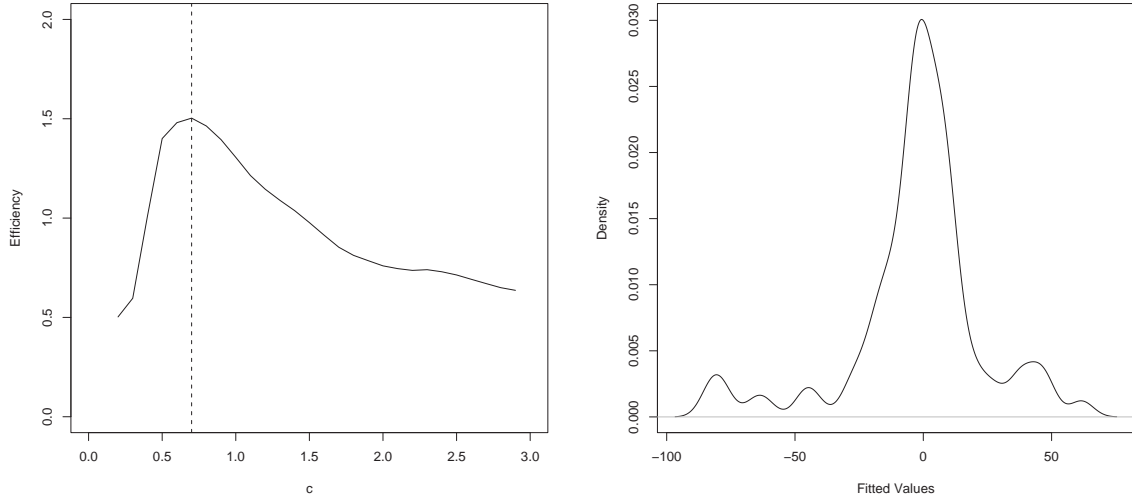


Figure 2: Left panel: Efficiency factor,  $\tau$ , at different  $c$  values based on the image calibration data from Australia. The vertical line indicates the optimal  $c$  value. Right panel: Kernel density function of the residuals obtained from the data-driven approach for the calibration data.

specific band, and  $\sigma$  is a scale parameter whose estimator is denoted by  $s$ . According to Campbell *et al.* (1994), the S-estimators of the gain and offset are chosen to minimize  $s$  subject to

$$\frac{1}{n-p} \sum_{i=1}^n \rho \left( \frac{y_i - X_i^T \hat{\beta}}{s} \right) = b_0, \quad (3)$$

where  $b_0$  is a constant. There are various choices for the objective function  $\rho$ . The calculation reported here are based on the Tukey function

$$\rho_c(x) = \begin{cases} \frac{x^2}{2} - \frac{x^4}{2c^2} + \frac{x^6}{6c^4} & \text{if } |x| \leq c, \\ \frac{c^2}{6} & \text{elsewhere,} \end{cases}$$

where  $c$  is the so-called tuning constant which reflects the resistance of the estimators to outliers. The constant  $b_0$  is often chosen as  $E_{\Phi}\{\rho_c(x)\}$ , the expectation of the objective function with respect to a standard normal distribution. For the Tukey function, we

have

$$b_0 = \frac{-c^6 - c^5 + c^4 + 4c^3 - 18c^2 + 15}{3c^4} \Phi(c) + \frac{2c^6 - 3c^4 + 9c^2 - 45}{6c^4},$$

where  $\Phi(c)$  denotes the cumulative distribution for standard normal.

Denote the residual  $r_i = y_i - X_i^T \hat{\beta}$ , the S-estimators for  $(\beta, \sigma)$ ,  $(\hat{\beta}, s)$ , are solutions of the following estimating equations (p. 140, Rousseeuw & Leroy, 1987),

$$\frac{1}{n} \sum_{i=1}^n \psi(r_i/s) X_i^T = 0, \quad (4)$$

$$\sum_{i=1}^n \rho(r_i/s) - (n-p)b_0 = 0, \quad (5)$$

where  $\psi$  function is the derivative of  $\rho$ -function

$$\psi(x) = \begin{cases} x \{1 - (\frac{x}{c})^2\}^2 & \text{if } |x| \leq c, \\ 0 & \text{elsewise.} \end{cases}$$

By denoting

$$w_i = \frac{\psi(r_i/s)}{r_i/s}, \quad (6)$$

we can reexpress  $\hat{\beta}$  as

$$\hat{\beta} = \frac{\sum_{i=1}^n w_i X_i^T y_i}{\sum_{i=1}^n w_i X_i^T X_i},$$

which is a version of the weighted least-squares estimation. Thus, for a given scale estimate  $s$ , the S-estimators for the coefficients are equivalent to the weighted least squares estimators with weights  $w_i$ .

The steps to calculate S-estimators for calibration are:

- Get initial value of  $\beta$  using `rlm` function in the MASS package for R under the objective function `psi.bisquare` and tuning constant  $c$ .
- Get the initial value of the scale  $s$  from the constraint (5).
- Calculate the coefficients using the weighted least squares with weight given by (6) for a given scale.

- Repeat step 2 and step 3 until converge.

### 3.1 Standard errors of estimators

Note that in equation (5) the constant  $b_0$  is pre-specified as  $E_{\Phi}[\rho_c(\epsilon_i)]$  by assuming  $\epsilon_i$  has a standard normal. If the true underlying distribution of  $\epsilon_i$  is not the standard normal, we would not expect  $s$  to converge to  $\sigma$ . For a given  $c$ , suppose the limiting value of  $s$  as  $n \rightarrow \infty$  is  $\sigma_0$ , which can be determined from

$$E\rho_c(\epsilon_i/\sigma_0) = b_0 \quad (7)$$

Suppose  $\sigma_{\psi}^2$  is the variance of  $\psi(\epsilon_i\sigma/\sigma_0)$ , and  $\lambda = E\{\psi'(\epsilon_i\sigma/\sigma_0)\}$ , where

$$\psi'(x) = \begin{cases} \{1 - (\frac{x}{c})^2\}\{1 - 5(\frac{x}{c})^2\} & \text{if } |x| \leq c, \\ 0 & \text{elsewise.} \end{cases}$$

The S-estimator  $\hat{\beta}$  has the following asymptotic properties

$$\sqrt{n}(\hat{\beta} - \beta) \sim N(0, \tau^{-1}V_n), \quad (8)$$

where

$$V_n = \left( n^{-1} \sum_{i=1}^n X_i X_i^T \right)^{-1},$$

and  $\tau^{-1} = \sigma_0^2 \sigma_{\psi}^2 / \lambda^2$  is a scalar.

Suppose we have obtained the S-estimators of  $\beta$ ,  $\hat{\beta}$ , one can obtain their standard errors by the following steps.

1. Calculate the residuals  $r_i = y_i - X_i \hat{\beta}$ .
2. Obtain  $s$  from (5) as an estimate of  $\sigma_0$ .
3. Obtain an estimate of  $\lambda$  as  $n^{-1} \sum_i \psi'(r_i/s)$ .
4. Obtain  $\sigma_{\psi}^2$  as  $n^{-1} \sum_i \psi^2(r_i/s)$ .
5. Obtain the variance-covariance matrix  $\tau V_n$ .

### 3.2 Effects of tuning constant $c$ in $S$ -estimation

Different  $c$  values give different levels of robustness against outliers. The larger the  $c$  value is, the less robustness the estimation procedure is, but more efficient if the data follow a normal distribution (without outliers). The least squares estimator is corresponding to  $c = \infty$ . Due to the possibility of high contamination in data, Rousseeuw & Leroy (1987, p.143) conservatively suggested estimators with a breakdown point greater than 25% to be used in practice which corresponds to  $c$  value less than 2.937. The default  $c$  value is 1.548 corresponding to 50% breakdown point (28.7% efficiency at the normal) in *lqs* function in MASS package for R. The tuning constant used by the CMIS Mathematics for Mapping and Monitoring Group for calibration is 2.15. Also,  $c = 4.685$  which gives 95% efficiency at the normal is a common choice in data analysis. Clearly, the choice of  $c$  values will affect the estimation efficiency for a given data set.

Based on the standard error expression (8), we can see that an efficient estimator should maximize the scalar value  $\tau$ . It is therefore sensible to regard  $\tau$  as an efficiency factor (Scharder & Hettmansperger, 1980), and we may choose the  $c$  value which produces the largest  $\tau$  value. This data-driven method selects an optimal  $c$  value gain the necessary resistance level for parameter estimation.

To investigate the performance of this data-driven method, we carry out simulations to compare the mean squared errors of estimators from least squares method, the  $S$ -estimators with a fixed tuning constant  $c = 2.15$  and our data-driven method. We search the best  $c$  from 1.548 to 5.948 by an increment 0.1. We consider a linear relationship with gain and offset  $\beta = (8.2, 1.05)$  for two sets of digital counts (sample size  $n = 150$ )

from reference image and raw image

$$Y_i = \beta_1 + \beta_2 X_i, \quad i = 1, 2, \dots, 150.$$

In the simulations, the covariates are generated from the uniform distribution on  $(0, 220)$ .

To account the measurement errors caused by the sensor, so that we no longer observe  $X_i$  directly but rather  $X_i$ , where

$$X_i = X_i + \mu_i.$$

Same for variable  $Y_i$ , we observe  $y_i$  instead of  $Y_i$ ,

$$y_i = Y_i + \nu_i.$$

Both  $\mu_i$  and  $\nu_i$  are random numbers generated from  $N(0, 4)$  and they are uncorrelated.

Three different contamination types are considered, covariates contaminated, responses contaminated and both contaminated. For first two one-sided contamination types, we allow counts to be contaminated by either  $\chi(30)$  (asymmetric contamination) or  $N(0, 20)$  (symmetric contamination). We only consider symmetric contamination for the two-sided contamination type which is contaminated by random numbers from  $N(0, 20)$ . Also for this type of contamination and a specific contamination rate, 30% contaminated values are assigned to responses and the rest 70% are assigned to covariates. This is due to covariates (digital counts from raw images) are more likely to be contaminated than responses (digital counts from reference images). Different contamination rates are considered, namely,  $\lambda = 5\%$ ,  $10\%$ ,  $20\%$ ,  $30\%$  and  $40\%$ . We obtain estimates using  $c = 2.15$  and  $\hat{c}$  – the tuning constant given by the data-driven method. The least squares estimator is also obtained. We evaluate the relative efficiency (RE) of

different  $\beta$  estimators based on their mean squared errors (MSE) using the S-estimators with  $c = 2.15$  as a benchmark, i.e.  $RE = \text{MSE}(\hat{\beta}_{c=2.15}) / \text{MSE}(\hat{\beta})$ . The larger RE value is, the more efficient the estimator is (relative to the estimator based on  $c = 2.15$ ).

Table 4 summarizes the results for normal measurement errors with or without contaminations based on 1000 simulations. As the contamination rate increases, the average  $\hat{c}$  decreases. This is consistent to our expectation because more outliers should be truncated as the contamination rate becomes larger. The data-driven approach outperforms the fixed tuning constant method in most of the cases. Occasionally, it has a trivial efficiency loss of 5% at most. In the case of one-sided contamination, the efficiency reaches as high as 5.62 (panel (a)) when  $\lambda$  is 40%. Another simulations which are not listed here show the more obvious the outliers are, the better the data-driven method are for all contamination types. The fixed c method with  $c = 2.15$  performs not as well as our expectation. In asymmetric one-sided contamination cases with high contamination ( $\lambda = 40\%$ , panels (a) and (c)), fixed c S-estimators and the least squares estimators both much more inefficient compared with our data-driven approach. The least squares method performs best when there is no contamination ( $\lambda = 0$ ), and it becomes worse in general as  $\lambda$  increases. In no contamination case, the data-driven method gives quite large  $\hat{c}$  indicating no resistance against outliers is necessary in these cases and its performance is nearly as good as the least squares.

Based on the simulation results, we recommend to obtain the tuning constant from the data instead of fixing it to maximize the estimate efficiency.

Table 3: Relative efficiency of two  $\beta$  estimators, least squares (LS) and data-dependent  $c$  (DD) when errors are contamination rate is  $\lambda$ . Panel (e) is for  $\lambda = 0$ . A value larger than one indicates more efficient than using  $c = 2.15$ .  $\bar{c}$  is the average  $\hat{c}$  values used in the DD method based on 1000 simulations, and  $\bar{\sigma}_c^2$  is the corresponding variance of the  $\hat{c}$  values.

$\lambda$	5%	10%	20%	30%	40%	$\lambda$	5%	10%	20%	30%	40%
	$\hat{\beta}_1$						$\hat{\beta}_2$				
(a). Asymmetric contamination in covariates											
LS	0.91	0.55	0.25	0.17	0.25		0.29	0.12	0.05	0.12	1.04
DD	1.43	1.25	0.99	1.33	1.72		1.31	1.11	0.97	3.01	5.62
$\bar{c}$	4.15	3.23	2.37	1.86	1.89	$\bar{\sigma}_c^2$	0.45	0.21	0.08	0.04	1.26
(b). Symmetric contamination in covariates											
LS	0.91	0.39	0.21	0.14	0.12		0.29	0.34	0.17	0.12	0.11
DD	1.43	1.13	1.04	1.01	0.97		1.31	1.11	1.03	1.02	0.97
$\bar{c}$	4.15	4.05	3.10	2.54	2.16	$\bar{\sigma}_c^2$	0.45	0.62	0.34	0.24	0.16
(c). Asymmetric contamination in responses											
LS	0.30	0.11	0.04	0.06	0.30		0.72	0.52	0.36	0.48	1.59
DD	1.25	1.06	0.98	2.27	3.93		1.26	1.15	1.00	1.40	3.47
$\bar{c}$	4.06	3.22	2.33	1.83	1.75	$\bar{\sigma}_c^2$	0.46	0.22	0.09	0.06	0.78
(d). Symmetric contamination in responses											
LS	1.21	0.88	0.63	0.54	0.49		1.22	0.87	0.62	0.57	0.49
DD	1.46	1.33	1.18	1.01	0.97		1.48	1.27	1.16	1.01	0.97
$\bar{c}$	4.88	4.07	3.18	2.65	2.28	$\bar{\sigma}_c^2$	0.72	0.67	0.35	0.27	0.22
(e). Symmetric contamination in both covariates and responses											
LS	0.89	0.51	0.27	0.20	0.16		0.80	0.45	0.25	0.17	0.13
DD	1.44	1.24	1.04	1.00	0.95		1.39	1.19	1.03	0.99	0.97
$\bar{c}$	4.89	4.12	3.19	2.62	2.26	$\bar{\sigma}_c^2$	0.73	0.62	0.36	0.26	0.18
(f). $\lambda = 0$ , no Contamination											
	$\hat{\beta}_1$						$\hat{\beta}_2$				
LS	1.72						1.74				
DD	1.60						1.62				
$\bar{c}$	5.80					$\bar{\sigma}_c^2$	0.37				

### **3.3 Application to Land Monitor Project 2005**

Four data sets from the Land Monitor Project 2005 are analysis in Table 3 to illustrate our proposed standard error calculation. The tuning constant  $c$  is fixed at 2.15 — the value used by MMM group. Details can be found in the Report by Zhu et al. (2005).

Table 4: Four data sets from the Land Monitor Project 2005 are analysed to illustrate our proposed standard error and  $R^2$ . The tuning constant  $c$  is fixed at 2.15 — the value used by CMIS Mathematics for Mapping and Monitoring Group.

Scene 11083												
	Band 1		Band 2		Band 3		Band 4		Band 5		Band 7	
Least squares												
$\hat{\beta}$	44.84	0.921	5.36	0.980	4.17	1.213	0.80	1.045	-3.25	1.156	-5.04	1.157
se	2.70	0.026	1.63	0.024	2.34	0.028	1.76	0.020	2.79	0.019	1.99	0.023
S-estimators												
$\hat{\beta}$	36.52	1.011	1.19	1.067	-1.11	1.298	-3.91	1.137	-7.55	1.178	-4.79	1.101
se	2.71	0.026	1.16	0.017	1.44	0.017	1.09	0.012	1.97	0.013	1.54	0.018
$R^2$	0.859		0.880		0.892		0.898		0.926		0.903	
Scene 11380												
Least squares												
$\hat{\beta}$	76.08	0.991	25.18	0.971	38.97	1.139	27.57	0.906	63.10	0.878	28.50	0.972
se	5.10	0.050	4.42	0.063	7.09	0.081	5.46	0.057	9.39	0.062	5.68	0.071
S-estimators												
$\hat{\beta}$	58.60	1.085	7.63	1.156	9.48	1.422	5.40	1.105	14.84	1.143	4.29	1.264
se	2.46	0.024	2.56	0.037	3.74	0.043	2.63	0.027	3.53	0.023	2.58	0.032
$R^2$	0.763		0.694		0.631		0.673		0.510		0.526	
Scene 11384												
Least squares												
$\hat{\beta}$	58.16	1.397	12.23	1.268	17.42	1.592	8.82	1.180	6.28	1.417	2.04	1.495
se	1.12	0.027	0.82	0.026	1.30	0.036	1.73	0.027	2.31	0.027	1.19	0.031
S-estimators												
$\hat{\beta}$	54.75	1.465	9.52	1.313	0.88	2.329	2.13	1.258	-9.55	1.680	-4.66	1.747
se	0.54	0.013	0.34	0.011	0.61	0.017	1.61	0.025	1.11	0.013	0.73	0.019
$R^2$	0.631		0.662		0.579		0.730		0.800		0.769	
Scene 11479												
	Band 1		Band 2		Band 3		Band 4		Band 5		Band 7	
Least squares												
$\hat{\beta}$	42.11	0.955	13.11	0.978	14.11	1.084	13.62	0.916	15.76	0.928	5.30	1.018
se	4.29	0.034	2.37	0.032	3.27	0.032	2.34	0.023	3.11	0.018	2.02	0.021
S-estimators												
$\hat{\beta}$	19.44	1.108	4.10	1.055	2.83	1.141	7.77	0.956	9.80	0.927	1.73	1.011
se	2.17	0.017	1.07	0.015	1.31	0.012	1.06	0.010	1.82	0.010	0.89	0.009
$R^2$	0.742		0.781		0.815		0.843		0.870		0.874	

Table 5:  $S$ -estimation for Scene 11479

	Band 1	Band 2	Band 3	Band 4	Band 5	Band 7						
$\hat{c}$	1.548	1.648	1.548	1.548	2.548	1.548						
$\hat{\beta}$	13.76	1.177	4.16	1.056	3.78	1.139	7.50	0.931	10.01	0.927	1.96	1.013
se	1.19	0.015	0.99	0.014	1.12	0.011	0.90	0.009	1.71	0.010	0.78	0.008
$R^2$	0.680		0.757		0.763		0.776		0.894		0.812	

Table 5 illustrate this data-driven method using the data from the Scene 11479.

One fundamental question is, should we use regression or inverse regression? It is not clear which is  $X$  and which is  $Y$  variable for the regression. In the method adopted by our group, the reference values (e.g. calibrated 1994 images) are treated as the responses, and overpass values (e.g. 2005 raw image) are treated as the covariate for the linear regression. This setup is mainly due to simplifying the calculation instead of the careful inspecting different response and covariate choices.

For satellite images, the raw data values recorded by the sensors on the satellite are not consistent over time or between scenes. Before images from different dates and sites in different scenes are comparable, they need to be calibrated to common reference values. This procedure is so-called image calibration. Calibration is an important step in satellite data analysis.

For any two sets of digital counts  $DN^1$  and  $DN^2$  coming from reference image and raw image each, a linear relationship is assumed in the image calibration,  $DN_k^1 = \alpha_k + \beta_k DN_k^2$ , where  $\alpha$  and  $\beta$  are the gain and offset, and  $k$  denotes the image band. In the method adopted by our group, the reference values (e.g. calibrated 1994 images) are treated as the responses, and overpass values (e.g. 2005 raw image) are treated as the covariate for the linear regression. This setup is mainly due to simplifying the

calculation instead of the careful inspecting different response and covariate choices.

In theory, responses and covariates cannot be arbitrarily destined for linear regression. Ideally, covariates in linear regression should be as less as possible contaminated. People often ignore this in practice. It is not straight forwards to determine  $X$  and  $Y$  for the linear regression in calibration procedure given the possibility that the two images may be contaminated due to cloud, misregistration, etc, and measure errors in digital counts due to sensors. But a ignored fact is that different results obtained by different choice of  $X$  and  $Y$ .

For a real data set coming from Landmonitor 2005 Project, the 2005 raw image of Newdegate needs to be calibrated to 1994 Newdegate calibrated image. There are at least two setup for regression:

- Regression with reference values chosen as response,

$$V_{1994} = \alpha_k + \beta_k V_{2005} + \epsilon$$

Denote the estimation of gain and offset to calibrate the raw data  $(\alpha_k, \beta_k)$  as  $(\hat{\alpha}_k, \hat{\beta}_k)$  for band  $k$ .

- ‘Inverse regression’ by assuming the true linear relationship is  $V_{2005} = \alpha'_k + \beta'_k V_{1994}$ . By denoting the estimation of  $(\alpha'_k, \beta'_k)$  as  $(\hat{\alpha}'_k, \hat{\beta}'_k)$ , after simple transition, we get new estimation of gain and offset to calibrate the raw data for band  $k$  as  $(-\frac{\hat{\alpha}'_k}{\hat{\beta}'_k}, \frac{1}{\hat{\beta}'_k})$ .

It is also worth investigating how to obtain more reliable predictions based on the measurement error models, or alternative structure models may be considered.

## 4 Rank Regression

Rank estimation is known to be robust and retains high efficiency; see for example Hettmansperger & McKean (1998, Ch.1-3). In this section we outline the estimation of multiple regression parameters. Observations are  $Y_i = X_i^T \beta + \epsilon_i$ , for  $i = 1, \dots, n$ , where  $\{\epsilon_i\}$  are independent and identically distributed errors,  $\beta$  is a regression parameter vector, and  $X_i^T$  is the  $i$ th row of a known design matrix  $X$ . The  $i$ th residual is  $e_i = Y_i - X_i^T \beta$ . The standard form of objective function for regression models whose errors are independent and identically distributed but not necessarily symmetric is the Jaeckel (1972) criterion

$$T(\beta) = \sum_{i=1}^n e_i \left\{ \text{Rank}(e_i) - \left( \frac{n+1}{2} \right) \right\} = \frac{1}{2} \sum_{i < j} |e_i - e_j|; \quad (9)$$

see Hettmansperger and McKean (1998).

Much work has been done for independent cases. A challenging problem is how to incorporate spatial correlations in this approach. The optimal linear estimating functions require knowledge on joint error distributions. As far as I know, there is no published work on this topic. In the context of clustered/longitudinal data analysis, Wang and Zhu (2006) and Wang and Zhao (2006) have made some contributions in incorporating correlations within each cluster (or block). For landsat data analysis, one idea is to we may divide the image into  $N$  blocks so that we can roughly assume there is negligible correlations between blocks. The approach of Wang and Zhu (2006) will be applicable to account for correlations within each blocks in statistical inferences. Some other approaches are worth investigating as well.

## 4.1 Independence Model

One simple idea is to minimize the independent sum  $T(\beta)$  as a vehicle for parameter estimation. It is easy to see the resultant ‘score’ functions are Fisher consistent, i.e., having 0 expectations. This native approach is to ‘bury your head in the sand’ and to pretend there is no correlations at all for parameter estimation. It is actually a valid approach. The only task is now to work out the asymptotic covariance for the corresponding parameter estimates. Such covariance, of course, in general will depend on the true underlying correlation patterns. The downside is this approach can be very inefficient when there is strong correlations.

## 4.2 Weighted Ranks

In general, if the errors are correlated, we can consider the following modified objective function,

$$T^*(\beta) = \frac{1}{2} \sum_{i < j} w_{ij} |r_i - r_j|, \quad (10)$$

where  $w_{ij}$  is 1 when  $\epsilon_i$  and  $\epsilon_j$  are uncorrelated. For example,  $w_{ij} = 1 - \rho_{ij}$ , where  $\rho_{ij}$  is the correlation between the ranks of  $\epsilon_i$  and  $\epsilon_j$ .

The score functions for  $\beta$  are

$$U(\beta) = \sum_{i=1}^N \sum_{j=1}^N w_{ij} (X_i - X_j) \text{sgn}(e_i - e_j),$$

It will be useful if some simple and efficient weighting  $w_{ij}$  can be constructed for

More applications will be introduced in Section (9).

## 5 Quantile Regression

### 5.1 Introduction

Quantile regression as a robust alternative to the least squares approach has been found useful in economics and many other areas since Koener and Bassett (1978; 1982). However, there are still computational issues on estimation of the regression parameters and especially calculation of the standard errors due to unsmoothness of the score functions (Chen and Wei, 2005). Koenker and Ng (2005) provided new developments using linear programming to gain computational efficiency in finding the estimates. They also advocated more research on developing an inference apparatus for quantile regression.

Suppose we are interested in the parameter vector  $\beta$  of dimension  $p$  which are related to observations  $Y$ . Let  $U(\beta) = U(Y, \beta)$  be a vector of estimating functions for  $\beta$  derived from some criterion such as the maximum likelihood or any other objective functions. The parameter estimator  $\hat{\beta}$  can then be obtained by solving  $U(\beta) = 0$ . In general, the estimator can be found easily by modern computational methods with the help of high-speed computing facilities.

A few authors have considered how to compute confidence intervals. Wu & Wei (2005) provided a good review. The major concern is computational complexity for both the rank-score method of Koenker (1994) and resampling methods of Parzen, Wei & Ying (1994) and Markov Chain Marginal bootstrap proposed by He & Hu (2002). In this section, we provide a simple method to obtain the asymptotic covariance of the regression estimators. The method is practically useful for both small and large data sets because it does not require intensive resampling or any subjective choices.

## 5.2 The Model

Assume that  $y_i$  ( $1 \leq i \leq n$ ) is the observed vector and  $X_i$  is the associated covariate vector of dimension  $p$ . The quantile regression model assumes that

$$y_i = X_i^T \beta + \epsilon_i,$$

where  $\epsilon_i$ ,  $1 \leq i \leq n$ , are random variables satisfying  $\Pr(\epsilon_i \leq 0) = \tau$ . Parameter estimate  $\hat{\beta}$  of  $\beta$  is obtained by minimizing the following loss function

$$L(\beta) = n^{-1} \sum_{i=1}^n \rho(y_i - X_i^T \beta), \quad (11)$$

where the loss function

$$\rho(u) = \begin{cases} (1 - \tau)|u| & \text{if } u \leq 0 \\ \tau|u| & \text{if } u > 0 \end{cases}$$

with the corresponding derivative  $\psi(u) = \text{sgn}(u) + \tau - 1$ , where  $\text{sign}(u)$  takes the sign of  $u$  (-1 or 1).

The corresponding ‘score’ function for  $\beta$  is

$$U(\beta) = n^{-1} \sum_i X_i \{\text{sgn}(y_i - X_i^T \beta) + \tau - 1\}.$$

It is easy to see that the covariance function of  $U(\beta)$  is given by

$$B = \tau(1 - \tau)n^{-2} \sum_i^n X_i X_i^T.$$

We assume that, as  $n \rightarrow \infty$ ,  $B$  converges to a positive definite matrix. However, it is not so easy to obtain an estimator for  $D$  because it involves the common distribution of  $\epsilon_i$ .

Taking  $\tau$  as 0.5, we obtain the median regression model. As we know, obtaining the asymptotic covariance using this simple linearization relies on  $U(\beta)$  being smooth in  $\beta$ , which is not the case here.

If  $\Lambda$  is the asymptotic variance of  $\hat{\beta}$ , we may express  $\hat{\beta}$  as  $\beta + Z$ , where  $Z$  follows the standard multivariate normal  $\sim N(0, \Lambda)$ , and  $\Lambda$  is  $O(1/n)$ . Define  $\sigma_i^2 = X_i^T \Lambda X_i$ ,  $a_i = y_i - X_i^T \beta$  and  $b_i = a_i / \sigma_i$ .

Let  $\beta_0$  be the true parameter value. Usually, the distribution of  $U(\beta_0)$  is approximately zero-mean normal with a variance-covariance matrix  $\text{cov}\{U(\beta_0)\}$ . We have, under certain regularity conditions, that the delta method can be applied to obtain the asymptotic covariance of  $\hat{\beta}$  as

$$\Lambda = D^{-1} B (D^{-1})^T, \quad (12)$$

where  $D = \partial E\{U(Y, \beta_0)\} / \partial \beta$ . If  $U$  were smooth, we could estimate  $D$  by  $\hat{D} = \partial U(Y, \hat{\beta}) / \partial \beta$  evaluated at  $\hat{\beta}$ . However, if  $U$  is not smooth and these derivative do not exist at certain points as in quantile regression and rank estimation,  $D$  will depend on the unknown underlying density function, which makes it difficult to obtain an estimate for the covariance matrix of  $\hat{\beta}$ . Recently, Zhu, Shao and Wang (2006) has eliminated this problem by applying the asymptotic induced smoothing. This makes minimization much easier computationally because it strictly convex and smooth. They have also used the food expenditure data to illustrate the estimation procedure, and in the process of applying the technique in analyzing a hydrological dataset. It is interest to explore this approach when analyzing the landsat data sets as an alternative robust approach.

### 5.3 More Details?

This approach is applied to frequent analysis in hydrology, which is fundamental in water resources research. Rao and Hamed (2000) and Reiss and Thomas (2001) provide detailed methodological development and references for flood frequency analysis. More

details on the data analysis and simulation results can be found in Zhu, Shao and Wang (2006).

## 6 Least Trimmed-squares for Image Calibration

### 6.1 Introduction

Suppose  $e_{(i)}$  are the ordered residuals from (the smallest to the largest). The least trimmed squares approach is to minimize the sum of the  $h$  smallest squares (excluding  $n - h$  residuals with larger squares),

$$T = \sum_{i=1}^h e_{[i]}^2,$$

where  $e_{(i)}^2$  are the ordered residual squares. The breakdown value is  $1 - h/n$ .

Satellite images are often collected from different dates in different scenes. In order to make these images comparable so that we can effectively monitoring changes through time in land conditions, it is essential to calibrating all images to a common image under standard conditions which as relatively small noises. This is also important when mosaicing images together to remove 'edging effects' due to the fact that different images are taken under different atmospheric conditions including solar zenith angle and atmospheric temperature, humidity and haze. Furby and Campbell (2001) provided a number of approaches to the calibration of images. In calibrating images from different dates, a training dataset consisting of invariant targets is created by the user. The invariant targets may consist of ocean, rock outcrops and sandy beach. A simple linear model

$$y = a + bx + \epsilon,$$

is assumed for calibrating the images. Here  $a$  is the offset and  $b$  is the gain for calibrating the raw images. When a number of targets including dark and bright objects are

chosen and their digital counts  $y$  and  $x$  are obtained in the reference image and the raw image, parameter estimates of  $(a, b)$  can be obtained from traditional regression method (preferably by a robust approach, see Furby and Campbell, 2001).

However, identifying potential targets for many different images across the whole continent each year only requires expert knowledge but also very time consuming. Also, different experts may come up with different sets of targets and hence resulting different calibrations. Note that such difference does **not** seem to diminish as the sample size becomes larger! To avoid this *ad hoc* approach, one may consider applying the least trimmed-squares, possibly least heavily trimmed-squares, because we will only select around 200 targets among millions of pixels.

There are six bands in Landsat TM data. All six bands trimmed-residuals can be pooled together to obtain six sets of parameters based on the same set of targets. That is, trimming is based on the the sum of six bands sum of squared residuals.

If the images are consist of a proportion of farmed area, and the reference image is taken when the crops are young and green while the raw image are taken when the crops are close to harvest. These areas will be mistakenly chosen as targets because a linear transformation can make their six band counts very similar. Such calibration will distort the crop pixels in the raw image so that they look like the young crops.

To avoid selecting such false targets, it is necessary for the selected targets to cover a range of different targets. We suggest to obtain 10-20 strata based on Band 5 values so that a fixed number of targets such as (10-20) are chosen within each strata.

It is also necessary to stratify the target selection across a range of band values. The following R code illustrates how calibration parameters can be estimated by choosing as et of targets across a range of band 5 values. Specifically, choose the same number of

targets in each band quartiles. Note that this is only intended as an illustration. Much more fine-tuning is necessary. The squared residuals can also be weighted in the overall objective function to reflect different random variations in each band. The following example uses  $(0,0.5,1,1,1,1)$  as the weighting for the six band.

## 6.2 Potential Targets

Ideally, remote sensed data are all calibrated based on sensor radiometric parameters taking account of the atmospheric conditions. However, such approach is not always practical due to the relatively high costs or lack of important ground information. In this paper, we will focus on regression based on band values from pseudoinvariant targets. As described by Furby and Campbell (2001), potential targets from the Western Australian wheatbelt include:

- Dark targets (Ocean Lakes, Water in dams and reservoirs)
- Mid-range targets (Rock outcrops, Airfields, Quarries, gravel scrapes and open mines)
- Bright targets (Roaded catchments, Beach sand and Bare ground).

In the past, the MMM group selects the targets for each scene manually, which is quite labor-intensive. Identifying potential targets for many different images across the whole continent each year not only requires expert knowledge but also very time consuming. Also, different experts may come up with different sets of targets and hence resulting different calibrations. Note that such difference does not seem to diminish as the sample size becomes larger! To avoid this *ad hoc* approach, one may consider applying the least trimmed-squares, possibly least heavily trimmed-squares, because we will only select around 200 targets among millions of pixels. There are six bands in Landsat TM

data. All six bands trimmed-residuals can be pooled together to obtain six sets of parameters based on the same set of targets. It is also necessary to stratify the target selection across a range of band values.

### 6.3 The Model and the Algorithm

In this section, we will present a new approach that automatically detects these pseudoinvariant targets. Let  $i$  index the sequence of selected targets, from 1 to  $n$ , and  $j$  index the band number, 1 to 6. Suppose  $x_{ij}$  and  $y_{ij}$  are the band  $j$  counts from target  $i$  in the raw image and the reference image, respectively. We shall consider the ordinary regression model (see Furby and Campbell, 2001),

$$y_{ij} = a_j + b_j x_{ij} + \epsilon_{ij},$$

where  $\beta_j = (a_j, b_j)$  are the offset and gain parameters for band  $j$ . The errors  $\epsilon_{ij}$  are independent of each other. Suppose we have already identified  $n$  invariant targets. Denote the vector  $(1, x_{ij})^T$  as  $X_{ij}$ , a  $2 \times 1$  vector. The least squares estimates are not robust against outliers although consistent under the assumption of  $E\epsilon_{ij} = 0$ . Pros and cons for manual selection

- Expert knowledge can be utilized
- The judgement can be subjective
- Outliers can present
- It is not obvious how to incorporate the uncertainty in human selection into standard errors of the parameter estimates, which are required in quantifying the calibrated maps.

- Labour intensive and time consuming
- Not easy to pass on such technology to others

Suppose  $N$  is the total number of pixels. For a given set of values for  $\beta_j$ , ( $j = 1, 2, \dots, 6$ ), we can calculate the residuals for all the pixels as  $e_{ij} = y_{ij} - \beta_j X_{ij}$ ,  $i = 1, 2, \dots, N$ . Let  $d_i = \sum_{j=1}^6 e_{ij}^2$  to represent the goodness of fit for pixel  $i$ , and  $[i]$  are the pixel numbers so that  $d_{[i]}$  are in order from the smallest to the largest. In calibrating images from different dates, a training dataset consisting of invariant targets is created by the user. The invariant targets may consist of ocean, rock outcrops and sandy beach. We propose to select the  $n$  pseuinvariant targets by minimizing the sum of the first smallest  $n$  squared residuals,

$$L = \sum_{i=1}^n d_{[i]} = \sum_{j=1}^6 \sum_{i=1}^n e_{[i]j}^2.$$

where  $d_{[i]} = \sum_{j=1}^6 e_{[i]j}^2$  is the residual when ordered by  $d_i$ . This approach is known as the least trimmed squares approach. The breakdown value is the proportion being trimmed,  $1 - n/N$ , which is often taken as 0.1 or 0.25 or up to 0.50 in practice (only a small proportion such as 10% or up to 50% of the data can be trimmed. In our case, we will trim more than 99.99% because targets are very rare. If the images consist of a proportion of farmed area, and the reference image is taken when the crops are young and green while the raw image are taken when the crops are close to harvest. These areas will be mistakenly chosen as targets because a linear transformation can make their six band counts very similar. Such calibration will distort the crop pixels in the raw image so that they look like the young crops. To avoid selecting such false targets, it is necessary for the selected targets to cover a range of different targets. We suggest using 10-20 strata based on Band 5 values so that a fixed number of targets such as (10-20) are chosen within each strata.

The computational algorithm can be summarized as follows.

1. Choose  $(0, 1)$  as the initial values for all  $\beta_j$  and obtain the residuals  $e_{ij} = y_{ij} - \beta_j X_{ij}$ .
2. Obtain  $d_i$  and hence sorted  $d_{[i]}$
3. Minimizing the sum of the first smallest  $n$  squared residuals,  $L$ , and obtain updated  $\beta_j$  estimates.
4. Repeat Step 1 to 3 and obtain the updated estimates of  $\beta_j$ .

There is no need to reiterate.

Weighted scheme Different band responses have different noise levels. To reflect this, we introduce a weighted version of the objective function,

$$L_w = \sum_{i=1}^n d_{[i]},$$

where  $d_i = \sum_{j=1}^6 e_{ij}^2 / \sigma_j^2$ .

where  $\sigma_j^2$  is the noise variance for band  $j$ . The squared residuals can be weighted in the overall objective function to reflect different random variations in each band. In lack of such parameter estimates, we can employ the following iterative algorithm.

1. Minimize  $L$  without weighting and obtain initial residuals for each band.
2. Calculate the residuals  $e_{ij}$  and obtain the sample variance  $\hat{\sigma}_j^2$ .
3. Minimize  $L_w$  using  $w_j = \hat{\sigma}_j^2$  and obtain initial residuals for each band.
4. Iterate between Step 3 and 4 until convergence.

In general, step is not necessary as further iteration will hardly change the estimates. When a number of targets including dark and bright objects are chosen and their digital counts  $y$  and  $x$  are obtained in the reference image and the raw image, parameter estimates of  $\beta$  can be obtained from traditional regression method (preferably by a robust approach, see Furby and Campbell, 2001).

The following R code illustrates how calibration parameters can be estimated by choosing a set of targets across a range of band 5 values. Specifically, choose the same number of targets in each band quartiles. Note that this is only intended as an illustration.

```
pn1=length(d05[,1])
pseq= seq(0, 1, 0.25)
sx=matrix(nrow=6,nc=length(pseq))
medy=NA*(1:6)

for (k in 1:6)
  sx[k,]=quantile(d94[,k], probs = pseq)
nstrata=length(pseq)-1

autarg<- function(theta)
{
  #print(theta)
  td05=t(t(d05[,1:6])*theta)
  dd=(d94[,1:6] - td05)
  intcpt=apply(dd,FUN=mean,2)
  dd= (t(t(dd)-intcpt))^2
  ## discount band 2 contributions
  dd[,2]=dd[,2]/2

  ##remove band 1 residual contributions;
  mse=apply(dd[,2:6],FUN=mean,1)
  ## stratify first by base data;

  trimsse=0 tarset94=tarset05=NULL
  sx[,nstrata+1]=sx[,nstrata+1]+1

  ## use band 5 only for strata
  sx=sx[5,]
```

```

dsx=abs(diff(sx)^1.5)/sum(abs(diff(sx)^1.5))
## total target is about 100
nsx=floor(dsx*100)+1 for (ii in 1:nstrata)
{
  qi= ((d94[,5]>= sx[ii]) & (d94[,5] < sx[ii+1]))
  subdata94=d94[qi,]
  subdata05=d05[qi,]
  smse=mse[qi]
##err=sdd*NA print(length(smse))
## band-weight may be considered in future

### excluding band 1 in selection;
  torder=order(smse)[1:nsx[ii]]
  tarset94= rbind(tarset94,subdata94[torder,])
  tarset05= rbind(tarset05,subdata05[torder,])
  trimsse=trimsse + sum(smse[torder])
}

## to save my target;
mytar <<- tarset05[,7:8]
par(mfrow=c(2,3))
for (i in 1:6)
{
  plot(tarset05[,i],tarset94[,i],xlab="Over Pass",ylab="Ref 94")
  title(paste("band",i))
}
trimsse
}

g0=c(1.1860728, 0.8669934, 1.0340040, 0.9714478, 0.9822812, 1.0585260)
g=optim(g0,autarg,
control=list(maxit=20, reltol=0.001))
print(g)

```

Application to land satellite data is now carried out. Preliminary results look very promising.

Results will be submitted for publication.

The following two scenarios will be considered for testing.

1. We use 1994 as the reference image. We now manually change the raw images by multiplying 0.95 (plus 5 as offset) to all band 1-3 counts and 1.05 (minus 5 as offset)

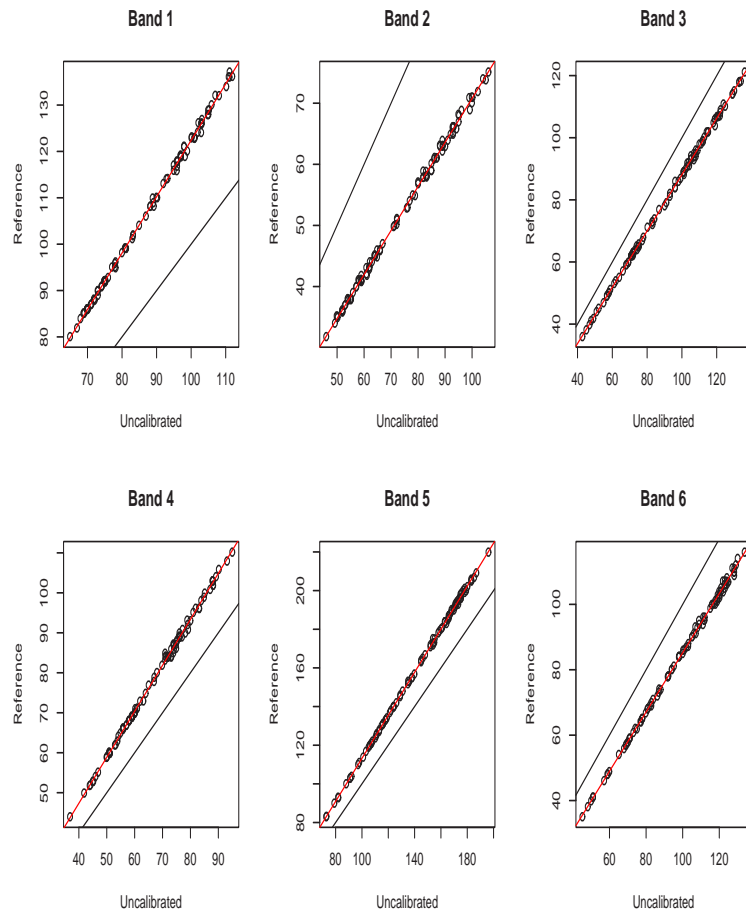


Figure 3: Regression for the automated target selection data (2000 vs 1994) for Ben-cubbin, WA.

to band 4-6 counts. In theory we should obtain the same calibrated image. The corresponding parameter estimates for band 1-3 are multiplied by 0.95 and for band 4-6 by 1.05. All values are then truncated to be between 0 and 255 to be realistic.

2. We now carry out a simulation study using 1994 as the reference image. The raw image is created by first perturbing each pixel with a random noise  $N(0,100)$  (except 250 selected targets), and then applying a linear transformation to all six bands with  $\beta_j = (5, 0.95)$  for  $j=1, 2, 3$ , and  $\beta_j = (-5, 1.05)$  for  $j=4, 5$  and 6. All values are then truncated to be between 0 and 255 to be realistic.

## 6.4 Generalization to Time Series Images

Suppose we have a sequence of raw images over time together with one reference image, and we wish to calibrate all the time series image to the reference one. Let  $d_i = \sum_{j=1}^6 \sum_{t=1}^s e_{ijt}^2$ , where  $e_{ijt}$  is the residual from pixel  $i$  band  $j$  and raw image  $t$ . Again, let  $d_{[i]}$  be the ordered  $d_i$ . Using obvious notation, we consider minimize the total

$$L_w = \sum_{i=1}^n d_{[i]} = \sum_{i=1}^n \sum_{j=1}^6 \sum_{t=1}^s \frac{e_{[i]jt}^2}{\sigma_j^2}.$$

For given targets, statistical inferences can be carried out using the traditional linear regression theory. However, the challenge here is how to incorporate the **uncertainties due to the fact that targets are estimated from the data instead of given**. We propose a resampling method for calculating the standard errors of the parameter estimates. This resampled method can be regarded as a smoothed version of bootstrapping (Wang and Zhu, 2006). Suppose  $w_{ijt}$  are independent random numbers sampled from a normal distribution with unit mean and variance,  $N(1, 1)$ . Each  $w_{ijt}$  is associated with each pixel. We obtain a set of ‘bootstrapped’ estimates of  $\hat{\beta}_{jt}$  for  $j = 1, 2, \dots, 6$  and

$t = 1, 2, s$ , by minimising the following objective function,

$$L_w^* = \sum_{i=1}^n d_{[i]} = \sum_{i=1}^n \sum_{j=1}^6 \sum_{t=1}^s w_{[i]jt} \frac{e_{[i]jt}^2}{\sigma_j^2}.$$

We repeat this process for a number of times ( $B$ , say) by using an independent set of  $w_{ijt}$  each time. These  $B$  copies of parameter estimates can be used for statistical inference.

## 7 Model Selection

Goodness of fit is essential in statistical analysis. As we know, ‘all models are wrong’, the question is whether our fitted model is ‘adequate’ for our purpose. Much work has been done on procedures/algorithms for selecting significant explanatory variables. Here I will focus on selection for the error models. For likelihood-based methods, the  $-2\log$ likelihood can be used as a criterion for selecting nested models. When one model is not a special case of the other, it may rely on modified criteria such as BIC, AIC or some kind of predictive errors.

For marginal models when we only specify the mean and variance/covariance functions, likelihood-based approaches cannot be used. Consider using the landsat data for environmental monitoring, we wish to incorporate spatial patterns in predicting the status of each pixel. One approach is to use the neighborhood information (Kiiveri, Caccetta and Evans 2001). The neighborhood information reflects the spatial correlation. If two different correlation functions are used, how do we tell which one is more appropriate? Criteria must be established based on rigorous scientific principles.

Here I provide two possible approaches for further investigation.

## 7.1 Introduction

A methodological imperative emerging from recent developments is clarification of how protections against infeasibility or severe inefficiency might be obtained without sacrificing a parsimonious agnosticism regarding true covariance structure. Positive remedies in the literature to date include suggestions by Crowder (1995) for always-feasible covariance parameter estimators (Chaganty, 1997). Robustness properties of an asymptotic nature were elaborated by Liang and Zeger (1995). In practice, the working correlation matrix is often parameterized by a vector of second-order parameters,  $\alpha$ , that are to be estimated jointly with  $\beta$ . Introduction of a second set of estimating functions for covariance matrix parameters has been proposed by various authors, including Zeger and Liang (1986), Lipsitz, Laird and Harrington (1991), Prentice and Zhao (1991), Liang, Zeger and Qaqish (1992), Hall and Severini (1998) and Wang and Carey (2004).

I will first discuss estimation and model selection methods for parametric families of covariance matrices approximating many specific models of interest in the context of GEE, and provide a variety of substantively interesting covariance structures not readily accommodated in standard approaches to GEE modeling. I then suggest a variety of working criteria for model selection. These criteria can be easily adopted in selecting models when analyzing landsat TM data even if spatial correlations are incorporated. In §(7.4), I will introduce a general method which is also applicable even when the predictive function is not smooth. It will also be of interest to see how it performs when applying to random forests.

## 7.2 Working Covariance Models

In the framework we are studying, there are  $K$  vector observations indexed by subscript  $i$ ; vector elements are double-subscripted. Outcome vectors  $Y_i = (y_{ij})$  are of dimension

$n_i \times 1$ , and constant covariate or design data are recorded in  $X_i$ ,  $n_i \times p$ , which will include a column of units if an intercept term is desired. Additional information on the proximities of outcome elements to one another are recorded in the  $n_i \times u$  “distance matrix”  $C_i$ . For longitudinal data,  $u = 1$  and  $C_i$  is the vector of observation times, which need not be integer-valued. For data obtained in a space of  $k$ -dimensions,  $u = k$  and the  $j^{\text{th}}$  row of  $C_i$  gives the coordinates of observation  $Y_{ij}$ . Finally, to accommodate heterogeneities of clustering, a further  $n_i \times q$  matrix  $B_i$  of subcluster indicators may be employed.

We let  $g(\cdot)$  denote a link function for a generalized linear model (McCullagh and Nelder, 1989) and  $\text{cov}(\alpha; C, h)$  be a function specifying a covariance matrix through a (potentially nonlinear) function  $h$  of a  $q$ -dimensional parameter  $\alpha$  and coordinates  $C$ . Statistical inference proceeds under the following parameterizations of the first two moments of  $Y_i$ :

$$E(Y_i|X_i) = g^{-1}(X_i\beta) = \mu_i(\beta),$$

and  $\text{Cov}(Y_i|C_i, X_i)$  may be specified as  $\text{Cov}_i(\alpha; C_i, h) = \sigma^2 A_i^{1/2} R_i A_i^{1/2}$ , where  $R_i$  is a correlation matrix. The univariate specification is  $E(y_{ij}) = \mu_{ij}$  and  $\text{var}(y_{ij}) = \sigma^2 \sigma_{ij}^2$ . The mean model is a familiar generalized linear model (GLM) specification, and the covariance factor  $A_i = (\sigma_{ij})$  is the GLM variance function or a generalized version with additional parameters to be more flexible (Wang and Zhao, 2006). The solution to the GEE

$$\sum_i \frac{\partial \mu_i^T(\beta)}{\partial \beta} W_i^{-1}(\alpha) \{Y_i - \mu_i(\beta)\} = 0,$$

where  $W_i(\alpha)$  is some working model for  $\text{cov}(Y_i|C_i, X_i)$ , will be denoted  $\hat{\beta}_G$ .

The choice of  $W_i$  is often based on the variance function in GLMs and serial correlation models in time series models. Covariance structures deduced from random coef-

ficients or random effects models are also very useful as they have easy interpretations. In the context of overdispersed Poisson model, Thall and Vail (1990) have suggested a few interesting covariance models. Their idea is to assume  $Y_{it}$  is Poisson with mean  $\gamma_i \xi_t \mu_{it}$  where  $\gamma_i$  and  $\xi_t$  represent between-subject and temporal random effects. The induced covariance  $\text{Cov}_i = \text{diag}(\mu_i) + \text{diag}(\mu_i) \{E(\gamma_i^2) \text{cov}(\xi) + \text{var}(\gamma_i) J\} \text{diag}(\mu_i)$ , where  $J$  is a square matrix of 1's. This approach is further elaborated by Jowaheer and Sutradhar (2002) and Henderson and Shimakura (2003). For likelihood induced covariance models, such as Poisson-gamma, incorporation of higher moments in estimating functions is possible. However, such mathematically extrapolated information should not be deemed reliable unless higher moment assumptions can be verified.

The variance function is often adopted from the GLMs (McCullagh and Nelder, 1989). For example, we often use  $\phi\mu$  for overdispersed Poisson data, and  $\sigma^2$  or  $\phi\mu^2$  for continuous data. Misspecification of variance function may lead to efficiency loss for the estimators of the regression parameter (Wang and Lin, 2005). As pointed out by Crowder (2001), the framework of generalized linear models, in which the variance is assumed to depend on the parameters only through  $\mu_i$ , is often too restrictive. Useful variance functions include  $\text{var}(y_{ij}) = \phi\mu^\gamma$ , where  $1 \leq \gamma \leq 2$ , and  $\gamma_1\mu + \gamma_2\mu^2$ , as suggested by a few authors (Carroll and Ruppert, 1988; Thall and Vail, 1990; Morton, 1987; Carroll, 2003). The power function of  $\mu$  is a generalization to the overdispersed Poisson model (Paul and Plackett, 1978) and it is also applicable to continuous data (see Davidian and Giltinan, 1995).

We now present a number of examples of correlation models, with motivational discussion of data structures from several real-world examples. Let  $[M]_{jk}$  denote the  $(j, k)$  element of a matrix  $M$ .

1. Independence,  $q = 0$ :  $R_i = I_{n_i}$  for all  $i$ .
2. Homogeneous exchangeable,  $q = 1$ :  $[R_i]_{jk} = 1_{\{j=k\}} + 1_{\{j \neq k\}}\alpha$  for all  $i$ . Such a matrix will be denoted  $R_i = \text{exch}(\alpha)$  in the sequel.
3. Serial models,  $q = 1$ . The first-order Markovian process for real-valued observation times  $C_{ij} = t_{ij}$  has  $[R_i]_{jk} = \alpha^{|t_{ij} - t_{ik}|}$ . The first order moving average model can be written  $[R_i]_{jk} = 1_{\{j=k\}} + 1_{\{|j-k|=1\}}\alpha$ . We denote such matrices by  $R_i = AR(1; \alpha)$  and  $R_i = MA(1; \alpha)$  respectively.
4. Heterogeneous exchangeable,  $q = 2$ . An example of this model is the cross-sectional twin study. In this case the coordinatization data simply indicate whether twins are identical or fraternal. We let

$$\begin{aligned}
h(\alpha, C_i) &= \alpha_1 && \text{if } C_i \text{ indicates that} \\
&&& \text{twinship } i \text{ is monozygotic (MZ)} \\
&= \alpha_2 && \text{if twinship } i \text{ is dizygotic (DZ);} \\
[R_i] &= \text{exch}(\alpha_1) && \text{if twinship } i \text{ is MZ} \\
&= \text{exch}(\alpha_2) && \text{if twinship } i \text{ is DZ.}
\end{aligned}$$

With these stipulations  $2(\alpha_1 - \alpha_2)$  is a traditional measure of heritability in the outcome measure.

5. Heterogeneous generalized serial correlation models,  $q = 4$ . For concreteness, we consider the analysis of longitudinal data from a comparative trial (p.73, Diggle et al., 2002). The data consist of growth traces for Sitka spruce trees under two cultivation conditions. The intervention appears to affect the mean level of size achieved over time, the typical shape of the size growth curve, and the “visit-to-visit” variability of size. We will investigate a flexible working correlation model with parameter vector  $\alpha = (\alpha_1, \alpha_2)^t = (\gamma_1, \theta_1, \gamma_2, \theta_2)^t$ , measurement coordinates

$C_i = (t_{i1}, \dots, t_{in_i})^t$  and

$$[R_i]_{kl} = h(\alpha_j, t_{ik}, t_{il}) = \gamma_j^{|t_{ik} - t_{il}|^{\theta_j}}$$

for tree  $i$  in treatment group  $j = j(i) \in \{1, 2\}$ . Note that  $\theta_j \equiv 0$  implies  $R_i = \text{exch}(\gamma_{j(i)})$ ,  $\theta_j \equiv 1$  implies  $R_i = \text{AR}(1)(\gamma_{j(i)})$ , and that  $0 < \theta_j < 1$  will in certain circumstances parsimoniously approximate members of the AR(2) or AR(1) plus random intercept covariance families (Muñoz et al. 1992). In this example, the time-scale is exponentially deformed by the damping parameter  $\theta_j$ . Another generalization of the first-order Markovian model involves Box-Cox transformation of the observation time scale (Nuñez-Antòn and Woodworth, 1994).

6. Spatial variogram (Albert and McShane, 1995),  $q = 3$ . Each sub-plot is an 88-cell scoring of a single transcranial slice of a CT scan, with X present in grid boxes of which at least 50% of the space is occupied by lesions. The probability of lesion occupancy (mean model) of image grid boxes for subject  $i$ , indexed by  $\mathbf{s}_{ijk} = (s_{i1j}, s_{i2k})^t$ , follows a GLM depending on location  $\mathbf{s}_{ijk}$ , covariates, and interactions among these. The spatial clustering of lesion occupancies is assumed to follow an isotropic exponential semivariogram. Let  $\alpha = (a, b, c)^t$ . The parametric semivariogram has the form

$$\gamma(d; \alpha) = \begin{cases} c + b\{1 - \exp(-d/a)\} & \text{if } d > 0, \\ 0 & \text{otherwise} \end{cases} \quad (13)$$

where

$$d_{lm} = \|s_l - s_m\|$$

and  $l$  and  $m$  index cells in the plane. Finally, the elements of  $R_i]_{jk}$  is  $1 - \gamma(d_{jk}, \alpha)$ .

7. Pairwise log odds ratio regression, general  $q$ . Here the working covariance model for a cluster of binary outcomes is derived from the  $O(n_i C_2)$  conditional logistic

regressions of every outcome in a cluster on all other outcomes in that cluster using in addition a  $q$ -dimensional design matrix. Details and a fully worked example in functional brain mapping are provided in Carey et al., (1993); extensions to ordinal outcomes are provided by Heagerty and Zeger, (1996).

8. Unstructured correlation or covariance,  $q = O(\max_i n_i^2)$ . The meaningfulness of this model depends upon discrete observation times. If the study satisfies this requirement, this is often a good start in pursuit of a more parsimonious and interpretable working model.

### 7.3 Selection Criteria

Testing hypotheses on the regression parameters is typically of primary importance (Barnhart and Williamson, 1998). For example, Breslow (1990) considered two types of test statistics (Wald and score) for evaluating the significance of added variables, i.e., the models to be compared are nested. We will focus on covariance model selection for a fixed set of regression variables in the mean function.

Our intent is to determine if quasi-likelihood-based or any other goodness of fit criteria can be used from a practical viewpoint. For a given estimation method, there may be a few ways of establishing the corresponding likelihood function for inference. All these likelihood functions can be regarded as some approximation to the true likelihood function. It is therefore important to quantify the possible differences between these ‘approximate inferences’ when used for hypothesis testing or any other likelihood-based evaluations.

We may consider quasi-likelihood functions from which the corresponding estimating functions (score functions) can be derived. For example, Zhao et al (1992) have shown that the following “partly exponential” distributions produce the score functions that

are identical to the GEE1,

$$f(y_i, \mu_i, \phi) = d_i^{-1} \exp\{y_i^T \theta_i + c_i(y_i, \phi)\},$$

in which  $\phi$  is unknown. Let us consider the following exponential model for multivariate discrete and continuous outcome

$$f(y_i, \mu_i, \alpha) = c_i^{-1} \exp\{(y_i - \Phi_i)^T \Omega_i (y_i - \mu_i) + \Lambda_i^T \Theta_i\},$$

where  $\Theta_i$  is a vector consisting of three- and possibly higher-way cross-products of  $y_{ij}$ , and  $c_i$  is a normalization constant. If we restrict the model up to 2-way cross-products only, i.e.,  $\Lambda_i = 0$ , it becomes the quadratic exponential model proposed by Prentice and Zhao (1991). For any specified  $\Lambda_i$ ,  $\mu_i$  and  $\Omega_i$  can be determined from given marginal mean and covariance, used in the GEE estimation. To find how estimation (parameter estimates, likelihood ratio statistics) might be sensitive to high order assumptions or likelihood specifications, we can plot the relevant statistics of interest for a range of  $\Lambda_i$  values. For example, in testing two different models (not necessarily nested), if the goodness of fit statistic for one model is uniformly better than the corresponding values in the other model over a range of  $\Lambda_i$  values, we would be more confident in recommending the better-fitting model. We can also carry out simulation studies to quantify the impact when the covariance matrix or higher moments are misspecified, and hence obtain properties of different approximate likelihood functions.

From a practical viewpoint, it is of great interest to evaluate some simple available criteria to see if they are sufficient despite the unavailability of optimality justifications. For example, if estimation is only based on mean and variance assumptions, we can use the form of multivariate normal likelihood function as a quasi-likelihood for inference.

Recently, Hall and Severini (1998) established asymptotical properties for the Gaussian

estimates. However, it is not clear how these constructed quasi-likelihood would perform when used for log-likelihood ratio test. If  $p+q$  are the total number of parameters in fitting the model, we propose the following Gaussian criteria for model selection.

$$L_G = \sum_i \{(y_i - \mu_i)^t V_i^{-1} (y_i - \mu_i) + \log(|V_i|)\} + 2(p+q).$$

One possible improvement is to take account of the loss of degrees of freedom due to estimating  $\beta$ . This penalized Gaussian likelihood can be simply regarded as a vehicle for model selection. Suppose  $\tilde{V}_i$  is the true covariance matrix, we have  $E(L_G) = \sum_i \text{trace}(V_i^{-1} \tilde{V}_i) + \sum_i \log(|V_i|) + 2(p+q)$ . From likelihood perspective,  $l(y) = l(y|\hat{\beta})l(\hat{\beta})$ . Assuming normality for  $\hat{\beta}$  would result in an extra term of  $(\beta - \hat{\beta})^T (\beta - \hat{\beta}) + \log[\det\{\text{cov}(\hat{\beta})\}]$  in the  $-2\log$ -likelihood function. We can therefore consider a “REML” version of  $L_G$  (conditional on  $\hat{\beta}$ ),

$$L_{RM} = L_G + \log[\det\{\sum_i D_i V_i^{-1} D_i\}],$$

where the last term is an approximation to  $-\log[\det\{\text{cov}(\hat{\beta})\}]$ .

However, it is not clear if these simple criteria are practically adequate, and the performance of these criteria and their derivatives has not been systematically assessed. In section 4, we will investigate their performance numerically and analytically.

We now adopt a block matrix notation to avoid subscripts, so that, e.g.,  $D^t V^{-1} D$  denotes  $\sum_i D_i^t V_i^{-1} D_i$ , with  $D_i = \partial \mu_i / \partial \beta^t$ . The sandwich covariance estimator for  $\hat{\beta}$  under working covariance model  $W$  is then similar to (12),

$$\hat{V}_\beta = (D^t W^{-1} D)^{-1} [D^t W^{-1} \text{cov}(Y) W^{-1} D] (D^t W^{-1} D)^{-1} = Q_0^{-1} Q_1 Q_0^{-1}.$$

The (working) model-based or “naive” covariance estimator is  $\hat{W}_\beta = Q_0^{-1}$ . In practice  $Q_1$  is evaluated as  $\sum_{i=1}^K D_i^T W_i^{-1} e_i e_i^T W_i^{-1} D_i$ , in which  $e_i = y_i - \mu_i$ .

The working Wald test statistic

$$T_w^* = (\hat{\beta}_G - \beta_0)^T \hat{W}_\beta^{-1} (\hat{\beta}_G - \beta_0).$$

Theorem 1 of Rotnitzky and Jewell (1990) gives the asymptotic distribution of the working Wald test statistic:  $T_w^* = \sum c_i \chi_j^2 + o_p(1)$ , where  $\chi_j^2$  are independent  $\chi_1^2$  variates and  $c_j$  are eigenvalues of  $Q = Q_0^{-1}Q_1$ . They noted that the  $c_j$  are all identically unity whenever the working covariance and true covariance coincide, and proposed that the two-dimensional criterion with components  $\bar{c}_1 = \sum c_j/p = \text{tr } Q/p$ ,  $\bar{c}_2 = \sum c_j^2/p = \text{tr } Q^2/p$  be inspected as a measure of working correlation model structure adequacy.

As a naive formalization of this proposal, we consider the criterion

$$\Delta_0 = \|(\bar{c}_1, \bar{c}_2)^T - (1, 1)^T\|^2,$$

which should be small for correct structure selection and large for incorrect structure selection. When  $V_i$  is the true underlying covariance matrix, all  $c_i$  should be close to unity. Note that if and only if all  $c_i = 1$ , we have  $Q_1 = Q_0$ . It is therefore appropriate to consider the following criteria for covariance model selection

(i)  $\Delta_1 = \sum_i (c_i - 1)^2/p = \bar{c}_2 - 2\bar{c}_1 + 1$ ;

(ii)  $\Delta_2 = \sum_i \{\log(c_i)\}^2$ .

## 7.4 Absolute Predictive error

If  $\hat{y}(\hat{\beta}, x)$  is the predicted value at  $x$ . One measure of goodness of a model is to evaluate the prediction error for a future response (testing datum),  $D = d(y, \hat{y})$ . The ‘apparent’ estimate of  $D$  is  $\sum_i d(y_i, \hat{y}_i)/n$ . The square error  $d(y, \hat{y}) = (y - \hat{y})^2$  is often used. A robust version is to use  $d(y, \hat{y}) = |y - \hat{y}|$ .

For cases with small sample sizes, the ‘apparent’ method usually underestimate  $D$ . Jackknife or  $K$ -fold cross validation may be used to estimate  $D$ . This criterion may be developed for a robust version of the random-forest approach. Such generalization will be highly valuable for analyzing remote sensing data because of frequent presence of outliers. Such criterion can also be investigated for model selection. For example, in predicting perennial vegetation coverage using the Landsat data, we may use develop a robust index for goodness of fit to select among traditional regression models (linear regression or generalized linear regression models) and nonparametric tree-classification such as (random forests) methods.

## 8 Monitoring Land Changes

### 8.1 Classification

Currently, remotely sensed Landsat data are used to predict perennial vegetation coverage. A statistical model can be fitted to the ground data (where percent of coverage is measured). Some work has been done in incorporating woody vegetation texture to increase the mapping accuracy (woody vs nonwoody).

Caccetta and Furby (technical Report, Monitoring Sparse Perennial Vegetation Cover) investigated how texture information can improve classification. They demonstrated how to use texture data ( $x$ ) at pixel sizes 2.5, 5m, 10m, 20m, 40m and 80m based on canonical analysis.

Conditional categorical regression is an alternative approach. According to their Figure, such approach will result in perfect separation. I briefly summarize the approach below.

Suppose  $p_1$ ,  $p_2$ , and  $p_3$  are the probability of being woody ( $\geq 20\%$  coverage), Sparse

(1-20% coverage) and Bare. First, a regression model is assumed for being Woody,

$$\text{logit}(p_1) = x^T \beta.$$

Conditional on being non-woody, we further split for Sparse and Bare,

$$\text{logit}\{p_2/(1 - p_1)\} = x^T \gamma.$$

This leads to

$$\begin{cases} p_1 &= \frac{\exp(x^T \beta)}{1 + \exp(x^T \beta)} \\ p_2 &= \frac{\exp(x^T \gamma)}{\{1 + \exp(x^T \beta)\} \{1 + \exp(x^T \gamma)\}} \\ p_3 &= \frac{1}{\{1 + \exp(x^T \beta)\} \{1 + \exp(x^T \gamma)\}} \end{cases} \quad (14)$$

This parametric approach can be compared with the ML classifier. Band information can be added as well as part of the explanatory variables ( $x$ ).

One of the fundamental approaches in land monitoring is to classify the pixels into a few different vegetation types and identify those with changed vegetation types. Different images are often taken at different dates. Calibration may be necessary if direct comparison is used. Suppose data for pixel  $i$  consist of  $y_i$ , the probability being woody in the reference image, and  $X_i$ , the derived variates from remote sensed data from the new image, where  $i$  is from 1 to  $N$ . The problem here is to establish a predictive model for being woody based on the data.

A natural model is to consider a regression model such as a logistic model  $g(X_i; \beta)$  or other threshold models for the probability being woody. The tricky part is how to obtain estimates for  $\beta$ . The most commonly used approach, the least squares, is to obtain the estimates of  $\beta$  by minimising

$$\sum_{i=1}^N \{y_i - g(X_i, \beta)\}^2$$

or its robust version of the least absolute deviate (LAD)

$$\sum_{i=1}^N |y_i - g(X_i, \beta)|.$$

This  $L_1$  norm is known to be robust against outliers in the responses. However, in this case, the responses are probabilities for the base image classification, which is between 0 and 1. So robust consideration in this regard is not necessary. Also, extra care may be needed in computation when  $L_1$  norm is used because of its unsmoothness at 0. But if computation is not any issue here, this objective function is fine. Otherwise I suggest using  $L_2$  norm (the least squares approach). However, such procedures become **inappropriate** in our case because the residuals are ‘contaminated’ by possible changes in the new image. Ideally, if we knew which pixels have changed, we could just exclude them in the analysis. On the other hand, if we knew which pixels were changes, we would not need to carry out such analysis any more! Perhaps the biggest challenge here lies in excluding the unknowns. Determination of the thresholds (parameters in the prediction model) is based on minimising the sum of the absolute differences in the probabilities or its squares.

I suggest the following iterative approach.

1. Let  $n = N/2$ . Obtain an estimate of  $\hat{\beta}^{(0)}$ , by minimising  $\sum_{i=1}^n r_{[i]}^2$ , where  $r_{[i]}$  is the ordered residuals,  $\{y_i - g(X_i, \beta)\}$ , the order is based on the  $r_i^2$  values. This trimmed approach only uses pixels with the  $n$  smallest residuals. We assume here no more than 50% pixels have changed. Otherwise,  $n$  should be made smaller.

2. Calculate the initial residuals,  $\{y_i - g(X_i, \hat{\beta}^{(0)})\}$ , and obtain a preliminary assessment on the proportion of pixels being changed ( $q$ ).

3. Repeat step 1 with  $n = [(0.95 - q)N]$ . Specifically, obtain the estimate of  $\beta$ ,  $\hat{\beta}$ , by minimising  $\sum_{i=1}^n r_{[i]}^2$ . I expect further iteration is practically unnecessary. Resampling

methods can be developed for standard errors. More details can be developed using the smooth bootstrap approach (Wang & Li, 2006, Call # 2057).

## 8.2 Some Statistics for Monitoring

There are a variety of methods in the literature on how to detect changes from a sequence of data,  $(y_i)$ . An interesting nonparametric method proposed by Pettitt (1979) is to consider

$$U_t = 2 \sum_{i=1}^t \text{Rank}(y_i) - t(N + 1),$$

and use  $U_{\max} = \max_{1 \leq t \leq N} |U_t|$  as the test statistic. The ranking is based on the whole sequence for that pixel. Analytical approximations can be derived under the i.i.d. assumption. It is unlikely that the band data from the same pixel are independent over time. However, we can establish the distribution of  $U_{\max}$  under the null by examining many pixels. Changes in most pixels are due to natural variation in reflection and weather conditions, and such changes are deemed to be random noises.

For each pixel,  $U_{\max}$  can be calculated, and a mixture distribution can be fitted to partition the pixels into those with significant changes and without significant changes. Such procedure can also be used for other statistics.

Changes in each pixel is unlikely to be independent. If pixel  $a$  is changed due to logging or bush fire, its neighborhood pixels are quite likely to have a similar change. Therefore, instead of calculating the statistic  $U_{\max}$  for each pixel, we can consider incorporating the clustering effects to increase the detection power.

Such approach can also be used to identify invariant targets when the statistic  $U_{\max}$  is minimum.

## 9 Rank Estimation For Linear Models

Adaptions of weighted rank regression to the accelerated failure time model for censored survival data have been successful in yielding asymptotically normal estimates and flexible weighting schemes to increase statistical efficiencies. However, for only one simple weighting scheme, Gehan or Wilcoxon weights, are estimating equations guaranteed to be monotone in parameter components, and even in this case are step functions, requiring the equivalent of linear programming for computation. The lack of smoothness makes standard error or covariance matrix estimation even more difficult. An induced smoothing technique in Brown and Wang (2005, *Biometrika* **92**) overcame these difficulties in various problems involving monotone but pure jump estimating equations, including conventional rank regression. This section applies induced smoothing to the Gehan-Wilcoxon weighted rank regression for the accelerated failure time model, for the more difficult case of survival time data subject to censoring, where the inapplicability of permutation arguments necessitates a new method of estimating null variance of estimating functions. Smooth monotone parameter estimation and rapid, reliable standard error or covariance matrix estimation is obtained.

### 9.1 Introduction

The accelerated failure time model is a formulation of linear regression for logarithms of survival time data which may be subject to censoring. It has attracted considerable attention in recent years as an alternative to the proportional hazards model. For background and current references see Kalbfleisch and Prentice (2002, p 44), Cox and Oakes (1984, pp 64-65), and Lawless (2003, p 270).

Assume that  $T_i$  ( $1 \leq i \leq n$ ) is the failure time for subject  $i$  and that  $x_i$  is the

associated covariate vector of dimension  $p$ . Under the usual accelerated failure time model,

$$y_i = \log(T_i) = \beta_0^T x_i + \varepsilon_i,$$

where  $\beta_0$  is the true value of a  $p \times 1$  vector of regression slopes, and error terms are  $\{\varepsilon_i\}$ . Usually,  $T_i$  is subject to censoring at  $C_i$ , a random variable independent of  $T_i$ . We assume that, conditional on  $\{x_i\}$  and  $\{C_i\}$ , the  $\{\varepsilon_i\}$  are independent and identically distributed. Note that the common conditional distribution of  $\{\varepsilon_i\}$  may depend on  $\{x_i\}$ . We will regard the covariates  $\{x_i\}$  as realisations of a random variable, possibly multivariate.

The observed data may be written as a triplet  $(\tilde{T}_i, \delta_i, x_i)$ , where  $\tilde{T}_i = \min(T_i, C_i) = T_i \wedge C_i$ , and  $\delta_i$  is an indicator function,  $I(T_i \leq C_i)$ . Write  $\tilde{y}_i = \log(\tilde{T}_i)$ . Then for a trial  $\beta$  value  $e_i = y_i - \beta^T x_i$  is an uncensored but possibly unobservable residual, while  $\tilde{e}_i = \log(\tilde{T}_i) - \beta^T x_i$  is an observable but possibly censored residual. Note that  $\tilde{e}_i = e_i \wedge C_i^T$ , where  $C_i^T = \log(C_i) - \beta^T x_i$ , so that the residuals are censored whenever the observed lifetimes are censored.

If the residual  $e_j$  is censored, its rank cannot be defined. However, if  $e_j$  is uncensored, indicated event  $e_i > e_j$  can be confirmed if  $e_j < \tilde{e}_i$  regardless  $e_i$  is censored or not. Therefore, to adapt rank regression methods for uncensored data to the censored case, we consider the following modified rank for residual  $e_j$

$$R(e_j) = \sum_i I(\tilde{e}_i \geq e_j), \tag{15}$$

which applies only to uncensored residuals  $e_j$ . Note that, this ‘reversed rank’ where the largest (smallest) residuals have the smallest (largest) rank, is introduced here for notational convenience, as we will see.

Considering rank regression schemes for the accelerated failure time model, for esti-

mating  $\beta$  using these modified ranks, there are three main issues (see Geha, 1965,; Jin, Ying and Wei, 2001; Jin et al. 2003):

- (i) asymptotic normality and efficiency of estimators;
- (ii) unique estimation of  $\beta$ ; and
- (iii) estimation of standard error and covariance matrix estimation for  $\hat{\beta}$ .

The most general weighted estimating equations are those proposed by Jin et al. (2003) They have considerable flexibility, allowing the adaptive formulation of weights to increase efficiency; in addition, the estimates are asymptotically normal. Thus it is reasonable to conclude that item (i) has seen substantial progress. However, concerning (ii) computation, among the simple rank-scoring schemes with weights not depending on trial  $\beta$  values, only the Gehan (1965) or Wilcoxon weights lead to estimating equations guaranteed to be monotone in all components of  $\beta$ . The paper by Fygenon and Ritov (1994) shows that monotonicity can hold for other, more complex weighting schemes with data-dependent weights. For the Gehan-Wilcoxon scheme, outlined in Section 2, these monotone equations for  $\hat{\beta}$  are pure jump functions, and have to be solved by linear programming techniques.

The difficulties are greater still for item (iii), standard error and covariance matrix estimation of  $\hat{\beta}$ . Because of lack of smoothness, estimation of asymptotic standard errors may appear to require density function estimation of the integral  $f$ -squared expression characteristic of rank methods, a further computational burden. Another method, proposed in Jin et al. (2002) and Jin et al. (2003), employed a re-sampling technique based on repeated perturbations of the objective function.

Recently, Brown and Wang (2005) introduced an induced smoothing method, suitable for rank regression and related problems where monotone estimating functions have

jumps. Induced smoothing simplifies standard error and covariance matrix calculations for estimates. The aim of the present section is to introduce a smooth method for rank regression when data are subject to censoring. However, while the induced smoothing method simplifies covariance matrix estimation, it cannot remove non-monotonicity of estimating equations. For this reason, and also because having smooth, monotone estimating equations is essential for the induced smoothing method, its application is made only to the case of Gehan weights, which are the censored rank regression analogues of Wilcoxon and Mann-Whitney procedures.

It is straightforward to formulate most elements of induced smoothing for the Gehan-Wilcoxon weighted scheme in the accelerated failure time rank regression model. However, one component, the null variance of the test statistic, is not so easy to obtain as for uncensored data, because the presence of censoring prevents use of permutation arguments. An alternative method for estimating the null variance is proposed. The resulting formulation yields smooth monotone estimating equations, and rapid iterative convergence of covariance matrix estimation, essentially ‘one-step’ convergence in large samples.

Section 2 summarises the Gehan-Wilcoxon rank regression scheme; Section 3 outlines the induced smoothing method, which is then applied to the Gehan-Wilcoxon scheme in Section 4. An example is discussed in Section 5. The focus of this section is on the accelerated failure time model, but the methods can be applied also to other rank regression applications with censored data.

## 9.2 The Gehan-Wilcoxon Scheme

For uncensored data, the rank regression scheme seeks to minimise the objective function [?]

$$L(\beta) = 2 \sum e_i \left\{ \text{rank}(e_i) - \frac{1}{2}(n+1) \right\} = \sum_{i < j} |e_i - e_j|, \quad (16)$$

a convex function of  $\beta$  (see Hettmansperger and McKean, 1998). Estimation is through setting the gradient function to zero, i.e. by solving the following estimating equations  $S(\beta) = 0$ , where

$$\begin{aligned} S(\beta) &= \sum_{i < j} (x_i - x_j) \text{sgn}(e_i - e_j) \\ &= \sum_j \sum_i (x_i - x_j) I(e_i > e_j). \end{aligned}$$

For censored data, the indicated event  $e_i > e_j$  is confirmed only if  $e_j$  is uncensored and  $e_j < \tilde{e}_i$ , ie if  $e_j < e_i \wedge C'_i \wedge C'_j$ . Therefore it is reasonable formulate the censored data estimating function as

$$\begin{aligned} S(\beta) &= \sum_{j,i} (x_i - x_j) I(e_j < e_i \wedge C'_i \wedge C'_j) \\ &= \sum_j \delta_j R(e_j) (\bar{x}_j - x_j), \\ &= \sum_j \delta_j \sum_i (x_i - x_j) I(\tilde{y}_i - \log T_j \geq \beta^T (x_i - x_j)), \end{aligned} \quad (17)$$

where  $R$  is the modified rank, as in (15),  $R(e_j)$  being the number of residuals  $\geq$  the uncensored residual  $e_j$ , and  $\bar{x}_j$  is the mean of all covariates  $\{x_i\}$  over  $\{i : \tilde{e}_i \geq e_j\}$ . It is easily seen from (17) that this estimating function is a monotone step function in the components of  $\beta$ .

The Gehan estimator of  $\beta$  can also be obtained by minimising the following dispersion function (see Jin et al. 2003),

$$L(\beta) = \sum_{i=1}^n \sum_{j=1}^n \delta_i (\tilde{e}_i - e_j)^-, \quad (18)$$

where  $x^- = |x|$  if  $x < 0$ , and 0 if  $x > 0$ . Note that in the absence of censoring, i.e., all  $\delta_i = 1$ , we have  $L(\beta) = \sum_{i,j}(e_i - e_j)^- = \sum_{i < j} |e_i - e_j|$ , which is the same as (16). The corresponding quasi score function is the same as (17).

Because  $S$  is a step function,  $S = 0$  must be solved by linear programming methods or their equivalent, as in Jin et al. (2003), and in particular, derivatives of  $S$  are not defined, so that standard error or covariance matrix estimation is problematical. As mentioned earlier, the asymptotic covariance depends on the density function. Thus, computationally intensive methods such as nonparametric density estimation or may need to be considered. Alternatively, we now introduce in the next section a simple method based on the induced smoothing idea for asymptotic inference.

### 9.3 Induced Smoothing

Suppose observations  $Y$  depend upon a parameter vector  $\theta$  of dimension  $p$ , and that  $S(Y, \theta)$  is a vector of estimating functions for  $\hat{\theta}$ , from solving  $S(Y, \theta) = 0$ . Let  $\theta_0$  be the true parameter value. Usually, the distribution of  $S(\theta_0)$  is approximately zero-mean normal with known variance: this is the case for many procedures related to rank methods.

If  $S$  is differentiable with respect to  $\theta$ , then under certain regularity conditions the delta method can be applied to obtain the asymptotic covariance matrix of  $\hat{\theta}$ , as given by (12),

$$D^{-1}B\{D^T\}^{-1}, \tag{19}$$

where  $D = [\partial E\{S(Y, \theta)\}/\partial\theta]_{\theta_0}$ , and  $B = \text{cov}\{S(Y, \theta_0)\}$ .

In practice,  $D$  is estimated from  $\partial S(Y, \theta)/\partial\theta$  evaluated at  $\hat{\theta}$  when these partial derivatives exist, and  $B$  is either known, as in rank methods, or estimated by a suitable function of residuals, as in Section 4, for example.

However, if the derivatives do not exist, it is difficult to obtain  $D$  and hence the covariance matrix of  $\hat{\theta}$ . The asymptotic form may depend on underlying density functions, which then need to be estimated. This is often the case in  $M$ -estimation, and rank-based regression or quantile estimation.

In the induced smoothing method proposed in Brown and Wang (2005), a simple, ‘pseudo-Bayesian’ approach to estimation viewed the estimation of a parameter  $\theta$  not only as making a choice  $\hat{\theta}$ , but also stating a covariance matrix  $\Gamma$ . Thus, assuming asymptotic normality as is usually appropriate, knowledge about the unknown  $\theta$  can be summarised by a distribution of  $N(\hat{\theta}, \Gamma)$  for  $\theta$ .

This is the view that is taken at the beginning of the inference process, before data is gathered. Then, the implied smoothing of estimating equations has two consequences. First, the estimating equations  $S = 0$  for  $\hat{\theta}$  are replaced by a smoothed version, whose solutions are asymptotically equivalent to the unsmoothed estimates. Thus, the smoothed estimates will inherit asymptotically normal distributions of the original, unsmoothed estimates. In the present application, smoothed estimates of the regression parameter  $\beta$  will inherit the asymptotic normality proved by Jin et al. (2003) for the unsmoothed estimates. Second, the smoothed estimating equations are differentiable, enabling the sandwich formula (19) to be applied, and solved for  $\Gamma$ .

By estimation of  $\theta$ , a pair  $(\hat{\theta}, \Gamma)$  is nominated with  $\theta = \hat{\theta} + \Gamma^{1/2}Z$ , where the random vector  $Z \sim N(0, I_p)$ , with  $I_p$  the identity matrix of dimension  $p$ . If there are  $n$  independent observations, the elements of  $\Gamma$  are expected to be  $O(n^{-1})$ . In one-parameter problems,  $\Gamma = s^2$ , where  $s$  is standard error, playing a role similar to the bandwidth in kernel methods.

Therefore we replace  $L(\theta)$ , the original objective function for estimation of  $\hat{\theta}$ , with

$\bar{L}(\hat{\theta}) = E\{L_Z(\hat{\theta} + \Gamma^{1/2}Z)\}$ , where expectation is over  $Z$ . The derivative is

$$\bar{S}(\theta) = \partial\bar{L}(\theta)/\partial\theta = E_Z\{S(\theta + \Gamma^{1/2}Z)\}, \quad (20)$$

and the smoothed  $\hat{\theta}$  is found by solving

$$\bar{S}(\hat{\theta}) = 0. \quad (21)$$

Asymptotically, the estimation of  $\hat{\theta}$  is unchanged in replacing the non-smooth  $S$  by the smooth  $\bar{S}$ , but  $\bar{S}$  will be smooth enough to use the sandwich formula (19). Thus, the asymptotic covariance matrix  $\Gamma$  satisfies  $\Gamma = D^{-1}\text{cov}\{\bar{S}(\theta_0)\}\{D^T\}^{-1}$ , where  $D = E\{\partial\bar{S}(\theta)/\partial\theta\}_{\theta_0}$ . Further modifications, with negligible asymptotic effect, are to replace  $\text{cov}\{\bar{S}(\theta_0)\}$  by  $B = \text{cov}\{S(\theta_0)\}$ , or an estimate, and to replace  $D$  by the estimate  $\hat{D} = \partial\bar{S}(\hat{\theta})/\partial\theta$ , giving

$$\Gamma = \hat{D}^{-1}\hat{\text{cov}}\{S(\theta_0)\}\{\hat{D}^T\}^{-1}, \quad (22)$$

Thus the induced method consists of solving (21) and (22) for  $\hat{\theta}$  and  $\Gamma$ .

## 9.4 Smoothed Rank Estimation for the AFT Model

Applying induced smoothing as in (21) and (22), the estimating equations (17) for  $\beta$  become

$$\bar{S}(\hat{\beta}) = \partial\bar{L}(\hat{\beta})/\partial\theta = E_Z\left[\sum_j \delta_j \sum_i (x_i - x_j) I\{\tilde{y}_i - \log(T_j) - (x_i - x_j)^T(\hat{\beta} + \Gamma^{1/2}Z) > 0\}\right]. \quad (23)$$

Let  $d_{ij} = x_i - x_j$ ,  $u_{ij} = d_{ij}^T \Gamma d_{ij}$ , and  $r_{ij} = r_i - r_j = \tilde{y}_i - y_j - d_{ij}^T \hat{\beta}$ . Then

$$\bar{S}(\hat{\beta}) = \sum_j \delta_j \sum_i d_{ij} \Phi\left(\frac{r_{ij}}{\sqrt{u_{ij}}}\right), \quad (24)$$

and the regression parameter estimate  $\hat{\beta}$  is found by solving  $\bar{S}(\hat{\beta}) = 0$ . The corresponding second derivative matrix is

$$\hat{D} = \partial \bar{S}(\hat{\beta}) / \partial \theta = - \sum_{i \neq j} \phi\left(\frac{r_{ij}}{\sqrt{u_{ij}}}\right) \frac{d_{ij} d_{ij}^{\Gamma}}{\sqrt{u_{ij}}}. \quad (25)$$

Clearly, this matrix is negative semi-definite, ensuring unique solutions to (24), and regular computation for  $\hat{\beta}$ . Here,  $\Phi, \phi$  are the cumulative distribution function and density function of  $N(0, 1)$ .

Next, a version of (22) is needed, to solve for the asymptotic covariance matrix  $\Gamma$ . The matrix  $\hat{D}$  comes from (25); therefore, recalling the asymptotic equivalence of smoothed and unsmoothed estimates, only an asymptotic covariance matrix expression for  $B = \text{cov}\{S(\beta_0)\}$  is needed. When there is no censoring, the null covariance matrix of  $S$ , asymptotically co-inciding with that of  $\bar{S}$ , can be found by permutation calculations, but this is no longer possible in the presence of censored observations.

The asymptotic null distribution of  $S$  is given in Jin et al. (2003), and its variance is a complicated function of both lifetime and censoring distributions. This formula could be used, along with ‘plug-in’ parameter estimates, to approximate to  $B$ , the large sample null covariance matrix of  $S$ .

Alternatively, there is a simpler approach, as follows.

First, take the case  $p = 1$ , when  $\{x_i\}$  are scalar covariates, and consider how to obtain an estimator for  $B = \text{var}(S)$ .

The modified rank test statistic, assuming the current  $\beta$  to be the true value of the regression parameter, is expressible from (17) as

$$S = \sum_{i \neq j} (x_i - x_j) \delta_j I(\varepsilon_j < \tilde{\varepsilon}_i) = \sum_{i \neq j} U_{ij},$$

where  $U_{ij}$  is the summand and  $\tilde{\varepsilon}_i = \varepsilon_i \wedge C_i$ . The required null variance of  $S$  can be found in Brown and Wang (2006).

The steps for implementing the induced smoothing method now are as follows.

- (i) select a trial value of  $\Gamma$ ;
- (ii) solve the monotone equation  $\bar{S}(\hat{\beta}) = 0$  for  $\hat{\beta}$ , where  $\bar{S}$  is given by (24);
- (iii) using the new  $\hat{\beta}$ , calculate  $\hat{D}$  from (25) and  $\hat{B}$ , and hence the updated  $\Gamma$ ;
- (iv) then return to step (i) and iterate until convergence to the required degree of accuracy is achieved.

As expected from the examples in Brown and Wang (2005) convergence is very rapid. This is illustrated by the example in the coming subsection.

## 9.5 A Simulation Study

I also carried out a small simulation study by generating two covariates  $\{x_i\}$  from standard normal distributions and  $y_i = X_i^T \beta + \varepsilon_i$ , where  $\beta = (-0.5, 0.3)$  and  $\varepsilon_i$  are standard normal. The  $y_i$  are subject to censoring at  $C_i$ , which is generated from a gamma distribution  $\Gamma(2, 1)$ , resulting in roughly 40% censored observations. The sample size is  $n = 65$ . Based on 1000 simulations, we obtained the mean of the Gehan estimates as  $(-0.539, 0.331)$  and the mean of the smoothed estimates as  $(-0.540, 0.332)$ . The mean of 1000 estimated standard errors are  $(0.126, 0.123)$ , while the true standard errors based on 1000 sets of the parameter estimates are  $(0.132, 0.128)$ . We further investigated the case when the sample size is 130. The corresponding means for the Gehan and the smoothed estimates are  $(-0.535, 0.324)$  and  $(-0.537, 0.3230)$ . The mean estimates of the smoothed standard errors are  $(0.091, 0.090)$  and the true standard errors (again, based on Monte Carlo) are  $(0.098, 0.105)$ .

## 9.6 Estimation of Perennial Vegetation Coverage

Rectified and calibrated Landsat TM data are used for predicting the vegetation density. A prediction model is established for each broad zone based on ground data from various training sites. We collected coverage information in percentage from 338 sites in NSW. Using 6 band data as covariates, we first fitted a simple linear model with  $B_1$  to  $B_6$  as six covariates. The  $R^2$  value is 0.47, and the residual standard error is 23.19. The residual plot (top panel) in Figure (9.6) shows that the variances are not homogenous. All predicted values are truncated to be between 0 and 100 in calculating the residuals. A weighted least-squares approach is therefore considered with weight  $w = 1/\{p * (100 - p)\}$ , where  $p$  is the predicted coverage from the simple model. The  $R^2$  dramatically increases to 0.86.

The rank estimation, which is robust, can also be applied. The covariance matrix for the estimating functions need to be derived. Because in this case, the variance of  $\epsilon$  appears to depend on  $x$ . Another approach is the random forests. As in the residual plot, the random forests provide much smaller residual variations indicating a much better fit. However, the random forests rely on many more parameters for producing the trees. More scientific comparisons based on reasonable statistical criteria are of interest to assess their performance in this context. Another issue is that predictions for 100% coverage can only be lower if not exact, leading to positive residuals. This censoring nature at 100% need special attention.

One approach is to use two-stage modeling. First fit a logistic regression to predict the pixels with 90% or more (heavy) coverage and apply random forests or regression methods if below 100%. For this dataset, there are 9 sites with 100% coverage, and 44 sites with more than 90% coverage. Applying GLM approach for the binary responses

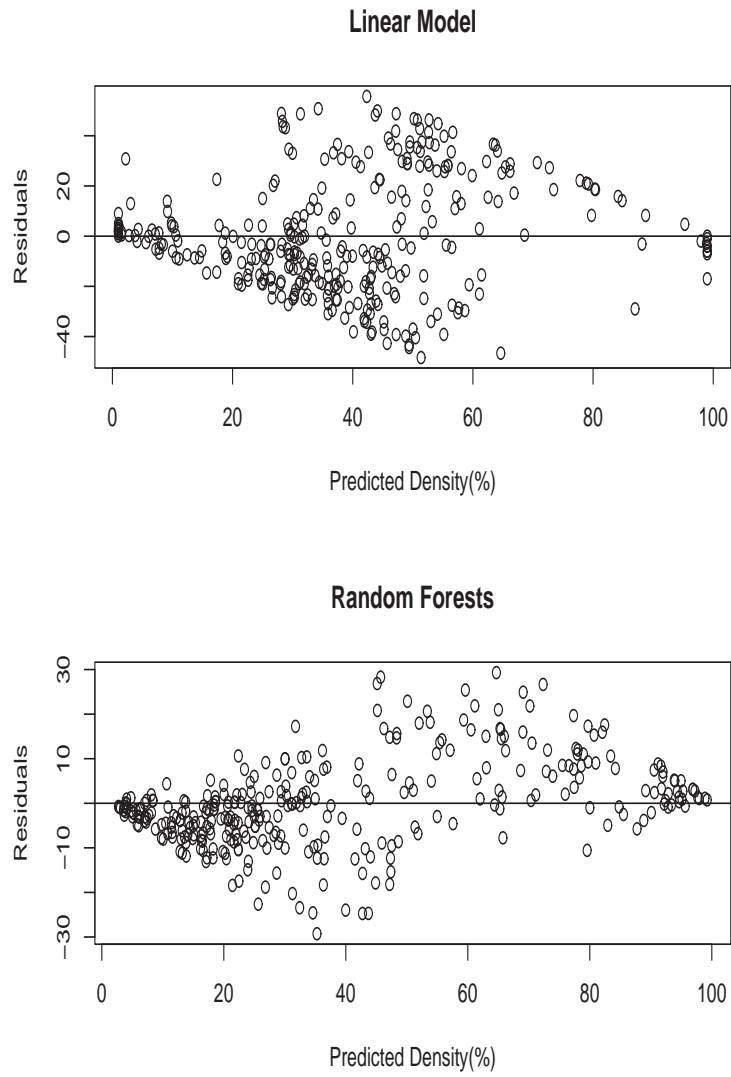


Figure 4: Residual plots for the linear model and the random forests approach.

(heavy coverage or not) results in the following fit. The parameter estimates including the intercept (3.842, 0.106, 0.148, -0.212, -0.0138, -0.0317, -0.132). There are 34 sites predicted as more than 90% coverage.

## 10 Other Potential Directions

### 10.1 Wavelets

There has been much development in wavelets in the past 20 years. Wavelets are functional generalizations to the Taylor series expansions, and have natural applications for image analysis. The wavelet technologies are quite mature and the user friendly softwares are available. As a layman on wavelets, I need more research before I can make specific recommendations for the MMM research. However, my understanding is research in mapping and monitoring can explore applications of wavelets.

As we know, responses from each band for an image (an area of interest) can be represented by a 2 -  $D$  function  $F(x, y)$ , which can be approximated (up to decomposition level  $J$ ) by three mother wavelets and one father wavelet,

$$F(x, y) \approx \sum_{j=1}^J \left\{ \sum_{m,n} d_{j,m,n}^{(1)} \Psi_{j,m,n}^{(1)}(x, y) + \sum_{m,n} d_{j,m,n}^{(2)} \Psi_{j,m,n}^{(2)}(x, y) + \sum_{m,n} d_{j,m,n}^{(3)} \Psi_{j,m,n}^{(3)}(x, y) \right\} + F_J(x, y),$$

where the first three components are the detail coefficients, and  $F_J(x, y)$  is the sum of coarse resolution at level  $J$  smooth coefficients,  $F_J(x, y) = \sum_{m,n} s_{J,m,n} \Phi_{J,m,n}(x, y)$ . These coefficients can be calculated using the pyramid algorithm efficiently. The initial image is  $F_0(x, y)$ , and the father component  $F_j(x, y)$  can be perfectly expanded by three mother coefficients at level  $j + 1$  and  $F_{j+1}(x, y)$ .

The potential applications include the following.

- (i) Noise removal

By removing the detail coefficients in the multiresolution analysis (wavelet shrinkage), an image can be reconstructed by the inverse wavelet transformation.

(ii). Land monitoring

There are two possible approaches to proceed. The first one is to apply wavelet analysis to both images at two different dates and then examine the differences in the transform coefficients to determine the level of changes. The other approach is to apply wavelet analysis to the difference data. The transform coefficients will represent changes at different resolutions.

(iii). Multisensor-data integration

If multiple data sets are available for the same area from different sensors possibly with different resolutions, they can be merged to increase accuracy and reduce variations of errors. This will increase enable us for better monitoring land covers, conditions and changes.

We will first geometrically register the low resolution image onto the fine resolution image (Land Sat TM) so that they possess the same pixel size, and then obtain wavelet representations for both images (band by band) up to the same coarser level ( $J$ ). The ‘merged’ wavelet representation is to use the father coefficient from the low resolution image and all the mother coefficients from the fine resolution image. The merged image can be constructed based on the merged wavelet representation.

## 10.2 Incorporating Spatial-Temporal Variations

Incorporation of spatial and temporal variations in environmental assessment and monitoring is of great statistical interest in the literature nowadays. Kiiveri, Caccetta and Evans (2001) proposed a framework of conditional probability networks (CPN) for monitoring salt in farmland. CPN was developed to incorporate spatial and temporal

variations in mapping and monitoring ten years ago. It has now built into the operational process. While it substantially improves the independence (native) model, it will be interesting to see how general space-time covariance functions can be built in (Stein, 2005). The other approach also published in JASA by Valpine (2004) can also be investigated.

### 10.3 Internal Collaboration and Outreaching

The stream of MMM (8.5 EFT) and Hydro-Climate Informatics (3.8 EFT) are based in Leeuwin Center together. The staff's interests are diverse. More internal communication and collaboration would be helpful. Geographically we are so isolated. Regular meetings and informal scientific talks may be helpful to keep members updated on literature and to identify scientific issues. Project-based work has taken too much time. To reach a new balance between external-earning (client projects) and scientific impacts (measured by national and international recognition), more junior staff (CSOF3, 4 and 5) are needed so that senior staff can focus on science aspects. International impacts can be greatly improved by focusing more on publications in *top* international journals. This will also have international recognition for the local client products and maybe increase the possibility of international engagement. It may be also a good idea to merge the two streams (MMM and HC). This will also further strengthen the statistical knowledge in MMM and provide more interactions.

In machine learning and image/signal processing, the statistical Department in Berkley is very (see <http://www.stat.berkeley.edu/faculty/index.html>).

MARK VAN DER LAAN Interests: Computational biology, optimal methods for censored data and survival analysis with applications in medical research, causal inference in longitudinal studies Office: 140 Warren Email: [laan@stat.berkeley.edu](mailto:laan@stat.berkeley.edu) (Joint

with BioStatistics).

BIN YU Interests: Machine learning (boosting and support vector machines), classification and unmixing in remote sensing, network tomography, Minimum Description Length (MDL) principle and information theory, and computational neuroscience Office: 429 Evans Email: binyu@stat.berkeley.edu

IAIN JOHNSTONE Interests: Nonparametric function estimation including wavelet shrinkage, uses of random matrix theory in statistics Email: imj@stat.berkeley.edu

Professor L.X. Zhu has invited me to visit the Department of Mathematics, Hong Kong Baptist University **at their cost**. The focus will be on nonparametric work. I will deliver a talk on the nonparametric work of Wang & Zhu (2006, *Biometrika*, in press) for correlated data and explore research opportunities. As robust methods are fundamental for MMM research, any progress on this approach will directly impact MMM research. The invitation (at almost their costs) is for mutual benefit. Their group has also done interesting work on theory and applications of Monte Carlo & quasi-Monte Carlo methods, a potential tool for MMM research to make computation feasible when dealing with large remote sensing data. I am also interested in talking to Dr. Gang Wei there who has done work on Neural Networks (NN) because NN is an interesting approach that is worth exploring and applying to classification (but I really know too little on this). Dr. C.S. Tong in HKBU has been working on image analysis for a number of years. His major interests are on image analysis and wavelets which are directly related to MMM group research. During the visit, we will discuss these ideas presented here with colleagues in HKBU and seek research collaborations for developing new statistical models and methods for remotely sensed data and building statistical models and develop inference procedures for multiple resolutions.

## References

- Atkinson, P.M., Foody, G.M., Curran, P. J. and Boyd, D.S.(2000). Assessing the ground data requirements for regional scale remote sensing of tropical forest biophysical properties. *Int. J. Remote Sensing***21** (13& 14), 2571-2587.
- Blanc, P. T. Blu, T., Ranchin, T. Wald, L. and Aloisi, R. (1998). Using iterated rational filter banks within the ARSIS concept for producing 10m Landsat multispectral images *International Journal of Remote Sensing* **19**, 2331 - 2343.
- Brown, B. and Wang, Y.-G. (2005). Standard errors and covariance matrices for smoothed rank estimators. *Biometrika* **92**, 149–158.
- Brown, B. and Wang, Y.-G. (2006). Induced smoothing for rank regression with censored survival times. *Statistics in Medicine*, in press.
- Campbell, N.A., Lopuhaä, H.P. and Rousseeuw, P.J. (1998). On the calculation of a robust  $S$ -estimator of a covariance matrix. *Statistics in Medicine* **17**, 2685–2695.
- Cantoni, E. and Ronchetti, E. (2001). Robust Inference for generalized linear models. *Journal of the American Statistical Association* **96**, 1022 - 1030.
- Chi, E.M. (1994).  $M$ -estimation in cross-over trials. *Biometrics* **50**, 486 - 493.
- Chen, C. (2002). Robust regression and outlier detection with the ROBUSTREG procedure, Paper 265-27, SUGI Conference Proceedings.
- Cox, D.R. (1972). Regression models and life-tables (with discussion). *Journal of the Royal Statistical Society, Series B, Methodological*, **34**, 1972.
- Cox, D.R. and Oakes, D. (1984). *Analysis of Survival Data*. Chapman and Hall: London.

- Furby, S.L. and Campbell, N.A. (2001). Calibration images from different dates to 'like-value' digital counts. *Remote Sensing of Environment* **77**, 186–196.
- Fygenson, M, Ritov, Y. (1994). Monotone estimating equations for censored data. *The Annals of Statistics* **22**, 732–746, 1994.
- Garguet-Duport, B., Girel, J., Chassery, J. and Pautou, G. (1996). Then use of multiresolution analysis and wavelets transform for merging SPOT panchromatic and multispectral image data. *Photogrammetric Engineering and Remote Sensing* **62**, 1057–1066.
- Gehan, E.A. (1965). A generalized Wilcoxon test for comparing arbitrarily single-censored samples. *Biometrika* 52: 203–223.
- Gill, P.S. (2000). A robust mixed linear model analysis for longitudinal data. *Statistics in Medicine* **19**, 975 – 987.
- Hettmansperger, T.P. and McKean, J.W. (1998). *Robust Nonparametric Statistical Methods*. Arnold: London.
- González-Audéana, M. , Otazu, X., Fors, O. and Seco, A. (2005). Comparison between Mallat's and the trous discrete wavelet transform based algorithms for the fusion of multispectral and panchromatic images. *International Journal of Remote Sensing* **26**, 595 - 614.
- Huber, P.J. (1981), *Robust Statistics*, New York: John Wiley.
- Huggins, R. M. (1993). A robust approach to the analysis of repeated measures. *Biometrics* **49**, 715 –720.
- Jaeckel L. (1972). Estimating regression coefficients by minimizing the dispersion of the residuals. *The Annals of Mathematical Statistics* **42**, 1540–1552, 1972.
- Jin Z, Lin DY, Wei LJ, Ying Z. (2003). Rank-based inference for the accelerated failure

- time model. *Biometrika*, 90:341–353.
- Jin, Z., Ying, Z. and Wei, L.J. (2001). A simple resampling method by perturbing the minimand. *Biometrika* **88**, 381–390.
- Kalbfleisch, J.D. and Prentice, R.L. (2002). *The Statistical Analysis of Failure Time Data, 2nd Edition*. Wiley: New York.
- Krall, J.M., Uthoff, V.A. and Harley, J.B. (1975). A step-up procedure for selecting variables associated with survival. *Biometrics* **31**, 49–57.
- Kumar, A. S. Kartikeyan, B. and Majumdar, K. L. (2000). Band sharpening of IRS-multispectral imagery by cubic spline wavelets *International Journal of Remote Sensing* **21**, 581 - 594.
- Lawless J. (2003) *Statistical Models and methods for Lifetime Data, 2nd Edition*. Wiley: New York.
- Pettitt, A.N. (1979). A non-parametric approach to the chang-point problem. *Applied Statistics* **28**, 126–135.
- Maronna, R.A. and Zamar, R.H. (2002). Robust estimates of a location and dispersion for high-dimensional datasets. *Technometrics* **44**, 307-317.
- Park, J. H. and Kang, M.G. (2004). Spatially adaptive multi-resolution multispectral image fusion. *International Journal of Remote Sensing* **25**, 5491 - 5508.
- Parzen, M. I., Wei, L. J. and Ying, Z. (1994). A resampling method based on pivotal estimating functions. *Biometrika* 81, 341–50
- Rao, A. R. and Hamed, K. H. (2000) *Flood Frequency Analysis*. CRC Press, Boca Raton, Florida, USA.
- Reiss, R. D. and Thomas, M. (2001) *Statistical Analysis of Extreme Values with Applications to Insurance, Finance, Hydrology and Other Fields*. Birkuser Verlag,

Boston, USA.

- Ray, S. S. (2004). Merging of IRS LISS III and PAN data—evaluation of various methods for a predominantly agricultural area. *International Journal of Remote Sensing* **25**, 2657 - 2664.
- Rousseeuw, P.J. (1984). Least median of squares regression. *Journal of the American Statistical Association* **79**, 871 - 881.
- Rousseeuw, P.J. and Leroy, A.M. (1987). *Robust Regression & outlier Detection*. New York: John Wiley and Sons.
- SAS Institute (1997). *SAS 6.12 Tech. Report*, Cary, NC.
- Schrader, R.M. and Hettmansperger, T.P. (1980). Robust analysis of variance based on upon a likelihood ratio criterion. *Biometrika* **67**, 93 - 101.
- Sobrino, J.A. and And Raissouni, N. (2000). Toward remote sensing methods for land cover application to Morocco. *Int. J. Remote Sensing* **21** (2), 353 – 366.
- Stein, M. (2005). Space-time covariance functions. *Journal of the American Statistical Association* **100**, 310–321.
- Street, J.O., Carroll, R.J. and Ruppert, D. (1988). A note on computing robust regression estimates via iterative reweighted least squares. *American Statistician* **42**, 152 – 154.
- Valpine, P.D. (2004). Monte Carlo state-space likelihoods by weighted posterior kernel density estimation. *Journal of the American Statistical Association* **99**, 523–536.
- Venables, W.N. and Ripley, B.D. (2002). *Modern Applied Statistics with S-PLUS*, Springer, 4th edition.
- Verburg, P.H., Schot, P., Dijst, M. and Veldkamp, A. (2004). Land use change modelling: current practice and research priorities. *Geojournal* **61** (4): 309324.

- Yohai, V., Stahel, W.A. and Zamar, R. H. (1991). A procedure for robust estimation and inference in linear regression. In *Directions in Robust Statistics and Diagnostics, Part II*, eds W.A. Stahel and S.W. Weisberg. New York: Springer-Verlag.
- Wang, Y.-G., Lin, X. and Zhu, M. (2005). Robust estimating functions and bias correction for longitudinal data analysis. *Biometrics* **61**: 684–691.
- Wang, Y.-G. Lin, X., Zhu, M. and Bai. Z.D. (2006). Robust estimation using the Huber function with a data-dependent tuning constant. Submitted for publication.
- Wang, Y.-G. and Zhu, M. (2006). Rank-based regression for analysis of repeated measures. *Biometrika*, to appear in the June issue.
- Yocky, D.A. (1996). Multiresolution wavelet decomposition image merger of landsat thematic mapper and SPOT panchromatic data. *Photogrammetric Engineering and Remote Sensing* **62**, 1067–1074.
- Zhang, Y., Guindon, B. and Cihlar, J. (2002). An image transform to characterize and compensate for spatial variations in thin cloud contamination of Landsat images. *Remote Sensing of Environment* **82**, 173-187.
- Zhou, J., Civco, D.L. and Silander, J. A. (1998). Wavelet transform method to merge Landsat TM and SPOT panchromatic data. *International Journal of Remote Sensing* **19**, 743–757.
- Zhu, M. Shao, QX. and Wang, Y.-G. (2006). Quantile regression without curse of unsmoothness. Manuscript.
- Zhu, M., Wallace, J., Caccetta, P. and Wang, Y.-G (2005). Asymptotic properties of S-estimators for calibration, CMIS internal report, 05/200.