Statistica Sinica **19** (2009), 1263-1280

ROBUST TIME SERIES ANALYSIS VIA MEASUREMENT ERROR MODELING

Qiong Wang[1], Leonard A. Stefanski[2], Marc G. Genton[3] and Dennis D. Boos[2]

[1]*GlaxoSmithKline,* [2]*North Carolina State University and* [3]*Texas A&M University*

Abstract: We describe an approach for robustifying inference in parametric models that is attractive for time series models. The key feature is that data from the postulated models are assumed to be measured with sporadic gross errors. We show that the tails of the error-contamination model kernel control the influence function properties (unbounded, bounded, redescending), with heavier tails resulting in greater robustness. The method is studied first in location-scale models with independent and identically distributed data, allowing for greater theoretical development. In the application to time series data, we propose a Bayesian approach and use Markov chain Monte Carlo methods to implement estimation and obtain outlier diagnostics. Simulation results show that the new robust estimators are competitive with established robust location-scale estimators, and perform well for ARMA(p, q) models.

Key words and phrases: Bayesian inference, error contamination model, influence function, measurement error, MCMC, robustness.

1. Introduction

Robust estimation for time series (Maronna, Martin, and Yohai (2006, Chap. 8)) is a challenging problem that has generated much interesting research; e.g., see Denby and Martin (1979), and Künsch (1984) for autoregressive models, and Bustos and Yohai (1986), Allende and Heiler (1992), and de Luna and Genton (2001) for autoregressive, moving-average processes. We develop a general approach that provides an attractive method of robust inference for time series models. It combines modern computing methods with the tried-and-true robustification strategy of using sporadic gross-error models. Specifically we use simulation-based likelihood construction and Bayesian Markov chain Monte Carlo methods (MCMC) to implement a robustification strategy that is readily adapted to time series models.

Suppose $\mathbf{X}_{n\times 1}$ is an outlier-free realization of a sample with assumed density $f_{\mathbf{x}}(\mathbf{x};\boldsymbol{\theta})$. Let X_i be the ith observation of $\mathbf{X}$. We postulate that the observed, outlier-prone version of $\mathbf{X}$ is $\mathbf{W}$, where,

$$W_i = X_i + U_i, \tag{1.1}$$

and where U_i is independent of X_i and has a sporadic, gross-error distribution depending on tuning parameters $\boldsymbol{\eta}$, and also possibly on components of $\boldsymbol{\theta}$. Assuming that U_i is independent of X_j for $i \neq j$ implies that the model for the observed-data density is

$$f_{\mathbf{W}}(\mathbf{w};\boldsymbol{\theta},\boldsymbol{\eta}) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f_{\mathbf{X}}(\mathbf{w}-\mathbf{u};\boldsymbol{\theta}) f_{\mathbf{U}}(\mathbf{u};\boldsymbol{\eta},\boldsymbol{\theta})\, d\mathbf{u} = E\left\{f_{\mathbf{X}}(\mathbf{w}-\mathbf{U};\boldsymbol{\theta})\right\}, \tag{1.2}$$

where $\mathbf{U} = (U_1,\ldots,U_n)^T$. The key idea is that inference on $\boldsymbol{\theta}$ via the log-likelihood $L(\boldsymbol{\theta}) = \log\{f_{\mathbf{W}}(\mathbf{w};\boldsymbol{\theta},\boldsymbol{\eta})\}$ is robust because the model incorporates a sporadic, gross-error component. In this sense our method is reminiscent of Lange, Little and Taylor (1989) for certain additive models, but there are fundamental differences between the two methods. Lange, Little and Taylor (1989) achieve robustness by basing their model for the data on heavy-tailed distributions (t distributions). The effectiveness of their approach is well-established. However, their approach loses the connection to a common, simple statistical model for the majority, non-outlying data, and thus the simple interpretations that accompany the more common time-series models. An advantage of our approach is flexibility in the contamination-error modeling due to the use of simulation-based likelihoods.

In some models both $\boldsymbol{\theta}$ and $\boldsymbol{\eta}$ are identified with observed data, and estimation of both is possible in theory. However, we adopt the strategy of taking $\boldsymbol{\eta}$ as a tuning parameter to achieve a desired efficiency at a central model. Nevertheless, we exhibit the dependence of $f_{\mathbf{W}}(\mathbf{w};\boldsymbol{\theta},\boldsymbol{\eta})$ on $\boldsymbol{\eta}$ for clarity as in (1.2), even though our interest lies in $\boldsymbol{\theta}$.

Apart from the special distribution for U_i, (1.1) is a structural measurement error model (Fuller (1987), Cheng and Van Ness (1999), Gustafson (2004) and Carroll, Ruppert, Stefanski and Crainiceanu (2006)). In measurement error models, Gaussian errors are often assumed to model in-control measurement methods. However, in application to robustness, we use error models that result in negligible errors much of the time, but produce occasional large outlying values, i.e., sporadic gross errors. Thus we assume that $f_{\mathbf{X}}(\mathbf{x};\boldsymbol{\theta})$ holds except for a random fraction of the data that are contaminated by additive random errors.

Although the gross-error model is conceptually simple, the likelihood based on (1.2) is usually unwieldy. For certain $f_{\mathbf{X}}(\mathbf{x};\boldsymbol{\theta})$ the likelihood can be estimated via simple Monte Carlo integration and this also works in principle for all $f_{\mathbf{X}}(\mathbf{x};\boldsymbol{\theta})$, but computation time is generally prohibitive. In such cases a Bayesian perspective, wherein the contaminating errors are parameters having a sporadic, gross-error model prior, allows for estimation via MCMC methods. The Bayesian approach has two additional advantages in that it provides natural outlier diagnostics via the posterior distribution of the contaminating errors, and it allows for inference on $\boldsymbol{\theta}$ that incorporates the uncertainty due to outlier estimation.

Although our interest is robustness for time series models, our method is also useful for independent data, and we first describe it in Section 2 in the context of independent data and location-scale models. This has the advantage of enabling a deeper theoretical treatment. Our study of the robustness properties of the estimators also sheds light on Gleason's (1993) finding that the contaminated normal mixture density is not heavy tailed. The simple Monte Carlo integration method that works well for independent-data models is less effectual in time series models because the integration is no longer one-dimensional. However, Markov chain Monte Carlo (MCMC) methods provide a viable alternative. For the time series models in Section 3 we start with theoretical influence function properties, then reformulate the approach from a Bayesian perspective, and apply MCMC methods for robust estimation and outlier detection. Our new method is illustrated and compared to other robust methods for time series data via simulation studies and examples.

2. Independent-Data Models

Our results apply with minor modification to non-identically distributed data. However, to minimize the notational burden we present only the identically distributed case.

2.1. Gross-error models

We consider the gross-error model,

$$U_i = \tau\sigma Z_i, \tag{2.1}$$

where σ is the scale parameter in $\boldsymbol{\theta}$ and Z_i has cumulative distribution

$$G_\epsilon(z) = (1-\epsilon)I(0 \le z) + \epsilon \int_{-\infty}^{z} g(x)\ dx, \tag{2.2}$$

with $g(\cdot)$ a standardized error density, e.g., normal, double exponential, or Cauchy, chosen to achieve desired robustness properties. When $g(\cdot)$ has a finite variance, standardization is to mean zero and variance one; for other cases, e.g., Cauchy, standardization is to median zero, and $\int_{-1}^{1} g(x)dx = \Phi(1) - \Phi(-1) \approx 0.6827$, where $\Phi(\cdot)$ is the standard normal distribution. We call $g(\cdot)$ the *kernel* of the sporadic, gross-error model (2.2). Here, $\boldsymbol{\eta} = (\epsilon, \tau)$, and $f_{U_i}(u; \boldsymbol{\eta}, \boldsymbol{\theta})$ depends on $\boldsymbol{\theta}$ through σ. This model is appropriate when the model for $\mathbf{X}$ contains a scale parameter. We study only the scale-invariant error model here because of its relevance to time series models. In the first author's PhD thesis (Wang (2005)),

non-scale-invariant error models in which $U_i = \tau Z_i$, relevant for robustifying certain non-time series models, are studied and shown to possess some interesting robustness properties.

2.2. Influence function properties

For $W_i = X_i + U_i$ and (2.1) and (2.2), the observed W_i has the density

$$f_{W_i}(w;\boldsymbol{\theta},\boldsymbol{\eta}) = (1-\epsilon) f_{X_i}(w;\boldsymbol{\theta}) + \epsilon \int_{-\infty}^{\infty} f_{X_i}(t;\boldsymbol{\theta}) \frac{1}{\tau\sigma} g\left(\frac{w-t}{\tau\sigma}\right) dt. \tag{2.3}$$

Define the likelihood score function

$$\boldsymbol{\psi}(w;\boldsymbol{\theta},\epsilon,\tau) = \frac{\partial}{\partial\boldsymbol{\theta}} \ln\left\{f_{W_i}(w;\boldsymbol{\theta},\boldsymbol{\eta})\right\}. \tag{2.4}$$

We write $\boldsymbol{\psi}(w;\boldsymbol{\theta},\epsilon,\tau)$ to emphasize the dependence of the score on the tuning parameters ϵ and τ. Properties of the influence function are determined by the tail behavior of the error-model kernel $g(\cdot)$ with greater robustness properties corresponding to heavier-tailed kernels. We now summarize how key robustness properties of (2.4) are determined by the tails of $g(\cdot)$. Outlines of proofs of the main results in this section appear in the Appendix. A more complete discussion with detailed proofs can be found in Wang (2005). With $\dot{f}_{X_i}(u;\boldsymbol{\theta}) = (\partial/\partial\boldsymbol{\theta}) f_{X_i}(u;\boldsymbol{\theta})$, consider the conditions:

(T1) $\lim_{|u|\to\infty} \; f_{X_i}(u;\boldsymbol{\theta}) \big/ g(u/(\tau\sigma)) = 0$;

(T2) $\lim_{|u|\to\infty} \; \dot{f}_{X_i}(u;\boldsymbol{\theta}) \big/ g(u/(\tau\sigma)) = 0$;

(T3) $\lim_{|u|\to\infty} \; g(u+b) \big/ g(u) = 1$;

(T4) $\lim_{u\to\infty} g(u+b) \big/ g(u) = \exp(c_1 b)$ and $\lim_{u\to-\infty} g(u+b) \big/ g(u) = \exp(c_2 b)$ for some constants $c_1 < 0$ and $c_2 > 0$.

(T1) and (T2) can be satisfied by choosing $g(\cdot)$ to have sufficiently heavy tails, i.e., heavier than those of $f_{X_i}(\cdot;\boldsymbol{\theta})$ and $\dot{f}_{X_i}(\cdot;\boldsymbol{\theta})$. (T3) holds when $g(\cdot)$ has polynomial-like tails, Cauchy densities, for example. (T4) is satisfied when $g(\cdot)$ has exponential-like tails as in Laplace or logistic densities. Normal densities $g(\cdot)$ do not satisfy (T3) or (T4), nor generally (T1) or (T2) unless $f_{X_i}(\cdot;\boldsymbol{\theta})$ has lighter-than-normal tails. The results below are derived under the assumptions that differentiation with respect to $\boldsymbol{\theta}$, integration, and limits can be interchanged in (2.3).

The qualitative robustness properties for the gross-error model (2.1) are specific to the underlying model. We illustrate this assuming a location-scale, true-data model with parameters μ and σ. All the influence function properties for the location parameter μ hold for other parameters in the model

except for the scale parameter σ. The μ-component of the score function is $\psi_\mu(w;\boldsymbol{\theta},\epsilon,\tau) = (\partial/\partial\mu)\ln\{f_{W_i}(w;\boldsymbol{\theta},\boldsymbol{\eta})\}$, and the σ-component of the score function is $\psi_\sigma(w;\boldsymbol{\theta},\epsilon,\tau) = (\partial/\partial\sigma)\ln\{f_{W_i}(w;\boldsymbol{\theta},\boldsymbol{\eta})\}$.

Polynomial-like tails. If $g(\cdot)$ satisfies (T1), (T2) and (T3), then

$$\lim_{|w|\to\infty}\psi_\mu(w;\boldsymbol{\theta},\epsilon,\tau) = 0, \quad \text{and} \quad \lim_{|w|\to\infty}\psi_\sigma(w;\boldsymbol{\theta},\epsilon,\tau) = \frac{p-1}{\sigma},$$

where p is the tail power, i.e., $g(u)\sim u^{-p}$; for Cauchy densities, $p=2$.

Exponential-like tails. If $g(\cdot)$ satisfies (T1), (T2) and (T4), then

$$\lim_{w\to\infty}\psi_\mu(w;\boldsymbol{\theta},\epsilon,\tau) = \frac{\partial}{\partial\mu}\ln\{m_f(-\frac{c_1}{\tau};\boldsymbol{\theta})\},$$
$$\lim_{w\to-\infty}\psi_\mu(w;\boldsymbol{\theta},\epsilon,\tau) = \frac{\partial}{\partial\mu}\ln\{m_f(-\frac{c_2}{\tau};\boldsymbol{\theta})\},$$

where $m_f(\cdot;\boldsymbol{\theta})$ is the moment generating function of $f_{X_i}(\cdot;\boldsymbol{\theta})$, assumed to exist for all real values of its argument. The score function for the location parameter is generally bounded, and $\lim_{|w|\to\infty}\psi_\sigma(w;\boldsymbol{\theta},\epsilon,\tau)=\infty$.

Gaussian tails. When $g(\cdot)$ is standard normal, neither (T3) nor (T4) holds and the tail behavior of the score function depends on the underlying model. For example, when the underlying true-data model in (2.3) is a Gaussian location-scale model with $\boldsymbol{\theta}=(\mu,\sigma)$, and $g(\cdot)$ in (2.3) is standard normal, the score functions for both location parameter and scale parameter are unbounded. Thus the normal contamination model *does not* result in qualitatively robust estimators. This finding is consistent with the results in Gleason (1993), who argued that the contaminated normal is not truly heavy tailed.

2.3. Location-scale models

Location-scale models and generalizations to regression, have been test beds for robustness studies since the Princeton Robustness Year (Gross and Tukey (1973)). We consider location-scale models to gain insight into tuning parameter selection and to study the influence function properties as a prelude to the time series models of Section 3. The central, contamination-free model has $f_{X_i}(x,\boldsymbol{\theta}) = \sigma^{-1}\phi\{(x-\mu)/\sigma\}$, where $\boldsymbol{\theta}=(\mu,\sigma)$ and $\phi(\cdot)$ is the standard normal density.

Our strategy is to tune the robust methods to achieve consistency and specified efficiency at the non-contaminated normal model. For given ϵ and τ, let $\boldsymbol{\psi}(w;\mu,\sigma,\epsilon,\tau)$ denote the 2×1 contaminated model score function. The equations in (μ,σ),

$$\int_{-\infty}^{\infty}\boldsymbol{\psi}(x;\mu,\sigma,\epsilon,\tau)\frac{1}{\sigma_0}\phi\Big\{\frac{x-\mu_0}{\sigma_0}\Big\}\,dx = (0,\ 0)^T,$$

Table 1. Tuning parameters (ϵ, τ), realized target efficiencies (μ_{EFF}, σ_{EFF}), scaling factors (c), and finite-sample, Monte Carlo-estimated likelihood efficiencies ($\widehat{\mu}_{\text{EFF}}$, $\widehat{\sigma}_{\text{EFF}}$) for normal, Laplace, and Cauchy contamination models. Standard errors for Monte Carlo efficiency ranged from 0.004 to 0.018, averaging approximately 0.01.

	Normal			Laplace			Cauchy		
	85%	90%	95%	85%	90%	95%	85%	90%	95%
ϵ	0.484	0.486	0.500	0.887	0.991	0.991	0.769	0.710	0.406
τ	1.937	1.689	1.386	2.148	1.825	1.145	0.652	0.407	0.295
μ_{EFF}	0.85	0.90	0.95	0.85	0.91	0.96	0.93	0.96	0.99
$\widehat{\mu}_{\text{EFF}}$	0.87	0.91	0.94	0.88	0.90	0.97	0.93	0.96	0.99
σ_{EFF}	0.85	0.90	0.95	0.85	0.89	0.94	0.81	0.88	0.94
$\widehat{\sigma}_{\text{EFF}}$	0.86	0.92	0.95	0.81	0.87	0.94	0.81	0.88	0.93
c	0.630	0.672	0.728	0.487	0.514	0.677	0.798	0.872	0.941

have solution $\mu = \mu_0$, $\sigma = c(\epsilon, \tau)\sigma_0$. For consistency at the $\text{N}(\mu, \sigma^2)$ model, scale estimates derived from the invariant error model likelihood are divided by $c(\epsilon, \tau)$. The correction factor does not depend on unknown parameters and need only be calculated once for a given (ϵ, τ).

Using numerical integration to evaluate the relevant likelihoods and asymptotic covariance matrices, we determined the correction factor $c(\epsilon, \tau)$ for $\text{N}(\mu, \sigma^2)$ data with $\mu = 0$ and $\sigma = 1$, and determined the asymptotic variances of the consistent estimators. We then determined pairs (ϵ, τ) that result in known asymptotic efficiencies for both μ and σ. It is not possible to achieve *specified target efficiencies* (e.g., 85%, 90% and 95%) for both μ and σ exactly, and so ϵ and τ were chosen to make the efficiencies as close to the target efficiencies as possible using a least-squares criterion. We call the efficiencies minimizing the least squares objective functions the *realized target efficiencies*. They are denoted as μ_{EFF} and σ_{EFF} in Table 1. The least-squares objective functions were relatively flat in many cases. Consequently the computed values of ϵ and τ are not precisely determined, but the computed values result in efficiencies close to the optimal. The realized target efficiencies (μ_{EFF}, σ_{EFF}), the corresponding tuning parameters (ϵ, τ), and the correction factor, $c(\epsilon, \tau)$, for specified target efficiencies 85%, 90% and 95%, appear in Table 1.

The large contamination proportions, ϵ, and scales, τ, in Table 1 were unexpected. When the contaminating component is unimodal and symmetric around zero, scale mixtures do not produce a high proportion of outliers unless the heavy-tailed component dominates the mixture. So in order for estimators derived from such models to have efficiencies significantly less than 100%, the heavy-tailed component must be substantial.

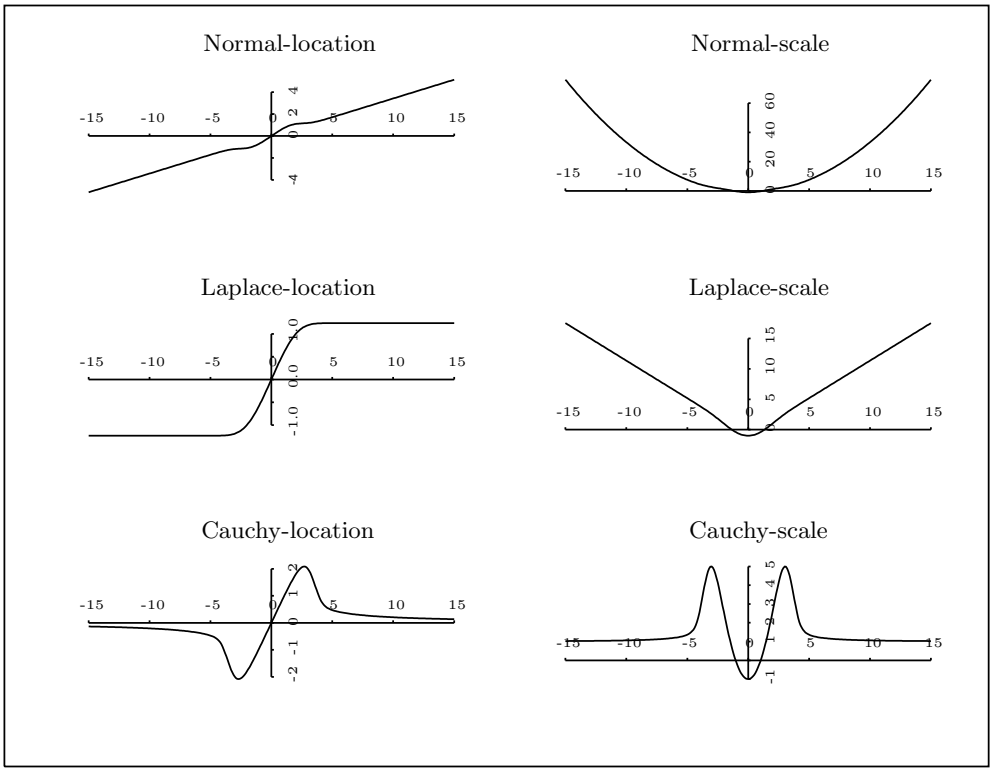

Figure 1. $\boldsymbol{\psi}(w;\mu,\sigma)$ for the invariant case when τ and ϵ are chosen to achieve the efficiency 95%.

The numerical evaluation of the tuning parameters worked with the exact (i.e., no Monte Carlo integration) observed-data model distribution (1.2), and is asymptotic in nature. We assessed the computed tuning parameters via a simulation study designed to estimate actual efficiencies when the likelihood was estimated via Monte Carlo integration with finite sample size, thus incorporating the added variability due to Monte Carlo averaging and finite sample sizes. In the simulation study 2,500 $\mathrm{N}(0,1)$ data sets of size $n = 200$ were generated and analyzed using the Monte Carlo likelihood with Monte Carlo size equal to 800. Table 1 displays Monte Carlo-estimated efficiencies ($\widehat{\mu}_{\mathrm{EFF}}$, $\widehat{\sigma}_{\mathrm{EFF}}$) for three contamination models (normal, Laplace, Cauchy). The agreement between the Monte Carlo-estimated ($\widehat{\mu}_{\mathrm{EFF}}$, $\widehat{\sigma}_{\mathrm{EFF}}$) and theoretical, realized target (μ_{EFF}, σ_{EFF}) efficiencies is very good.

2.4. Location-scale score function characteristics

Location and scale score functions, $\boldsymbol{\psi}(w;\mu,\sigma,\epsilon,\tau)$, from the three contamination models are plotted in Figure 1 when τ and ϵ are from Table 1 for 95%

efficiency. The baseline model is $\mathrm{N}(\mu, \sigma^2)$ with $\mu = 0$, $\sigma^2 = 1$. The theoretical properties of the score functions are readily observed from the graphs. The noteworthy features are the unboundedness of the score functions for the normal (location and scale) and Laplace (scale) contamination models, and the non-redescending of the Cauchy (scale).

3. Time Series Models

We now consider time series models for which the true-data density $f_{\mathbf{x}}(\mathbf{x}; \boldsymbol{\theta})$ is multivariate normal. In Section 3.1 it is shown that the exact (i.e., no Monte Carlo averaging) robust estimating equations for the time series data models possess the same qualitative robustness properties (redescending, bounded, unbounded) as the corresponding robust estimating equations for the independent and identically distributed (i.i.d.) location-scale models. That is, qualitatively our robustification strategy works the same for time series models as it does for simpler i.i.d. location-scale models, *in theory*. However, the simple Monte Carlo likelihood estimator is not viable for time series models and thus another approach is required. The primary focus of this section is on a Bayesian formulation of the sporadic, gross-error approach to robustification.

3.1. Theoretical robustness

The theoretical robustness results derived for independent and identically distributed data in Section 2.2 can be directly extended to the case where $X_1, \ldots, X_n$ are correlated. The likelihood score function becomes

$$\boldsymbol{\psi}(\mathbf{w}; \boldsymbol{\theta}, \epsilon, \tau) = \frac{\partial}{\partial \boldsymbol{\theta}} \ln \{ f_{\mathbf{w}}(\mathbf{w}; \boldsymbol{\theta}, \boldsymbol{\eta}) \}. \tag{3.1}$$

Throughout the section we assume that limits, differentiation, and integration are interchangeable.

Polynomial-like tails. If $f_{\mathbf{x}}(\mathbf{x}; \boldsymbol{\theta})$ is multivariate normal and $g(\cdot)$ satisfies (T3), then

$$\lim_{\min |u_i| \to \infty} \boldsymbol{\psi}_{\boldsymbol{\theta}_{(\sigma)}}(\boldsymbol{u}; \boldsymbol{\theta}, \epsilon, \tau) = 0,$$

where $\boldsymbol{\theta}_{(\sigma)}$ denotes all parameters except σ. The σ component of the score function is different from others due to the fact that the density of the added errors $\tau\sigma Z_1, \ldots, \tau\sigma Z_n$ depends on σ. In the Appendix, we show that

$$\lim_{\min |u_i| \to \infty} \psi_{\sigma}(\boldsymbol{u}; \boldsymbol{\theta}, \epsilon, \tau) = \frac{n(p-1)}{\sigma}, \tag{3.2}$$

where $g(u) \sim u^{-p}$. So the Cauchy error contamination model ($p = 2$) yields a redescending score function except for scale estimation, which is bounded by n/σ.

Exponential-like tails. Now assume that $g(u)$ satisfies assumption (T4) in Section 2.2. Set N_1 to be any subset of $N = \{1, \ldots, n\}$, and let N_2 be the set difference $N - N_1$. Then

$$\lim_{\substack{u_k \to \infty,\, k \in N_1 \\ u_j \to -\infty,\, j \in N_2}} \int \cdots \int f(\boldsymbol{s};\boldsymbol{\theta}) \prod_{i=1}^{n} \left\{ \frac{g((u_i - s_i)/(\tau\sigma))}{g(u_i/(\tau\sigma))}\, ds_i \right\}$$

$$= \int \cdots \int f(\boldsymbol{s};\boldsymbol{\theta}) \prod_{k \in N_1} \lim_{u_k \to \infty} \frac{g((u_k - s_k)/(\tau\sigma))}{g(u_k/(\tau\sigma))} \prod_{j \in N_2} \lim_{u_j \to -\infty} \frac{g((u_j - s_j)/(\tau\sigma))}{g(u_j/(\tau\sigma))}\, d\boldsymbol{s}$$

$$= \int f(\boldsymbol{s};\boldsymbol{\theta}) \exp\left(-\frac{c_1}{\tau\sigma} \sum_{k \in N_1} s_k - \frac{c_2}{\tau\sigma} \sum_{j \in N_2} s_j \right) d\boldsymbol{s}. \tag{3.3}$$

It follows that

$$\lim_{\substack{u_k \to \infty,\, k \in N_1 \\ u_j \to -\infty,\, j \in N_2}} \boldsymbol{\psi}_{\boldsymbol{\theta}_{(\sigma)}}(\boldsymbol{u};\boldsymbol{\theta},\epsilon,\tau) = \frac{\partial}{\partial \boldsymbol{\theta}_{(\sigma)}} \ln\{m_f(\boldsymbol{v};\boldsymbol{\theta})\},$$

where $v_k = -c_1/(\tau\sigma)$ for $k \in N_1$, $v_j = -c_2/(\tau\sigma)$ for $j \in N_2$, and $m_f(\cdot,\boldsymbol{\theta})$ is the moment generating function of $f(\cdot,\boldsymbol{\theta})$, assumed to exist. So the error contamination model with exponential-like tails results in a bounded score function except for σ. In Section 2.2 we showed that for the invariant error model with exponential-like tails, the score function for σ is unbounded in the case of independent and identically distributed data. The proof in the Appendix shows that it is unbounded in the case of correlated data as well.

Gaussian tails. When $f_{\mathbf{x}}(\mathbf{x};\boldsymbol{\theta})$ is multivariate normal and $g(\cdot)$ is normal, the score functions for both location and scale parameters are unbounded.

3.2. Bayesian inference for robust models

The key idea in the Bayesian implementation of the robust error contamination model is to treat the errors $Z_1, \ldots, Z_n$ as unknown parameters together with $\boldsymbol{\theta}$, and to formulate prior distributions for them. A natural choice of the prior for $Z_1, \ldots, Z_n$ is the sporadic, gross-error model G_ϵ in (2.2). Let $p(\boldsymbol{\theta})$ denote the prior density for $\boldsymbol{\theta}$. The posterior density of $\boldsymbol{\theta}, Z_1, \ldots, Z_n$ is

$$\pi(\boldsymbol{\theta}, \mathbf{z}|\mathbf{w}) = \frac{f(\mathbf{w}|\boldsymbol{\theta}, \mathbf{z}) \prod_{i=1}^{n} G_\epsilon(dz_i) p(\boldsymbol{\theta})}{\int\int \ldots \int f(\mathbf{w}|\boldsymbol{\theta}, \mathbf{z}) \prod_{i=1}^{n} G_\epsilon(dz_i) p(\boldsymbol{\theta}) d\boldsymbol{\theta}}, \tag{3.4}$$

where $f(\mathbf{w}|\boldsymbol{\theta}, \mathbf{z}) = f_{\mathbf{x}}(\mathbf{w} - \tau\sigma\mathbf{z};\boldsymbol{\theta})$.

The marginal posterior density of the parameter of interest, $\boldsymbol{\theta}$, is then given by

$$\int \ldots \int \pi(\boldsymbol{\theta}, \mathbf{z}|\mathbf{w}) dz_1 \ldots dz_n = \frac{\int \ldots \int f(\mathbf{w}|\boldsymbol{\theta}, \mathbf{z}) \prod_{i=1}^{n} G_\epsilon(dz_i) p(\boldsymbol{\theta})}{\int\int \ldots \int f(\mathbf{w}|\boldsymbol{\theta}, \mathbf{z}) \prod_{i=1}^{n} G_\epsilon(dz_i) p(\boldsymbol{\theta}) d\boldsymbol{\theta}}. \tag{3.5}$$

The posterior distribution (3.5) is identical to the one obtained by starting with the data distribution (1.2) which, in the notation of this section, is

$$f_{\mathbf{w}}(\mathbf{w}|\boldsymbol{\theta}) = \int \dots \int f(\mathbf{w}|\boldsymbol{\theta}, \mathbf{z}) \prod_{i=1}^{n} G_{\epsilon}(dz_i),$$

and using the prior $p(\boldsymbol{\theta})$. In this case the posterior density of $\boldsymbol{\theta}$ is given by

$$\pi(\boldsymbol{\theta}|\mathbf{w}) = \frac{f_{\mathbf{w}}(\mathbf{w}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{\int f_{\mathbf{w}}(\mathbf{w}|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}}, \tag{3.6}$$

which is the marginal posterior density of $\boldsymbol{\theta}$ in (3.5). It follows that, in the case that $p(\boldsymbol{\theta})$ is non-informative (proportional to a constant), the posterior mode of $\boldsymbol{\theta}$ in (3.6) is equivalent to the maximum likelihood estimate derived from $f_{\mathbf{w}}(\mathbf{w}; \boldsymbol{\theta}, \boldsymbol{\eta})$ given in (1.2). This is the sense in which our Bayesian approach is the natural Bayesian formulation of the frequentist approach based on model (1.2).

By taking the Bayesian perspective (3.4), we obtain the posterior joint density of the parameters of interest $\boldsymbol{\theta}$ and the errors $Z_1, \dots, Z_n$, thereby providing a means of inference for both the parameter and the errors. Thus a very attractive feature of the Bayesian approach is that it can be used to detect and estimate outliers, and therefore implement outlier identification along with robust estimation.

We use MCMC methods to estimate the posterior joint density of $\boldsymbol{\theta}$ and the unobserved errors $Z_1, \dots, Z_n$, and refer to the estimators so obtained as BMEM-estimators, where BMEM stands for Bayesian measurement error model. Having the posterior joint density of $\boldsymbol{\theta}$ and the unobserved errors $Z_1, \dots, Z_n$ means that inference on $\boldsymbol{\theta}$ will naturally reflect uncertainty due to outlier identification and estimation. Thus the Bayesian approach offers a distinct advantage over outlier identification and deletion followed by model fitting. The latter approach, while popular and convenient, does not properly account for uncertainty related to outlier detection and deletion. Our robust Bayesian approach treats gross errors and the parameters of interest on an equal footing.

3.3. Simulation study of Bayesian estimators

An autoregressive, moving-average (ARMA) model with order (p, q) is defined by

$$X_i - \mu = \sum_{j=1}^{p} \alpha_j (X_{i-j} - \mu) + \sum_{k=1}^{q} \beta_k e_{i-k} + e_i, \tag{3.7}$$

where $\{e_i\}$ are i.i.d. with mean zero and finite variance δ^2, selected to make $\mathrm{Var}(X_i) = \sigma^2$. When $\{e_i\}$ is normal, $\{X_i\}$ has a $\mathrm{N}(\mu, \sigma^2)$ distribution. We

Table 2. Monte Carlo average and mean squared error of the posterior means of σ and α_1 for 100 replicate data sets presented in the manner of (σ, α_1). Data Type: 0%, 5%, 10%, 20%: AR(1) with 0%, 5%, 10%, 20% contamination; AR(1)/N, AR(1)/L, AR(1)/C: AR(1) time series model with normal, Laplace, Cauchy contamination, respectively.

True Value: ($\sigma = 1$, $\alpha_1 = 0.75$)					
Data Type		MLE	AR(1)/N	AR(1)/L	AR(1)/C
(0%)	Mean	(0.91, 0.68)	(0.94, 0.70)	(0.90, 0.74)	(0.99, 0.72)
	MSE	(0.04, 0.02)	(0.03, 0.01)	(0.03, 0.01)	(0.02, 0.01)
(5%)	Mean	(2.39, 0.10)	(1.92, 0.24)	(1.48, 0.79)	(1.06, 0.75)
	MSE	(2.25, 0.45)	(1.03, 0.28)	(0.28, 0.01)	(0.03, 0.01)
(10%)	Mean	(2.73, 0.07)	(2.19, 0.20)	(1.83, 0.80)	(1.24, 0.78)
	MSE	(3.26, 0.49)	(1.59, 0.32)	(0.76, 0.01)	(0.10, 0.01)
(20%)	Mean	(4.12, -0.01)	(3.35, 0.07)	(3.10, 0.78)	(2.80, 0.87)
	MSE	(10.23, 0.60)	(5.85, 0.47)	(4.62, 0.04)	(3.61, 0.12)

assume that (3.7) is causal and invertible (Fuller (1996)). We are interested in the BMEM-estimator for ARMA(p, q) models in the presence of additive outliers; see Bustos and Yohai (1986).

We burned in $m = 5,000$ points and used $M = 10,000$ points after burn-in as the Markov chain samples used to obtain posterior distribution summaries. We used the posterior distribution mean as the parameter estimate.

We calculated the ARMA model maximum likelihood estimator (MLE) to compare to the BMEM-estimators. For the sporadic, gross-error model prior, we used the (ϵ, τ) values that were calculated for the location-scale model with i.i.d. data corresponding to 90% efficiency. We also used the scale correction factors calculated for the independent and identically distributed data to achieve approximate consistency at the central normal model.

Data were generated from ARMA(p, q) time series models contaminated by a fixed number κ (κ=0, 0.05n, 0.1n and 0.2n) of points, where the sample size $n = 50$, and the contamination points were generated from χ^2_8 or $-\chi^2_8$ with probability 1/2. We investigated the performance of the robust estimators for the AR(1) model ($p = 1$, $q = 0$) with $\alpha_1 = -0.5$, 0, 0.25, 0.5, and 0.75; the AR(2) model ($p = 2$, $q = 0$) with $\alpha_1 = 0.6$ and $\alpha_2 = 0.3$, $\alpha_1 = -0.6$ and $\alpha_2 = 0.3$, $\alpha_1 = 0.3$ and $\alpha_2 = -0.6$, and $\alpha_1 = -0.3$ and $\alpha_2 = -0.6$; and the MA(1) model ($p = 0$, $q = 1$) with $\beta_1 = 0.5$ and $\beta_1 = 0.8$. Due to space considerations, we do not display all results. Table 2 displays results of the simulation for the AR(1) model with $\alpha_1 = 0.75$, while Table 3 displays results of the simulation for the MA(1) model with $\beta_1 = 0.8$. The number of simulated data sets was 100. We present Monte Carlo averages of the estimators and their mean squared errors (MSE).

Table 3. Monte Carlo average and mean squared error of the posterior means of σ and α_1 for 100 replicate data sets presented in the manner of (σ, α_1). Data Type: 0%, 5%, 10%, 20%: MA(1) with 0%, 5%, 10%, 20% contamination; MA(1)/C: MA(1) time series model with Cauchy contamination.

True Value: ($\sigma = 1$, $\alpha_1 = 0.8$)			
Data Type		MLE	MA(1)/C
(0%)	Mean	(0.99, 0.80)	(0.99,0.80)
	MSE	(0.01, 0.01)	(0.01, 0.00)
(5%)	Mean	(2.36, 0.10)	(1.04, 0.77)
	MSE	(2.14, 0.52)	(0.02, 0.01)
(10%)	Mean	(2.98, 0.04)	(1.07, 0.75)
	MSE	(4.28, 0.61)	(0.02, 0.01)
(20%)	Mean	(4.02, 0.04)	(1.29, 0.66)
	MSE	(9.53, 0.60)	(0.14, 0.04)

Table 2 displays results for the AR(1) model with $\alpha_1 = 0.75$. The BMEM-estimators generally perform better than the maximum likelihood estimator for the contaminated data, although they tend to break down at the highest level of contamination (20% contamination, 10 points). However, the Cauchy BMEM-estimators are less biased and show better performance for moderate levels of contamination.

In all the simulations for AR processes, including those not reported here, the Cauchy BMEM-estimators exhibited greatest resistance to outliers when contamination is no more than 10%. However, all of the BMEM-estimators break down when contamination is as high as 20%. Due to the generally better performance of the Cauchy BMEM-estimator, we focused on it in the simulation study of MA(1) models.

Table 3 displays results of the simulation for the MA(1) model with $\beta_1 = 0.8$. The superiority of the Cauchy BMEM-estimators over maximum likelihood estimators is greater than that in the cases of AR(1) and AR(2) models. The Cauchy BMEM-estimator does not break down, even with contamination as high as 20%.

Both maximum likelihood estimates and robust estimates for location parameters are unbiased under all the situations. As in the case that data are i.i.d., the robust estimates for the scale parameter σ are divided by a correction factor in order to achieve consistency at the non-contaminated model. Although we used the correction factors calculated for location-scale models with i.i.d. data, the robust scale estimates are mostly close to 1 (true value) or as good as maximum likelihood estimates for the clean data, suggesting that the strategy of using the correction factor and tuning parameters calculated for the independent and identically distributed case is reasonable.

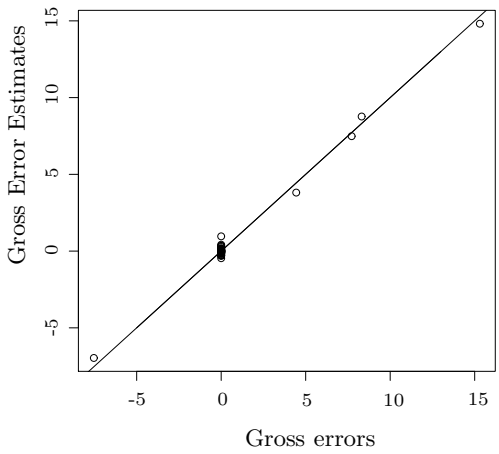

Figure 2. The Cauchy BMEM-estimates of the gross errors against the true values for the AR(1) model with $\alpha_1 = 0$ and 10% contamination ($\kappa = 5$).

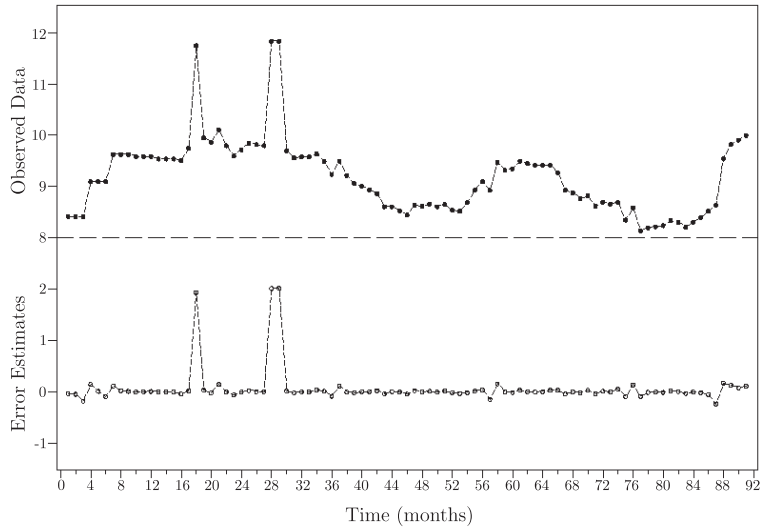

Figure 3. Monthly interest rates of an Austrian bank during 91 months. Top plot: observed data for the monthly interest rates; bottom plot: the Cauchy BMEM-estimates of the gross errors.

A useful feature of the Bayesian approach is the information provided by the posterior distribution of the sporadic, gross errors, $Z_1, \ldots, Z_n$. We expect most of these to be essentially equal to zero, with only contaminated observations giving rise to significantly-non-zero estimated Z_i. We illustrate this in Figure 2,

Table 4. Parameter estimates obtained by fitting an AR(1) model to the data of monthly interest rates. AR(1)/N, AR(1)/L, AR(1)/C: AR(1) time series model with normal, Laplace, Cauchy contamination; Künsch: range of Künsch's estimates for various tuning parameters.

data		Estimation Method				
		MLE	AR(1)/N	AR(1)/L	AR(1)/C	Künsch
real data	μ	9.19	9.01	9.10	9.13	(9.11 : 9.18)
	δ	0.443	0.143	0.119	0.169	(0.133 : 0.154)
	α_1	0.789	0.959	0.889	0.911	(0.958 : 0.959)
adjusted data	μ	9.12	9.08	9.10	9.13	(9.17 : 9.23)
	δ	0.212	0.127	0.104	0.162	(0.125 : 0.137)
	α_1	0.923	0.920	0.884	0.907	(0.958 : 0.965)

displaying a representative plot of the Cauchy BMEM-estimates $\tau\hat{\sigma}\hat{Z}_i$ against the true gross errors, $\tau\sigma Z_i$, $i = 1, \ldots, 50$, in the model (2.1) for a single AR(1) data set from the simulation with $\alpha_1 = 0$ and five contamination points. The solid line is the 45 degree line. The Cauchy BMEM-estimates of the gross errors lie very close to the 45 degree line, indicating that the Cauchy BMEM-estimator estimates the true errors very well for this data set. In all of our simulation studies we kept track of the median correlation between the BMEM-estimates of the errors and their true values. These were all close to 1. In other words, Figure 2 is representative of the quality of the Bayesian estimators of $Z_1, \ldots, Z_n$. Thus the BMEM-estimators of $Z_1, \ldots, Z_n$ are very useful for identifying outliers, as is also evident in the examples that follow.

4. Example

4.1. Example I: Austrian interest rates

We calculated the MLE and our new robust estimator for the time series consisting of 91 monthly interest rates of an Austrian bank, displayed in Figure 3. The data were analyzed previously by Künsch (1984), who suggested that the underlying model is AR(1), and that the data contain three large outliers at months 18, 28 and 29.

Table 4 presents our estimators and Künsch's estimators for μ, α_1, and the innovation standard deviation $\delta = \sigma\sqrt{1-\alpha_1^2}$. For Künsch's method, the tuning parameters were chosen to achieve acceptable efficiencies, and we report the estimates in the format (lowerbound:upperbound) for various tuning parameters as done by Künsch. We replaced the three outliers by 9.85, as suggested by Künsch, and call these the adjusted data. We then calculated the various estimators on the adjusted data. The BMEM-estimates are quite close for the unadjusted and adjusted data sets, demonstrating the BMEM-estimators resistance to outliers.

TAble 5. Parameter estimates obtained by fitting an ARMA(1,2)-process to the data of saving rates. Six interventions: Pankratz's approach with six outliers identified and adjusted; one intervention: Pankratz's approach with the 82nd outlier identified and adjusted; IGM: de Luna and Genton's estimator based on the indirect inference and the GM-estimator; ARMA(1,2)/C: the Cauchy BMEM-estimator with the central ARMA(1,2) model; Bayes intervals: the 2.5% and 97.5% quantiles of the sample posterior distribution from the MCMC output are given in parentheses.

Estimation Method	α_1	β_2	μ	δ^2
MLE	0.74	0.34	6.11	0.44
six interventions	0.80	0.38	6.16	0.23
one intervention	0.81	0.25	6.07	0.34
IGM	0.82	0.40	6.15	0.34
ARMA(1,2)/C	0.82	0.27	5.90	0.34
Bayes intervals	(0.69, 0.94)	(0.02, 0.49)	(4.78, 6.93)	(0.25, 0.46)

Künsch's estimates for α_1 are larger than maximum likelihood estimates and the BMEM-estimates. The fact that the BMEM-estimates are closer to the maximum likelihood estimate of α_1 for the adjusted data supports their use.

Figure 3 displays the data and the gross error estimates, $\tau\hat{\sigma}\hat{\mathbf{Z}}$, obtained via the Cauchy contamination model. Outliers are indicated by $\tau\hat{\sigma}\hat{\mathbf{Z}}$ components that are far from zero. The three outliers identified by Künsch are clearly apparent from the index plot of $\tau\hat{\sigma}\hat{\mathbf{Z}}$.

4.2. Example II: US saving rates

We applied the Cauchy BMEM-estimator to the time series on saving rates (saving as percent of income) in the United States from the start of 1955 to the end of 1979. The series has been analyzed in Pankratz (1991, Chap. 8) using an outlier detection procedure. de Luna and Genton (2001) used this time series to illustrate their robust estimation method, which combines the indirect inference method with the generalized M-estimator.

Pankratz (1991) recommended an ARMA(1,2) model with $\beta_1 = 0$, and identified six purported outliers (observations 82, 43, 62, 55, 89, 100). But it has also been recognized that the outlier detection procedure was too sensitive and led to a deflated estimate of the scale parameter. The 82nd observation was exceptionally large due to Congress having passed a law granting a one-time tax rebate. Pankratz thus fitted the ARMA(1,2) model with only the 82nd observation adjusted.

Table 5 presents the Cauchy BMEM-estimates, Pankratz's estimates, and de Luna and Genton's indirect generalized M-estimates (IGM). For the Cauchy BMEM-estimator, we burned in $5,000$ points and drew an additional $5,000$ points

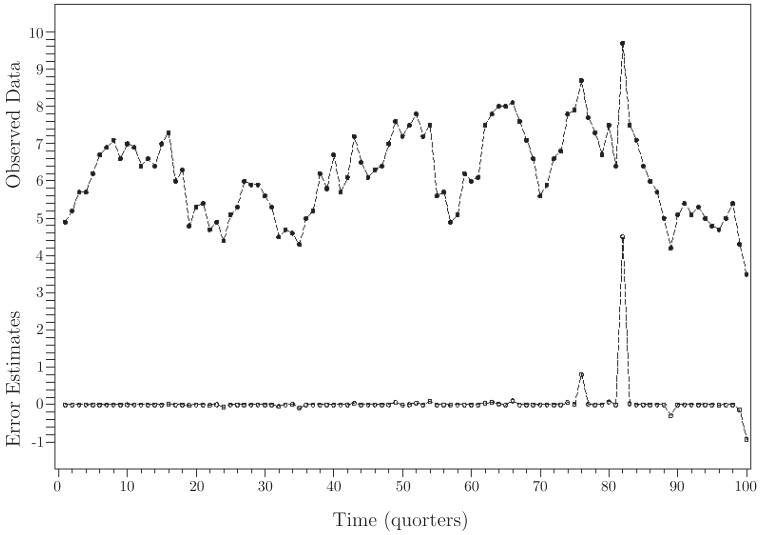

Figure 4. Saving rates in the United States from the first quarter of 1955 to the fourth quarter of 1979. Top plot: the observed data for the saving rates; bottom plot: the Cauchy BMEM-estimates of the gross errors.

as a sample for the posterior distribution. The Cauchy BMEM-estimates are fairly close to Pankratz's estimates when only the 82nd observation is adjusted in the model.

Figure 4 displays the plot of the data and the gross error estimates. The Cauchy BMEM-estimates found one exceptionally large outlier at the 82nd observation, which explains why the Cauchy BMEM-estimates are close to the Pankratz's estimates with only 82nd observation adjusted. The Cauchy BMEM-estimator also identified three other significantly-non-zero errors at the 76th, 89th and 100th observations. In contrast to Pankratz's outlier identification, the 76th observation was detected by the Cauchy BMEM-estimator as an outlier but not by Pankratz, while Pankratz's other three purported outliers (43rd, 62nd, 55th) were not identified by the Cauchy BMEM-estimator.

5. Summary

The sporadic, gross error measurement error model (2.2) provides a natural means of robustifying many statistical models. Desired levels of robustness are attained by a suitable choice of the kernel, with heavier-tailed kernels resulting in greater qualitative robustness. With i.i.d. data, simple Monte Carlo averaging provides a viable means of estimating the robustified model likelihood, whereas for time series data, MCMC methods are required. In either case, the model can be generally applied as it is only necessary to generate observations from

the sporadic gross error model kernel density. The simulation results presented herein are representative of more extensive simulation studies reported in the first author's PhD thesis (Wang (2005)) establishing the method as a viable alternative for i.i.d. location-scale modeling, and a generally more useful competitor to existing robust methods for time series data. The method is more general than portrayed herein, and can be readily adapted to regression modeling with independent data and to certain non-time series, dependent-data models.

Computer programs for calculating the estimators described in this paper are available upon request from the first author. The Appendix is provided as an online supplement at the following URL:
`http://www.stat.sinica.edu.tw/statistica.`

References

Allende, H. and Heiler, S. (1992). Recursive generalized M estimates for autoregressive moving-average models. *J. Time Ser. Anal.* **13**, 1-18.

Bustos, O. H. and Yohai, V. J. (1986). Robust estimates for ARMA models. *J. Amer. Statist. Assoc.* **81**, 155-168.

Carroll, R. J., Ruppert, D., Stefanski, L. A. and Crainiceanu, C. M. (2006). *Measurement Error in Nonlinear Models, A Modern Perspective*, 2nd Edition. Monographs on Statistics and Applied Probability 63. Chapman and Hall.

Cheng, C. and Van Ness, J. (1999). *Statistical Regression with Measurement Error*. Oxford University Press, Oxford.

de Luna, X. and Genton, M. G. (2001). Robust simulation based estimation of ARMA models. *J. Comput. Graph. Statist.* **10**, 370-387.

Denby, L. and Martin, R. D. (1979). Robust estimation of the first order autoregressive parameter. *J. Amer. Statist. Assoc.* **74**, 140-146.

Fuller, W. (1987). *Measurement Error Models*. Wiley, New York.

Fuller, W. (1996). *Introduction to Statistical Time Series*. Wiley, New York.

Gleason, J. R. (1993). Understanding elongation: The scale contaminated normal family. *J. Amer. Statist. Assoc.* **88**, 327-337.

Gross, A. M. and Tukey, J. W. (1973). The estimators of the Princeton robustness study. Tech. Rep. 38, Ser. 2, Dept. of Statistics, Princeton University, Princeton, N.J.

Gustafson, P. (2004). *Measurement Error and Misclassification in Statistics and Epidemiology*. Chapman and Hall/CRS, Boca Raton.

Hwang, J. T. and Stefanski, L. A. (1994). Monotonicity of regression functions in structural measurement error models. *Statist. Probab. Lett.* **20**, 113-116.

Künsch, H. (1984). Infinitesimal robustness for autoregressive models. *Ann. Statist.* **12**, 119-126.

Lange, K. L., Little, R. J. and Taylor, J. M. (1989). Robust statistical modeling using the t-distribution. *J. Amer. Statist. Assoc.* **84**, 881-896.

Maronna, R. A., Martin, D. R. and Yohai, V. J. (2006). *Robust Statistics: Theory and Methods*. Wiley, New York.

Pankratz, A. (1991). *Forecasting with Dynamic Regression Models*. Wiley, New York.

Wang, Q. (2005). Robust estimation via measurement error modeling. Ph.D. Thesis, NCSU.

GlaxoSmithKline, 1250 S. Collegeville Rd, Collegeville, PA 19426, U.S.A.
E-mail: qiong.z.wang@gsk.com

Department of Statistics, North Carolina State University, Box 8203, Raleigh, NC 27601, U.S.A.
E-mail: stefanski@stat.ncsu.edu

Department of Statistics, Texas A&M University, College Station, TX 77843, U.S.A.
E-mail: genton@stat.tamu.edu

Department of Statistics, North Carolina State University, Box 8203, Raleigh, NC 27601, U.S.A.
E-mail: boos@stat.ncsu.edu

(Received January 2007; accepted February 2008)