INFLUENCE FUNCTION APPROXIMATION TO ESTIMATION IN LINEAR MODELS WITH GENERALIZED t-DISTRIBUTION NOISE

VU HOANG DUNG
(M.Eng., Ho Chi Minh City University of Technology)

A THESIS SUBMITTED

FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

DEPARTMENT OF ELECTRICAL & COMPUTER ENGINEERING

NATIONAL UNIVERSITY OF SINGAPORE

2014

DECLARATION

I hereby declare that this thesis is my original work and it has been written by me in its entirely.

I have duly acknowledged all the sources of information which have been used in this thesis.

This thesis has also not been submitted for any degree in any university previously.

VU HOANG DUNG
10 January 2014

Acknowledgments

First of all, I would like to express my sincere gratitude to my supervisor, Professor Ho Weng Khuen, who always graciously guide me throughout the course of my research. His kind concern and valuable advices offered me an inspiration and confidence in overcoming all the hurdles in the completion of the proposal. I also acknowledge Professor Ling Keck Voon for his kind consideration and detailed guidance to my work.

I also want to thank my friend, Mr Chao Yu, in Advanced Control Technology Laboratory and Ms Le Ngoc Lieu for their sharing and support. They're always there for me whenever I need help or advice.

Last but not least, I would like to thank my parents for their tremendous support, understanding and encouragement throughout my 4 years study.

Contents

Summary

Commonly made assumption of Gaussian noise is an approximation to reality. The occurrence of outliers, transient data in steady-state measurements, instrument failure, human error, process nonlinearity, etc. can all induce non-Gaussian process data. Indeed whenever the central limit theorem is invoked - the central limit theorem being a limit theorem can at most suggest approximate normality for real data. However, even high-quality process data may not fit the Gaussian distribution and the presence of a single outlier can spoil the statistical analysis completely. In this thesis, instead of assuming Gaussian distributed noise, we use the genelized t-distribution as noise model. By being a distribution superset encompassing Gaussian, uniform, t and double exponential distributions, the generalized t-distribution has the flexibility to characterize data with non-Gaussian statistical properties. We also use the influence function, a robust statistic tool, to analyze the proposed estimator. Specifically how it can predict the change in the estimate due to outliers and the variance of the estimate. Moreover, the influence function is also used to formulate a recursive algorithm that gives an approximate solution making it suitable for real-time and on-line implementation. The proposed theory is verified by simulation and experiments.

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

Commonly made assumption of Gaussian noise is an approximation to reality. The occurrence of outliers, transient data in steady-state measurements, instrument failure, human error, process nonlinearity, etc. can all induce non-Gaussian process data [1]. Indeed whenever the central limit theorem is invoked — the central limit theorem being a limit theorem can at most suggest approximate normality for real data [2]. However, even high-quality process data may not fit the Gaussian distribution and the presence of a single outlier can spoil the statistical analysis completely. For instance, most of works on industrial measurement statistics usually carry out under the assumption of Gaussian distribution noise. This raises an important question of effect of gross error presenting in measurement on the estimate. The presence of gross error may significantly cause the estimate to be biased, which leads to the inconsistency of statistics analysis.

In literature, Least-Squares (LS) and Kalman filter are the most common methods to reconcile the data measurement thanks to their simplicity and ease to implement. However, as well known in robust statistics, LS is very sensitive to gross error and outliers. A common method to robustify LS is to introduce preliminary outlier test before applying LS, e.g. detecting gross error based on statistical tests and constraint tests of residuals [3–7], using a serial compensation strategy to delete suspicious measurements [8] or employing a detection method based on the bounds of measurements [9]. However, a major drawback of all the above works is that their techniques were still derived based on the assumption of Gaussian distribution noise which may not be plausible in real life situations. Another disadvantage is that statistic tests can only be conducted before and/or after estimating, which require extra computational time to detect gross errors [10]. Hence, it is better to derive a versatile estimator that can efficiently reconcile data with the present of outlier without any preliminary test.

Real time on-line optimization is a key requirement for industrial process. Hence, it is necessary to derive an efficient estimator that can estimate and detect outlier at the same time. Recently, Robust Statistics has become more and more popular [1, 2, 10, 11] thanks to the property of simultaneously estimating and rejecting gross error. This has inspired other researchers to applied robust statistics in data reconciliation problem [12], parameter estimation [13], etc. However, there is a disadvantage in their methods as they may fail to address the question of theoretical evaluation of their estimators. Monte-Carlo simulation evaluation method is limited to some certain circumstances only. Therefore, it is clear that a theoretical analysis

of estimator evaluation is needed to analyze the performance of the estimator. Also experiments might be needed to validate the proposed theory. Another problem when applying robust statistics is the non-linear nature of the robust estimators. Hence, it is necessary to develop a recursive algorithm by approximating the robust estimator that makes it applicable to practical situations.

1.2 Contribution of the thesis

The main contributions of this thesis are summarized as below:

In this thesis, instead of using normal distribution to model noise, we took another approach of modeling noise by employing the Generalized t-distribution (GT). Thanks to the flexibility of the GT distribution, various types of measurement noise can be modeled, hence the estimator gives better results in the case of non-Gaussian noise compared to the Least-Squares and Kalman Filter methods.

The influence function (IF), a powerful robust statistics tool [2], is used to analyze the estimate from the estimator designed with GT noise model. Specifically how it can predict the change in the estimate due to outliers and the variance of the estimate. These equations enable us to compute the sample size needed by the estimator to meet specified variance or tune the estimator to limit the impact of outliers. Alternatively, these equations allow us to calculate the variances of the estimates and hence their precisions if the number of data points used is given. The theory is verified through simulations and an experiment on the thickness measurements in the chemical-mechanical polishing of semiconductor wafers.

We also use IF approximation to derive a recursive solution for the maximum likelihood estimation of the ARMAX Process with GT noise. We also show how the IF can be used to analyze the filter, specifically how it can predict the filter output due to outliers and the variance of the output. It will be shown through an example that if the noise is Gaussian then the proposed ARMAX filter is equivalent to the Kalman filter. Experiment on liquid-level couple tank is conducted to highlight the

advantage of our proposed method over the conventional Kalman filter.

We also extend the use of GT noise model and IF to data reconciliation framework where multiple inputs are treated simultaneously. Moreover, the cases of skewed and correlated noise are also be considered. The IF is also derived to make use of its advantages. Simulation are conducted to verify the theory.

1.3 Scope of the thesis

This thesis is organized as follows. In the next chapter, we use the influence function to analyze the estimator behavior towards outlier. Chapter 3 is the extension of work done in the previous chapter to the ARMAX problem. Moreover, instead of using the IF as an analysis tool to analyze a given estimator, this chapter makes use of the IF to construct a recursive estimator that can be used for real-time situation. In chapter 4, we expand our maximum likelihood estimator to the data reconciliation problem where multiple outputs with different noise characters are needed to be corrected. The follow up chapter will discuss the case of multivariate GT distribution with correlated noise and skew GT distribution where noise distribution is no longer symmetric which may cause bias on the conventional estimators. In the end, conclusion and future work of this thesis are given in Conclusion chapter.

Chapter 2

Parameter Estimation

2.1 Introduction

Commonly made assumption of Gaussian noise is an approximation to reality. The occurrence of outliers, transient data in steady-state measurements, instrument failure, human error, process nonlinearity, etc. can all induce non-Gaussian process data [14]. Indeed whenever the central limit theorem is invoked — the central limit theorem being a limit theorem can at most suggest approximate normality for real data [2]. However, even high-quality process data may not fit the Gaussian distribution and the presence of a single outlier can spoil the statistical analysis completely.

Take the example of the chemical-mechanical polishing of semiconductor wafers [15, 16]. The histogram of the distribution of 576 thickness measurements (see Figure 2.1) after chemical-mechanical polishing of twenty-four 200mm semiconductor wafers and after subtracting the mean are plotted in Figure 2.2. Using the maxi-

mum likelihood criterion, a Gaussian distribution was fitted to the histogram. It is evident in Figure 2.2 that the Gaussian curve does not give a good fit. The Generalized t-distribution (GT), by being a distribution superset encompassing Gaussian, uniform, t and double exponential distributions, has the flexibility to characterize data with non-Gaussian statistical properties [17–20]. It is evident in Figure 2.2 that the GT distribution fit the experimental data better than the Gaussian curve.

GT distribution was employed in the data reconciliation problem to model random noise [14, 21–24]. It was shown [24] that the influence function (IF) [2, 25–27] in robust statistics was useful in analyzing the data reconciliation problem with GT noise. GT distribution was also used in econometrics [17, 20, 28, 29] to model random noise in the parameter estimation problem. In this study, we use the IF to analyze the parameter estimation problem with GT noise. The analysis is further generalized to the case where the estimator designed with probability density

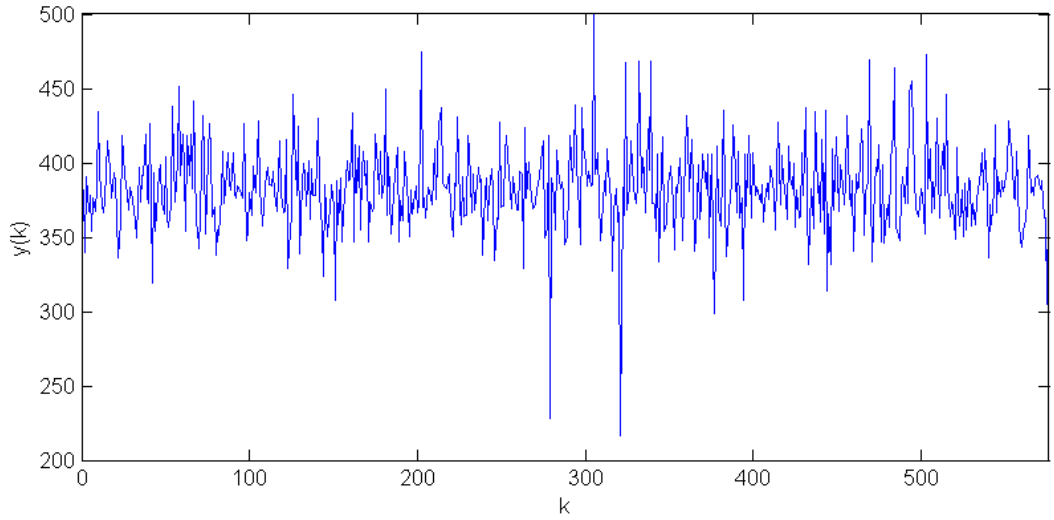


Figure 2.1: Thickness measurements on 24 semiconductor wafers after Chemical Mechanical Polishing

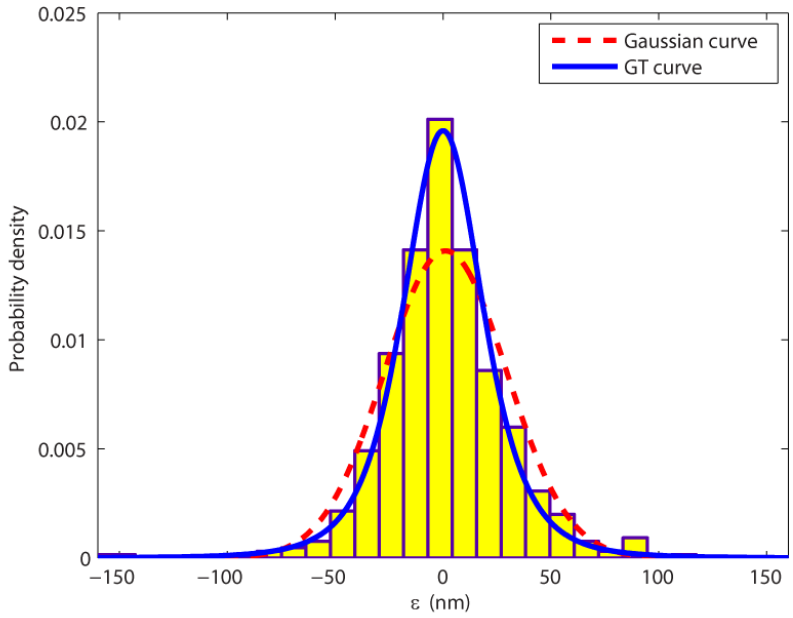


Figure 2.2: The maximum likelihood criterion was used to fit a Gaussian distribution (dotted-line, $\mu = 0$, $\sigma = 28.5$nm) and GT distribution (solid-line, $p = q = 2$, $\sigma = 29.5$nm) to the thickness measurement distribution.

function $f(\varepsilon)$ is applied to noise with different probability density function $g_k(\varepsilon)$ at different sampling instance, k, to provide a framework for the analysis of outliers.

In Section 2.2, we first describe the parameter estimator design problem, modeling noise with the GT distribution instead of the usual Gaussian distribution. With proper choice of the parameters, the GT distribution reduces to the Gaussian distribution and the estimator reduces to the well-known batch least-squares estimator. Hence within the more general framework of the parameter estimator with GT distributed noise model is the least-squares estimator if the noise is Gaussian distributed. If the noise is not Gaussian then the GT distribution has extra degrees of freedom to model the noise. The proposed estimator is not applicable to non-stationary noise time series. Other approaches [30–32] for handling non-Gaussian

noise include particle filters which is based on point mass or particle representation of probability densities.

The main contribution of this chapter is in Sections 2.3 and 2.4 where we show how the IF [2, 25–27] can be used to analyze the estimate from the parameter estimator designed with GT noise model, specifically how it can predict the change in the estimate due to outliers and the variance of the estimate. These equations enable us to compute the sample size needed by the estimator to meet specified variance or tune the estimator to limit the impact of outliers. Alternatively, these equations allow us to calculate the variances of the estimates and hence their precisions if the number of data points used is given. The theory is verified through simulations and an experiment on the thickness measurements in the chemical-mechanical polishing of semiconductor wafers.

2.2 Estimator Design

In this section, we discuss the estimator design problem using the GT noise model which includes the Gaussian noise as a special case.

Consider the linear in the parameter model

$$y(k) = \phi(k)^T\theta + \varepsilon(k) \tag{2.1}$$

where the vector $\phi(k) = \Big[\phi_1(k), \ldots, \phi_n(k)\Big]^T$ are known, the parameters $\theta = \Big[\theta_1, \ldots, \theta_n\Big]^T$ are to be estimated and $k = 1, \ldots, N$ is the sampling instance.

2.2.1 GT Distributed Noise

Let the noise $\varepsilon(k)$ be modeled by the zero mean GT probability density function [17, 18, 20]

$$f(\varepsilon) = \frac{p}{2\sigma q^{1/p}\beta(1/p,q)\left(1+\frac{|\varepsilon|^p}{q\sigma^p}\right)^{q+1/p}} \tag{2.2}$$

where σ is the scale parameter, p and q are the shape parameters. The beta function is given by $\beta(a,b) = \int_0^1 z^{a-1}(1-z)^{b-1}dz$. By different choices of p and q, GT can represent a wide range of distributions [17, 18]. The relationships between GT distribution, Gaussian, uniform, t, double exponential distributions are shown in Figure 2.3 [17, 18].

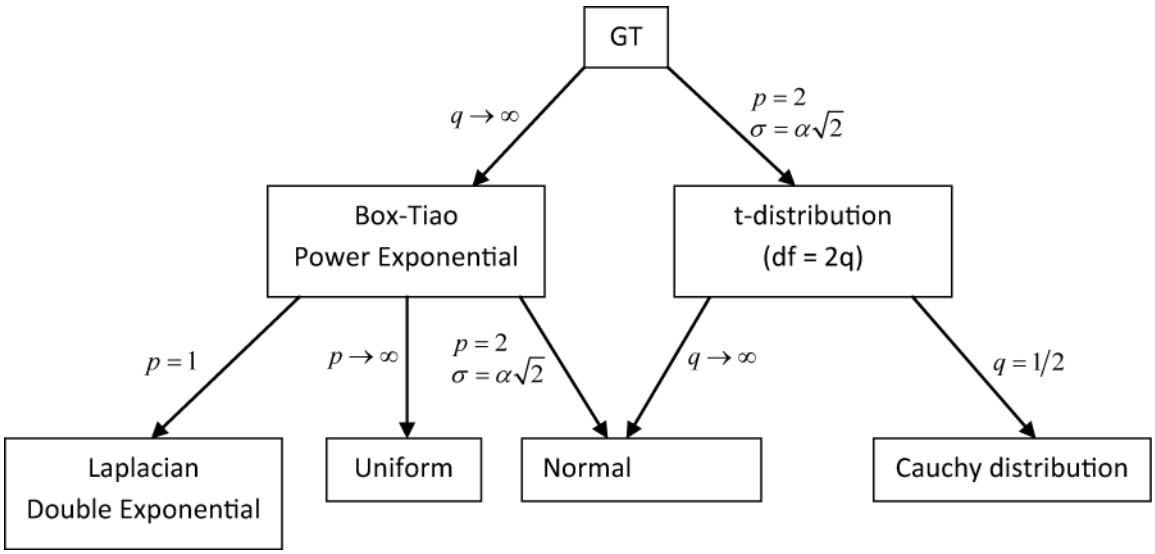


Figure 2.3: Different choices of the GT distribution shape parameters p and q can give different well-known distributions.

To obtain the maximum likelihood estimate $\hat{\theta}$, we minimize the following cost function [14, 18, 33]

$$J = -\sum_{k=1}^{N} \ln\left(f(y(k) - \phi(k)^T\theta\right) = -\sum_{k=1}^{N} \ln(f(\varepsilon(k)) \tag{2.3}$$

by differentiating

$$\frac{\partial J}{\partial \theta} = \psi(\varepsilon) = -(pq+1)\begin{bmatrix} \sum_{k=1}^{N} \frac{\phi_1(k)\varepsilon(k)|\varepsilon(k)|^{p-2}}{q\sigma^p+|\varepsilon(k)|^p} \\ \vdots \\ \sum_{k=1}^{N} \frac{\phi_n(k)\varepsilon(k)|\varepsilon(k)|^{p-2}}{q\sigma^p+|\varepsilon(k)|^p} \end{bmatrix} \tag{2.4}$$

if $p > 1$ and setting

$$\psi(\varepsilon) = 0 \tag{2.5}$$

Equation (2.5) can be solved for $\hat{\theta}$ numerically using the Newton Raphson or the Expectation Maximization algorithm [34].

2.2.2 Gaussian Distributed Noise

To see things in perspective, we now show that by choosing $p = 2$, $q = \infty$, the estimator reduces to the well-known least-squares estimator.

Consider the GT probability density function, $f(\varepsilon)$, in Equation (2.2) with $p = 2$ and $q = \infty$,

$$\begin{aligned} \frac{p}{2\sigma q^{1/p}\beta(1/p,q)\left(1+\frac{|\varepsilon|^p}{q\sigma^p}\right)^{1/p}} &= \frac{1}{\sqrt{\pi}\sigma} \\ \frac{1}{\left(1+\frac{|\varepsilon|^p}{q\sigma^p}\right)^{q}} &= \exp\left(-\frac{\varepsilon^2}{\sigma^2}\right) \end{aligned}$$

and Equation (2.2) reduces to

$$f(\varepsilon) = \frac{1}{\sqrt{2\pi}\Lambda}\exp\left(-\frac{\varepsilon^2}{2\Lambda^2}\right)$$

the Gaussian probability density function with standard deviation $\Lambda = \frac{\sigma}{\sqrt{2}}$. Thus Equation (2.3) reduces to

$$\begin{aligned} J &= -\sum_{k=1}^{N} \ln \frac{1}{\sqrt{2\pi}\Lambda} \exp\left(-\frac{\varepsilon(k)^2}{2\Lambda^2}\right) \\ &= \frac{1}{2\Lambda^2} \sum_{k=1}^{N} \left(y(k) - \phi(k)^T\theta\right)^2 - N \ln \frac{1}{\sqrt{2\pi}\Lambda} \end{aligned}$$

Since the second term in the cost function J is independent of θ, minimizing J with respect to θ reduces to the well-known least-squares optimization problem. Equation (2.5) reduces to

$$\psi(\varepsilon) = -\frac{1}{\Lambda^2} \begin{bmatrix} \sum_{k=1}^{N} \phi_1(k) \left(y(k) - \phi(k)^T\theta\right) \\ \vdots \\ \\ \sum_{k=1}^{N} \phi_n(k) \left(y(k) - \phi(k)^T\theta\right) \end{bmatrix} = 0$$

and the well-known least-squares solution [35]

$$\hat{\theta} = \left(\Phi^T\Phi\right)^{-1} \Phi^T Y \tag{2.6}$$

where

$$\Phi = \begin{bmatrix} \phi(1)^T \\ \phi(2)^T \\ \vdots \\ \phi(N)^T \end{bmatrix} \qquad Y = \begin{bmatrix} y(1) \\ y(2) \\ \vdots \\ y(N) \end{bmatrix}$$

2.3 Influence Function Analysis of the Estimate

In Section 2.2, the estimator was designed with $f(\varepsilon)$, the GT noise model of Equation (2.2). In this section and the next section, we use IF to analyze the estimate

when the estimator designed with $f(\varepsilon)$ is applied to actual noise with probability density function $g(\varepsilon)$ which is not necessarily equal to $f(\varepsilon)$.

Recall that the first-order Taylor series expansion

$$y = y_0 + \left.\frac{dy}{dx}\right|_{x=x_0} (x - x_0)$$

make use of the gradient $\left.\frac{dy}{dx}\right|_{x=x_0}$ at $x = x_0$ to give the approximate value of y at x. Consider $\bar{\theta}$, the asymptotic value of the estimate. Let $\bar{\theta}$ be associated with the probability density function of $(1-h)f(\varepsilon) + h\delta(\varepsilon)$. Likewise, the Taylor series expansion

$$\bar{\theta} \;=\; \bar{\theta}_0 + \left.\frac{\partial\bar{\theta}}{\partial h}\right|_{h=0} (1-0) \tag{2.7}$$

make use of the gradient $\left.\frac{\partial\bar{\theta}}{\partial h}\right|_{h=0}$ at $h = 0$ to give the approximate value of $\bar{\theta}$ at $h = 1$.

The term for the gradient in Equation (2.7) is known as the influence function (IF) defined as [2, 25]

$$\mathrm{IF}(\varepsilon) \;=\; \left.\frac{\partial\bar{\theta}}{\partial h}\right|_{h=0} = -\left(\int_{-\infty}^{\infty} \frac{\partial\psi(\varepsilon)}{\partial\bar{\theta}} f(\varepsilon)d\varepsilon\right)^{-1} \psi(\varepsilon) \tag{2.8}$$

where

$$\int_{-\infty}^{\infty} \frac{\partial\psi(\varepsilon)}{\partial\bar{\theta}} f(\varepsilon)d\varepsilon \;=\; \begin{bmatrix} \int_{-\infty}^{\infty} \frac{\partial\psi_1(\varepsilon)}{\partial\bar{\theta}_1} f(\varepsilon)d\varepsilon & \dots & \int_{-\infty}^{\infty} \frac{\partial\psi_1(\varepsilon)}{\partial\bar{\theta}_n} f(\varepsilon)d\varepsilon \\ \vdots & \vdots & \vdots \\ \int_{-\infty}^{\infty} \frac{\partial\psi_n(\varepsilon)}{\partial\bar{\theta}_1} f(\varepsilon)d\varepsilon & \dots & \int_{-\infty}^{\infty} \frac{\partial\psi_n(\varepsilon)}{\partial\bar{\theta}_n} f(\varepsilon)d\varepsilon \end{bmatrix} \tag{2.9}$$

For the linear in the parameter model of Equation (2.1)

$$\frac{\partial\psi_i(\varepsilon)}{\partial\bar{\theta}_j} \;=\; \sum_{k=1}^{N} \frac{\phi_i(k)\phi_j(k)(pq+1)[(p-1)q\sigma^p - |\varepsilon(k)|^p]|\varepsilon(k)|^{p-2}}{(q\sigma^p + |\varepsilon(k)|^p)^2} \tag{2.10}$$

The derivation of Equation (2.8) is given in the Appendix A.

Using the definition of IF(ε) in Equation (2.8), Equation (2.7) gives

$$\bar{\theta} = \bar{\theta}_0 + \text{IF}(\varepsilon) \tag{2.11}$$

In Equation (2.11), IF(ε) is the change in the estimate, $\bar{\theta} - \bar{\theta}_0$. When ε is associated with probability density function $g(\varepsilon)$ then the mean is used in the following first-order von Mises expansion [36, 37] to give $\bar{\theta}$.

$$\bar{\theta} = \bar{\theta}_0 + \int_{-\infty}^{\infty} \text{IF}(\varepsilon) g(\varepsilon) d\varepsilon \tag{2.12}$$

Let $\Delta\bar{\theta} = \bar{\theta} - \bar{\theta}_0$ and rewrite Equation (2.12) as

$$\Delta\bar{\theta} \;\; = \;\; \int_{-\infty}^{\infty} \text{IF}(\varepsilon) g(\varepsilon) d\varepsilon \tag{2.13}$$

and the variance is given by [2]

$$\text{Var } \bar{\theta} \;\; = \;\; \int_{-\infty}^{\infty} \text{IF}(\varepsilon) \text{IF}^T(\varepsilon) g(\varepsilon) d\varepsilon \tag{2.14}$$

Equations (2.13) and (2.14) are useful in analyzing the estimate when the actual noise has probability density function $g(\varepsilon)$ which is not necessarily equal to $f(\varepsilon)$ the noise model in the design of the estimator.

The assumption that $g(\varepsilon)$ is the same for all k is commonly made. In this study, we extend to the case where $g(\varepsilon)$ could be different for different sample k denoted as $g_k(\varepsilon)$. The case where $g(\varepsilon)$ could be different for different sample k is useful for the analysis of outliers (see Example 1). Hence, instead of integrating IF(ε) with $g(\varepsilon)$ in Equation (2.13), we first substitute Equation (2.4) into Equation (2.8) and

then integrate IF(ε) with different $g_k(\varepsilon)$ for different k giving

$$\Delta\bar{\theta} = \left(\int_{-\infty}^{\infty} \frac{\partial\psi(\varepsilon)}{\partial\bar{\theta}} f(\varepsilon)d\varepsilon\right)^{-1} (pq+1) \begin{bmatrix} \sum_{k=1}^{N} \int_{-\infty}^{\infty} \frac{\phi_1(k)\varepsilon(k)|\varepsilon(k)|^{p-2}}{q\sigma^p+|\varepsilon(k)|^p} g_k(\varepsilon)d\varepsilon \\ \vdots \\ \sum_{k=1}^{N} \int_{-\infty}^{\infty} \frac{\phi_n(k)\varepsilon(k)|\varepsilon(k)|^{p-2}}{q\sigma^p+|\varepsilon(k)|^p} g_k(\varepsilon)d\varepsilon \end{bmatrix} \quad (2.15)$$

Equation (2.15) is useful in the analysis of outliers.

Like Section 2.2, we can connect with the well-known least-squares estimator if we let the parameters of $f(\varepsilon)$ of Equation (2.2) be $p = 2$, $q = \infty$. If we do this then Equations (2.4), (2.8) to (2.10) reduce to

$$\psi(\varepsilon) = -\frac{1}{\Lambda^2} \begin{bmatrix} \sum_{k=1}^{N} \phi_1(k)\varepsilon(k) \\ \vdots \\ \sum_{k=1}^{N} \phi_n(k)\varepsilon(k) \end{bmatrix}$$

$$\text{IF}(\varepsilon) = -\Lambda^2 \left(\Phi^T\Phi\right)^{-1} \psi(\varepsilon) \quad (2.16)$$

$$\int_{-\infty}^{\infty} \frac{\partial\psi(\varepsilon)}{\partial\theta} f(\varepsilon)d\varepsilon = \frac{1}{\Lambda^2}\Phi^T\Phi \quad (2.17)$$

$$\frac{\partial\psi_i(\varepsilon)}{\partial\theta_j} = \sum_{k=1}^{N} \frac{\phi_i(k)\phi_j(k)}{\Lambda^2}$$

respectively. If we also let $g(\varepsilon) = f(\varepsilon)$ in Equation (2.14) then

$$\begin{aligned} \text{Var}\ \bar{\theta} &= \Lambda^2 \left(\Phi^T\Phi\right)^{-1} \left(\Phi^T\Phi\right) \Lambda^{-2} \left(\Phi^T\Phi\right)^{-1} \Lambda^2 \\ &= \Lambda^2 \left(\Phi^T\Phi\right)^{-1} \end{aligned} \quad (2.18)$$

and Equation (2.18) is the well-known variance formula for the least-square estimate [35]. If it is given that the distribution of the the noise is normal, then we should set $p = 2$ and $q = \infty$. The proposed estimator in Equation (2.5) then reduces

to Equation (2.6), the least-squares estimator, and the variance of the estimates in Equation (2.14) reduces to Equation (2.18), the variance of the least-squares estimates.

2.4 Examples

Equations derived in Section 2.3 are useful in determining the variance of the estimates and the effect of outliers. This is illustrated through the 3 examples below where the two estimators are also compared i.e. Equation (2.5) and Equation (2.6). For easy reference, the parameters used to generate the figures for the results in the 3 examples are summarized in Table 4.1. As shown in Table 4.1, the actual noise probability density function $g(\varepsilon)$ is not necessarily equal to $f(\varepsilon)$ the noise model used in the estimator design. Note that the least-squares estimator of Equation (2.6) may be considered as a special case of Equation (2.5) with $p = 2$ and $q = \infty$ for $f(\varepsilon)$.

2.4.1 Example 1: Outlier

In this example, we first do 1000 simulation runs and then show how the IF can be used to predict the effect of an outlier in the simulation result where the probability density function of the actual noise $g(\varepsilon)$ and noise model $f(\varepsilon)$ in the estimator design are not the same.

Consider the autoregressive (AR) model

$$y(k) = ay(k-1) + \varepsilon(k) \tag{2.19}$$

Table 2.1: Parameters used in Figures 2.4 – 2.12 of Examples 2.4.1 – 2.4.3.

Example Number	Figure Number	Estimator Equation	Line	N	$f(\varepsilon)$ p	q	σ	$g(\varepsilon)$
1	2.5	5	black & white	2 to 10	2	1.5	$0.1\sqrt{2}$	$\begin{cases}\delta(\varepsilon_1) & k=3\\ f(\varepsilon) & k\neq 3\end{cases}$
	2.6	6	black & white	2 to 10	2	∞	$0.1\sqrt{2}$	$\begin{cases}\delta(\varepsilon_1) & k=3\\ f(\varepsilon) & k\neq 3\end{cases}$
	2.7	5	dashed	∞	2	1.5	$0.1\sqrt{2}$	$\begin{cases}\delta(\varepsilon_1) & k=2\\ f(\varepsilon) & k\neq 2\end{cases}$
	2.7	5	solid	∞	2	1.5	$0.1\sqrt{2}$	$\begin{cases}\delta(\varepsilon_1) & k=3\\ f(\varepsilon) & k\neq 3\end{cases}$
	2.7	6	dotted	∞	2	∞	$0.1\sqrt{2}$	$\begin{cases}\delta(\varepsilon_1) & k=2\\ f(\varepsilon) & k\neq 2\end{cases}$
	2.7	6	dashed-dotted	∞	2	∞	$0.1\sqrt{2}$	$\begin{cases}\delta(\varepsilon_1) & k=3\\ f(\varepsilon) & k\neq 3\end{cases}$
2	2.9	5	solid	127	2	1.5	$0.1\sqrt{2}$	$f(\varepsilon)$
	2.10	6	solid	127	2	∞	$0.1\sqrt{2}$	$f(\varepsilon)$
3	2.11	5	solid	3	2	2	29.5	$\frac{1}{576}\delta(\varepsilon_i)\ \ i=1,..,576$
	2.12	6	solid	3	2	∞	29.5	$\frac{1}{576}\delta(\varepsilon_i)\ \ i=1,..,576$

where $y(1) = 1$, $a = 0.6$ and $\varepsilon(k)$ belongs to the t_3 distribution with zero mean and scale 0.1 except for an outlier of magnitude ε_1 at $k = k_1$. Compare with the linear in the parameters model of Equation (2.1) gives $\phi^T(k) = y(k-1)$ and $\theta = a$.

Simulation

One thousand runs of the signal $y(k)$ for $k_1 = 3$ and $\varepsilon_1 = 1$ is shown in Figure 2.4. The average values for the 1000 runs is given by the white curve.

Figure 2.5 shows the solution of Equation (2.5) for the 1000 runs in Figure 2.4 for batch size N from 2 to 10. A batch size of $N = 10$ means that 10 data points were used to give 1 estimate. Equation (2.5) assumes a GT noise model and according to Figure 2.3, $f(\varepsilon)$ of Equation (2.2) with $p = 2$, $q = 1.5$ and $\sigma = 0.1\sqrt{2}$ can be used to model the t_3 noise.

Figure 2.6 shows the solution of Equation (2.6), the least-squares estimator where the average is given by the white curve. Notice that for batch size $N = 10$ which included the outlier $\varepsilon_1 = 1$ at $k = 3$, the white curve in Figure 2.6 gives $\hat{a} = 0.74 \neq a = 0.6$ for $N = 10$, not robust to even a single outlier.

On the other hand, the outlier is largely rejected by the proposed estimator and the estimate hardly affected by the outlier as shown by the white curve in Figure 2.5.

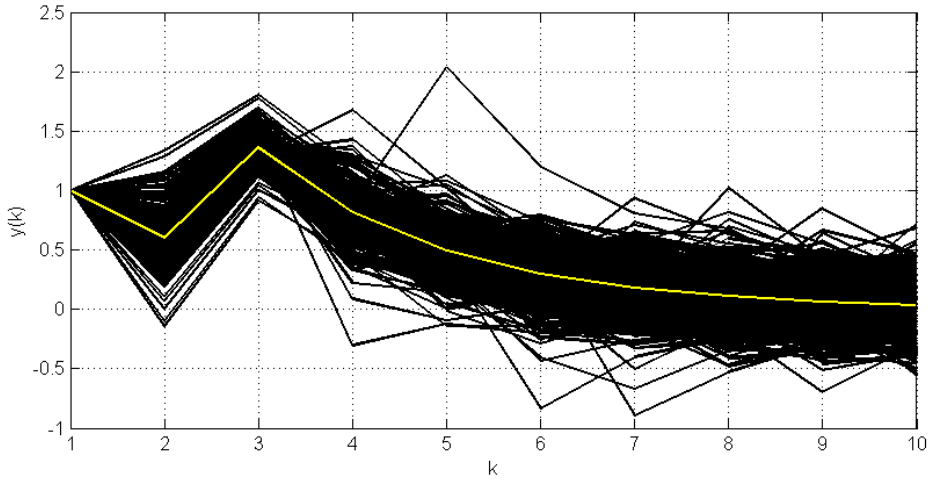


Figure 2.4: AR model $y(k) = 0.6y(k-1) + \varepsilon(k), y(1) = 1$, $\varepsilon(k)$ belongs to t_3 distribution for $k \neq 3$ and $\varepsilon(k) = 1$ for $k = 3$.

IF Analysis

Instead of simulation, Equation (2.15) can be used to give analytical results for $\Delta\bar{a}$. The case where $g(\varepsilon)$ could be different for different sample k is useful for the analysis of outliers. Let

$$g_k(\varepsilon) = \begin{cases} \delta(\varepsilon_1) & k = k_1 \\ f(\varepsilon) & k \neq k_1 \end{cases} \tag{2.20}$$

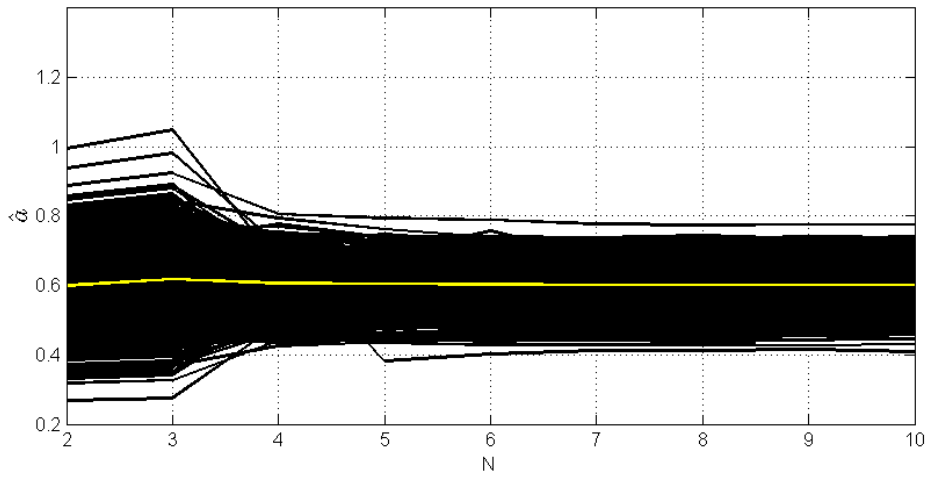


Figure 2.5: Estimate $\hat{a}$ from Equation (2.5) with GT noise assumption for 1000 runs with different batch-size, N (white: mean, black: individual run).

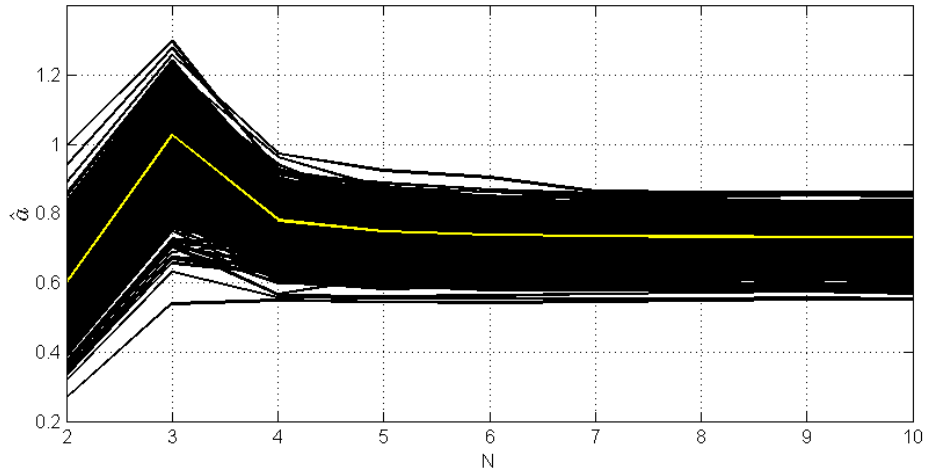


Figure 2.6: Least-Squares Estimate of $\hat{a}$ from Equation (2.6) for 1000 runs with different batch-size, N (white: mean, black: individual run).

where $\delta(\varepsilon_1)$ is an impulse at ε_1 to model the outlier of ε_1 at $k = k_1$. For the model of Equation (2.19) and $g_k(\varepsilon)$ of Equation (2.20) with $N \geq k_1$, Equation (2.15) gives

$$\begin{aligned} \Delta\bar{a} &= -\left(\Gamma \sum_{k=1}^{N} \phi^2(k)\right)^{-1} (pq+1) \\ &\quad \times \left[\int_{-\infty}^{\infty} \frac{\phi(k_1)\varepsilon(k_1)|\varepsilon(k_1)|^{p-2}}{q\sigma^p + |\varepsilon(k_1)|^p} \delta(\varepsilon_1) d\varepsilon + \sum_{k=1,\neq k_1}^{N} \int_{-\infty}^{\infty} \frac{\phi(k)\varepsilon(k)|\varepsilon(k)|^{p-2}}{q\sigma^p + |\varepsilon(k)|^p} f(\varepsilon) d\varepsilon\right] \end{aligned} \tag{2.21}$$

where

$$\begin{aligned}\phi(k) &= \begin{cases} 0 & k=1 \\ a^{k-2}y(1) & 2\le k\le k_1 \\ a^{k-1}\left[a^{-1}y(1)+a^{-k_1}\varepsilon_1\right] & k\ge k_1+1 \end{cases} \\ \Gamma &= (pq+1)\int_{-\infty}^{+\infty}\frac{[(p-1)q\sigma^p-|\varepsilon|^p)]|\varepsilon|^{p-2}}{(q\sigma^2+|\varepsilon|^p)^2}f(\varepsilon)d\varepsilon \end{aligned} \tag{2.22}$$

Note that the second term in the square bracket of Equation (2.21) is zero because the expectation of $\varepsilon(k)$ for $k \neq k_1$ is zero. Hence

$$\Delta\bar{a} = -\left(\Gamma\sum_{k=1}^{N}\phi^2(k)\right)^{-1}(pq+1)\left[\frac{\phi(k_1)\varepsilon_1|\varepsilon_1|^{p-2}}{q\sigma^p+|\varepsilon_1|^p}\right] \tag{2.23}$$

Substitute Equation (2.22) into (2.23) gives

$$\Delta\bar{a} = \frac{-(1-a^2)a^{k_1-2}y(1)(pq+1)\varepsilon_1|\varepsilon_1|^{p-2}}{\Gamma\left[\left(1-a^{2k_1-2}\right)y(1)^2+\left(a^{2k_1}-a^{2N}\right)\left(a^{-1}y(1)+a^{-k_1}\varepsilon_1\right)^2\right]\left(q\sigma^p+|\varepsilon_1|^p\right)} \tag{2.24}$$

If we substitute $a = 0.6$, $p = 2$, $q = 1.5$, $\sigma = 0.1\sqrt{2}$, $k_1 = 3$, $\varepsilon_1 = 1$ in Equation (2.24) and plot $\Delta\bar{a} + a$ versus N then the white curve in Figure 2.5 for $N \geq k_1$ is obtained. If instead of $q = 1.5$ we substitute $q = \infty$ then the white curve in Figure 2.6 for $N \geq k_1$ is obtained. Equation (2.24) allows us to study the impact of an outlier on the estimate.

To study the estimate when it has reached steady-state, substitute $N = \infty$, $a = 0.6$, $p = 2$, $q = 1.5$, $\sigma = 0.1\sqrt{2}$ in Equation (2.24) to obtain the dashed-line for $k_1 = 2$ and the solid-line for $k_1 = 3$ in Figure 2.7. For the least-squares estimator, instead of $q = 1.5$, substitute $q = \infty$ in Equation (2.24) to give the dotted-line for $k_1 = 2$ and dashed-dotted-line for $k_1 = 3$ in Figure 2.7.

Some trends can be observed in Figure 2.7. Firstly, $\Delta\bar{a}$ increases with the outlier ε_1 for the least-squares estimator (see the dotted-line and dashed-dotted-line) giving unacceptable $\Delta\bar{a}$ for large ε_1. Whereas for the estimator with GT noise model, $\Delta\bar{a}$ is small for large outlier ε_1 (see solid-line and dashed-line). Notice in Equation (2.24) that the term $a^{2N} \approx 0$ for $N \geq 10$. Hence Figure 2.7 may be used to predict the results for $N = 10$. At $\varepsilon_1 = 1$, $k_1 = 3$, Figure 2.7 predicted that $\Delta\bar{a} = 0.01$ (solid-line) and $\Delta\bar{a} = 0.14$ (dashed-dotted-line) giving $\bar{a} = a + \Delta\bar{a} = 0.61$ (white-line at $N = 10$ in Figure 2.5) and 0.74 (white-line at $N = 10$ in Figure 2.6). Hence Equation (2.24) can be used to select p, q and σ to limit the effect of outlier on the estimation results.

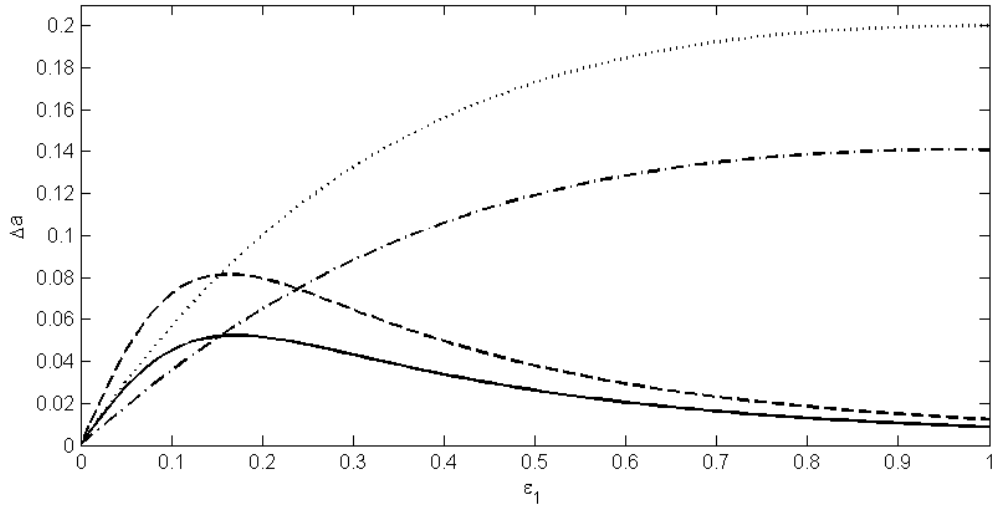


Figure 2.7: Change in estimate $\Delta\bar{a}$ for batch-size $N = \infty$ and outlier ε_1 (Equation 2.24). Least-square estimator for outlier at $k_1 = 2$ (dotted-line), $k_1 = 3$ (dashed-dotted-line). Estimator with GT noise model for outlier at $k_1 = 2$ (dashed-line) and $k_1 = 3$ (solid-line).

2.4.2 Example 2: Variance

In this example, we first do 1000 simulation runs and then show how the IF can be used to predict the variance of the simulation results where the probability density function $g(\varepsilon)$ of the actual noise and the noise model $f(\varepsilon)$ in the estimator design are the same.

Consider the following autoregressive with exogenous input (ARX) model which is commonly used to model first-order dynamics encountered in chemical processes such as thermal processes or liquid-level systems:

$$y(k+1) = ay(k) + bu(k) + \epsilon(k+1) \tag{2.25}$$

where $a = 0.6$, $b = 0.4$ and $\epsilon(k)$ belongs to the t_3 distribution with zero mean and scale 0.1. Comparing with the linear in the parameters model of Equation (2.1) gives $\phi^T(k) = [y(k) \quad u(k)]$ and $\theta = [a \quad b]^T$. The input signal is the Pseudo-random Binary Sequences (PRBS)[38].

Simulation

An example of a simulation run with sample size $N = 127$ is shown in Figure 2.8 with $\hat{a}$ and $\hat{b}$ estimated using Equations (2.5) and (2.6) where Equation (2.5) assumes a GT noise model and from Figure 2.3, the parameters of $f(\varepsilon)$ of Equation (2.2) to model the t_3 distribution are $p = 2$, $q = 1.5$ and $\sigma = 0.1\sqrt{2}$. Equation (2.6) is the Least-Square estimator. A total of 1000 simulation runs were conducted. The estimates $\hat{a}$ and $\hat{b}$ at the end of each run were recorded in Figure 2.9 (GT noise model) and Figure 2.10 (least-squares estimation) with their variances computed in

Column 2 and 4 of Table 2.2 which clearly shows that the variance of the estimate with the GT noise model is smaller and hence more precise.

Table 2.2: Variance in Example 2

Estimate	Est w GT Noise Model, (Eq. 2.5)		Least-Squares Est (Eq. 2.6)	
	Variance (Fig. 2.9 simulation)	Variance (Eqn. 2.27)	Variance (Fig. 2.10 simulation)	Variance (Eqn. 2.29)
$\hat{a}$	0.71×10^{-3}	0.69×10^{-3}	1.26×10^{-3}	1.35×10^{-3}
$\hat{b}$	0.62×10^{-3}	0.58×10^{-3}	1.17×10^{-3}	1.14×10^{-3}

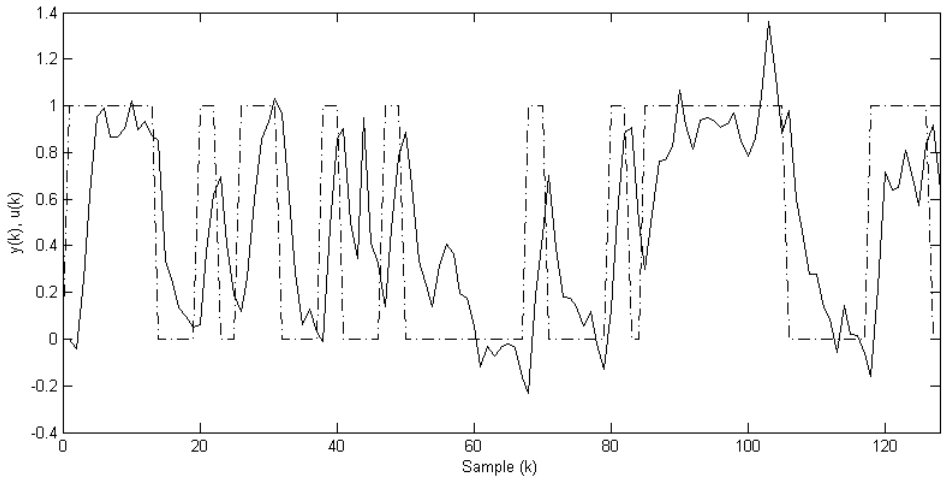


Figure 2.8: Response of the ARX model to pseudo random binary control signal (dashed-line: $u(k)$, solid-line: $y(k)$).

IF Analysis: GT Noise Model

Instead of simulation, Equation (2.14) can be used to give analytical result for the variance of the estimate. Consider the estimator of Equation (2.5) which assumed

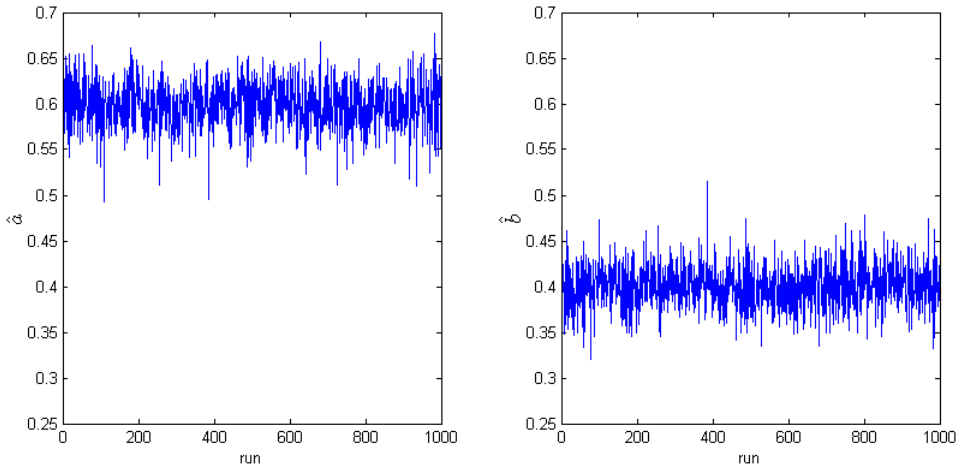


Figure 2.9: Estimates $\hat{a}$ and $\hat{b}$ using Equation (2.5) assuming the GT noise model.

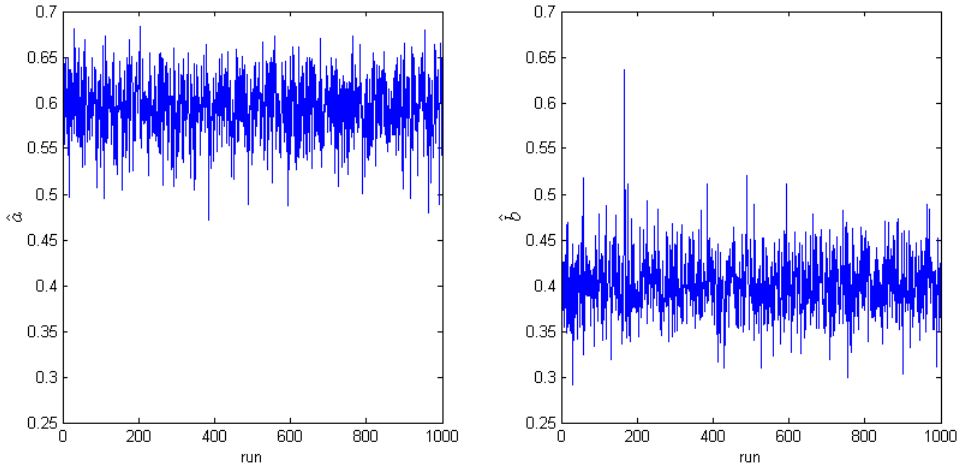


Figure 2.10: Estimates $\hat{a}$ and $\hat{b}$ using Equation (2.6), the least-squares estimator.

the GT noise model $f(\varepsilon)$ of Equation (2.2). From Equation (2.8),

$$\begin{aligned} \mathrm{IF}(\varepsilon) &= (\Phi^T\Phi)^{-1}\left(\int_{-\infty}^{\infty}\frac{0.03-\varepsilon^2}{(0.03+\varepsilon^2)^2}f(\varepsilon)d\varepsilon\right)^{-1}\begin{bmatrix}\sum_{k=1}^{127}\frac{y(k)\varepsilon(k)}{0.03+\varepsilon(k)^2}\\ \sum_{k=1}^{127}\frac{u(k)\varepsilon(k)}{0.03+\varepsilon(k)^2}\end{bmatrix} \\ &= \frac{3}{50}(\Phi^T\Phi)^{-1}\begin{bmatrix}\sum_{k=1}^{127}\frac{y(k)\varepsilon(k)}{0.03+\varepsilon(k)^2}\\ \sum_{k=1}^{127}\frac{u(k)\varepsilon(k)}{0.03+\varepsilon(k)^2}\end{bmatrix} \end{aligned} \tag{2.26}$$

where

$$\Phi^T\Phi = \begin{bmatrix} \sum_{k=1}^{127} y(k)^2 & \sum_{k=1}^{127} y(k)u(k) \\ \sum_{k=1}^{127} u(k)y(k) & \sum_{k=1}^{127} u(k)^2 \end{bmatrix}$$

Substituting the IF(ε) from Equation (2.26) into Equation (2.14) gives

$$\text{Var } \bar{\theta} = \left(\frac{3}{50}\right)^2 (\Phi^T\Phi)^{-1} \int_{-\infty}^{\infty} \left(\begin{bmatrix} \sum_{k=1}^{127} \frac{y(k)\varepsilon(k)}{0.03+\varepsilon(k)^2} \\ \sum_{k=1}^{127} \frac{u(k)\varepsilon(k)}{0.03+\varepsilon(k)^2} \end{bmatrix} \begin{bmatrix} \sum_{k=1}^{127} \frac{y(k)\varepsilon(k)}{0.03+\varepsilon(k)^2} & \sum_{k=1}^{127} \frac{u(k)\varepsilon(k)}{0.03+\varepsilon(k)^2} \end{bmatrix} \right) g(\varepsilon)d\varepsilon(\Phi^T\Phi)^{-1}$$

Because ε is assumed to be a zero mean independent random variable,

$$\int_{-\infty}^{\infty} \frac{\varepsilon(j)\varepsilon(k)}{(0.03+\varepsilon(j)^2)(0.03+\varepsilon(k)^2)} g(\varepsilon)d\varepsilon = 0 \quad \text{for } j \neq k$$

and

$$\text{Var } \bar{\theta} \;=\; \left(\frac{3}{50}\right)^2 (\Phi^T\Phi)^{-1} \left(\int_{-\infty}^{\infty} \frac{\varepsilon^2}{(0.03+\varepsilon^2)^2} g(\varepsilon)d\varepsilon \right) \tag{2.27}$$

The variances in Column 3 of Table 2.2 were computed from Equation (2.27) which is close to the variances obtained from simulation in Column 2. In Equation (2.27), $y(k)$ is taken from the first run and $g(\varepsilon) = f(\varepsilon)$.

IF Analysis: Least-Squares Estimate

Consider the least-squares estimator of Equation (2.6). From Equation (2.16)

$$\text{IF}(\varepsilon) = (\Phi^T\Phi)^{-1} \begin{bmatrix} \sum_{k=1}^{127} y(k)\varepsilon(k) \\ \sum_{k=1}^{127} u(k)\varepsilon(k) \end{bmatrix} \tag{2.28}$$

Substituting the IF(ε) from Equation (2.28) into Equation (2.14) gives

$$\text{Var } \bar{\theta} = (\Phi^T\Phi)^{-1} \int_{-\infty}^{\infty} \left(\begin{bmatrix} \sum_{k=1}^{127} y(k)\varepsilon(k) \\ \sum_{k=1}^{127} u(k)\varepsilon(k) \end{bmatrix} \begin{bmatrix} \sum_{k=1}^{127} y(k)\varepsilon(k) & \sum_{k=1}^{127} u(k)\varepsilon(k) \end{bmatrix} \right) g(\varepsilon)d\varepsilon(\Phi^T\Phi)^{-1}$$

Because ε is assumed to be a zero mean independent random variable, $\int_{-\infty}^{\infty} \varepsilon(j)\varepsilon(k)$ $g(\varepsilon)d\varepsilon = 0$ for $j \neq k$ and

$$\text{Var } \bar{\theta} = (\Phi^T\Phi)^{-1} \int_{-\infty}^{\infty} \varepsilon^2 g(\varepsilon)d\varepsilon \tag{2.29}$$

The variances in Column 5 of Table 2.2 were computed from Equation (2.29) which is close to the variances obtained from simulation in Column 4. In Equation (2.29), $y(k)$ is taken from the first run and $g(\varepsilon) = f(\varepsilon)$.

Equation (2.14) allows us to calculate the variances of the estimates and hence their precisions if the number of data points, N, used is given. Alternatively, it enables us to compute the sample size, N, needed by the estimator to meet specified variance.

2.4.3 Example 3: Chemical-Mechanical-Polishing Experiment

The linear in the parameter model with GT noise of Equation (2.1) can also be used to estimate the states. Consider the chemical-mechanical polishing of twenty-four 200mm wafers where the thickness at 576 points (24 points per wafer) were measured after polishing. Figure 2.1 shows the data points. The process can be modeled by Equation (2.1) where $y(k)$ is the measurement and $\phi(k) = 1$.

The maximum likelihood criterion can be used to find the parameters of the GT probability density function [14, 17]. In this example we fixed $p = 2$ and then use the maximum likelihood criterion to find the other parameters q, σ and μ of the GT probability density function $f(\varepsilon)$ of Equation (2.2) by maximizing the objective

function

$$J_f = \sum_{k=1}^{576} \ln \frac{p}{2\sigma q^{1/p}\beta(1/p,q)\left(1+\frac{|y(k)-\mu|^p}{q\sigma^p}\right)^{q+1/p}}$$

This gives $q = 2$, $\sigma = 29.5$nm, $\mu = 383$nm and the resultant GT distribution is superimposed on the data distribution in Figure 2.2. The estimate $\hat{\theta}$ which is also the estimate $\hat{y}$ obtained from Equations (2.5) and (2.6) for $N = 3$ are shown in Figures 2.11 and 2.12 respectively.

IF Analysis: GT Noise Model

Consider the estimator of Equation (2.5). From Equation (2.8)

$$\begin{aligned} \text{IF}(\varepsilon) &= \frac{1}{3}\left(\int_{-\infty}^{\infty} \frac{1740-\varepsilon^2}{(1740+\varepsilon^2)^2} f(\varepsilon)d\varepsilon\right)^{-1} \sum_{k=1}^{3} \frac{\varepsilon(k)}{1740+\varepsilon(k)^2} \\ &= 1020 \sum_{k=1}^{3} \frac{\varepsilon(k)}{1740+\varepsilon(k)^2} \end{aligned} \tag{2.30}$$

Equation (2.14) can be used to calculate the variance by using the empirical discrete distribution (the histogram of data distribution in Figure 2.2) from the 576 experimental data points, ε, given as

$$g(\varepsilon) = \frac{1}{576}\delta(\varepsilon_i), \qquad i = 1, 2, \ldots, 576 \tag{2.31}$$

Substituting IF(ε) and $g(\varepsilon)$ from Equations (2.30) and (2.31) into Equation (2.14) gives

$$\text{Var } \bar{y} = \frac{1020^2}{576} \int_{-\infty}^{\infty} \left(\sum_{k=1}^{3} \frac{\varepsilon(k)}{1740+\varepsilon(k)^2}\right)\left(\sum_{k=1}^{3} \frac{\varepsilon(k)}{1740+\varepsilon(k)^2}\right)\delta(\varepsilon_i)d\varepsilon$$

Because ε is assumed to be a zero mean independent random variable

$\int_{-\infty}^{\infty} \frac{\varepsilon(j)\varepsilon(k)}{(1740+\varepsilon(j)^2)(1740+\varepsilon(k)^2)}\delta(\varepsilon_i)d\varepsilon = 0$ for $j \neq k$ and

$$\text{Var } \bar{y} = 5419 \sum_{i=1}^{576} \frac{\varepsilon_i^2}{(1740+\varepsilon_i^2)^2} \tag{2.32}$$

Variances obtained from the experimental results in Figure 2.11 and Equation (2.32) are both 217nm. Hence Equation (2.32) can be used to predict the variance of the experimental results.

IF Analysis: Least-Squares Estimate

Consider the least-squares estimator of Equation (2.6). From Equation (2.16)

$$\text{IF}(\varepsilon) \quad = \quad \frac{1}{3}\sum_{k=1}^{3}\varepsilon(k) \tag{2.33}$$

Substituting the $g(\varepsilon)$ and $\text{IF}(\varepsilon)$ of Equations (2.31 and (2.33) into Equation (2.14) gives

$$\text{Var } \bar{y} = \frac{1}{3^2 \times 576}\int_{-\infty}^{\infty}\left(\sum_{k=1}^{3}\varepsilon(k)\right)\left(\sum_{k=1}^{3}\varepsilon(k)\right)\delta(\varepsilon_i)d\varepsilon$$

Because ε is assumed to be a zero mean, independent and identical distributed random variable $\int_{-\infty}^{\infty}\varepsilon(j)\varepsilon(k)\delta(\varepsilon_i)d\varepsilon = 0$ for $j \neq k$ and

$$\text{Var } \bar{y} = \frac{1}{1728}\sum_{i=1}^{576}\varepsilon_i^2 \tag{2.34}$$

Variances obtained from the experimental results in Figure 2.12 and Equation (2.34) are both 264nm. Hence Equation (2.34) can be used to predict the variance of the experimental results.

The variances show that using the GT distribution to model the noise reduces the variance by 18% $\left(\frac{264-217}{264}\right)$. Notice the 2 outlier measurements around $k = 300$ in Figure 2.1 gives rise to a large change in the least-squares estimate $\hat{y}$ in Figure 2.12 at around Batch 100 but not the estimate $\hat{y}$ in Figure 2.11 when the GT noise model was used.

In manufacturing, time and cost are incurred when measurements are taken especially if these measurements have to be done separately, off-line from the manufacturing process such as the thickness measurements here. If a desired measurement variance is specified then Equation (2.14) enable us to calculate the number of measurements (N) needed and not take more measurements than is needed.

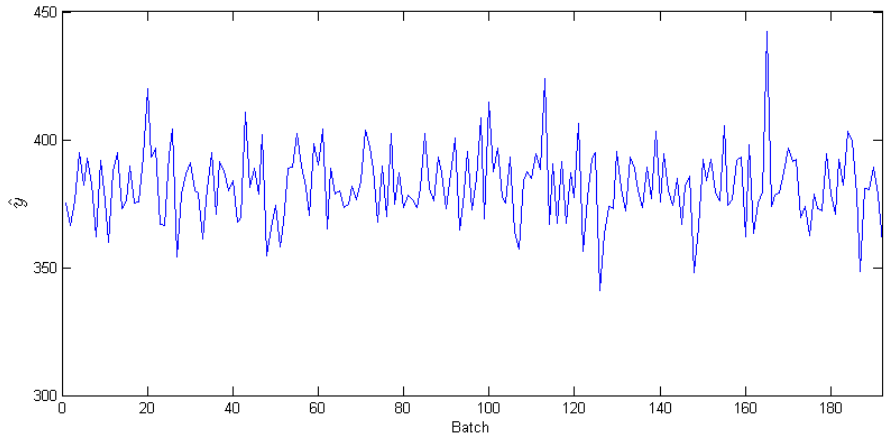


Figure 2.11: Estimate of the thickness measurements, $\hat{y}$, with GT noise model.

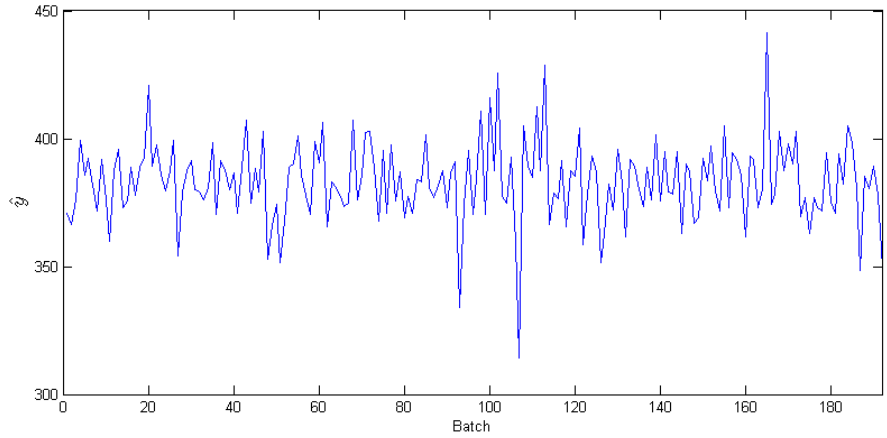


Figure 2.12: Estimate of the thickness measurements, $\hat{y}$, using least-squares estimation.

2.5 Conclusion

In this chapter, we used IF to analyze the estimate from the parameter estimator designed with the GT noise model instead of the usual Gaussian noise model. The analysis is extended to the case where the estimator designed with probability density function $f(\varepsilon)$ is applied to noise with different probability density function $g_k(\varepsilon)$ at different sampling instance, k, to provide a framework for analysis of outliers. Equations derived are useful in determining the variance of the estimates and the impact of outliers. If the noise is modeled by the Gaussian distribution then the proposed estimator reduces to the least-square estimator. Otherwise, the GT distribution has the extra degree of freedom to model non-Gaussian noise. If we do not know the distribution of the noise then one can use use the least-squares estimator. However, if there is information on the distribution then it can be used gainfully in the GT distribution framework to model non-Gaussian noise giving rise to a smaller variance for the estimates and robustness to outliers.

Chapter 3

Filtering of the ARMAX process

3.1 Introduction

In the statistical analysis of time series, the autoregressive-moving-average with exogenous inputs model (ARMAX) with Gaussian noise is commonly used. However, the Gaussian noise assumption is an approximation to reality. The occurrence of outliers, transient data in steady-state measurements, instrument failure, human error, model nonlinearity, etc. can all induce non-Gaussian data [14]. Indeed whenever the central limit theorem is invoked — the central limit theorem being a limit theorem can at most suggest approximate normality for real data [2]. However, even high-quality model data may not fit the Gaussian distribution and the presence of a single outlier can spoil the statistical analysis completely for the case of least-squares estimation [39] including the Kalman filter [40].

As stated in previous chapter, the GT distribution is applied in many applications. However, the problem of estimation with GT noise was usually solved nu-

merically using the Newton Raphson or the Expectation Maximization algorithm [14, 17, 20, 21, 24, 28, 29]. Unlike recursive algorithm such as the recursive least-squares estimator, it is not suitable for real-time applications.

In this chapter, Influence Function (IF), an analysis tool in robust statistics [2, 25], is used to formulate a recursive algorithm that give an approximate solution making it suitable for real-time and on-line implementation. Specifically the problem is formulated as the filtering of the ARMAX process with GT noise. Other well-known approaches [30–32] for handling non-Gaussian noise include the approach of particle filters which is based on point mass or particle representation of probability densities.

The IF was used in chapter 2 to analyze parameter estimation with GT noise. Instead of using the IF as an analysis tool to analyze a given estimator this thesis make use of the IF to synthesize or construct an estimator. The other difference is that while chapter 2 studied the estimation of the parameters in the transfer function, this study estimates the states or output of the transfer function.

The main contribution of this chapter is in Sections 3.3 and 3.4 where we use IF approximation to derive a recursive solution for the maximum likelihood estimation of the ARMAX Process with GT noise. We also show how the IF can be used to analyze the filter, specifically how it can predict the filter output due to outliers and the variance of the output. To put things in perspective, it will be shown through an example that if the noise is Gaussian then the proposed ARMAX filter is equivalent to the Kalman filter [40]. Otherwise the ARMAX filter has the extra degrees of freedom to model the noise.

3.2 Maximum Likelihood Estimation of the ARMAX Process with GT Noise

The ARMAX process and maximum likelihood estimation with GT distribution [14, 17, 20, 21, 24, 28, 29] are already given in the literature. In this section we only give the equations necessary for the derivation of the recursive algorithm using IF approximation in the next section.

3.2.1 The ARMAX Process

Consider the single-input single-output ARMAX process:

$$A(q^{-1})y(k) = B(q^{-1})u(k) + C(q^{-1})\varepsilon(k) \tag{3.1}$$

where

$$A(q^{-1}) = 1 + a_1 q^{-1} + \cdots + a_n q^{-n}$$

$$B(q^{-1}) = b_1 q^{-1} + b_2 q^{-2} + \cdots + b_{n_B} q^{-n_B}$$

$$C(q^{-1}) = 1 + c_1 q^{-1} + \cdots + c_n q^{-n}$$

$k = 1, \ldots, N$ is the sampling instance, $n_B \leq n$ and q^{-1} is the backward shift operator, i.e., $q^{-1}y(k) = y(k-1)$. The polynomial C may be multiplied by an arbitrary power of q as this does not change the correlation structure of $C(q^{-1})$. This is used to normalized C so that deg C = deg A = n. The input and output are given by $u(k)$ and $y(k)$ respectively. Let the noise $\varepsilon(k)$ be modeled by the zero-mean GT probability density function (2.2) [17].

3.2.2 The Diohpantine Equation

The Diophantine Equation [35, 41, 42] or Identity can be used to isolate the noise term in the ARMAX process. The Diophantine Equation [35, 41, 42] is given as

$$C\left(q^{-1}\right) = E(q^{-1})A(q^{-1}) + q^{-j}F(q^{-1}) \tag{3.2}$$

where $C\left(q^{-1}\right)$ is asymptotically stable and

$$\begin{aligned} E(q^{-1}) &= 1 + e_1 q^{-1} + \cdots + e_{j-1} q^{-j+1} \\ F\left(q^{-1}\right) &= f_0 + f_1 q^{-1} + \cdots + f_{n_F} q^{-n_F} \\ n_F &= n - 1 \end{aligned}$$

Using Equation (3.2) for $j = 1$, Equation (3.1) becomes

$$y(k+1) = \frac{F(q^{-1})}{C(q^{-1})} y(k) + \frac{B(q^{-1})}{C(q^{-1})} u(k+1) + \varepsilon(k+1) \tag{3.3}$$

Multiplying by q^{-1}, the current measurement can be obtained from Equation (3.3) as

$$y(k) = \frac{F(q^{-1})}{C(q^{-1})} y(k-1) + \frac{B(q^{-1})}{C(q^{-1})} u(k) + \varepsilon(k) \tag{3.4}$$

As it was found to be more convenient to work in the state-space, expressing Equations (3.4) in the state-space form gives

$$x(k+1) = \Phi x(k) + \Gamma u(k) + \Omega y(k) \tag{3.5}$$

$$y(k) = H x(k) + \varepsilon(k) \tag{3.6}$$

where

$$\Phi = \begin{bmatrix} -c_1 & 1 & 0 & \dots & \dots & 0 \\ -c_2 & 0 & 1 & 0 & \dots & 0 \\ \vdots & & & & & \\ -c_{n-1} & 0 & \dots & \dots & 0 & 1 \\ -c_n & 0 & \dots & \dots & \dots & 0 \end{bmatrix}$$

$$\Gamma = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_{N_B} \\ 0 \end{bmatrix}$$

$$\Omega = \begin{bmatrix} c_1 - a_1 \\ c_2 - a_2 \\ \vdots \\ c_n - a_n \end{bmatrix}$$

$$H = \begin{bmatrix} 1 & 0 & \dots & 0 \end{bmatrix}$$

Iterating from the initial value $x(1)$, Equations (3.5) and (3.6) gives

$$\begin{aligned} x(2) &= \Phi x(1) + \Gamma u(1) + \Omega y(1) \\ x(3) &= \Phi^2 x(1) + \Phi\Gamma u(1) + \Gamma u(2) + \Phi\Omega y(1) + \Omega y(2) \\ &\vdots \\ x(N) &= \Phi^{N-1} x(1) + \bar{x}(N) \qquad (3.7) \\ y(N) &= H\Phi^{N-1} x(1) + H\bar{x}(N) + \varepsilon(N) \qquad (3.8) \end{aligned}$$

where

$$\bar{x}(N) = \sum_{k=1}^{N-1} \Phi^{k-1}\Gamma u(N-k) + \sum_{k=1}^{N-1} \Phi^{k-1}\Omega y(N-k) \qquad (3.9)$$

3.2.3 Maximum Likelihood Estimation

Given N measurements $y(k)$, $k = 1, \ldots N$, the initial condition, $x(1)$, can be estimated using Equation (3.8) in the minimization of the following maximum likelihood cost function

$$J = -\sum_{k=1}^{N} \ln f(\varepsilon(k)) = -\sum_{k=1}^{N} \ln f\left(y(k) - H\Phi^{k-1}x(1) - H\bar{x}(k)\right)$$

This can be done by differentiating wrt $x(1)$

$$\frac{\partial J}{\partial x(1)} = \psi(\varepsilon) = -(pq+1)\sum_{k=1}^{N} (H\Phi^{k-1})^T \frac{\varepsilon(k)|\varepsilon(k)|^{p-2}}{q\sigma^p - |\varepsilon(k)|^p} \qquad (3.10)$$

where $p > 1$ and setting

$$\psi(\varepsilon) = 0 \qquad (3.11)$$

Equation (3.11) can be solved for $x(1)$ numerically using the Newton Raphson or the Expectation Maximization algorithm [34, 43]. Unlike recursive algorithm such

as the recursive least-squares estimator, Equation (3.11) is not suitable for real-time applications. For example in real-time control, the information is used by the controller to calculate the control signal for the next sampling instance. The number of iterations required by Equation (3.11) to converge to a solution can be different for different sample and hence there is no guarantee that the information is available before the next sampling instance.

3.3 Influence Function Approximation

In this Section, we introduce the influence function to approximate and solve Equation (3.11) recursively.

Consider the function $x = f(h)$. The first-order Taylor series expansion

$$x = \left.\frac{dx}{dh}\right|_{h=0} h$$

makes use of the gradient $\left.\frac{dx}{dh}\right|_{h=0}$ to give the approximate value of x at h. Consider $\hat{x}(1)$, the asymptotic value of the estimate of $x(1)$. Let $\hat{x}(1)$ be associated with the probability density function of $(1-h)f(\varepsilon) + h\delta(\varepsilon)$. Likewise the Taylor series expansion

$$\hat{x}(1) = \left.\frac{d\hat{x}(1)}{dh}\right|_{h=0} h \tag{3.12}$$

makes use of the gradient $\left.\frac{d\hat{x}(1)}{dh}\right|_{h=0}$ to give the approximate value of $\hat{x}(1)$ at h. The gradient term in Equation (3.12) known as the Influence Function (IF) is defined in [2, 25] as

$$\mathrm{IF}(\varepsilon) \;=\; \left.\frac{\partial\hat{x}(1)}{\partial h}\right|_{h=0} = -\left(\int_{-\infty}^{\infty}\frac{\partial\psi(\varepsilon)}{\partial\hat{x}(1)}f(\varepsilon)d\varepsilon\right)^{-1}\psi(\varepsilon)\Big|_{x(1)=0} \tag{3.13}$$

where

$$\frac{\partial\psi(\varepsilon)}{\partial\hat{x}(1)} = \sum_{k=1}^{N}(H\Phi^{k-1})^T H\Phi^{k-1}\frac{(pq+1)[(p-1)q\sigma^p - |\varepsilon(k)|^p]|\varepsilon(k)|^{p-2}}{(q\sigma^p+|\varepsilon(k)|^p)^2}$$

Derivation of Equation (3.13) is given in the Appendix A. When $h = 0$, the associated probability density function of $\hat{x}(1)$ is $f(\varepsilon)$ and the usual assumption of zero initial condition for the ARMAX transfer function is made i.e. $x(1) = 0$.

3.3.1 The Recursive Algorithm

The solution for $\hat{x}(1)$ can be written in the form of a recursive algorithm. Substituting Equations (3.10), (3.13) and $h = 1$ into Equation (3.12) gives

$$\hat{x}(1|N) = \mathrm{IF}(\varepsilon) = \left(\sum_{k=1}^{N}(H\Phi^{k-1})^T H\Phi^{k-1}\right)^{-1}\left(\sum_{k=1}^{N}(H\Phi^{k-1})^T z(k)\right) \tag{3.14}$$

where

$$z(k) = \left(\int_{-\infty}^{+\infty}\frac{[(p-1)q\sigma^p - |\varepsilon|^p]|\varepsilon|^{p-2}}{(q\sigma^2+|\varepsilon|^p)^2}f(\varepsilon)d\varepsilon\right)^{-1}\frac{\varepsilon(k)|\varepsilon(k)|^{p-2}}{q\sigma^p+|\varepsilon(k)|^p}\bigg|_{x(1)=0} \tag{3.15}$$

and $\hat{x}(1|N)$ denotes the estimate of $x(1)$ at sample N.

Notice that Equation (3.14) gives the well known least-squares estimates $\hat{x}(1|N)$ from the minimization of the least-squares loss function

$$V = \frac{1}{2}\sum_{k=1}^{N}\left(z(k) - H\Phi^{k-1}\hat{x}(1|N)\right)^2$$

and the recursive version in Equations (3.17) and (3.18) with the covariance matrix

$$P(1|N) = \left(\sum_{k=1}^{N}(H\Phi^{k-1})^T H\Phi^{k-1}\right)^{-1} \tag{3.16}$$

are given in many textbooks that discuss least-squares [35]. Equations (3.5) and (3.8) are then used to obtain $\bar{x}(N)$ and $\hat{y}(N|N)$ in Equations (3.19) and (3.20) respectively.

The derivation is complete and the recursive ARMAX filter algorithm for $N = 1, 2, 3 \ldots$ is summarized below.

ARMAX filter:

$$P(1|N) = P(1|N-1) - \frac{P(1|N-1)(H\Phi^{N-1})^T H\Phi^{N-1} P(1|N-1)}{1 + H\Phi^{N-1} P(1|N-1)(H\Phi^{N-1})^T} \quad (3.17)$$

$$\hat{x}(1|N) = \hat{x}(1|N-1) + P(1|N)(H\Phi^{N-1})^T \times \left[z(N) - H\Phi^{N-1}\hat{x}(1|N-1)\right] \quad (3.18)$$

$$\bar{x}(N+1) = \Phi\bar{x}(N) + \Gamma u(N) + \Omega y(N) \quad (3.19)$$

$$\hat{y}(N|N) = H\Phi^{N-1}\hat{x}(1|N) + H\bar{x}(N) \quad (3.20)$$

The covariance of $\hat{x}(1)$ and estimate $\hat{y}(N)$ at sample N are denoted by $P(1|N)$ and $\hat{y}(N|N)$ respectively. For initialization, $P(1|0)$ can be set as an identity matrix multiplied by some large number and $x(1) = \bar{x}(1) = 0$. The derivation of the ARMAX Filter is given in Appendix B.

3.3.2 Mean, Variance and Outlier

Let the actual noise be associated with probability density function $g(\varepsilon)$ which is not necessarily equal to $f(\varepsilon)$ the noise model used in the design of the filter. The mean is then given by the following first-order von Mises expansion [36, 37]

$$\hat{x}(1|N) = \int_{-\infty}^{\infty} \mathrm{IF}(\varepsilon) g(\varepsilon) d\varepsilon \quad (3.21)$$

and the variance

$$\text{var } \hat{x}(1|N) = \int_{-\infty}^{\infty} \text{IF}^T(\varepsilon)\text{IF}(\varepsilon)g(\varepsilon)d\varepsilon \tag{3.22}$$

The assumption that $g(\varepsilon)$ is the same for all k is commonly made. Here we extend to the case where $g(\varepsilon)$ could be different for different sample k denoted by $g_k(\varepsilon)$. The case where $g(\varepsilon)$ could be different for different sample k is useful for the analysis of outliers (see Example 3). Hence, instead of integrating IF(ε) with $g(\varepsilon)$ in Equation (3.21), we first substitute Equation (3.10) into Equation (3.13) and then integrate IF(ε) with different $g_k(\varepsilon)$ for different k giving

$$\begin{aligned}\hat{x}(1|N) &= \left(\int_{-\infty}^{\infty} \frac{\partial\psi(\varepsilon)}{\partial\hat{x}(1)} f(\varepsilon)d\varepsilon\right)^{-1} (pq+1) \\ &\quad \times \sum_{k=1}^{N} (H\Phi^{k-1})^T \int_{-\infty}^{\infty} \frac{\varepsilon(k)|\varepsilon(k)|^{p-2}}{q\sigma^p + |\varepsilon(k)|^p} g_k(\varepsilon)d\varepsilon \Big|_{x(1)=0}\end{aligned} \tag{3.23}$$

In the next section examples will be given to illustrate the properties of the ARMAX filter such as the equivalence to the Kalman filter if we design with Gaussian noise in mind by choosing $p = 2$, $q = \infty$ (see Example 1) and the variance of the filter output (see Examples 2 and 4).

3.4 Examples

Four examples are given to illustrate the properties of the ARMAX filter and the IF analysis. For easy reference, the parameters of the ARMAX process and ARMAX filter are summarized in Table 3.1. The parameters Φ, H, Γ and Ω depend on A, B and C of the ARMAX process. In Examples 1, 2 and 3 the parameters p, q and σ are chosen according to Figure 2.3 to give $f(\varepsilon)$. In Example 4, p, q and σ are

Table 3.1: Parameters of the ARMAX process and ARMAX filter in the Examples

Example		1	2	3	4
Figure Number		–	2(a)	3	8
ARMAX Model	A	$1+aq^{-1}$	$1+aq^{-1}$, $a=-0.9$	$1+aq^{-1}$, $a=-0.6$	$1+aq^{-1}$, $a=-0.987$
	B	bq^{-1}	bq^{-1}, $b=0.1$	bq^{-1}, $b=0.4$	bq^{-1}, $b=0.037$
	C	$1+cq^{-1}$	$1+cq^{-1}$, $c=a$	$1+cq^{-1}$, $c=-0.8$	$1+cq^{-1}$, $c=a$
	$f(\varepsilon)$	$N(0,\alpha)$	$t_3(0,0.1)$	$t_3(0,0.1)$	Equation (2.2)
ARMAX Filter	Φ	$-c$	$-c$	$-c$	$-c$
	H	1	1	1	1
	Γ	b	b	b	b
	Ω	$c-a$	$c-a$	$c-a$	$c-a$
	p	2	2	2	2
	q	∞	1.5	1.5	3.433
	σ	$\alpha\sqrt{2}$	$0.1\sqrt{2}$	$0.1\sqrt{2}$	0.1636
Actual Noise	$g(\varepsilon)$	$f(\varepsilon)$	$f(\varepsilon)$	$\begin{cases}\delta(\varepsilon_1) & k=k_1=2;\varepsilon_1=-1\\ f(\varepsilon) & k\neq k_1\end{cases}$	$f(\varepsilon)$

obtained by fitting the GT-distribution of Equation (2.2) to the experimental data. The ARMAX filter is designed with $p=2$, $q=\infty$ for Gaussian noise in Example 1, $p=2$, $q=1.5$ for t_3 noise in Examples 2 and 3 and $p=2$, $q=3.433$ for the noise in Example 4. Note that the distribution $f(\varepsilon)$ used for the filter design need not be the same as $g(\varepsilon)$, the distribution of the actual noise in the last row of the table. One thousand simulation runs were conducted in Examples 2 and 3 and one hundred experimental runs were conducted in Example 4 to give the variance of the estimate. The simulation is started with $P(1|0)=1000I$ and $\hat{x}(1|0)=0$.

3.4.1 Example 1: The Kalman Filter Connection

This example shows that if the ARMAX filter is designed with Gaussian noise in mind then it is equivalent to the Kalman filter although it was formulated through maximum likelihood estimation with GT noise and IF approximation.

$$\begin{aligned}
\hat{x}(N|N) &= \hat{x}(N|N-1) + \frac{p(N|N-1)}{1+p(N|N-1)}[y(N) - \hat{x}(N|N-1)] && (3.25)\\
\hat{x}(N+1|N) &= bu(N) - ay(N) + \frac{c}{1+p(N|N-1)} \\
&\quad \times \Big[y(N) - \hat{x}(N|N-1)\Big] && (3.26)\\
p(N+1|N) &= \frac{p(N|N-1)c^2}{1+p(N|N-1)} && (3.27)\\
\hat{y}(N|N) &= \hat{x}(N|N) && (3.28)
\end{aligned}$$

The Kalman Filter

Consider the ARMAX process with Gaussian noise in Example 1 of Table 3.1. A state-space representation is given by

$$\begin{aligned}
x(k+1) &= -ax(k) + bu(k) + (c-a)\varepsilon(k) \\
y(k) &= x(k) + \varepsilon(k)
\end{aligned} \qquad (3.24)$$

The Kalman filter [35] for the above state-space model of Equation (3.24) is given in Equations (3.25) to (3.28) below.

The ARMAX Filter

The ARMAX filter is designed for the ARMAX process with Gaussian noise of standard deviation α. According to Figure 2.3, the GT parametes to model the Gaussian noise are $p = 2$, $q = \infty$, $\sigma = \alpha\sqrt{2}$ as shown in Table 3.1. Equations (3.15) and (3.8) give $z(N) = \varepsilon(N) = y(N) - \bar{x}(N)$ and Equations (3.17) to (3.20) give the ARMAX filter Equations (3.29) to (3.32) below.

$$
\begin{aligned}
p(1|N) &= \frac{p(1|N-1)}{1+p(1|N-1)c^{2(N-1)}} && (3.29)\\
\hat{x}(1|N) &= \hat{x}(1|N-1) + p(1|N)(-c)^{N-1} \\
&\quad \times\Big[y(N) - \bar{x}(N) - (-c)^{N-1}\hat{x}(1|N-1)\Big] && (3.30)\\
\bar{x}(N+1) &= -c\bar{x}(N) + bu(N) + (c-a)y(N) && (3.31)\\
\hat{y}(N|N) &= (-c)^{N-1}\hat{x}(1|N) + \bar{x}(N) && (3.32)
\end{aligned}
$$

where from Equation (3.16)

$$p(1|N) = \left(\sum_{k=1}^{N} c^{2(k-1)}\right)^{-1} \tag{3.33}$$

The Connection

To connect the ARMAX filter with the Kalman filter, we will now show that the Kalman filter Equations (3.25) to (3.28) can be obtained from the ARMAX filter Equations (3.29) to (3.32).

Multiplying Equation (3.30) by $(-c)^{N-1}$ and then add $\bar{x}(N)$ to both sides of the equation gives

$$
\begin{aligned}
(-c)^{N-1}\hat{x}(1|N) + \bar{x}(N) &= (-c)^{N-1}\hat{x}(1|N-1) + \bar{x}(N) + p(1|N)(-c)^{2(N-1)} \\
&\quad \times[y(N) - \bar{x}(N) - (-c)^{N-1}\hat{x}(1|N-1)] && (3.34)
\end{aligned}
$$

Multiplying Equation (3.30) by $(-c)^{N}$ and then add Equation (3.31) gives

$$
\begin{aligned}
(-c)^{N}\hat{x}(1|N) + \bar{x}(N+1) = &(-c)^{N}\hat{x}(1|N-1) + p(1|N)(-c)^{2N-1} \\
&\times \Big[y(N) - \bar{x}(N) - (-c)^{N-1}\hat{x}(1|N-1)\Big] \\
&- c\bar{x}(N) + bu(N) + (c-a)y(N) && (3.35)
\end{aligned}
$$

Note that from Equation (3.7), $x(N) = (-c)^{N-1}x(1) + \bar{x}(N)$ and so Equations (3.32), (3.34) and (3.35) can be written as

$$\hat{y}(N|N) = \hat{x}(N|N) \tag{3.36}$$

$$\hat{x}(N|N) = \hat{x}(N|N-1) + p(1|N)c^{2(N-1)}[y(N) - \hat{x}(N|N-1)] \tag{3.37}$$

$$\begin{aligned}\hat{x}(N+1|N) &= bu(N) - ay(N) + [c + p(1|N)(-c)^{2N-1}] \\ &\quad \times [y(N) - \hat{x}(N|N-1)]\end{aligned} \tag{3.38}$$

Substitute $\hat{x}(1|N) = \frac{\hat{x}(N+1|N) - \bar{x}(N+1)}{(-c)^N}$ from Equation (3.7) into Equation (3.14) to give

$$\hat{x}(N+1|N) - \bar{x}(N+1) = \left(\frac{1}{c^{2N}}\sum_{k=1}^{N} c^{2(k-1)}\right)^{-1} \left(\frac{1}{(-c)^N}\sum_{k=1}^{N}(-c)^{k-1}z(k)\right)\Bigg|_{x(1)=0} \tag{3.39}$$

and corresponding to Equation (3.16) the covariance matrix

$$P(N+1|N) = \left(\frac{1}{c^{2N}}\sum_{k=1}^{N} c^{2(k-1)}\right)^{-1} \tag{3.40}$$

Using Equation (3.33), Equation (3.40) becomes

$$p(1|N) = \frac{p(N+1|N)}{c^{2N}} \tag{3.41}$$

Substituting Equation (3.41) into Equations (3.37), (3.38), (3.29) and (3.36) gives the Kalman filter Equations (3.25) to (3.28) respectively. For simplicity, we have used the first-order ARMAX process as an example. It can be shown that in general, the ARMAX filter is equivalent to the Kalman filter if the GT parameters p and q in the ARMAX filter design are chosen as 2 and ∞ respectively to model Gaussian noise.

3.4.2 Example 2: Variance

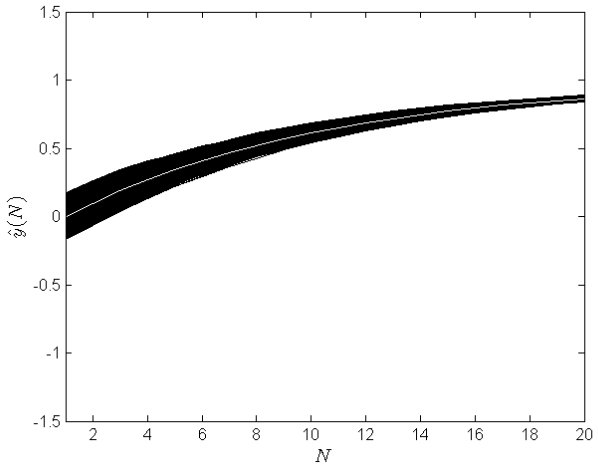
In this example, the ARMAX filter is designed for the ARMAX process with t_3 noise. According to Figure 2.3, to model the t_3 noise, the GT parameters are $p = 2$, $q = 1.5$ and $\sigma = 0.1\sqrt{2}$ as shown in Table 3.1.

Simulation

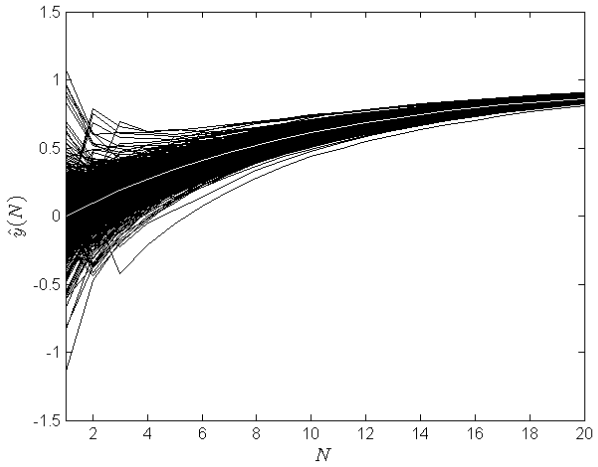
We conducted 1000 simulation runs using the ARMAX filter (Equations 3.17 to 3.20) and Kalman filter (Equations 3.25 to 3.28). The results are shown in Figure 3.1 where the mean value for the 1000 runs is given by the white curve. The mean and variance at $N = 1, 5, \ldots, 20$ are tabulated in Table 3.2 under the Column "Eq. (3.17) to (3.20)" and "Eq. (3.25) to (3.28)". The result of solving Equation (3.11) numerically for $\hat{x}(1|N)$ and then $\hat{y}(N|N)$ from Equation (3.20) is also given under the Column "Eq. (3.11) and (3.20)". Figure 3.1 and Table 3.2 show clearly that the variance from the Kalman filter is larger than the ARMAX filter. The Kalman filter assumes Gaussian and not t_3 noise. This example shows that the GT parameters in the ARMAX filter can be chosen gainfully to give smaller variance.

Table 3.2: Mean and Variance of $\hat{y}(N)$ in Figure 3.1

	ARMAX Filter					Kalman Filter		
	Eq. (3.11) and (3.20) Numerical Solution		Eq. (3.17) to (3.20) Recursive Solution		Eq. (3.44)	Eq. (3.25) to (3.28) Recursive Solution		Eq. (3.48)
N	Mean	Var (10^{-2})	Mean	Var (10^{-2})	Var (10^{-2})	Mean	Var (10^{-2})	Var (10^{-2})
1	0.00	3.062	0.00	1.460	1.500	0.00	3.062	3.000
5	0.34	0.268	0.34	0.190	0.188	0.34	0.399	0.377
10	0.61	0.051	0.61	0.049	0.049	0.61	0.102	0.097
15	0.77	0.016	0.77	0.015	0.016	0.77	0.032	0.031
20	0.86	0.006	0.86	0.005	0.005	0.86	0.011	0.011



(a) ARMAX filter output $\hat{y}(N)$.



(b) Kalman filter output $\hat{y}(N)$.

Figure 3.1: Simulation results of Example 2.

IF Analysis: ARMAX Filter

The IF can be used to derive an equation to calculate the variance in Table 3.1. Using Equation (3.14)

$$\mathrm{IF}(\varepsilon) = \left(\sum_{k=1}^{N} c^{2(k-1)}\right)^{-1} \left(\int_{-\infty}^{+\infty} \frac{0.03-\varepsilon^2}{(0.03+\varepsilon^2)^2} f(\varepsilon) d\varepsilon\right)^{-1} \left(\sum_{k=1}^{N} \frac{(-c)^{k-1}\varepsilon(k)}{0.03+\varepsilon(k)^2}\right)$$

Using Equation (3.22)

$$\begin{aligned}\text{var } \hat{x}(1|N) &= \left(\sum_{k=1}^{N} c^{2(k-1)}\right)^{-1} \left(\int_{-\infty}^{+\infty} \frac{0.03-\varepsilon^2}{(0.03+\varepsilon^2)^2} f(\varepsilon)d\varepsilon\right)^{-2} \\ &\quad \times \left(\int_{-\infty}^{+\infty} \frac{\varepsilon^2}{(0.03+\varepsilon^2)^2} g(\varepsilon)d\varepsilon\right) \\ &= \frac{3}{200}\frac{1-c^2}{1-c^{2N}} \end{aligned} \tag{3.42}$$

where $\int_{-\infty}^{+\infty} \frac{0.03-\varepsilon^2}{(0.03+\varepsilon^2)^2} f(\varepsilon)d\varepsilon = \frac{50}{3}$, $\int_{-\infty}^{+\infty} \frac{\varepsilon^2}{(0.03+\varepsilon^2)^2} g(\varepsilon)d\varepsilon = \frac{25}{6}$ and since ε is assumed to be a zero mean independent random variable, $\int_{-\infty}^{+\infty} \varepsilon(j)\varepsilon(k)/((0.03+\varepsilon(j)^2)(0.03+\varepsilon(k)^2))g(\varepsilon)d\varepsilon = 0$ for $j \neq k$. From Equation (3.14) $\hat{x}(1|N)$ is zero-mean as $\varepsilon(k)$ is zero-mean and since $\bar{x}(N)$ is not a function of the random variable ε, Equation (3.20) gives

$$\text{var } \hat{y}(N|N) = c^{2(N-1)}\text{var } \hat{x}(1|N) \tag{3.43}$$

Substituting Equation (3.42) into Equation (3.43) gives

$$\text{var } \hat{y}(N|N) = \frac{3}{200}\frac{c^{2(N-1)}-c^{2N}}{1-c^{2N}} \tag{3.44}$$

Equation (3.44) is used to calculate the variance in the Column "Eq. (3.44)" and Table 3.2 shows that it is close to the values obtained from simulation in Column "Eq. (3.17) to (3.20)". The table also show that for $N \geq 10$, the variances in the Columns "Eq. (3.44)" and "Eq. (3.17) to (3.20)" from IF approximation are close to the variance in Column "Eq. (3.11) and (3.20)" obtained by solving Equation (3.11) numerically without approximation.

IF Analysis: Kalman Filter

Although the Kalman filter assumes Gaussian noise, the IF can also be used to derive an equation to calculate the variance of the Kalman filter when the actual noise is t_3. Example 1 shows that the ARMAX filter designed with $p = 2$, $q = \infty$ is equivalent to the Kalman filter and so from Equation (3.14)

$$\text{IF}(\varepsilon) = \left(\sum_{k=1}^{N} c^{2(k-1)}\right)^{-1} \left(\sum_{k=1}^{N} (-c)^{k-1}\varepsilon(k)\right) \tag{3.45}$$

Using Equation (3.22)

$$\text{var } \hat{x}(1|N) = \left(\sum_{k=1}^{N} c^{2(k-1)}\right)^{-1} \left(\int_{-\infty}^{\infty} \varepsilon^2 g(\varepsilon) d\varepsilon\right) \tag{3.46}$$

where $\int_{-\infty}^{\infty} \varepsilon(j)\varepsilon(k)g(\varepsilon)d\varepsilon = 0$ for $j \neq k$ and $g(\varepsilon)$ is the GT probability density function of Equation (2.2) with $p = 2$, $q = 1.5$ and $\sigma = 0.1\sqrt{2}$ to model the actual t_3 noise. This gives

$$\text{var } \hat{x}(1|N) = \frac{3}{100}\frac{1-c^2}{1-c^{2N}} \tag{3.47}$$

since $\int_{-\infty}^{\infty} \varepsilon^2 g(\varepsilon)d\varepsilon = \frac{3}{100}$. Using Equation (3.4.2)

$$\text{var } \hat{y}(N|N) = c^{2(N-1)}\text{var } \hat{x}(1|N) = \frac{3}{100}\frac{c^{2(N-1)} - c^{2N}}{1-c^{2N}} \tag{3.48}$$

Equation (3.48) is used to calculate the variance in the Column "Eq. (3.48)" of Table 3.2. It is clear that it matched the variance from the simulation in the Column "Eq. (3.25) to (3.28)".

3.4.3 Example 3: Outlier

This example shows how the IF can be used to calculate the ARMAX and Kalman filter output in the presence of an outlier.

Simulation

Consider the ARMAX process in Example 3 of Table 3.1. It has an outlier of $\varepsilon_1 = -1$ at $k = k_1 = 2$. Here $g(\varepsilon)$ will be different at each sample k and is given as

$$g_k(\varepsilon) = \begin{cases} \delta(\varepsilon_1), & k = k_1 \\ f(\varepsilon) & k \neq k_1 \end{cases} \tag{3.49}$$

where $\delta(\varepsilon_1)$ is an impulse at ε_1 to model the outlier of ε_1 at the sample $k = k_1$. The ARMAX filter is design with $p = 2$, $q = 1.5$ and $\sigma = 0.1\sqrt{2}$ to model the t_3 noise. Unlike Example 2, here $f(\varepsilon) \neq g(\varepsilon)$. The output $\hat{y}(N)$ of the ARMAX filter (Equations 3.17 to 3.20) and Kalman filter (Equations 3.25 to 3.28) are shown in Figures 3.2 and 3.3 respectively. The mean value of the 1000 runs is given by the white curve. It is clear that the Kalman filter output is greatly affected by the outlier at $k = 2$ unlike the ARMAX filter. It is known that a single outlier can spoil the statistical analysis completely for the case of least-squares estimation [2] including the Kalman filter [39].

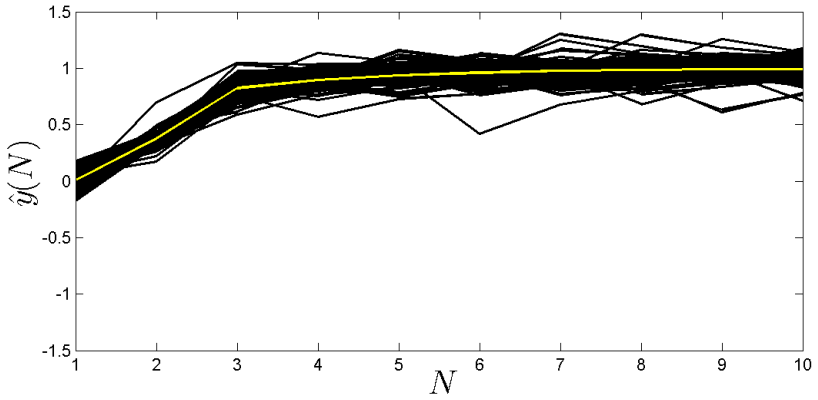


Figure 3.2: ARMAX filter output $\hat{y}(N)$.

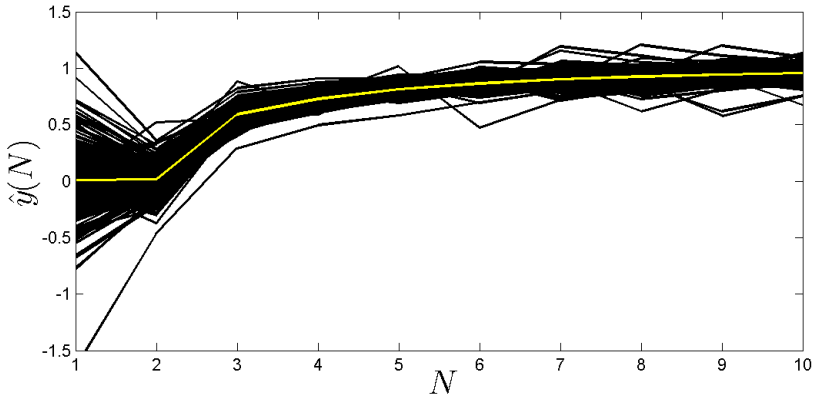


Figure 3.3: Kalman filter output $\hat{y}(N)$.

IF Analysis

Equation (3.23) can be used to draw the white or mean curves in Figures 3.2 and 3.3. With $g_k(\varepsilon)$ of Equation (3.49) and $p = 2$, Equation (3.23) gives

$$\begin{aligned}\hat{x}(1|N) &= \left(\sum_{k=1}^{N} c^{2(k-1)}\right)^{-1} \left(\int_{-\infty}^{+\infty} \frac{q\sigma^2 - \varepsilon^2}{(q\sigma^2+\varepsilon^2)^2} f(\varepsilon)d\varepsilon\right)^{-1} \\ &\quad \times \left[\int_{-\infty}^{+\infty} \frac{(-c)^{k_1-1}\varepsilon(k_1)}{q\sigma^2+\varepsilon(k_1)^2}\delta(\varepsilon_1)d\varepsilon \right. \\ &\quad \left. + \sum_{k=1,\neq k_1}^{N} \int_{-\infty}^{+\infty} \frac{(-c)^{k-1}\varepsilon(k)}{q\sigma^2+\varepsilon(k)^2} f(\varepsilon)d\varepsilon\right], \quad N \geq k_1 \end{aligned} \tag{3.50}$$

Note that the second term in the square bracket of Equation (3.50) is zero because the expectation of $\varepsilon(k)$ for $k \neq k_1$ is zero. Simplifying gives

$$\hat{x}(1|N) = \frac{(-c)^{k_1-1}(1-c^2)}{(1-c^{2N})} \left(\frac{\varepsilon_1}{q\sigma^2+\varepsilon_1^2}\right) \left(\int_{-\infty}^{+\infty} \frac{q\sigma^2-\varepsilon^2}{(q\sigma^2+\varepsilon^2)^2} f(\varepsilon)d\varepsilon\right)^{-1}, \quad N \geq k_1 \tag{3.51}$$

From Equation (3.20),

$$\begin{aligned}\hat{y}(N|N) &= \frac{(-c)^{N+k_1-2}(1-c^2)}{(1-c^{2N})} \left(\frac{\varepsilon_1}{q\sigma^2+\varepsilon_1^2}\right) \\ &\quad \times \left(\int_{-\infty}^{+\infty} \frac{q\sigma^2-\varepsilon^2}{(q\sigma^2+\varepsilon^2)^2} f(\varepsilon)d\varepsilon\right)^{-1} + \bar{x}(N), \quad N \geq k_1 \end{aligned} \tag{3.52}$$

Note that according to Figure 2.3, the t_3 and Gaussian distribution are modeled by setting $q = 1.5$ and $q = \infty$ respectively. So the white curve in Figure 3.2 is obtained by substituting $q = 1.5$ and $\sigma = 0.1\sqrt{2}$ into Equation (3.52) to give

$$\hat{y}(N|N) = \frac{(-c)^{N+k_1-2}(1-c^2)}{300(1-c^{2N})}\left(\frac{\varepsilon_1}{0.03+\varepsilon_1^2}\right) + \bar{x}(N), \quad N \geq k_1$$

where $\int_{-\infty}^{+\infty} \frac{q\sigma^2-\varepsilon^2}{(q\sigma^2+\varepsilon^2)^2} f(\varepsilon)d\varepsilon = 300$. The white curve in Figure 3.3 is obtained by substituting $q = \infty$ into Equation (3.52) to give

$$\hat{y}(N|N) = \frac{(-c)^{N+k_1-2}(1-c^2)}{1-c^{2N}}\varepsilon_1 + \bar{x}(N), \quad N \geq k_1$$

3.4.4 Example 4: Liquid Level Estimation Experiment

Consider the liquid-level estimation problem commonly encountered in chemical processes in the coupled tank of Figure 3.4. The transfer function between the liquid level in Tank 1, $y(k)$, and the control voltage, $u(k)$, at sampling interval of 1 second is given as

$$y(k) = \frac{0.037q^{-1}}{1-0.987q^{-1}}u(k) + \varepsilon(k) \tag{3.53}$$

The polynomials A, B and C in the ARMAX model can be obtained by comparing Equations (3.53) and (3.1) and is given in Table 3.1 under Column "Example 4".

Experiment

One thousand measurements of the liquid level $y(k)$ were collected as shown in Figure 3.5 when the control voltage $u(k)$ was held constant at $2V$. The histogram of the measurements $y(k)$ after subtracting the mean are plotted in Figure 3.6 and is considered as the noise, $\varepsilon(k)$, distribution.

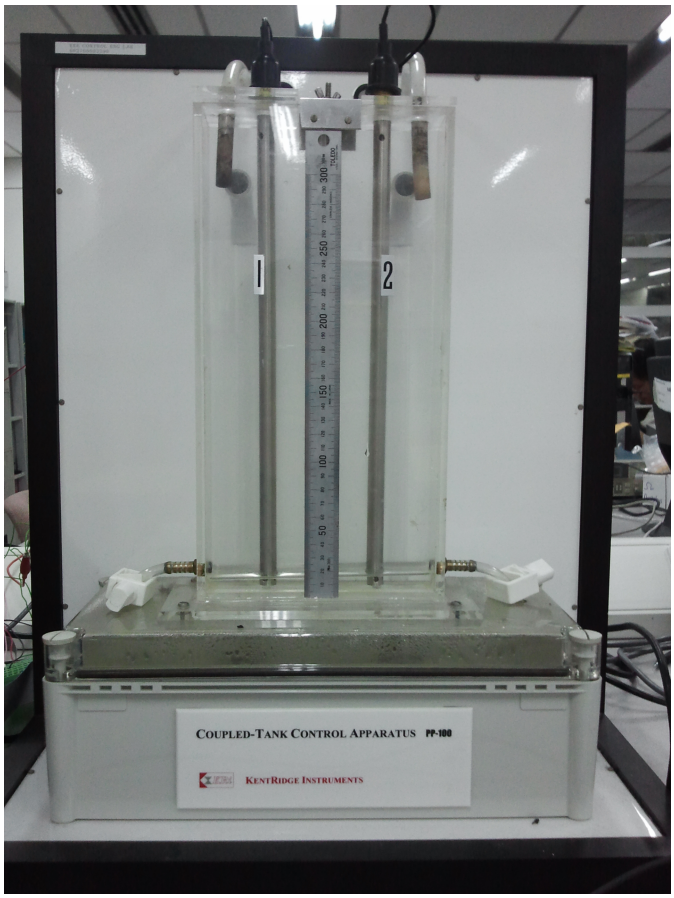


Figure 3.4: Photo of the coupled-tank

The maximum likelihood criterion can be used to find the parameters of the GT probability density function. In this example we fixed $p = 2$ and then use the maximum likelihood criterion to find the other parameters q and σ of the GT probability density function $f(\varepsilon)$ of Equation (2.2) by maximizing the objective function

$$J_f = \sum_{k=1}^{1000} \ln \frac{p}{2\sigma q^{1/p} \beta(1/p, q) \left(1 + \frac{|\varepsilon(k)|^p}{q\sigma^p}\right)^{q+1/p}}$$

This gives $q = 3.433$, $\sigma = 0.1636$ and the resultant GT distribution is superimposed on the distribution in Figure 3.6. Using maximum likelihood, a Gaussian distribution was also fitted to the histogram. It is evident in Figure 3.6 that the GT

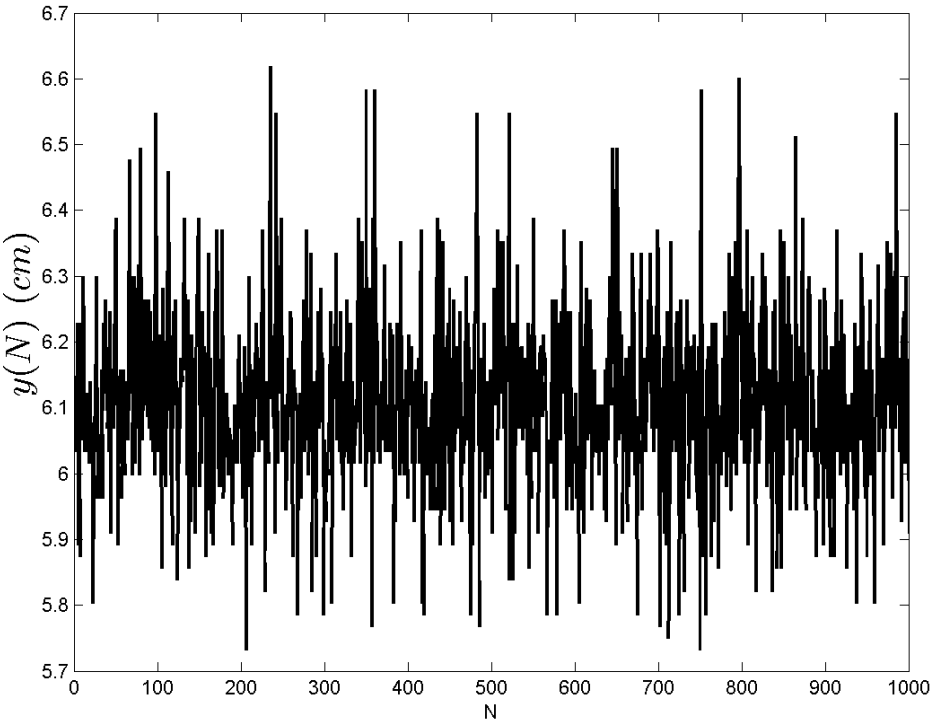


Figure 3.5: Measurement $y(N)$ for the liquid level estimation experiment.

distribution gives a better fit. In this example, p was not obtained by maximizing the objective function J_f but simply chosen as 2 gives an indication that the results do not depend critically on the value of p. The papers [14, 21] give further detailed discussion on the choice and determination of p, q and σ.

With the control voltage $u(k) = 2\text{V}$, we estimated the liquid level $y(k)$ for 10 samples using the ARMAX filter (Equations 3.17 to 3.20) and Kalman filter (Equations 3.25 to 3.28). This was repeated 100 times. The results are shown in Figures 3.7 and 3.8 and the variances are tabulated in Table 3.3 in the rows labeled as "Experimental Value". Figures 3.7 and 3.8 and Table 3.3 show that the variance from the ARMAX filter is about 10% smaller than the Kalman filter. This example

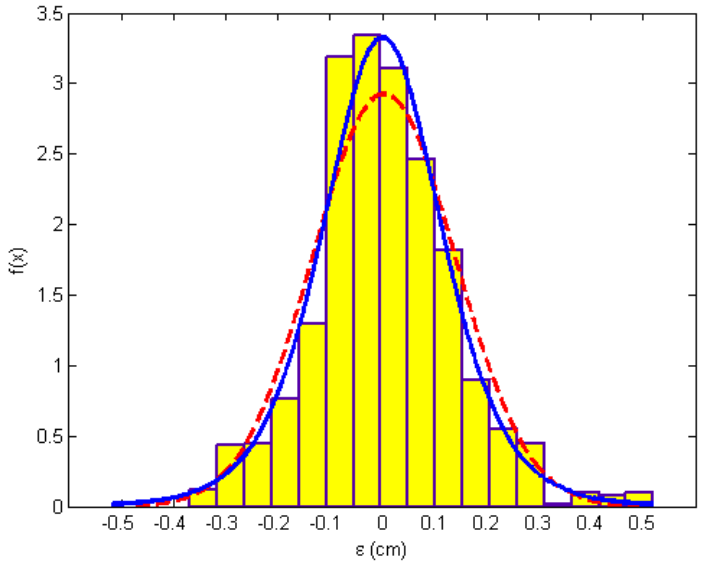


Figure 3.6: The maximum likelihood criterion was used to fit a GT distribution (solid-line) and Gaussian distribution (dashed-line) to the noise distribution

shows that the GT parameters in the ARMAX filter can be chosen gainfully to give smaller variance.

Table 3.3: Variance ($\times 10^{-3}$) of $\hat{y}(N)$ in Figures 3.7 and 3.8.

	N	1	2	3	4	5	6	7	8	9	10
ARMAX Filter	Experimental Value	17.7	8.2	4.8	4.1	3.1	2.5	2.2	1.8	1.7	1.5
(Figure 3.7)	Equation (3.56)	16.8	8.3	5.5	4.0	3.2	2.6	2.2	1.9	1.7	1.5
Kalman Filter	Experimental Value	20.0	9.6	5.8	4.9	3.6	3.1	2.6	2.2	1.9	1.7
(Figure 3.8)	Eqn (3.60)	18.7	9.2	6.1	4.5	3.6	2.9	2.5	2.1	1.9	1.7

IF analysis for ARMAX Filter

The IF can be used to derive an equation to calculate the variance in Table 3.3. Using Equation (3.14)

$$\mathrm{IF}(\varepsilon) = \left(\sum_{k=1}^{N} c^{2(k-1)}\right)^{-1} \left(\int_{-\infty}^{+\infty} \frac{0.0919 - \varepsilon^2}{(0.0919 + \varepsilon^2)^2} f(\varepsilon) d\varepsilon\right)^{-1} \left(\sum_{k=1}^{N} \frac{(-c)^{k-1} \varepsilon(k)}{0.0919 + \varepsilon(k)^2}\right)$$

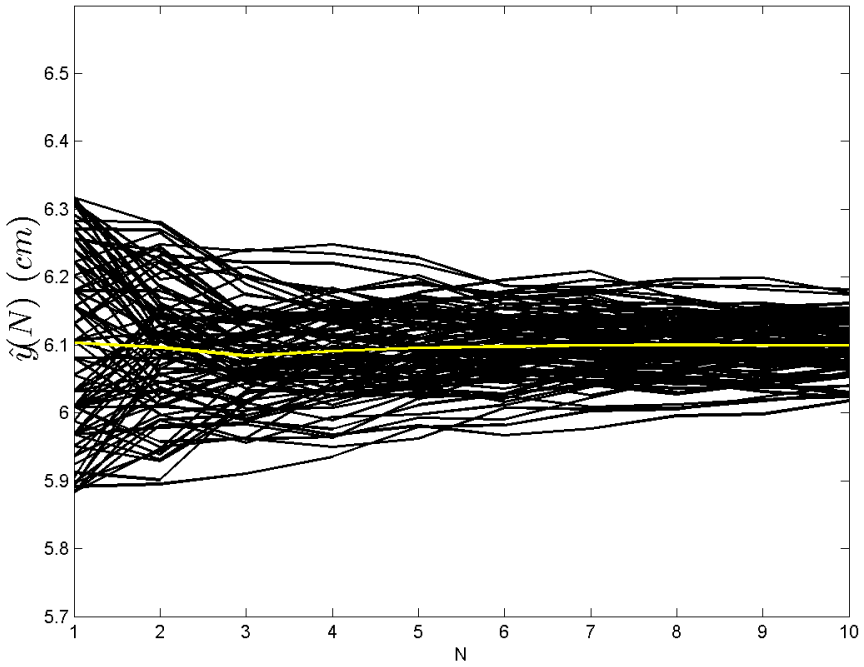


Figure 3.7: ARMAX filter estimate $\hat{y}(N)$.

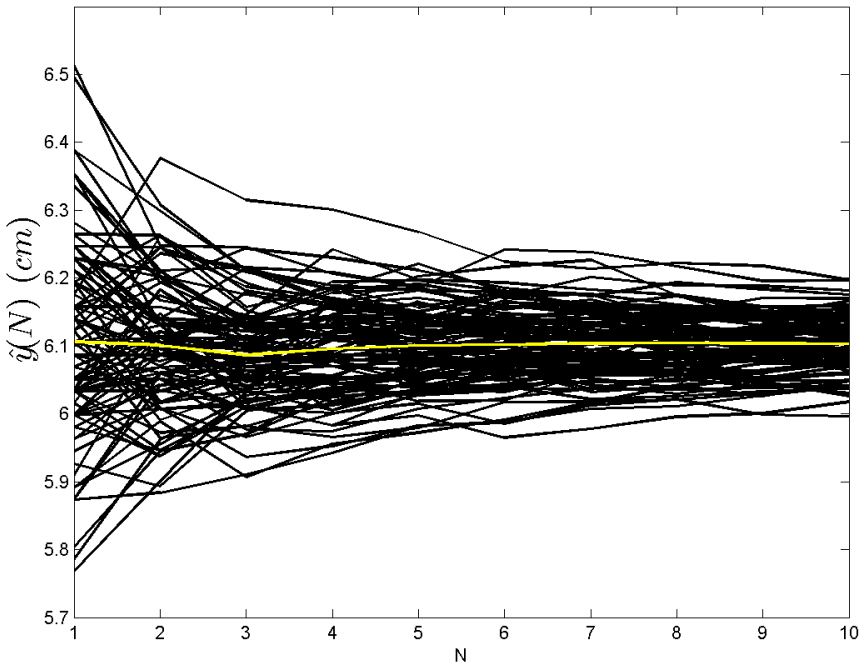


Figure 3.8: Kalman filter estimate $\hat{y}(N)$.

Using Equation (3.22)

$$\begin{aligned}\text{var } \hat{x}(1|N) &= \left(\sum_{k=1}^{N} c^{2(k-1)}\right)^{-1} \left(\int_{-\infty}^{+\infty} \frac{0.0919-\varepsilon^2}{(0.0919+\varepsilon^2)^2} f(\varepsilon)d\varepsilon\right)^{-2} \\ &\quad \times \left(\int_{-\infty}^{+\infty} \frac{\varepsilon^2}{(0.0919+\varepsilon^2)^2} g(\varepsilon)d\varepsilon\right) \\ &= 0.0168\frac{1-c^2}{1-c^{2N}} \end{aligned} \tag{3.54}$$

where $\int_{-\infty}^{+\infty} \frac{0.0919-\varepsilon^2}{(0.0919+\varepsilon^2)^2} f(\varepsilon)d\varepsilon = 7.57$, $\int_{-\infty}^{+\infty} \frac{\varepsilon^2}{(0.0919+\varepsilon^2)^2} g(\varepsilon)d\varepsilon = 0.963$ and since ε is assumed to be a zero mean independent random variable, $\int_{-\infty}^{+\infty} \varepsilon(j)\varepsilon(k)/((0.0919+\varepsilon(j)^2)(0.0919+\varepsilon(k)^2))g(\varepsilon)d\varepsilon = 0$ for $j \neq k$. From Equation (3.14) $\hat{x}(1|N)$ is zero-mean as $\varepsilon(k)$ is zero-mean and since $\bar{x}(N)$ is not a function of the random variable ε, Equation (3.20) gives

$$\text{var } \hat{y}(N|N) = c^{2(N-1)}\text{var } \hat{x}(1|N) \tag{3.55}$$

Substituting Equation (3.54) into Equation (3.55) gives

$$\text{var } \hat{y}(N|N) = 0.0168\frac{c^{2(N-1)} - c^{2N}}{1-c^{2N}} \tag{3.56}$$

Equation (3.56) is used to calculate the variance in Row "Equation (3.56)" of Table 3.3. It is close to the experimental values given in the row just above.

IF analysis for Kalman Filter

Although the Kalman filter assumes Gaussian noise, the IF can also be used to derive an equation to calculate the variance of the Kalman filter when the actual noise is not Gaussian. Example 1 shows that the ARMAX filter designed with $p = 2$, $q = \infty$ is equivalent to the Kalman filter and so from Equation (3.14)

$$\text{IF}(\varepsilon) = \left(\sum_{k=1}^{N} c^{2(k-1)}\right)^{-1} \left(\sum_{k=1}^{N} (-c)^{k-1}\varepsilon(k)\right) \tag{3.57}$$

Using Equation (3.22)

$$\text{var } \hat{x}(1|N) = \left(\sum_{k=1}^{N} c^{2(k-1)}\right)^{-1} \left(\int_{-\infty}^{\infty} \varepsilon^2 g(\varepsilon) d\varepsilon\right) \tag{3.58}$$

where $\int_{-\infty}^{\infty} \varepsilon(j)\varepsilon(k)g(\varepsilon)d\varepsilon = 0$ for $j \neq k$ and $g(\varepsilon)$ is the GT probability density function of Equation (2.2) with $p = 2$, $q = 3.433$ and $\sigma = 0.1636$ to model the noise. This gives

$$\text{var } \hat{x}(1|N) = 0.0189\frac{1 - c^2}{1 - c^{2N}} \tag{3.59}$$

since $\int_{-\infty}^{\infty} \varepsilon^2 g(\varepsilon)d\varepsilon = 0.0189$. Using Equation (3.55)

$$\text{var } \hat{y}(k|N) = c^{2(N-1)}\text{var } \hat{x}(1|N) = 0.0189\frac{c^{2(N-1)} - c^{2N}}{1 - c^{2N}} \tag{3.60}$$

Equation (3.60) is used to calculate the variance in Row "Equation (3.60)" of Table 3.3. It is close to the experimental values given in the row just above.

The values in Table 3.3 show that the variance of the estimate from the ARMAX filter is about 10% smaller than the one from Kalman filter. If the noise is non-Gaussian and can be modeled by the GT distribution then the ARMAX filter with GT noise model can produce more a accurate estimate of the process output $y(k)$ because of the more accurate noise model. The ARMAX filter can be used gainfully in control systems. For example in adaptive control, the process output $y(k)$ is fed back to the adaptive controller which make use of the information to adapt itself to meet performance criteria. The more accurate estimate of $y(k)$ from the ARMAX filter can then be used gainfully by the same adaptive controller to enhanced performance.

3.5 Conclusion

The IF is employed to give an approximate solution to the maximum likelihood estimation problem in the ARMAX filter. The solution is recursive making it suitable for on-line and real-time implementation. We also used the IF to analyze the output of the filter designed with the GT noise model instead of the usual Gaussian noise model. Equations derived are useful in determining the variance of the estimates and the impact of outliers. If the noise is modeled by the Gaussian distribution then the proposed filter reduces to the Kalman filter. Otherwise the GT distribution has the extra degree of freedom to model non-Gaussian noise. If we do not know the distribution of the noise then one can use the Kalman filter but if there is information then it can be used gainfully in the GT distribution framework to take into account the non-Gaussian noise.

Chapter 4

Data Reconciliation for Estimation of Film Thickness

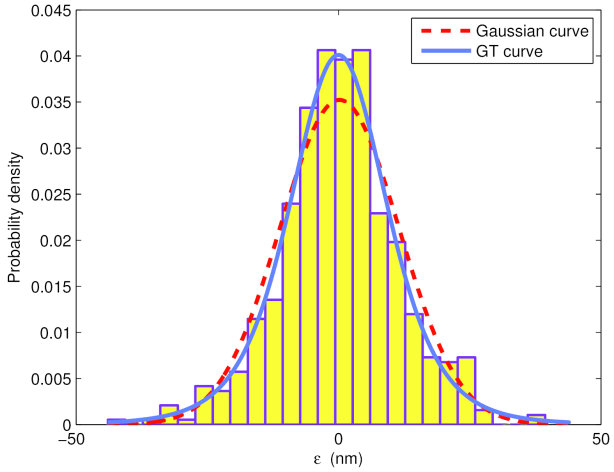
4.1 Introduction

Chemical-Mechanical Polishing is common and critical in semiconductor manufacturing. Surfaces uniformities is an important issue with stringent specifications and significant impact on the structure of integrated circuits. To obtain uniform and planar surface, chemical-mechanical polishing is applied to flatten the surface [15, 44]. Run-to-run control for the chemical-mechanical polishing process is considered in [45]. If the oxide layer has not been sufficiently thinned or the desired degree of planarity has not been obtained during chemical-mechanical polishing then the wafer may have to be reworked or even scrapped. Hence the measurement of the thickness must be precise and variance is a common measure of precision, the smaller the variance the more precise it is.

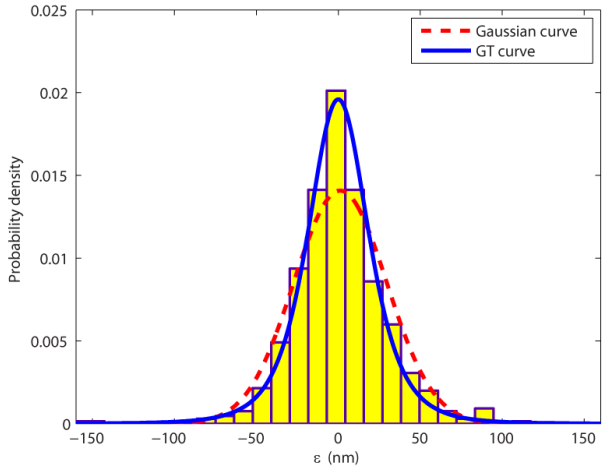
The measurement, however, is usually corrupted by noise and the noise needs not to be distributed from Gaussian distribution. Take the example of chemical-mechanical polishing of semiconductor wafers. The data distribution of 1,152 film thickness measurements after chemical-mechanical polishing of twenty-four 200mm blanket oxide semiconductor wafers and after subtracting the mean are plotted in Figure 4.1. Using the maximum likelihood method, a Gaussian distribution was fitted to the data but is evident in Figure 4.1 that the Gaussian curve does not give a good fit. The averages are 353 nm and 382 nm for the inner and outer zone respectively and that the distributions are different is evident in Figure 4.1. The measurement points for the inner (center) and outer (edge) zones are shown in Figure 4.2.

There are three main proposals in this chapter. The first proposal is to take into account the measurements of both zones irrespective of whether we estimate the thickness of the inner or outer zone. In process engineering, this technique is known as data reconciliation [46–49]. The dispersion of the measurements would relate to how well the measurement were made. Their average would provide an estimate of the thickness that generally would be more reliable when more data points were averaged. In this chapter, equations are derived to take into account the measurements of both zones. Since more data points (two zones instead of one) were taken into consideration, the variance of the estimate is expected to decrease.

The second proposal is to model the measurement distribution with the Generalized t-distribution (GT) instead of the usual Gaussian distribution. We will show that with proper choice of parameters, the GT distribution reduces to the



(a)



(b)

Figure 4.1: GT and Gaussian approximation of the experimental data distribution for (a) inner zone and (b) outer zone.

Gaussian distribution and the proposed estimation scheme reduces to the well-known least-squares estimation which is basically simple averaging if only one zone is considered. Hence within the more general framework of the estimation with GT distribution model is the least-squares estimation if the distribution is Gaussian. If the measurement is not Gaussian then the GT distribution has the extra degree of freedom to model the measurement. GT distribution has previously been employed

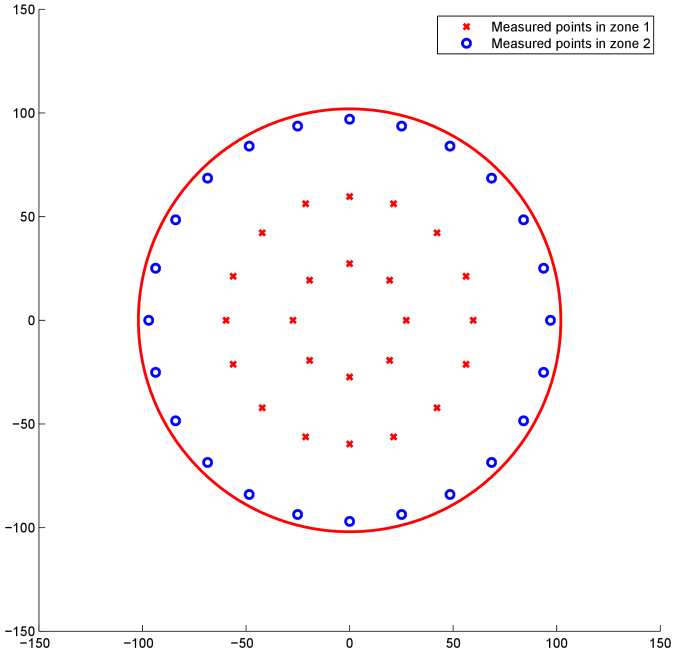


Figure 4.2: Measurement points on the inner zone (crosses) and outer zone (circles) of a wafer.

in econometrics [20] to model random residuals in regression parameter estimation. By being a distribution superset encompassing Gaussian, uniform, T and double exponential distributions [17], GT distribution has the flexibility to characterize data with non-Gaussian statistical properties. Data reconciliation, i.e. equations to take into account both zones are also derived for the GT model. Three zones were considered in [16] but the Gaussian assumption and least-squares estimation were used to estimate film thickness.

The third proposal is to make use of the Influence Function (IF) [2] to relate variance of the estimates to sample size. These equations enable us to compute the sample size needed by the estimator to achieve a desired variance. It would be shown that using GT, instead of Gaussian distribution, to characterize measurement data gives rise to estimate with a smaller variance. The theoretical results were verified

experimentally. Figure 4.1 provides an intuitive explanation: the flexibility of GT gives a distribution curve that is a better description of the experimental data represented by the histogram.

4.2 Maximum Likelihood Estimation of Film Thickness

Assume that the thickness measurement can be divided into n zones. The measurement vector $y(k)$ can be related to the thickness vector x- by

$$\begin{aligned} y(k) &= x + \varepsilon(k) \\ s.t. \qquad Ax &= 0 \end{aligned} \tag{4.1}$$

where y, x and ε are all $n \times 1$ vectors whose elements are y_i, x_i, ε_i, $i = 1, \ldots, n$. The sampling instance is given by $k = 1, \ldots, N$ and A is a $m \times n$ matrix.

4.2.1 GT Distribution

Consider the i^{th} zone. Let the noise ε_i be modelled by the GT distribution (2.2)

$$f(\varepsilon_i) = \frac{p_i}{2\sigma^{1/p_i}\beta(1/p_i, q_i)\Big(1 + \frac{|\varepsilon_i|^p}{q_i\sigma_i^p}\Big)^{q_i+1/p_i}} \tag{4.2}$$

In the framework of GT, the maximum-likelihood method is used to estimate

the optimal thickness value which minimizes

$$J = -\sum_{i=1}^{n}\sum_{k=1}^{N} \ln f(\varepsilon_i(k)) \tag{4.3}$$

$$s.t. \qquad Ax = 0 \tag{4.4}$$

$$\varepsilon_i(k) = y_i(k) - x_i, \qquad i = 1, \ldots, n \tag{4.5}$$

with respect to x. The solution to the optimization in Equations (4.3) and (4.4) can be obtained using Lagrange multipliers. The objective function for the optimization problem is then

$$L = -\sum_{i=1}^{n}\sum_{k=1}^{N} \ln f(\varepsilon_i(k)) - \lambda^T Ax \tag{4.6}$$

where λ is the vector of Lagrange multipliers. To minimize J, equate all relevant partial derivatives of L in Equation (4.6) to zero

$$\frac{\partial L}{\partial x_i} = \psi(\varepsilon_i) + \lambda^T a_i = 0, \qquad i = 1, ..., n \tag{4.7}$$

$$\frac{\partial L}{\partial \lambda} = Ax = 0 \tag{4.8}$$

where a_i is the i^{th} column vector of A and the term $\psi(\varepsilon_i)$ is given as

$$\psi(\varepsilon_i) = \sum_{k=1}^{N} \frac{\partial}{\partial x_i} \ln f(\varepsilon_i(k)) = \sum_{k=1}^{N} \frac{(p_i q_i + 1)\mathrm{sgn}(\varepsilon_i(k))|\varepsilon_i(k)|^{p_i - 1}}{q_i \sigma_i^{p_i} + |\varepsilon_i(k)|^{p_i}} \tag{4.9}$$

The solution for x in Equations (4.7) and (4.8) can be determined numerically [**?**] and the covariance matrix can be approximated from the Influence Function (IF) [2]:

$$\begin{aligned} \mathrm{Cov}\ x &\approx \mathrm{E}\left\{\mathrm{IF}(\varepsilon) \times \mathrm{IF}^T(\varepsilon)\right\} \\ &\approx [\Gamma - \Gamma A^T (A\Gamma A)^{-1} A\Gamma]\mathrm{E}\{\psi(\varepsilon)\psi^T(\varepsilon)\}[\Gamma - \Gamma A^T (A\Gamma A)^{-1} A\Gamma]^T \end{aligned} \tag{4.10}$$

where

$$\begin{aligned}
\mathrm{IF}(\varepsilon) &= (\Gamma - \Gamma A^T (A\Gamma A^T)^{-1} A\Gamma)\psi(\varepsilon) && (4.11)\\
\Gamma^{-1} &= \mathrm{diag}\left\{ -\int_{-\infty}^{+\infty} \frac{d}{dx_1}\psi(\varepsilon_1) f(\varepsilon_1)\mathrm{d}\varepsilon_1,\ \ldots\ , \right.\\
&\qquad \left. -\int_{-\infty}^{+\infty} \frac{d}{dx_n}\psi(\varepsilon_n) f(\varepsilon_n)\mathrm{d}\varepsilon_n \right\} && (4.12)\\
\frac{d\psi(\varepsilon_i)}{dx_i} &= -\sum_{k=1}^{N} \frac{(p_i q_i + 1)\left((p_i - 1) q_i \sigma_i^{p_i} - |\varepsilon_i(k)|^{p_i}\right)|\varepsilon_i(k)|^{p_i - 2}}{(q_i \sigma_i^{p_i} + |\varepsilon_i(k)|^{p_i})^2}\\
\psi(\varepsilon) &= \left[\ \psi(\varepsilon_1),\ \ \ldots,\ \ \psi(\varepsilon_n)\ \right]^T\\
\mathrm{E}\{\psi(\varepsilon)\psi^T(\varepsilon)\} &= \mathrm{diag}\left\{\ \int_{-\infty}^{+\infty}\psi^2(\varepsilon_1) f(\varepsilon_1) d\varepsilon_1,\ \ \ldots,\ \ \int_{-\infty}^{+\infty}\psi^2(\varepsilon_n) f(\varepsilon_n) d\varepsilon_n\ \right\}
\end{aligned}$$

as the off-diagonal elements are zeros because for $i \neq j$, ε_i and ε_j are independent. Expressions (4.10) and (4.11) are derived in the Appendix C.

4.2.2 Gaussian Distribution

In this section, we show that if ε is Gaussian distributed then by choosing $p = 2$, $q = \infty$, our estimator reduces to the well-known least-squares estimator.

Consider the GT distribution in Equation (4.2) with $p_i = 2$ and $q_i = \infty$,

$$\begin{aligned}
\frac{p_i}{2\sigma_i q_i^{1/p_i}\beta(1/p_i, q_i)\left(1 + \frac{|\varepsilon_i|^{p_i}}{q_i \sigma_i^{p_i}}\right)^{1/p_i}} &= \frac{1}{\sqrt{\pi}\sigma_i}\\
\frac{1}{\left(1 + \frac{|\varepsilon_i|^{p_i}}{q_i \sigma_i^{p_i}}\right)^{q_i}} &= \exp\left(-\frac{\varepsilon_i^2}{\sigma_i^2}\right)
\end{aligned}$$

and Equation (4.2) reduces to

$$f(\varepsilon_i) = \frac{1}{\sqrt{2\pi}\Lambda_i}\exp\left(-\frac{\varepsilon_i^2}{2\Lambda_i^2}\right)$$

the Gaussian distribution with standard deviation $\Lambda_i = \frac{\sigma_i}{\sqrt{2}}$.

Thus Equations (4.3) and (4.4) reduce to

$$\begin{aligned} J &= \frac{1}{2}\sum_{i=1}^{n}\sum_{k=1}^{N}\left(\frac{\varepsilon_i(k)}{\Lambda_i}\right)^2 - \sum_{i=1}^{n} N \ln \frac{1}{\sqrt{2\pi}\Lambda_i} \\ \text{s.t. } Ax &= 0 \end{aligned}$$

Since the second term in the cost function J is independent of x, minimizing J with respect to x reduces to the well-known least-squares optimization with constraint.

Equation (4.12) reduces to

$$\Gamma = \operatorname{diag}\left\{ \frac{\Lambda_1^2}{N}, \quad \ldots, \quad \frac{\Lambda_n^2}{N} \right\} \tag{4.13}$$

Equations (4.7) and (4.8) reduce to

$$\Gamma^{-1}\frac{1}{N}\sum_{k=1}^{N}\varepsilon(k) + A^T\lambda = 0 \tag{4.14}$$

$$Ax = 0 \tag{4.15}$$

Equations (4.1), (4.14) and (4.15) yields

$$x = \frac{1}{N}\sum_{k=1}^{N} y(k) - \Gamma A^T(A\Gamma A^T)^{-1}A\frac{1}{N}\sum_{k=1}^{N} y(k) \tag{4.16}$$

the well-known least-squares solution [46, 48].

As

$$\mathrm{E}\{\psi(\varepsilon)\psi^T(\varepsilon)\} = \int_{-\infty}^{\infty}\psi(\varepsilon)\psi^T(\varepsilon)f(\varepsilon)d\varepsilon = \Gamma^{-1}$$

Expression (4.10) reduces to

$$\operatorname{Cov} x = \Gamma - \Gamma A^T(A\Gamma A^T)^{-1}A\Gamma \tag{4.17}$$

the well-known covariance matrix of the least-squares solution [48].

4.3 Sample Calculations

To illustrate the calculations, four sets of three ($N = 3$) data points consisting of only -1, 0 and 1 are given in Table 4.1 while their corresponding averages and variances are given in Table 4.2. The parameters were chosen as $p_1 = p_2 = q_1 = q_2 = 2$ and $\sigma_1 = \sigma_2 = 1$. Example of calculations to obtain the values for x_1 in Table 4.2 from data Set 1 in Table 4.1 are as follows.

GT model without data reconciliation

Without data reconciliation means that if we consider $y_1(k)$ then $y_2(k)$ will not be taken into account. In this case, $n = 1$, $A = 0$, $\lambda = 0$ and Equation (4.7) can be written as

$$\frac{5(0-x_1)}{2+(0-x_1)^2} + \frac{5(1-x_1)}{2+(1-x_1)^2} + \frac{5(0-x_1)}{2+(0-x_1)^2} = 0$$

giving $x_1 = 0.29$ in entry (3,d) of Table 4.2. The average and variance for x_1 are calculated as

$$\begin{aligned} \text{Ave } x_1 = \frac{0.29 - 0.29 + 0 + 0}{4} &= 0 \\ \text{Var } x_1 = \frac{0.29^2 + 0.29^2 + 0 + 0}{4} &= 0.042 \end{aligned}$$

as shown in entries (3,h) and (3,i) respectively. *GT model with data reconciliation*

With data reconciliation means that if we consider $y_1(k)$ then $y_2(k)$ will be taken into account. In this case $n = 2$. Let $A = [1 \quad -1]$ and (4.7) gives for $i = 1$

$$\frac{5(0-x_1)}{2+(0-x_1)^2} + \frac{5(1-x_1)}{2+(1-x_1)^2} + \frac{5(0-x_1)}{2+(0-x_1)^2} + \lambda = 0 \tag{4.18}$$

for $i = 2$,

$$\frac{5(0-x_2)}{2+(0-x_2)^2} + \frac{5(0-x_2)}{2+(1-x_2)^2} + \frac{5(0-x_2)}{2+(0-x_2)^2} - \lambda = 0 \tag{4.19}$$

and Equation (4.8),

$$x_1 - x_2 = 0 \tag{4.20}$$

Since there are 3 equations (4.18), (4.19) and (4.20) with 3 unknowns, they can be solved to give $x_1 = x_2 = 0.13$, and $\lambda = -0.95$ as shown in entries (5,d) and (6,d) of Table 4.2.

Gaussian model without data reconciliation

In this case, $n = 1$, $A = 0$ and Equation (4.16) gives

$$x_1 = \frac{1}{3}(0 + 1 + 0) = 0.33$$

in entry (7,d) of Table 4.2.

Gaussian model with data reconciliation

In this case $n = 2$. Equation (4.13) gives

$$\Gamma = \text{diag}\left\{\frac{1}{3\sqrt{2}\sqrt{2}}, \quad \frac{1}{3\sqrt{3}\sqrt{3}}\right\}$$

Let $A = [1 \quad -1]$ and Equation (4.16) gives

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \frac{1}{3}\begin{bmatrix} 0+1+0 \\ 0+0+0 \end{bmatrix} - \Gamma A^T (A\Gamma A^T)^{-1} A \times \frac{1}{3}\begin{bmatrix} 0+1+0 \\ 0+0+0 \end{bmatrix} = \begin{bmatrix} 0.17 \\ 0.17 \end{bmatrix}$$

in entries (9,d) and (10,d).

	Set 1	Set 2	Set 3	Set 4
$y_1(k)$, $k = 1, 2, 3$	0, 1, 0	0, −1, 0	0, 0, 0	0, 0 , 0
$y_2(k)$, $k = 1, 2, 3$	0, 0, 0	0, 0 , 0	0, 1, 0	0, −1, 0

Table 4.1: Data Sets for Sample Calculation example

	a	b	c	d	e	f	g	h	i
1	Distribution	Reconciliation	States	Estimates				$\bar{\hat{x}}_i$	var($\hat{x}_i$)
2	Model			Set 1	Set 2	Set 3	Set 4		
3	GT	Without	x_1	0.29	−0.29	0	0	0	0.042
4			x_2	0	0	0.29	−0.29	0	0.042
5		With	x_1	0.13	−0.13	0.13	−0.13	0	0.017
6			x_2	0.13	−0.13	0.13	−0.13	0	0.017
7	Gaussian	Without	x_1	0.33	−0.33	0	0	0	0.055
8			x_2	0	0	0.33	−0.33	0	0.055
9		With	x_1	0.17	−0.17	0.17	−0.17	0	0.029
10			x_2	0.17	−0.17	0.17	−0.17	0	0.029

Table 4.2: Comparison of Estimation with and without Reconciliation for the Data Sets in Table 4.1.

4.4 Experimental Verification

The data were collected from the film thickness measurement after chemical-mechanical polishing of 24 semiconductor wafers. The wafer is divided into inner and outer zone with 24 measurement points each, giving a total of 1,152 data points. Since there are two zones, $n = 2$. The average thicknesses of the inner ($i = 1$) and outer ($i = 2$) zones were 353 nm and 382 nm respectively and hence $A = \begin{bmatrix} 1 & -\frac{353}{382} \end{bmatrix}$. The distributions of the inner and out zones are shown in Figure 4.1.

For simplicity, we fixed $p_1 = p_2 = 2$ and then the rest of the GT parameters were obtained by fitting the GT probability density function to the data distribution using the maximum likelihood criteria giving $q_1 = 2.7$, $q_2 = 2$, $\sigma_1 = 98.5$, $\sigma_2 = 205.9$. They are shown in Figures 4.1a and 4.1b.

The experimental data for $N = 3$ are tabulated in Table 4.3. The calculations in Table 4.4 mirror that of Table 4.2 except that the data $y_1(k)$ and $y_2(k)$, ($k =$

1, 2, 3) are taken from the 1,152 measurements on the wafers giving 192 sets (see Table 4.3) instead of only 4 sets of data in Table 4.1. The values in the last column of Table 4.4, Theory (Variance), were obtained from Equations (4.10) and (4.17) for the GT and Gaussian model respectively. Notice that they are close to the sample variance calculated from the experimental data in the second last column Experiment (Variance). Hence, before doing the experiment, one can use Equations (4.10) and (4.17) to predict the variance and decide on the sample size N to achieve a desired variance.

We will restrict our discussion to x_1 as the analysis for x_2 is similar. Notice the average values for x_1 are all equal to 353 i.e. Gaussian or GT model with or without reconciliation gives the same result. Consider column Experiment (Variance). Notice that without reconciliation i.e. when we only consider $y_1(k)$, the variance of the standard Least-Square Estimation (Gaussian model) has the largest variance of 43.8. As expected, the variance is reduced when more data i.e. $y_2(k)$ data is taken into account through reconciliation giving 34.8. Similarly, for the GT model, the variances reduced from 40.5 to 33.8 with reconciliation. We can also compare the variances using Gaussian and GT models. For the case without reconciliation, the variance reduced from 43.8 to 40.5 when GT model is used. For the case with reconciliation, the variance reduced from 34.8 to 33.8. In conclusion, the thickness estimate with the smallest variance is obtained when the GT model is used with data reconciliation.

	Set 1	...	Set 192
$y_1(k)$, $k = 1, 2, 3$	377, 351, 346	...	356, 377, 348
$y_2(k)$, $k = 1, 2, 3$	363, 372, 396	...	370, 384, 361

Table 4.3: Experiment Data Sets (nm).

Noise model	DR	State	Estimates			Average	Variance	
			Set 1	...	Set 192		Experiment	Theory
GT	Without	x_1	355	...	358	353	40.5	42.0
		x_2	376	...	372	382	212.2	219.7
	With	x_1	353	...	354	353	33.8	34.3
		x_2	382	...	383	382	39.6	40.2
Gaussian	Without	x_1	358	...	360	353	43.8	45.4
		x_2	377	...	372	382	251.1	282.1
	With	x_1	356	...	357	353	34.8	38.2
		x_2	386	...	387	382	40.7	44.7

Table 4.4: Theoretical and Experimental Results (nm) for $N = 3$

4.5 Conclusion

It was observed that the thickness after chemical-mechanical polishing can be divided into zones. The thickness estimate with the smallest variance is obtained when the GT model is used with data reconciliation. The use of GT distribution model can give smaller variance is because of the extra degree of freedom in the model such that it could give equal or better result than the Gaussian model and the use of data reconciliation can give smaller variance is because more measurement data (both inner and outer zones) were taken into account. The equations derived for computing the variance were verified by the experimental results. These equations enable us to compute the sample size needed by the estimator to achieve a desired variance.

Chapter 5

Skewed and Multivariate Generalized t-distributions

In previous chapters, measurement noise is always assumed as symmetric, independent and identically distributed. However, in practice, these assumptions might not always be satisfied. In this chapter, we present some preliminary works about the situation where noise is asymmetric and dependent to each other. Equations are derived for parameter and data reconciliation cases. However, further investigation and validation are required in the future work to complete the theory.

5.1 Parameter Estimation using skew Generalized t-distribution noise model for asymmetric measurement noise

5.1.1 Introduction

In Chapter 2, parameter estimation problem is investigated extensively where various types of noise are modeled by the generalized t-distribution. However, measurement noise is assumed to be symmetric which might not always be true in practice. When the noise is asymmetric, the conventional estimator will produce a bias result. Jaeckel [50] has provided a solution by modeling noise to be asymmetric with small data size but become symmetric when the sample size grows bigger. MacCullagh and Nelder [51] developed a generalized linear model that can be used with any type of noise belonging to the exponential distribution family. Bianco et al. [52] extended the MM-estimator for the generalized linear model to achieve high efficiency and high break down point. In this chapter, we propose a method as an extension for the conventional estimator to eliminate the bias when the noise is asymmetric while still maintain acceptable result in case of symmetric noise. We also introduce a new estimator based on the skew Generalized t-distribution (SGT) to achieve the optimal result when the shape of the noise is known. Equations for predicting estimate variance are also derived for performance comparison and batch size choosing. The simulation case is presented to verify the proposed theory.

5.1.2 Parameter Estimation

The parameter estimation problem (2.1) is re-stated here

$$y(k) = \phi(k)^T\theta + \varepsilon(k) \tag{5.1}$$

where the vector $\phi(k) = \Big[\phi_1(k), \ldots, \phi_n(k)\Big]^T$ are known, the parameters $\theta = \Big[\theta_1, \ldots, \theta_n\Big]^T$ are to be estimated and $k = 1, \ldots, N$ is the sampling instance.

There are several techniques to solve the problem (5.1), such as the popular LS (2.6) which is the solution of the Equation

$$\psi(\varepsilon) = \sum_{k=1}^{N} \phi(k)^T \Big(y(k) - \phi(k)^T\theta\Big) = 0 \tag{5.2}$$

In Chapter 2, we also proposed the GT Estimator (2.5) of which the estimate is obtained by solving the following Equation

$$\psi(\varepsilon) = \sum_{k=1}^{N} \phi(k)^T \frac{(pq+1)\text{sign}(\varepsilon(k))|\varepsilon(k)|^{p-1}}{q\sigma^p + |\varepsilon(k)|^p} = 0 \tag{5.3}$$

where $\varepsilon(k) = y(k) - \phi(k)^T\theta$. Both equations (5.2) and (5.3) can be rewritten in the form of

$$\sum_{k=1}^{N} \phi(k)^T\psi(\varepsilon(k)) = 0 \tag{5.4}$$

where

$$\phi(\varepsilon(k)) = \varepsilon(k) \qquad \text{in case of equation (5.2).}$$

$$\phi(\varepsilon(k)) = \frac{(pq+1)\text{sign}(\varepsilon)|\varepsilon|^{p-1}}{q\sigma^p + |\varepsilon|^p} \qquad \text{in case of equation (5.3).}$$

5.1.3 Asymmetric noise distribution

It is well-known that when $\varepsilon(k)$ is asymmetric or $\text{E}\varepsilon(k) \neq 0$ the estimator (5.4) will provide bias results [53].

Let $\alpha = \mathrm{E}\varepsilon(k)$ be the expectation value of $\varepsilon(k)$, Equation (5.1) can be rewritten as

$$\begin{aligned} y(k) &= \phi(k)^T\theta + \alpha - \alpha + \varepsilon(k) \\ &= [\phi(k) \quad 1]\begin{bmatrix}\theta \\ \alpha\end{bmatrix} + (\varepsilon(k) - \alpha) \\ &= \phi_n(k)x_n + (\varepsilon(k) - \alpha) \end{aligned} \tag{5.5}$$

where

$$x_n = \begin{bmatrix}\theta^T & \alpha\end{bmatrix}^T$$

$$\phi_n(k) = [\phi(k) \quad 1]$$

Because $\mathrm{E}(\varepsilon(k) - \alpha) = \mathrm{E}\varepsilon(k) - \alpha = 0$, hence with noise model (5.5) Estimator (5.4) will be unbiased. By using the noise model (5.5), the LS estimator (5.2) becomes

$$\sum_{k=1}^{N} \phi_n(k)^T \left(y(k) - \phi_n(k)x_n\right) = 0 \tag{5.6}$$

and the GT estimator (5.3) becomes

$$\sum_{k=1}^{N} \phi_n(k)^T \frac{(pq+1)\mathrm{sign}(y(k) - \phi_n(k)x_n)|y(k) - \phi_n(k)x_n|^{p-1}}{q\sigma^p + |y(k) - \phi_n(k)x_n|^p} = 0 \tag{5.7}$$

5.1.4 Parameter Estimation with skew GT noise model

Skewed Generalized t-distribution

The Skewed Generalized t-distribution (SGT) was first introduced by Theodossiou in 1998 [28]. Since then, it has been developed by many researchers and been mostly used in econometric problems [20, 29, 54–57]. The probability density function (pdf)

of the SGT distribution is given by

$$f(x;\,\mu,\lambda,\sigma,p,q) = \frac{p}{2\sigma p q^{1/p}\beta(1/p,q)\left(1+\frac{|x-\mu|^p}{(1+\lambda\mathrm{sign}(x-\mu))^p q\sigma^p}\right)^{q+1/p}} \tag{5.8}$$

where x is the random variable, μ is the mean, σ is the scale, λ is the skew factor and p, q are the shape parameters. By changing $\{p, q, \sigma, \lambda\}$, the SGT distribution can transform into many kinds of distributions, e.g the skew-elliptical distribution [58] and skew-t distribution [59] etc., as shown in Fig. 5.1.

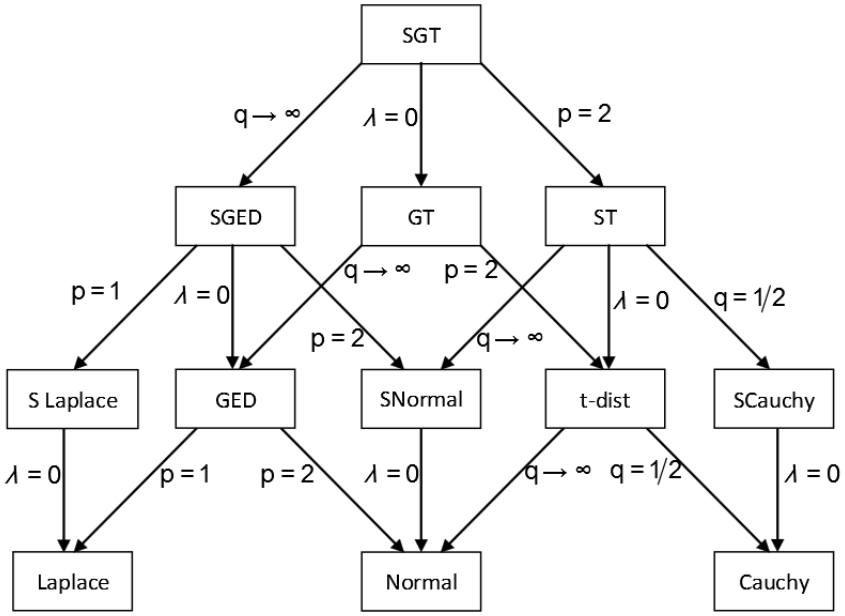


Figure 5.1: The Skewed Generelized T distribution family tree.

Parameter Estimation using SGT noise model

The Parameter Estimator using SGT noise model can be derived using the maximum likelihood estimation which minimizes the following cost function

$$J = -\sum_{k=1}^{N} \ln f\left(y(k) - \phi(k)^T\theta\right) \tag{5.9}$$

where $f(.)$ is defined in (5.8). Differentiating (5.9) with respect to θ and equating it to zero give

$$\sum_{k=1}^{N} \phi(k)^T \frac{(pq+1)|y(k)-\phi(k)^T\theta|^{p-1}\text{sign}(y(k)-\phi(k)^T\theta)}{(1+\lambda\text{sign}(y(k)-\phi(k)^T\theta))^p\, q\sigma^p + |y(k)-\phi(k)^T\theta|^p} = 0 \tag{5.10}$$

The Estimator (5.10) is very useful when the skewness of noise, i.e. λ, is known. However, if the noise skewness is different than λ in (5.10), the Estimator (5.10) will tend to be biased. A remedy to this problem can be achieved by estimating both λ and θ simultaneously, that is

$$\begin{cases} \sum_{k=1}^{N} \phi(k)^T \frac{(pq+1)|y(k)-\phi(k)^T\theta|^{p-1}\text{sign}(y(k)-\phi(k)^T\theta)}{(1+\lambda\text{sign}(y(k)-\phi(k)^T\theta))^p q\sigma^p+|y(k)-\phi(k)^T\theta|^p} & = 0 \\ \sum_{k=1}^{N} \frac{(pq+1)\text{sign}(y(k)-\phi(k)^T\theta)|y(k)-\phi(k)^T\theta|^p}{q\sigma^p[1+\lambda\text{sign}(y(k)-\phi(k)^T\theta)]^{p+1}+[1+\lambda\text{sign}(y(k)-\phi(k)^T\theta)]|y(k)-\phi(k)^T\theta|^p} & = 0 \end{cases} \tag{5.11}$$

where the second equation of (5.11) is derived by differentiating (5.9) with respect to λ and equating it to zero. Because Estimator (5.11) is Maximum Likelihood Estimation, the estimate $\hat{\theta}$ and $\hat{\lambda}$ will tend to the true values of θ and λ when $N \to \infty$.

5.1.5 Analysis of Estimators

In this section, we will derive the Influence Function (IF) [2, 25] of the above estimators. The derived Influence Function then will be used to compare the performance of those estimators. From [2, 25], by using IF, one can predict the estimate variances calculated by Equation (2.14). Equation (2.14) will also provide us useful information to choose the suitable batch size N for suitable circumstances.

In chapter 2, we have derived the Influence Function of the Estimator (5.2) which is given by

$$\mathrm{IF}(z) = \left(\Phi^T\Phi\right)^{-1}\Phi z \tag{5.12}$$

We also provided the Influence Functions of Estimator (5.3)

$$\mathrm{IF}(z) = M^{-1}\Phi^T\frac{(pq+1)|z|^{p-1}\mathrm{sign}(z)}{q\sigma^p+|z|^p} \tag{5.13}$$

where M is

$$M = \Phi^T\Phi\int_{-\infty}^{+\infty}\frac{(pq+1)\left[(p-1)(q\sigma^p+\varepsilon^p)\varepsilon^{p-2}-p\varepsilon^{2p-2}\right]}{\left(q\sigma^p+\varepsilon^p\right)^2}f(\varepsilon)d\varepsilon$$

Influence Functions of Estimators with noise model (5.5)

Let $\Phi_n = [\phi_n(1) \quad \cdots \quad \phi_n(N)]^T$. Substituting Φ by Φ_n into (5.12) gives the Influence Function of the Least-squares estimator with noise model (5.5)

$$\mathrm{IF}(z) = \left(\Phi_n^T\Phi_n\right)^{-1}\Phi_n z \tag{5.14}$$

The Influence Function of the GT-estimator with noise model (5.5) is also given by substituting Φ by Φ_n into (5.13)

$$\mathrm{IF}(z) = M_n^{-1}\Phi_n^T\frac{(pq+1)|z|^{p-1}\mathrm{sign}(z)}{q\sigma^p+|z|^p} \tag{5.15}$$

where M_n is

$$M_n = \Phi_n^T\Phi_n\int_{-\infty}^{+\infty}\frac{(pq+1)\left[(p-1)(q\sigma^p+\varepsilon^p)\varepsilon^{p-2}-p\varepsilon^{2p-2}\right]}{\left(q\sigma^p+\varepsilon^p\right)^2}f(\varepsilon)d\varepsilon$$

Note that the IF in (5.14) and (5.15) are $((m+1)\times 1)$ vectors as the bias term is an estimate together with the parameter θ.

Influence Function of SGT Estimator

The Influence Function of Estimator (5.10) is given by

$$\mathrm{IF}(z) = M^{-1}\Phi^T \frac{(pq+1)|z|^{p-1}\mathrm{sign}(z)}{(1+\lambda\mathrm{sign}(z))^p\, q\sigma^p + |z|^p} \tag{5.16}$$

where M is

$$M = \Phi^T\Phi \int_{-\infty}^{+\infty} \frac{\partial\psi(\varepsilon)}{\partial\varepsilon} f(\varepsilon)d\varepsilon$$

$\frac{\partial\psi(\varepsilon)}{\partial\varepsilon}$ is the derivative of $\psi(\varepsilon)$ and is defined as

$$\frac{\partial\psi(\varepsilon)}{\partial\varepsilon} = (pq+1)\frac{(p-1)|\varepsilon|^{p-2}\,(1+\lambda\mathrm{sign}(\varepsilon))^p\, q\sigma^p + (p-1)|\varepsilon|^{2p-2} - p|\varepsilon|^{2p-2}}{\left((1+\lambda\mathrm{sign}(\varepsilon))^p q\sigma^p + |\varepsilon|^p\right)^2}$$

The derivation of (5.16) is given in Appendix D.

Since the Estimator (5.11) involves the skew factor λ, the Influence Function becomes more complicated. It is defined as

$$\mathrm{IF}(z) = \begin{bmatrix} \int_{-\infty}^{+\infty} \sum_{k=1}^N \phi(k)^T \frac{\partial\psi_1(\varepsilon)}{\partial\varepsilon} f(\varepsilon)\phi(k)d\varepsilon & -\int_{-\infty}^{+\infty} \sum_{k=1}^N \phi(k)^T \frac{\partial\psi_1(\varepsilon)}{\partial\lambda} f(\varepsilon)d\varepsilon \\ \int_{-\infty}^{+\infty} \sum_{k=1}^N \frac{\partial\psi_2(\varepsilon)}{\partial\varepsilon} f(\varepsilon)\phi(k)d\varepsilon & -\int_{-\infty}^{+\infty} \sum_{k=1}^N \frac{\partial\psi_2(\varepsilon)}{\partial\lambda} f(\varepsilon)d\varepsilon \end{bmatrix}^{-1} \times \begin{bmatrix} \sum_{k=1}^N \phi(k)^T\psi_1(z) \\ \sum_{k=1}^N \psi_2(z) \end{bmatrix} \tag{5.17}$$

where

$$\psi_1(\varepsilon,\lambda) = \frac{(pq+1)|\varepsilon|^{p-1}\text{sign}(\varepsilon)}{(1+\lambda\text{sign}(\varepsilon))^p\, q\sigma^p + |\varepsilon|^p}$$

$$\psi_2(\varepsilon,\lambda) = \frac{(pq+1)\text{sign}(\varepsilon)|\varepsilon|^p}{q\sigma^p\left[1+\lambda\text{sign}(\varepsilon)\right]^{p+1} + \left[1+\lambda\text{sign}(\varepsilon)\right]|\varepsilon|^p}$$

$$\frac{d\psi_1(\varepsilon,\lambda)}{d\varepsilon} = (pq+1)\frac{(p-1)|\varepsilon|^{p-2}\left(1+\lambda\text{sign}(\varepsilon)\right)^p q\sigma^p - |\varepsilon|^{2p-2}}{\left((1+\lambda\text{sign}(\varepsilon))^p q\sigma^p + |\varepsilon|^p\right)^2}$$

$$\frac{d\psi_1(\varepsilon,\lambda)}{d\lambda} = -(pq+1)\frac{pq\sigma^p|\varepsilon|^{p-1}\left(\lambda\text{sign}(\varepsilon)+1\right)^{p-1}}{\left(|\varepsilon|^p + q\sigma^p\left(\lambda\text{sign}(\varepsilon)+1\right)^p\right)^2}$$

$$\frac{d\psi_2(\varepsilon,\lambda)}{d\varepsilon} = (pq+1)\left(\frac{p|\varepsilon|^{2p-1}\left(\lambda\text{sign}(\varepsilon)+1\right)}{\left(q\sigma^p\left[1+\lambda\text{sign}(\varepsilon)\right]^{p+1} + \left[1+\lambda\text{sign}(\varepsilon)\right]|\varepsilon|^p\right)^2} + \frac{p|\varepsilon|^{p-1}}{q\sigma^p\left[1+\lambda\text{sign}(\varepsilon)\right]^{p+1} + \left[1+\lambda\text{sign}(\varepsilon)\right]|\varepsilon|^p}\right)$$

$$\frac{d\psi_2(\varepsilon,\lambda)}{d\lambda} = -(pq+1)\frac{|\varepsilon|^{2p} + (p+1)q\sigma^p|\varepsilon|^p\left(\lambda\text{sign}(\varepsilon)+1\right)^p}{\left(q\sigma^p\left[1+\lambda\text{sign}(\varepsilon)\right]^{p+1} + \left[1+\lambda\text{sign}(\varepsilon)\right]|\varepsilon|^p\right)^2}$$

The derivation of Equation (5.17) is given in Appendix E.

5.1.6 Simulation Case Study

In this section, we give a simulation to show the effectiveness of the new derived estimators when the noise is asymmetric. The simulation is conducted with the above mentioned estimators, those are

- E1 : The origin Least-squares (5.2).
- E2 : The GT Estimator (5.3).
- E3 : The modified Least-squares Estimator (5.6).
- E4 : The modified GT Estimator (5.7).
- E5 : The skew-GT Estimator (5.10).

- E6 : The simultaneous skew-GT Estimator (5.11).

The estimate bias and the estimate variance are used to justify the performance of each estimator.

Consider a linear in parameter system (5.1) with $\phi(k) = 0.9^{k-1}$ and $\varepsilon(k)$ is drawn from the asymmetric t_3-distribution defined in [60] with location of 0, scale of 1 and skewness of 0.1. The histogram of $\varepsilon(k)$ is shown in Figure 5.2.

The shape parameters of $\{p = 2,\ q = 1.5,\ \sigma = \sqrt{2}\}$ are used for Estimator E2, E4, E5 and E6. The skew parameter $\lambda = 0.05$ is used for Estimator E5. Let $\theta = 1$, we run 2000 simulations with batch size of $N = 60$ then use the estimate variances for comparison. The simulation results are shown in Table 5.1. Column "Ave" is the average of the estimate; Column "Bias" is the absolute value of the average estimate minus the true value "1"; Column "Var Exp" is the variance of the estimate obtained from Simulation; and Column "Var Thr" is the variance calculated using the Influence Functions and Equation (2.14).

As can be seen in Table 5.1, when the noise is slightly asymmetric ,e.g. skewness of 0.1, the Estimator E1 and E2 provide bias results (0.19 for E1 and 0.16 for E2). The Estimator E3 and E4 provide a remedy for bias problem by greatly reducing the bias term (0.04 compare with 0.19 and 0.01 compare with 0.16); however, there's also a trade-off that is the estimate variances become greater. When the noise is well-fitted by the skew GT, i.e. the λ is known, the estimator E5 gives the best results with small estimate variance and almost no bias. When the skewness λ is not well-defined, the Estimator E6 still gives good result which is about 11%

improvement over the Estimator E4 which is the modified GT estimator.

	Estimator E1 (5.2)				Estimator E2 (5.3)			
	Ave	Bias	Var Exp	Var Thr	Ave	Bias	Var Exp	Var Thr
$\hat{\theta}$	1.19	0.19	0.53	0.53	1.16	0.16	0.29	0.28
	Estimator E3 (5.6)				Estimator E4 (5.7)			
	Ave	Bias	Var Exp	Var Thr	Ave	Bias	Var Exp	Var Thr
$\hat{\theta}$	0.96	0.04	0.79	0.77	0.99	0.01	0.41	0.41
	Estimator E5 (5.10)				Estimator E6 (5.11)			
	Ave	Bias	Var Exp	Var Thr	Ave	Bias	Var Exp	Var Thr
$\hat{\theta}$	1.02	0.02	0.28	0.28	1.00	0.00	0.38	0.36

Table 5.1: Simulation Results.

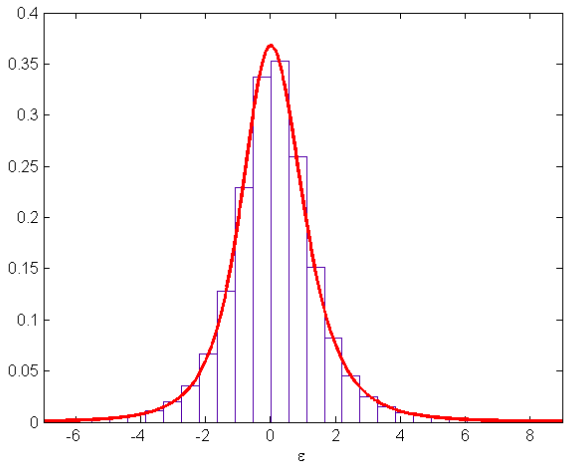


Figure 5.2: The histogram of $\varepsilon(k)$ in Simulation Example.

5.1.7 Conclusion

In this section, we proposed an extension for the conventional estimators to deal with the asymmetric error in the parameter estimation problem. We also introduce two new estimators based on the skew Generalized t-distribution to achieve better performance. Equations to predict the estimate variance are also derived. These

Equations are useful for one to choose a suitable batch size N for a selected estimator to meet a specific requirement. The proposed theory are verified by simulation case study. However, Arslan and Genç [57] pointed out that the estimation of the skewness λ in (5.11) is still affected by large outlier. A suggestion is to use robust measure of skewness [61, 62] to facilitate the problem. Influence Function will still need to be derived to ensure the performance of the estimator.

5.2 Data Reconciliation using Multivariate Generalized t-distribution noise model

5.2.1 Introduction

In Chapter 4, we have proposed a DR estimator using the GT noise model. The Influence Function was also derived to give the asymptotic variance of the estimate. Robust statistics have been applied to DR problems to robustify the DR framework [12, 13, 63–66]. However, DR problem requires reconcile data from multiple inputs/nodes simultaneously and the measurements might be cross correlated [48]. In this chapter, a new DR estimator is proposed to address the noise correlation problem by employing the multivariate Generalized t-distribution (MGT) [67, 68]. By being a natural extension of the univariate Generalized t-distribution [17], the MGT contains all the characteristics of the GT distribution, moreover, the noise correlation is also be considered. Hence, more accurate estimate can be achieved by this new proposed estimator. An industry simulation case study is presented

to support the proposed theory. However, the case of unknown covariance matrix Σ might need to be further investigated to make it possible to apply the proposed theory to practical systems.

5.2.2 The Multivariate Maximum Likelihood for Data Reconciliation

The Maximum Likelihood Estimation (MLE) is one of the most widely used methods of statistical estimation. It is reported to be best under most circumstances [33]. It is also the start point for various kinds of estimator [69]. The objective of the MLE is to maximize the likelihood function, this is equivalent to

$$\begin{aligned} \min\ J &= -\sum_{k=1}^{N} \log f(y(k) - x) \\ s.t.\ Ax &= 0 \end{aligned} \tag{5.18}$$

where $y(k)$, x and A are the ($\mathscr{p}\times 1$) measurement vector, ($\mathscr{p}\times 1$) true value vector and the ($m \times \mathscr{p}$) constraint matrix, respectively. $\varepsilon(k) = y(k) - x$ is the measurement noise which is assumed as zero-mean noise. N is the batch size, i.e., collect N measurements for one estimate.

To solve (5.18), the Lagrange Multipliers method is used as follows

$$L = -\sum_{k=1}^{N} \log f(\varepsilon(k)) - \gamma Ax = \sum_{k=1}^{N} \rho(\varepsilon(k)) - \gamma Ax \tag{5.19}$$

where γ is the ($1 \times m$) Lagrange multiplier and $\rho(\varepsilon(k)) = -\log f(\varepsilon(k))$.

In the multivariate framework, $f(\varepsilon(k))$ in (5.19) is usually in form of $f(s(k))$ with $s(k) = \varepsilon(k)^T \Sigma^{-1} \varepsilon(k)$ where Σ is the covariance matrix. Taking derivative of

(5.19) with respect to x gives

$$\begin{cases} \frac{\partial L}{\partial x} & = \sum_{k=1}^{N} \frac{\partial \rho(\varepsilon(k))}{\partial x} - A^T \gamma^T = 0 \\ \frac{\partial L}{\partial \gamma} & = Ax = 0 \end{cases} \tag{5.20}$$

where

$$\begin{aligned} \frac{\partial \rho(\varepsilon(k))}{\partial x} &= \frac{\partial \rho(\varepsilon(k))}{\partial f(\varepsilon(k))} \frac{\partial f(\varepsilon(k))}{\partial s(k)} \frac{\partial s(k)}{\partial \varepsilon(k)} \frac{\partial \varepsilon(k)}{\partial x} \\ &= -\frac{1}{f(\varepsilon(k))} \frac{\partial f(\varepsilon(k))}{\partial s(k)} 2\Sigma^{-1} \varepsilon(k) \end{aligned}$$

Denote

$$\psi(\varepsilon(k)) = -\varepsilon(k) \frac{1}{f(\varepsilon(k))} \frac{\partial f(\varepsilon(k))}{\partial s(k)} \tag{5.21}$$

Equation (5.20) becomes

$$\begin{cases} \frac{\partial L}{\partial x} & = \sum_{k=1}^{N} 2\Sigma^{-1} \psi(\varepsilon(k)) - A^T \gamma^T = 0 \\ \frac{\partial L}{\partial \gamma} & = Ax = 0 \end{cases} \tag{5.22}$$

(5.22) is the multivariate maximum likelihood DR estimator. By solving (5.22), the estimated value $\hat{x}$ can be acquired. By similar technique as in Chapter 4, the Influence Function of Estimator (5.22) is derived as follows

$$\mathrm{IF}(z) = \left(M^{-1} - M^{-1} A \left(A M^{-1} A^T \right)^{-1} A M^{-1} \right) 2\Sigma^{-1} \psi(z) \tag{5.23}$$

where

$$M = \int_{-\infty}^{+\infty} 2\Sigma^{-1} \frac{\partial \psi(\varepsilon)}{\partial \varepsilon} f(\varepsilon) d\varepsilon \tag{5.24}$$

Derivation of Equation (5.23) is given in Appendix F.

By using the Influence Function, [24] showed that with batch size N, the estimate variance can be approximated by the following

$$\mathrm{Var}(\hat{x}) = \frac{1}{N}\int_{-\infty}^{\infty} \mathrm{IF}(z) \times \mathrm{IF}(z)^T f(z) dz \tag{5.25}$$

5.2.3 The multivariate GT-based Data Reconciliation

The univariate GT distribution is first introduced by McDonald and Newey [17]. Later, Reinaldo B. Arellano-Valle [67] proposed a family of multivariate GT distribution which includes multivariate t-distribution as a special case. A more general family of multivariate GT distribution is later introduced by Arslan [68] of which the probability density function is

$$f(x;\, \mu, \Sigma, \lambda, \beta, \mathfrak{q}) = C\lambda^{-\mathfrak{p}/2}|\Sigma|^{-1/2}\frac{1}{\left\{\mathfrak{q} + \left(\frac{s}{\lambda}\right)^{\beta}\right\}^{\mathfrak{q}+\frac{\mathfrak{p}}{2\beta}}} \tag{5.26}$$

where

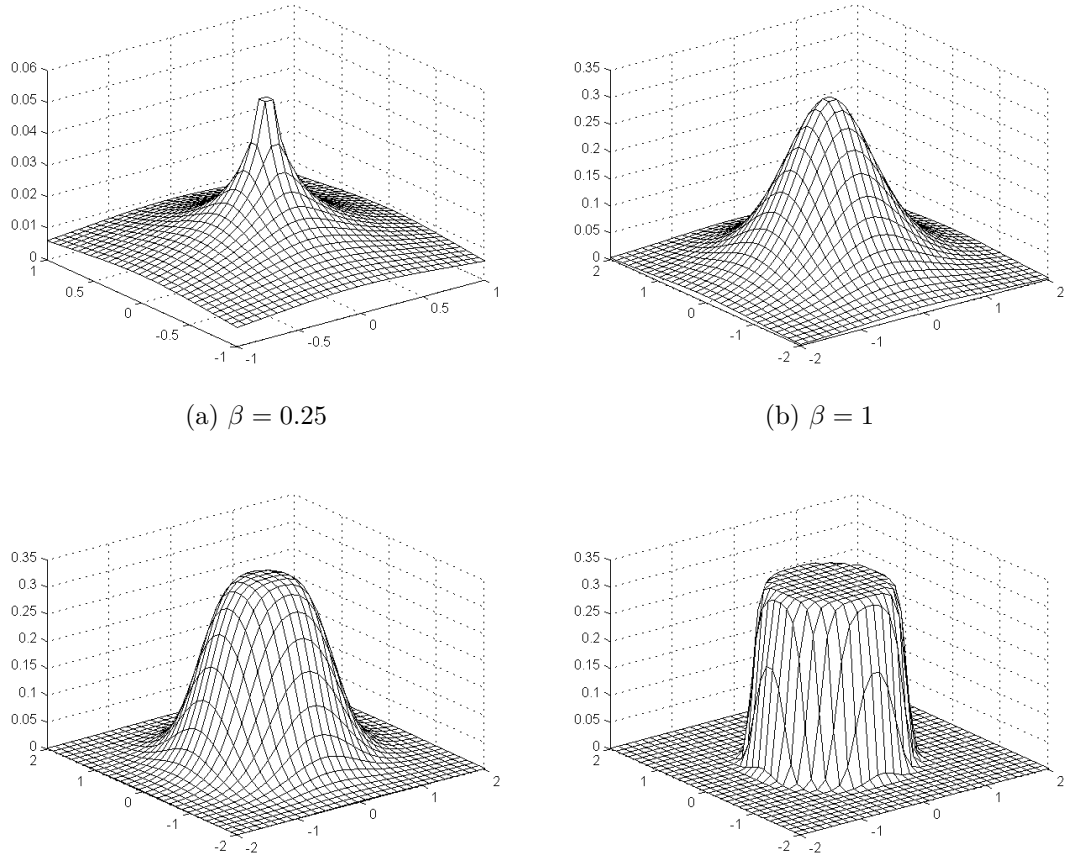
$$s = (x-\mu)^T\Sigma^{-1}(x-\mu)$$

and

$$C = \frac{\beta\Gamma(\mathfrak{p}/2)\mathfrak{q}^{\mathfrak{q}}}{\pi^{\mathfrak{p}/2}\mathcal{B}(\mathfrak{q}, \frac{\mathfrak{p}}{2\beta})}$$

$\mathfrak{p}$ is the size of random variable vector x (one should note that this $\mathfrak{p}$ is not the p in previous chapter which is one of the shape parameters). $\mathcal{B}$ and Γ are the Beta function and Gamma function, respectively. The constants $\{\lambda,\, \beta,\, \mathfrak{q}\}$ are the shape-parameters. Set $\mathfrak{p} = 1$, (5.26) will turn into the univariate GT distribution [17]. By varying the shape-parameters $\{\lambda,\, \beta,\, \mathfrak{q}\}$, the multivariate GT distribution can be transformed to various types of distribution including the multivariate skew t distribution [70] and the multivariate normal distribution [71]. Some visual illus-

trations of the MGT distribution with different values of $\{\lambda, \beta, \varphi\}$ are shown in Figure 5.3.



(a) $\beta = 0.25$ (b) $\beta = 1$

(c) $\beta = 3$ (d) $\beta = 10$

Figure 5.3: PDF Plots of $\mathrm{MGT}_2(0, I_2, 1, \beta, 2)$ with different values of β.

Substituting $f(\varepsilon(k))$ in (5.26) into (5.22) gives the multivariate GT-based state-estimator

$$\begin{cases} \sum_{k=1}^{N} \left(\varphi + \frac{p}{2\beta}\right) 2\Sigma^{-1}\varepsilon(k)\frac{s(k)^{\beta-1}}{\varphi\lambda^{\beta}+s(k)^{\beta}} - A^T\gamma^T = 0 \\ Ax = 0 \end{cases} \tag{5.27}$$

(5.27) is the multiple non-linear equations which can be solved by Newton-Raphson method.

5.2.4 The multivariate Least-Square Estimator

It is known that the multivariate GT distribution tranforms to the multivariate Gaussian distribution when $\beta = 1$, $\lambda = 2$ and $q \to \infty$ [68]. Consider the multivariate GT probability density function in Equation (5.26) with $\beta = 1$ and $\lambda = 1$

$$\begin{aligned} f(x) &= |\Sigma|^{-1/2} \frac{\Gamma(p/2) q^q}{\pi^{p/2} \mathcal{B}\left(q, p/2\right)} \times \frac{1}{q^q q^{p/2} \left(1 + \frac{s}{q}\right)^{q+p/2}} \\ &= \frac{|\Sigma|^{-1/2}}{\pi^{p/2}} \times \frac{\Gamma(p/2)}{\mathcal{B}\left(q, p/2\right) q^{p/2}} \times \frac{1}{\left(1 + \frac{s}{q}\right)^{q+p/2}} \end{aligned} \tag{5.28}$$

Note that

$$\lim_{q \to \infty} \left(1 + \frac{s}{q}\right)^{q+p/2} = \exp\left(s\right)$$

$$\lim_{q \to \infty} \mathcal{B}\left(q, p/2\right) q^{p/2} = \Gamma(p/2)$$

Substituting $q \to \infty$ and the above equations into (5.28) gives

$$f(x) = \pi^{-p/2} |\Sigma|^{-1/2} \exp(-s) \tag{5.29}$$

Let $\Sigma = 2\Lambda$, hence

$$\begin{aligned} \Sigma^{-1} &= \frac{1}{2}\Lambda^{-1} \\ |\Sigma|^{-1/2} &= 2^{-p/2} |\Lambda|^{-1/2} \\ s &= (x-\mu)^T \Sigma^{-1} (x-\mu) \\ &= \frac{1}{2}(x-\mu)^T \Lambda^{-1} (x-\mu) \end{aligned}$$

Equation (5.29) then becomes

$$f(x) = (2\pi)^{-p/2} |\Lambda|^{-1/2} \exp\left(-\frac{1}{2}(x-\mu)^T \Lambda^{-1} (x-\mu)\right) \tag{5.30}$$

the multivariate Gaussian probability density function [71] with the co-variance matrix $\Lambda = \Sigma/2$. As (5.29) is same as (5.30), substituting (5.29) into (5.21) and (5.22) gives

$$\begin{aligned} &\psi(\varepsilon(k)) = \varepsilon(k) = y(k) - x \\ &\begin{cases} \frac{\partial L}{\partial x} &= \sum_{k=1}^{N} 2\Sigma^{-1}\Big(y(k) - x\Big) - A^T\gamma^T = 0 \\ \frac{\partial L}{\partial \gamma} &= Ax = 0 \end{cases} \end{aligned} \tag{5.31}$$

Solving (5.31) gives

$$\begin{aligned} \lambda^T &= 2\left(A\Sigma A^T\right)^{-1} A\sum_{k=1}^{N} y(k) \\ \hat{x} &= \frac{1}{N}\sum_{k=1}^{N} y(k) - \Sigma A^T\left(A\Sigma A^T\right)^{-1} A\frac{1}{N}\sum_{k=1}^{N} y(k) \\ &= \frac{1}{N}\sum_{k=1}^{N} y(k) - \Lambda A^T\left(A\Lambda A^T\right)^{-1} A\frac{1}{N}\sum_{k=1}^{N} y(k) \end{aligned} \tag{5.32}$$

If vector noise $\varepsilon(k)$ is independent, i.e. Λ is diagonal, (5.32) becomes the univariate LS estimator (4.16) [46, 48].

5.2.5 Analysis of Estimator Performance

In this section, we derive the Influence Function (IF) [2] to predict the variances of estimate for each estimator. The IF will also tell one about the robust characteristic of one estimator. The IF is well-studied in [2], it has been shown to be very useful for realistic situation. IF formulas for the multivariate state-estimator (5.22) is derived in Appendix F, in this section, the IF for the multivariate GT-based and LS estimators are then determined.

IF of the Multivariate GT-based DR Estimator

Substituting $f(\varepsilon(k))$ in (5.26) into (5.21) gives

$$\psi(\varepsilon(k)) = -\frac{1}{f(\varepsilon(k))}\frac{\partial f(\varepsilon(k))}{\partial s(k)}\varepsilon(k) = \left(\mathcal{q} + \frac{\mathcal{p}}{2\beta}\right)\varepsilon(k)\frac{s(k)^{\beta-1}}{\mathcal{q}\lambda^{\beta} + s(k)^{\beta}}$$

where $s(k) = \varepsilon(k)^T\Sigma^{-1}\varepsilon(k)$. For the sake of brevity, we drop the suffix (k). It then follows that

$$\frac{\partial\psi}{\partial\varepsilon} = \frac{\mathcal{p} + 2\beta\mathcal{q}}{2\beta}\left(\frac{s^{\beta-1}}{\mathcal{q}\lambda^{\beta} + s^{\beta}}I_{\mathcal{p}} + 2\varepsilon\varepsilon^T\Sigma^{-1}\frac{s^{\beta-2}\left((\beta-1)\mathcal{q}\lambda^{\beta} - s^{\beta}\right)}{\left(\mathcal{q}\lambda^{\beta} + s^{\beta}\right)^2}\right)$$

where $I_{\mathcal{p}}$ is the identity matrix with size of $\mathcal{p}$. The matrix M (5.24) is calculated as

$$M = \frac{\mathcal{p} + 2\beta\mathcal{q}}{\beta}\Sigma^{-1}\int_{-\infty}^{\infty}\left(\frac{s^{\beta-1}}{\mathcal{q}\lambda^{\beta} + s^{\beta}}I_{\mathcal{p}} + 2\varepsilon\varepsilon^T\Sigma^{-1}\frac{s^{\beta-2}\left((\beta-1)\mathcal{q}\lambda^{\beta} - s^{\beta}\right)}{\left(\mathcal{q}\lambda^{\beta} + s^{\beta}\right)^2}\right)f(\varepsilon)d\varepsilon$$

Substituting the above Equations into Equation (5.23) gives the IF for the estimator (5.27). Using the obtained IF, the estimate variance of estimator (5.27) can also be approximated by Equation (5.25).

IF of the Multivariate LS DR Estimator

The IF of the Multivariate LS DR Estimator is calculated by substituting $\psi(\varepsilon) = \varepsilon$ into Equations (5.23) and (5.24)

$$M = 2\Sigma^{-1}$$

$$\mathrm{IF}(z) = \left(I_{\mathcal{p}} - \Sigma A\left(A\Sigma A^T\right)^{-1}A\right)z = \left(I_{\mathcal{p}} - \Lambda A\left(A\Lambda A^T\right)^{-1}A\right)z$$

Equation (5.25) then becomes

$$\mathrm{Var}\ \hat{x} = \frac{1}{N}\left(\Lambda - \Lambda A(A\Lambda A^T)^{-1}A\Lambda\right)$$

With the unit batch size and diagonal matrix Λ, the above equation becomes the one given in [46, 48].

5.2.6 Simulation Case Study

In this section, we will give some examples to show the advantage of the multivariate GT-based estimator (5.27) over the conventional LS estimators (4.16), (5.32) and the univariate GT-based estimator (4.7).

Consider the chemical reactor with four flows in [48]. The elemental balances define the following constraint matrix

$$A = \begin{bmatrix} 0.1 & 0.6 & -0.2 & -0.7 \\ 0.8 & 0.1 & -0.2 & -0.1 \\ 0.1 & 0.3 & -0.6 & -0.2 \end{bmatrix}$$

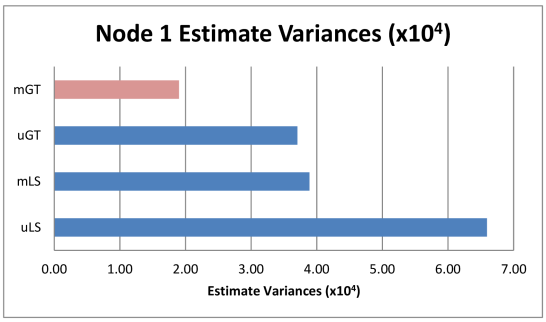
The simulated measurement is generated from the multivariate t-distribution with 3 degree of freedom and the covariance matrix

$$\Sigma_t = \begin{bmatrix} 1 & 0.5 & 0.5 & 0.5 \\ 0.5 & 1 & 0.5 & 0.5 \\ 0.5 & 0.5 & 1 & 0.5 \\ 0.5 & 0.5 & 0.5 & 1 \end{bmatrix}$$

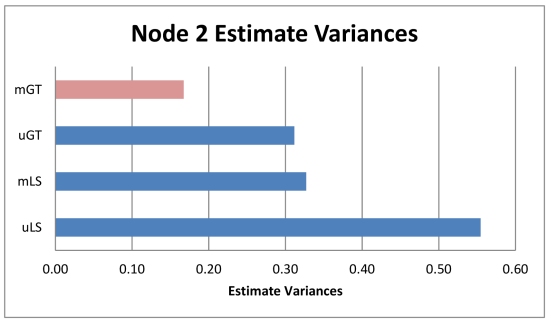
The parameters for the univariate and multivariate LS (uLS and mLS) are chosen to be the identity matrix and the variance of the generated data respectively. The parameters for the univariate GT (uGT) are chosen based on Figure 2.3 which gives $p_1 = \ldots = p_4 = 2$, $q_1 = \ldots = q_4 = 1.5$ and $\sigma_1 = \ldots = \sigma_4 = \sqrt{2}$. The parameters for the multivariate GT (mGT) is chosen from [68] which gives $\lambda = \beta = 1$, $\mathscr{q} = 1.5$ and $\Sigma = 2\Sigma_t$. The simulation is conducted with the batch size of $N = 5$ using 2000 data. The results are presented on Table 5.2 and Figure 5.4.

Table 5.2: Estimate variances in simulation case study section

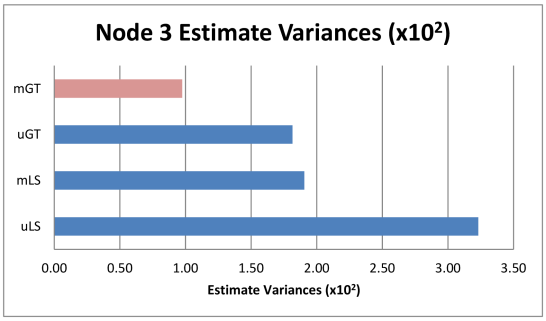
Flow #	uLS (4.16)	mLS (5.32)	uGT (4.7)	mGT (5.27)	Eqn (5.25)
1	6.59×10^{-4}	3.89×10^{-4}	3.70×10^{-4}	1.90×10^{-4}	1.84×10^{-4}
2	0.55	0.33	0.31	0.17	0.15
3	3.23×10^{-2}	1.91×10^{-2}	1.82×10^{-2}	0.97×10^{-2}	0.90×10^{-2}
4	0.35	0.21	0.20	0.10	0.10



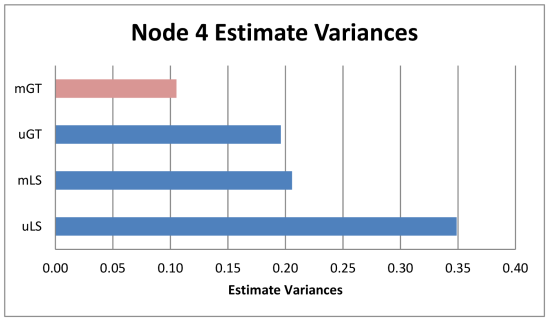
(a) Flow #1



(b) Flow #2



(c) Flow #3



(d) Flow #4

Figure 5.4: Comparison based on estimate variances of the four estimators for the simulation case study (lower is better).

As can be seen in Table 5.2, the univariate LS gives the worst result as it not only does not handle the long-tail distribution but also the noise correlation. The multivariate LS aids the noise correlation problem, which gives much better results. And the univariate GT estimate aids the non-Gaussian noise. The proposed

multivariate GT estimator takes the best of both methods resulting a significant improvement in term of estimate variance hence estimate accuracy. A clear visual representation of the results are shown in Figure 5.4 for illustrative purpose. Moreover, Equation (5.25) gives the variance values close to the experiment (see the last two columns in Table 5.2). This indicates that Equation (5.25) can be used to predict the estimate variance with a chosen batch size N, hence N can be calculated for a required estimate variance specification.

5.3 Conclusion

In this chapter, we first proposed an extension for the conventional estimators to deal with the asymmetric error in the parameter estimation problem. We also introduce two new estimators based on the skew Generalized t-distribution to achieve better performance. Equations to predict the estimate variance are also derived. These Equations are useful for one to choose a suitable batch size N for a selected estimator to meet a specific requirement. The proposed theory are verified by simulation case study. However, Arslan and Genç [57] pointed out that the estimation of the skewness λ in (5.11) is still affected by large outlier. A suggestion is to use robust measure of skewness [61, 62] to facilitate the problem. Influence Function will still need to be derived to ensure the performance of the estimator.

We also introduced a new DR estimator using the multivariate GT distribution as noise model. By using the MGT distribution, the noise correlation has been taken into account and the achieved estimate is more accurate than the one from the

univariate estimators. Influence Function of the proposed estimator is also derived to give an approximated estimate variance. The theory is verified by an industry simulation example. In this work, however, there are some drawbacks. First, the covariance matrix Σ must be known before applying the proposed estimator. Some techniques, e.g. [72] and [73], can be used to estimate the noise covariance matrix. Second, the Influence Function for the proposed estimator is only applicable for the linear constraint case. For non-linear systems, the Influence Function might be very complicated and requires further investigation.

Chapter 6

Conclusion

In this thesis, instead of using Gaussian distribution as noise model, we take a different approach by modeling noise using the Generalized t-distribution. This resulted new estimators for the parameter estimation, ARMAX filtering and data reconciliation problems. The performance of our proposed estimator was validated by Monte Carlo simulation and by experiment on couple-tank liquid level and Chemical Mechanical Polishing (CMP) process thickness measurements. It was found that the proposed estimator is very reliable in handling various kinds of noise. It is also very robust against outlier and gross error. This is because the Generalized-T distribution is very flexible in transforming into many other distributions. By being a superset encompassing Gaussian, uniform, t and double exponential distributions, GT distribution has the flexibility to characterize data with non-Gaussian statistical properties. The simulation and experiment results also suggest that a good understanding of measurement noise may give high accuracy of estimate.

In Chapter 2, The Influence Function of the parameter estimator using GT

noise model was derived for the analysis purpose. With the Influence Function, the estimate variance can be predicted, hence the batch size N can be calculated for a require specification. We also extend the use of Influence Function to predict the change of estimate due to one outlier. This is useful for choosing estimator parameters to minimize the effect of outlier. The proposed theory is verified by both simulation and experiment on CMP wafer thickness measurement.

In Chapter 3, a novel filter for ARMAX model is derived. It has been shown that the proposed filter is very robust against the outlier. We also make use of the influence function to establish a recursive algorithm for the proposed filter. The derived recursive algorithm significantly reduces the computational time and data storage, which makes it applicable for practical processes. The proposed algorithm is verified by Monte Carlo simulation and experiment on couple-tank liquid measurement model.

Chapter 4 extends the use of GT distribution to multivariate framework with constraint where multiple input measurements need to be treated simultaneously and the estimate must follow some physical constraints. The Influence Function for the framework is also derived to make use of the estimate variance prediction capability. As shown in Chapter 3, by using IF, one might find it possible to approximate a batch estimator to a recursive estimator. The proposed estimator is validated by simulation and experiment on CMP wafer thickness measurement.

Chapter 5 presents some preliminary works on asymmetric and correlated noise by making use of the skewed GT distribution and multivariate GT distribution for the parameter estimation and data reconciliation respectively. The simulation

results suggest that with the knowledge of noise correlation, the estimate accuracy would be vastly improved. The Influence Function is also to derived. However, experiment verification might be needed to validate the proposed theory.

The experimental results on the Chemical Mechanical Processing wafer thickness and the couple-tank liquid level suggests that real-life measurement noise needs not to be Gaussian or any other known distributions; hence our proposed estimator can easily handle them. Another point in the experiment is that even with slightly asymmetric measurement noise, the derived estimator can still be able to handle it with almost no loss of accuracy.

In this work, however, there are some limitations for our proposed estimator. Firstly, only the case of known measurement noise is considered. If measurement noise is poorly modeled, the reconciled data can be badly estimated, which will critically affect the whole statistic analysis. Further research is needed to deal with the case of unknown noise or un-fixed shape noise to ensure the reliability and stability of the estimator. The main challenge is to develop an adaptive system in which the noise model can adapt to the change of measurement noise.

Another limitation is that our proposed estimator and framework only consider noise that is independent and identically distributed. In case of correlated or time-variant noise, the proposed estimator may not give optimal solution. However, in case of correlated noise, the proposed estimator still can give acceptable result, even if it is not optimal. To address the problem of correlation noise, future studies should include noise correlation in consideration of expanding our framework. In the future work chapter, the multivariate DR estimator is proposed by employing

the multivariate GT distribution as noise model. However, experiment is needed to validate the theory, and also the estimation of covariance matrix might need further investigation.

Our studies can also be extended to the state estimation problem as ARMAX filtering and state estimation are similar and both are optimization problems. However, the main difficulty is that in state estimation there exists state noise which is usually unmeasurable.

Bibliography

[1] Peter J. Huber and Elvezio M. Ronchetti. *Robust Statistics*, volume 15 of *Wiley Series in Probability and Statistics*. John Wiley & Sons, Inc., Hoboken, NJ, USA, 2 edition, January 2009. ISBN 9780470434697. doi: 10.1002/9780470434697. URL http://doi.wiley.com/10.1002/9780470434697.

[2] Frank R. Hampel, Elvezio M. Ronchetti, Peter J. Rousseeuw, and Werner A. Stahel. *Robust Statistics: The Approach Based on Influence Functions*. Wiley, 1986. ISBN 978-0-471-73577-9.

[3] Gedeon A Almasy and T. Sztano. Checking and Correction of Measurements on the Basis of Linear System Model. *Problems of Control and Information Theory*, 4:57–69, 1975.

[4] John Wilder Tukey. *Exploratory data analysis*. 1977.

[5] Františěk Madron, Vladimir Veverka, and Vojtěch Vaněček. Statistical analysis of material balance of a chemical reactor. *AIChE Journal*, 23(4):482–486, July 1977. ISSN 0001-1541. doi: 10.1002/aic.690230412. URL http://doi.wiley.com/10.1002/aic.690230412.

[6] C. M. Crowe, Y. A. Garcia Campos, and A. Hrymak. Reconciliation of process flow rates by matrix projection. Part I: Linear case. *AIChE Journal*, 29(6): 881–888, November 1983. ISSN 0001-1541. doi: 10.1002/aic.690290602. URL http://doi.wiley.com/10.1002/aic.690290602.

[7] R. W. Serth and W. a. Heenan. Gross error detection and data reconciliation in steam-metering systems. *AIChE Journal*, 32(5):733–742, May 1986. ISSN 0001-1541. doi: 10.1002/aic.690320503. URL http://doi.wiley.com/10.1002/aic.690320503.

[8] S. Narasimhan and R. S. H. Mah. Generalized likelihood ratio method for gross error identification. *AIChE Journal*, 33(9):1514–1521, September 1987. ISSN 0001-1541. doi: 10.1002/aic.690330911. URL http://doi.wiley.com/10.1002/aic.690330911.

[9] P. Harikumar and S. Narasimhan. A method to incorporate bounds in data reconciliation and gross error detectionâĂŤII. Gross error detection strategies. *Computers & Chemical Engineering*, 17(11):1121–1128, November 1993. ISSN 00981354. doi: 10.1016/0098-1354(93)80093-3. URL http://linkinghub.elsevier.com/retrieve/pii/0098135493800933.

[10] Ricardo A Maronna, R Douglas Martin, and Víctor J. Yohai. *Robust Statistics: Theory and Methods*. Wiley Series in Probability and Statistics. John Wiley & Sons, Ltd, Chichester, UK, March 2006. ISBN 9780470010945. doi: 10.1002/0470010940. URL http://doi.wiley.com/10.1002/0470010940.

[11] Rand R Wilcox. *Introduction to robust estimation and hypothesis testing*. Academic Press, 2012. ISBN 0-12-751542-9.

[12] Moustapha Alhaj-Dibo, Didier Maquin, and José Ragot. Data reconciliation: A robust approach using a contaminated distribution. *Control Engineering Practice*, 16(2):159–170, February 2008. ISSN 09670661. doi: 10.1016/j.conengprac.2007.01.003. URL http://linkinghub.elsevier.com/retrieve/pii/S0967066107000226.

[13] Nikhil Arora and Lorenz T. Biegler. Redescending estimators for data reconciliation and parameter estimation. *Computers & Chemical Engineering*, 25(11-12):1585–1599, November 2001. ISSN 00981354. doi: 10.1016/S0098-1354(01)00721-9. URL http://linkinghub.elsevier.com/retrieve/pii/S0098135401007219.

[14] D Wang and J A Romagnoli. A Framework for Robust Data Reconciliation Based on a Generalized Objective Function. *Industrial & Engineering Chemistry Research*, 42(13):3075–3084, June 2003. ISSN 0888-5885. doi: 10.1021/ie0206655. URL http://pubs.acs.org/doi/abs/10.1021/ie0206655.

[15] Joseph M Steigerwald, Shyam P Murarka, and Ronald J Gutmann. *Chemical Mechanical Planarization of Microelectronic Materials*. Wiley-VCH Verlag GmbH, 2004. ISBN 9783527617746. doi: 10.1002/9783527617746. URL http://dx.doi.org/10.1002/9783527617746.

[16] Sheng-Jyh Shiu, Cheng-Ching Yu, and Shih-Haur Shen. Multivariable control

of multizone chemical mechanical polishing. *Journal of Vacuum Science & Technology B: Microelectronics and Nanometer Structures*, 22(4):1679, 2004. ISSN 0734211X. doi: 10.1116/1.1761483. URL http://link.aip.org/link/JVTBD9/v22/i4/p1679/s1&Agg=doi.

[17] James B McDonald and Whitney K. Newey. Partially adaptive estimation of regression models via the generalized t distribution. *Econometric Theory*, 4 (3):428–457, 1988.

[18] James B. McDonald. Partially adaptive estimation of ARMA time series models. *International Journal of Forecasting*, 5(2):217–230, January 1989. ISSN 01692070. doi: 10.1016/0169-2070(89)90089-7. URL http://linkinghub.elsevier.com/retrieve/pii/0169207089900897.

[19] Olcay Arslan and Ali Äř. Genç. Robust Location and Scale Estimation Based on the Univariate Generalized t (GT) Distribution. *Communications in Statistics - Theory and Methods*, 32(8):1505–1525, January 2003. ISSN 0361-0926. doi: 10.1081/STA-120022242. URL http://www.tandfonline.com/doi/abs/10.1081/STA-120022242.

[20] Christian Hansen, James B. McDonald, and Whitney K. Newey. Instrumental Variables Estimation With Flexible Distributions. *Journal of Business and Economic Statistics*, 28(1):13–25, January 2010. ISSN 0735-0015. doi: 10.1198/jbes.2009.06161. URL http://pubs.amstat.org/doi/abs/10.1198/jbes.2009.06161.

[21] D Wang and J.A. Romagnoli. Generalized T distribution and its applications to process data reconciliation and process monitoring. *Transactions of the Institute of Measurement and Control*, 27(5):367–390, December 2005. ISSN 01423312. doi: 10.1191/0142331205tm155oa. URL `http://tim.sagepub.com/cgi/doi/10.1191/0142331205tm155oa`.

[22] Han Yan. *Temperature Sensing and Control in Multi-zone Semiconductor Thermal Processing*. PhD thesis, National University of Singapore, 2009.

[23] Yen Yen Joe. *Distributed Data Reconciliation and Bias Estimation with non-Gaussian noise for Sensor Network*. PhD thesis, National University of Singapore, 2009.

[24] Han Yan, Weng Khuen Ho, Keck Voon Ling, and Khiang Wee Lim. Multi-Zone Thermal Processing in Semiconductor Manufacturing: Bias Estimation. *IEEE Transactions on Industrial Informatics*, 6(2):216–228, May 2010. ISSN 1551-3203. doi: 10.1109/TII.2010.2040285. URL `http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=5405070`.

[25] Frank R. Hampel. The Influence Curve and its Role in Robust Estimation. *Journal of the American Statistical Association*, 69(346):383–393, June 1974. ISSN 0162-1459. doi: 10.1080/01621459.1974.10482962. URL `http://www.tandfonline.com/doi/abs/10.1080/01621459.1974.10482962`.

[26] Richard A. Groeneveld. An Influence Function Approach to Describing the Skewness of a Distribution. *The American Statistician*, 45(2):97–102, May

1991. ISSN 0003-1305. doi: 10.1080/00031305.1991.10475777. URL http://www.tandfonline.com/doi/abs/10.1080/00031305.1991.10475777.

[27] Mario Romanazzi. Influence Function of Halfspace Depth. *Journal of Multivariate Analysis*, 77(1):138–161, April 2001. ISSN 0047259X. doi: 10.1006/jmva.2000.1929. URL http://linkinghub.elsevier.com/retrieve/pii/S0047259X00919298.

[28] Panayiotis Theodossiou. Financial Data and the Skewed Generalized T Distribution. *Management Science*, 44(12):1650–1661, 1998.

[29] Turan G. Bali and Panayiotis Theodossiou. A conditional-SGT-VaR approach with alternative GARCH models. *Annals of Operations Research*, 151(1):241–267, December 2006. ISSN 0254-5330. doi: 10.1007/s10479-006-0118-4. URL http://www.springerlink.com/index/10.1007/s10479-006-0118-4.

[30] Genshiro Kitagawa. A Self-Organizing State-Space Model. *Journal of the American Statistical Association*, 93(443):1203–1215, September 1998. ISSN 0162-1459. doi: 10.1080/01621459.1998.10473780. URL http://www.tandfonline.com/doi/abs/10.1080/01621459.1998.10473780.

[31] M.S. Arulampalam, S. Maskell, N. Gordon, and T. Clapp. A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking. *IEEE Transactions on Signal Processing*, 50(2):174–188, 2002. ISSN 1053587X. doi: 10.1109/78.978374. URL http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=978374.

[32] Dan Crisan and Boris Rozovskii. *The Oxford handbook of nonlinear filtering*. Oxford University Press, 2011. URL http://econpapers.repec.org/RePEc:oxp:obooks:9780199532902.

[33] L. Le Cam. Maximum likelihood: An introduction. *International Statistical Review*, 58(2):153–171, January 1990. ISSN 1557-170X. doi: 10.1109/IEMBS.2010.5626552. URL http://www.ncbi.nlm.nih.gov/pubmed/21096203.

[34] Geoffrey J. McLachlan and Thriyambakam Krishnan. *The EM Algorithm and Extensions*. Wiley Series in Probability and Statistics. John Wiley & Sons, Inc., Hoboken, NJ, USA, 2 edition, February 2008. ISBN 9780470191613. doi: 10.1002/9780470191613. URL http://doi.wiley.com/10.1002/9780470191613.

[35] Karl Johan Å ström and Björn Wittenmark. *Computer-controlled systems (3rd ed.)*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1997. ISBN 0-13-314899-8.

[36] R. v. Mises. On the Asymptotic Distribution of Differentiable Statistical Functions. *The Annals of Mathematical Statistics*, 18(3):309–348, 1949.

[37] Luisa Turrin Fernholz. On multivariate higher order von Mises expansions. *Metrika*, 53(2):123–140, May 2001. ISSN 0026-1335. doi: 10.1007/s001840000097. URL http://www.springerlink.com/openurl.asp?genre=article&id=doi:10.1007/s001840000097.

[38] Paul H Bardell, William H McAnney, and Jacob Savir. *Built-in test for VLSI: pseudorandom techniques*. Wiley-Interscience, 1987.

[39] H W Sorenson. Least-squares estimation: from Gauss to Kalman. *IEEE Spectrum*, 7(7):63–68, July 1970. ISSN 0018-9235. doi: 10.1109/MSPEC.1970.5213471. URL http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=5213471.

[40] RE Kalman. A new approach to linear filtering and prediction problems. *Journal of basic Engineering*, 82(Series D):35–45, 1960. URL http://wl-s191-122.resnet.ucla.edu/book/basar_control/09.pdf.

[41] Ioan Doré Landau, Rogelio Lozano, Mohammed M'Saad, and Alireza Karimi. *Adaptive Control*. Communications and Control Engineering. Springer London, London, 2011. ISBN 978-0-85729-663-4. doi: 10.1007/978-0-85729-664-1. URL http://www.springer.com.libproxy1.nus.edu.sg/engineering/control/book/978-0-85729-663-4http://www.springerlink.com/index/10.1007/978-0-85729-664-1.

[42] Eduardo Camacho and Carlos Bordons. *Model Predictive Control*. Springer Verlag, 1998. ISBN 3-540-76241-8.

[43] Chuanhai Liu and D.B. Rubin. ML estimation of the t distribution using EM and its extensions, ECM and ECME. *Statistica Sinica*, 5(1):19–39, 1995. URL http://www3.stat.sinica.edu.tw/statistica/oldpdf/A5n12.pdf.

[44] Wang Zhengfeng, Yin Ling, Ng Sum Huan, and Teo Phaik Luan. Chemical mechanical planarization. Technical Report 1, SIMTech, January 2010. URL http://www.ncbi.nlm.nih.gov/pubmed/19928828.

[45] D S Boning, W P Moyne, T H Smith, J Moyne, R Telfeyan, A Hurwitz, S Shellman, and J Tayor. Run by run control of chemical-mechanical polishing. *Components, Packaging, and Manufacturing Technology, Part C, IEEE Transactions on*, 19(4):307–314, 1996. ISSN 1083-4400 VO - 19. doi: 10.1109/3476.558560.

[46] D. R. Kuehn and H. Davidson. Computer control. II. Mathematics of control. *Chemical Engineering Progress*, 57:44–47, 1961.

[47] Cameron M. Crowe. Data reconciliation âĂŤ Progress and challenges. *Journal of Process Control*, 6(2-3):89–98, April 1996. ISSN 09591524. doi: 10.1016/0959-1524(96)00012-1. URL http://linkinghub.elsevier.com/retrieve/pii/0959152496000121.

[48] Jose A Romagnoli and Mabel Cristina Sanchez. *Data processing and reconciliation for chemical process operations*, volume 2. Academic Press, 1999.

[49] Shankar Narasimhan and Cornelius Jordache. *Data reconciliation and gross error detection : an intelligent use of process data*, volume 195. June 1999. ISBN 0884152553. doi: 10.1111/j.1749-6632.1972.tb54778.x. URL http://doi.wiley.com/10.1111/j.1749-6632.1972.tb54778.x.

[50] L.A. Jaeckel. Robust estimates of location: Symmetry and asymmetric contamination. *The Annals of Mathematical Statistics*, 42(3):1020–1034, 1971. URL http://www.jstor.org/stable/10.2307/2240248.

[51] Peter MacCullagh and John A Nelder. *Generalized linear models*. Chapman & Hall, Boca Raton [u.a.], 2 edition, 1999. ISBN 0412317605 9780412317606.

[52] Ana M. Bianco, Marta Garcia Ben, and Víctor J. Yohai. Robust estimation for linear regression with asymmetric errors. *Canadian Journal of Statistics*, 33(4):511–528, December 2005. ISSN 03195724. doi: 10.1002/cjs.5550330404. URL http://doi.wiley.com/10.1002/cjs.5550330404.

[53] Victor J. Yohai and Ricardo A. Maronna. Asymptotic Behavior of M-Estimators for the Linear Model. *The Annals of Statistics*, 7(2):258–268, March 1979. ISSN 0090-5364. doi: 10.1214/aos/1176344610. URL http://projecteuclid.org/euclid.aos/1176344610.

[54] Richard D. F. Harris and C. Coskun Kucukozmen. The Empirical Distribution of UK and US Stock Returns. *Journal of Business Finance & Accounting*, 28 (5-6):715–740, June 2001. ISSN 0306-686X. doi: 10.1111/1468-5957.00391. URL http://doi.wiley.com/10.1111/1468-5957.00391.

[55] Michalis Ioannides, Watson Wyatt Llp, and Harry M Kat. Fitting Distributions to Hedge Fund Index Returns. 2002.

[56] T Bali, H Mo, and Y Tang. The role of autoregressive conditional skewness and kurtosis in the estimation of conditional VaR. *Journal of Banking*

& Finance, 32(2):269–282, February 2008. ISSN 03784266. doi: 10.1016/j.jbankfin.2007.03.009. URL http://linkinghub.elsevier.com/retrieve/pii/S0378426607001823.

[57] Olcay Arslan and Ali Äř. Genç. The skew generalized t distribution as the scale mixture of a skew exponential power distribution and its applications in robust estimation. *Statistics*, 43(5):481–498, October 2009. ISSN 0233-1888. doi: 10.1080/02331880802401241. URL http://tandfprod.literatumonline.com/doi/abs/10.1080/02331880802401241.

[58] M Branco and Dipak K. Dey. A General Class of Multivariate Skew-Elliptical Distributions. *Journal of Multivariate Analysis*, 79(1):99–113, October 2001. ISSN 0047259X. doi: 10.1006/jmva.2000.1960. URL http://linkinghub.elsevier.com/retrieve/pii/S0047259X00919602.

[59] M C Jones and M. J. Faddy. A skew extension of the t-distribution, with applications. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(1):159–174, February 2003. ISSN 1369-7412. doi: 10.1111/1467-9868.00378. URL http://doi.wiley.com/10.1111/1467-9868.00378.

[60] Adelchi Azzalini and Antonella Capitanio. Distributions generated by perturbation of symmetry with emphasis on a multivariate skew t-distribution. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65 (2):367–389, May 2003. ISSN 1369-7412. doi: 10.1111/1467-9868.00391. URL http://doi.wiley.com/10.1111/1467-9868.00391.

[61] G Brys, M Hubert, and a Struyf. A Robust Measure of Skewness. *Journal of Computational and Graphical Statistics*, 13(4):996–1017, December 2004. ISSN 1061-8600. doi: 10.1198/106186004X12632. URL http://pubs.amstat.org/doi/abs/10.1198/106186004X12632.

[62] Guy Brys, Mia Hubert, and Anja Struyf. Goodness-of-fit tests based on a robust measure of skewness. *Computational Statistics*, 23(3):429–442, August 2007. ISSN 0943-4062. doi: 10.1007/s00180-007-0083-7. URL http://www.springerlink.com/index/10.1007/s00180-007-0083-7.

[63] Derya B. Özyurt and Ralph W. Pike. Theory and practice of simultaneous data reconciliation and gross error detection for chemical processes. *Computers & Chemical Engineering*, 28(3):381–402, March 2004. ISSN 00981354. doi: 10.1016/j.compchemeng.2003.07.001. URL http://linkinghub.elsevier.com/retrieve/pii/S0098135403001960.

[64] Kamal Morad, Brent R. Young, and William Y. Svrcek. Rectification of plant measurements using a statistical framework. *Computers & Chemical Engineering*, 29(5):919–940, April 2005. ISSN 00981354. doi: 10.1016/j.compchemeng.2004.07.019. URL http://linkinghub.elsevier.com/retrieve/pii/S0098135404002029.

[65] Diego Martinez Prata, Marcio Schwaab, Enrique Luis Lima, and José Carlos Pinto. Nonlinear dynamic data reconciliation and parameter estimation through particle swarm optimization: Application for an industrial polypropy-

lene reactor. *Chemical Engineering Science*, 64(18):3953–3967, September 2009. ISSN 00092509. doi: 10.1016/j.ces.2009.05.028. URL http://linkinghub.elsevier.com/retrieve/pii/S0009250909003480.

[66] Eduardo Damianik Valdetaro and Roberto Schirru. Simultaneous Model Selection, Robust Data Reconciliation and Outlier Detection with Swarm Intelligence in a Thermal Reactor Power calculation. *Annals of Nuclear Energy*, 38(9):1820–1832, September 2011. ISSN 03064549. doi: 10.1016/j.anucene.2011.06.001. URL http://linkinghub.elsevier.com/retrieve/pii/S030645491100212X.

[67] Heleno Bolfarine Reinaldo B. Arellano-Valle. On some characterizations of the t-distribution. *Statistics & Probability Letters*, 25(1):79–85, 1995.

[68] O Arslan. Family of multivariate generalized t distributions. *Journal of Multivariate Analysis*, 89(2):329–337, May 2004. ISSN 0047259X. doi: 10.1016/j.jmva.2003.09.008. URL http://linkinghub.elsevier.com/retrieve/pii/S0047259X03001647.

[69] E L Lehmann. *Theory of point estimation*. Wiley, New York, 1983. ISBN 0471058491 9780471058496.

[70] A. K. Gupta. Multivariate skew t-distribution. *Statistics*, 37(4):359–363, July 2003. ISSN 0233-1888. doi: 10.1080/715019247. URL http://linkinghub.elsevier.com/retrieve/pii/S0047259X03001313http:

//www.informaworld.com/openurl?genre=article&doi=10.1080/715019247&magic=crossref||D404A21C5BB053405B1A640AFFD44AE3.

[71] Y.L. Tong. *The multivariate normal distribution*. New York : Springer-Verlag, 1990. ISBN 0387970622.

[72] NA Campbell. Robust procedures in multivariate analysis I: Robust covariance estimation. *Applied statistics*, 29(3):231–237, 1980. URL http://www.jstor.org/stable/10.2307/2346896.

[73] R.G. Reynolds. Robust estimation of covariance matrices. *IEEE Transactions on Automatic Control*, 35(9):1047–1051, May 1990. ISSN 00189286. doi: 10.1109/9.58534. URL http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=1420804http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=58534.

Author's Publication

The author has contributed to the following papers

[1] Weng Khuen Ho, **Hoang Dung Vu**, and Keck Voon Ling. Influence Function Analysis of Parameter Estimation with Generalized t Distribution Noise Model. *Industrial & Engineering Chemistry Research*, 52(11):4168–4177, February 2013. ISSN 0888-5885. doi: 10.1021/ie303139k.

[2] **Hoang Dung Vu**. Iterative Algorithms for Data Reconciliation Estimator Using Generalized t -Distribution Noise Model. *Industrial & Engineering Chemistry Research*, 53(4):1478–1488, January 2014. ISSN 0888-5885.

[3] Weng Khuen Ho, Keck Voon Ling, **Hoang Dung Vu**, and Xiaoqiong Wang. Filtering of the ARMAX Process with Generalized t -Distribution Noise: The Influence Function Approach. *Industrial & Engineering Chemistry Research*, 53 (17):7019–7028, April 2014. ISSN 0888-5885. doi: 10.1021/ie401990x.

[4] Weng Khuen Ho, Keck Voon Ling, and **Hoang Dung Vu**. A batch State Estimator with Generalized t-distribution noise model. In *ASEAN Symposium on Automatic Control*, Ho Chi Minh city, Viet Nam, 2011.

Appendix A

Derivation of the Influence Function (2.8) and (3.13)

By taking expectation, Equation (2.5) can be written as

$$\int_{-\infty}^{+\infty} \psi(\varepsilon) f(\varepsilon) d\varepsilon = 0 \tag{A.1}$$

To study the change $\Delta\bar{\theta}$ when the distribution changes from $f(\varepsilon)$ to a new distribution $f_1(\varepsilon)$, replace $f(\varepsilon)$ in Equation (A.1) by $(1-h)f(\varepsilon)+hf_1(\varepsilon)$ where $0 \leq h \leq 1$, giving

$$\int_{-\infty}^{+\infty} \psi(\varepsilon)((1-h)f(\varepsilon)+hf_1(\varepsilon))d\varepsilon = 0$$

Differentiating with respect to h gives

$$\begin{gathered} \frac{\partial}{\partial h}\left(\int_{-\infty}^{+\infty} \psi(\varepsilon)((1-h)f(\varepsilon)+hf_1(\varepsilon))d\varepsilon\right) = 0 \\ \int_{-\infty}^{+\infty} \psi(\varepsilon)(-f(\varepsilon)+f_1(\varepsilon))d\varepsilon + \left(\int_{-\infty}^{+\infty} \frac{\partial\psi(\varepsilon)}{\partial\bar{\theta}}((1-h)f(\varepsilon)+hf_1(\varepsilon))d\varepsilon\right)\frac{\partial\bar{\theta}}{\partial h} = 0 \end{gathered} \tag{A.2}$$

Let $h = 0$ and using Equation (A.1), Equation (A.2) reduces to

$$\left.\frac{\partial\bar{\theta}}{\partial h}\right|_{h=0} = -\left(\int_{-\infty}^{+\infty}\frac{\partial\psi(\varepsilon)}{\partial\bar{\theta}}f(\varepsilon)d\varepsilon\right)^{-1}\int_{-\infty}^{+\infty}\psi(\varepsilon)(f_1(\varepsilon))d\varepsilon \tag{A.3}$$

Let $f_1(\varepsilon) = \delta(\varepsilon)$ an impulse function at ε and (A.3) reduces to the influence function

$$\mathrm{IF}(\varepsilon) = \left.\frac{\partial\bar{\theta}}{\partial h}\right|_{h=0} = -\left(\int_{-\infty}^{\infty}\frac{\partial\psi(\varepsilon)}{\partial\bar{\theta}}f(\varepsilon)d\varepsilon\right)^{-1}\psi(\varepsilon) \tag{A.4}$$

This gives Equations (2.8) in Chapter 2 and (3.13) in Chapter 3.

Appendix B

Derivation of the Recursive Algorithm

To derive the equations, let

$$\Phi(N) = \begin{bmatrix} \phi(1)^T \\ \vdots \\ \phi(N)^T \end{bmatrix}$$

$$Z(N) = \begin{bmatrix} z(1) \\ \vdots \\ z(N) \end{bmatrix}$$

It is assume that the matrix $\Phi(N)^T\Phi(N)$ is regular for all N. The estimate $\hat{x}$ is then given by Equation (3.14)

$$\hat{x} = \left(\Phi(N)^T\Phi(N)\right)^{-1}\Phi(N)^T Z(N) = \left(\sum_{k=1}^{N}\phi(k)\phi(k)^T\right)^{-1}\left(\sum_{k=1}^{N}\phi(k)z(k)\right)$$

Introduce the matrix

$$P(N) = \left(\sum_{k=1}^{N} \phi(k)\phi(k)^T \right)^{-1} = \left(\Phi(N)^T \Phi(N) \right)^{-1}$$

A recursive updating is given by

$$\begin{aligned} P(N)^{-1} &= P(N-1)^{-1} + \phi(N)^T \phi(N) \\ \hat{x} &= P(N) \left(\sum_{k=1}^{N-1} \phi(k) z(k) + \phi(N) z(N) \right) \\ &= P(N) \left(P(N-1)^{-1} \hat{x} + \phi(N) z(N) \right) \\ &= \bar{x} + P(N)\phi(N) \left(z(N) - \phi(k)^T \bar{x} \right) \end{aligned} \tag{B.1}$$

where $\bar{x}$ is the previous estimate of $\hat{x}$. Notice that

$$P(N) = \left(\Phi(N-1)^T \Phi(N-1) + \phi(N)^T \phi(N) \right)^{-1} = \left(P(N-1)^{-1} + \phi(N)^T \phi(N) \right)^{-1} \tag{B.2}$$

Apply of the matrix inversion relation

$$(D + EF)^{-1} = D^{-1} - D^{-1} E \left(I + F D^{-1} E \right)^{-1} F D^{-1}$$

to the expression in Equation (B.2) gives

$$\begin{aligned} P(N) &= \left(P(N-1)^{-1} + \phi(N)^T \phi(N) \right)^{-1} \\ &= P(N-1) - P(N-1)\phi(N) \left(1 + \phi(N)^T P(N-1) \phi(N) \right)^{-1} \phi(N)^T P(N-1) \end{aligned} \tag{B.3}$$

Using Equations (B.1) and (B.3), the derivation for the recursive algorithm in Chapter 3 is complete.

Appendix C

Derivation of Equations (4.10) and (4.11)

Take expectation and Equation (4.7) can be written as

$$a_i^T \lambda + \int_{-\infty}^{+\infty} \psi(\varepsilon_i) f(\varepsilon_i)\, \mathrm{d}\varepsilon_i = 0 \qquad i = 1 \ldots n \tag{C.1}$$

To study the change Δx_i when the distribution changes from $f(\varepsilon_i)$ to $g(\varepsilon_i)$, replace $f(\varepsilon_i)$ by $(1-h)f(\varepsilon_i) + hg(\varepsilon_i)$, $0 \leq h \leq 1$, in Equation (C.1) giving

$$a_i^T \lambda + \int_{-\infty}^{+\infty} \psi(\varepsilon_i) \left((1-h)f(\varepsilon_i) + hg(\varepsilon_i)\right) \mathrm{d}\varepsilon_i = 0 \qquad i = 1 \ldots n \tag{C.2}$$

Differentiating Equations (C.2) and (4.8) with respect to h gives

$$\begin{aligned} 0 &= \int_{-\infty}^{+\infty} \psi(\varepsilon_i) \left(-f(\varepsilon_i) + g(\varepsilon_i)\right) \mathrm{d}\varepsilon_i \\ &\quad + \left(\int_{-\infty}^{+\infty} \frac{d\psi(\varepsilon_i)}{dx_i} \left((1-h)f(\varepsilon_i) + hg(\varepsilon_i)\right) \mathrm{d}\varepsilon_i \right) \frac{dx_i}{dh} \qquad i = 1 \ldots n \end{aligned} \tag{C.3}$$

$$0 = A \begin{bmatrix} \frac{dx_1}{dh} & \ldots & \frac{dx_n}{dh} \end{bmatrix}^T \tag{C.4}$$

Let $h = 0$, $g(\varepsilon_i) = \delta(\varepsilon_i)$ and substract $a_i^T \lambda$ from both sides of Equation (C.3) gives

$$\psi(\varepsilon_i) + \left(\int_{-\infty}^{+\infty} \frac{d}{dx_i} \psi(\varepsilon_i) f(\varepsilon_i) \mathrm{d}\varepsilon_i \right) \mathrm{IF}(\varepsilon_i) = -a_i^T \lambda \qquad i = 1 \ldots n \tag{C.5}$$

$$A \left[\begin{array}{ccc} \mathrm{IF}(\varepsilon_1) & \ldots & \mathrm{IF}(\varepsilon_n) \end{array} \right]^T = 0 \tag{C.6}$$

where

$$\mathrm{IF}(\varepsilon_i) = \left. \frac{dx_i}{dh} \right|_{h=0, g(\varepsilon_i)=\delta(\varepsilon_i)} \tag{C.7}$$

Rewriting Equations (C.5) and (C.6) in matrix form

$$\left[\begin{array}{cc} -\Gamma^{-1} & A^T \\ A & 0 \end{array} \right] \left[\begin{array}{c} \mathrm{IF}(\varepsilon) \\ \lambda \end{array} \right] = \left[\begin{array}{c} -\psi(\varepsilon) \\ 0 \end{array} \right]$$

where

$$\begin{aligned} \Gamma^{-1} &= \mathrm{diag} \left\{ -\textstyle\int_{-\infty}^{+\infty} \frac{d}{dx_1} \psi(\varepsilon_1) f(\varepsilon_1) \mathrm{d}\varepsilon_1, \quad \ldots, \quad -\int_{-\infty}^{+\infty} \frac{d}{dx_n} \psi(\varepsilon_n) f(\varepsilon_n) \mathrm{d}\varepsilon_n \right\} \\ \mathrm{IF}(\varepsilon) &= \left[\mathrm{IF}(\varepsilon_1), \quad \ldots, \quad \mathrm{IF}(\varepsilon_n) \right]^T \end{aligned}$$

and solving it gives

$$\begin{aligned} \mathrm{IF}(\varepsilon) &= (\Gamma - \Gamma A^T (A \Gamma A^T)^{-1} A \Gamma) \psi(\varepsilon) \\ \lambda &= -(A \Gamma A^T)^{-1} A \Gamma \psi(\varepsilon) \end{aligned}$$

From Equation (C.7) using Taylor series expansion

$$\Delta x_i \approx \mathrm{IF}(\varepsilon_i) \Delta h$$

To obtain the expression for the covariance matrix in Equation (4.10) under distribution $g(\varepsilon)$, set $\Delta h = 1$ giving

$$\begin{aligned} \mathrm{E}\{\Delta x \Delta x^T\} &\approx \mathrm{E}\left\{\mathrm{IF}(\varepsilon) \times \mathrm{IF}^T(\varepsilon)\right\} \\ &\approx [\Gamma - \Gamma A^T (A\Gamma A)^{-1} A\Gamma]\mathrm{E}\{\psi(\varepsilon)\psi^T(\varepsilon)\}[\Gamma - \Gamma A^T (A\Gamma A)^{-1} A\Gamma]^T \end{aligned}$$

where

$$\mathrm{E}\{\psi(\varepsilon)\psi^T(\varepsilon)\} = \mathrm{diag}\left\{ \int_{-\infty}^{+\infty} \psi^2(\varepsilon_1) g(\varepsilon_1) d\varepsilon_1, \quad \ldots, \quad \int_{-\infty}^{+\infty} \psi^2(\varepsilon_n) g(\varepsilon_n) d\varepsilon_n \right\}$$

as the off-diagonal elements are zeros because for $i \neq j$, ε_i and ε_j are independent.

Appendix D

Derivation of Equation (5.16)

Take Expectation of (5.10)

$$\int_{-\infty}^{+\infty} \sum_{k=1}^{N} \phi(k)^T \frac{(pq+1)|\varepsilon(k)|^{p-1}\text{sign}(\varepsilon(k))}{(1+\lambda\text{sign}(\varepsilon(k)))^p q\sigma^p + |\varepsilon(k)|^p} f(\varepsilon)d\varepsilon = 0 \tag{D.1}$$

Replacing $f(\varepsilon)$ in (D.1) by $(1-h)f(\varepsilon) + hg(\varepsilon)$ gives

$$\int_{-\infty}^{+\infty} \sum_{k=1}^{N} \phi(k)^T \frac{(pq+1)|\varepsilon(k)|^{p-1}\text{sign}(\varepsilon(k))}{(1+\lambda\text{sign}(\varepsilon(k)))^p q\sigma^p + |\varepsilon(k)|^p} ((1-h)f + hg)d\varepsilon = 0 \tag{D.2}$$

Denote

$$\psi(\varepsilon) = \frac{(pq+1)|\varepsilon(k)|^{p-1}\text{sign}(\varepsilon(k))}{(1+\lambda\text{sign}(\varepsilon(k)))^p q\sigma^p + |\varepsilon(k)|^p}$$

Differentiating (D.2) with respect to h gives

$$\left(\int_{-\infty}^{+\infty} \sum_{k=1}^{N} \phi(k)^T \frac{\partial\psi(\varepsilon)}{\partial\theta}((1-h)f + hg)d\varepsilon\right)\frac{\partial\theta}{\partial h} + \int_{-\infty}^{+\infty} \sum_{k=1}^{N} \phi(k)^T\psi(\varepsilon)(-f+g)d\varepsilon = 0 \tag{D.3}$$

Note that $\varepsilon(k) = y(k) - \phi(k)^T\theta$, hence $\frac{\partial\psi}{\partial\theta} = -\frac{\partial\psi}{\partial\varepsilon}\phi(k)$. At $h = 0$ and $g = \delta(z)$, (D.3) becomes

$$\text{IF}(z) = \left(\int_{-\infty}^{+\infty} \sum_{k=1}^{N} \phi(k)^T \frac{\partial\psi(\varepsilon)}{\partial\varepsilon} f(\varepsilon)d\varepsilon\right)^{-1} \sum_{k=1}^{N} \phi(k)^T\psi(z) \tag{D.4}$$

where

$$\mathrm{IF}(z) = \left.\frac{\partial \theta}{\partial h}\right|_{h=0, g=\delta(z)}$$

Rewriting (D.4) in matrix form gives Equation (5.16)

$$\mathrm{IF}(z) = M^{-1}\Phi^T \frac{(pq+1)|z|^{p-1}\mathrm{sign}(z)}{(1+\lambda \mathrm{sign}(z))^p\, q\sigma^p + |z|^p}$$

where M is

$$M = \Phi^T \Phi \int_{-\infty}^{+\infty} \frac{\partial \psi(\varepsilon)}{\partial \varepsilon} f(\varepsilon) d\varepsilon$$

Appendix E

Derivation of Equation (5.17)

Take expectation of Equation (5.11)

$$\begin{cases} \int_{-\infty}^{+\infty} \sum_{k=1}^{N} \phi(k)^T \psi_1(\varepsilon, \lambda) f(\varepsilon) d\varepsilon & = 0 \\ \int_{-\infty}^{+\infty} \sum_{k=1}^{N} \psi_2(\varepsilon, \lambda) f(\varepsilon) d\varepsilon & = 0 \end{cases} \tag{E.1}$$

where

$$\psi_1(\varepsilon, \lambda) = \frac{(pq+1)|\varepsilon|^{p-1}\text{sign}(\varepsilon)}{(1+\lambda\text{sign}(\varepsilon))^p q\sigma^p + |\varepsilon|^p}$$

$$\psi_2(\varepsilon, \lambda) = \frac{(pq+1)\text{sign}(\varepsilon)|\varepsilon|^p}{q\sigma^p [1+\lambda\text{sign}(\varepsilon)]^{p+1} + [1+\lambda\text{sign}(\varepsilon)] |\varepsilon|^p}$$

Substituting $f(\varepsilon) = (1-h)f(\varepsilon)+hg(\varepsilon)$ into (E.1) and differentiating it with respect to h give

$$\begin{cases} \int_{-\infty}^{+\infty} \sum_{k=1}^{N} \phi(k)^T \psi_1(-f+g)d\varepsilon + \left(\int_{-\infty}^{+\infty} \sum_{k=1}^{N} \phi(k)^T \frac{\partial \psi_1}{\partial \theta}((1-h)f+hg)d\varepsilon \right) \frac{\partial \theta}{\partial h} \\ \qquad + \left(\int_{-\infty}^{+\infty} \sum_{k=1}^{N} \phi(k)^T \frac{\partial \psi_1}{\partial \lambda}((1-h)f+hg)d\varepsilon \right) \frac{\partial \lambda}{\partial h} = 0 \\ \int_{-\infty}^{+\infty} \sum_{k=1}^{N} \psi_2(\varepsilon,\lambda) f(\varepsilon)d\varepsilon + \left(\int_{-\infty}^{+\infty} \sum_{k=1}^{N} \frac{\partial \psi_2}{\partial \theta}((1-h)f+hg)d\varepsilon \right) \frac{\partial \theta}{\partial h} \\ \qquad + \left(\int_{-\infty}^{+\infty} \sum_{k=1}^{N} \frac{\partial \psi_2}{\partial \lambda}((1-h)f+hg)d\varepsilon \right) \frac{\partial \lambda}{\partial h} = 0 \end{cases} \tag{E.2}$$

Substituting $h = 0$ and $g = \delta(z)$ into (E.2) and rewriting it in matrix form give

$$\begin{bmatrix} \int_{-\infty}^{+\infty} \sum_{k=1}^{N} \phi(k)^T \frac{\partial \psi_1}{\partial \varepsilon} \phi(k) f(\varepsilon)d\varepsilon & -\int_{-\infty}^{+\infty} \sum_{k=1}^{N} \phi(k)^T \frac{\partial \psi_1}{\partial \lambda} f(\varepsilon)d\varepsilon \\ \int_{-\infty}^{+\infty} \sum_{k=1}^{N} \frac{\partial \psi_2}{\partial \varepsilon} \phi(k) f(\varepsilon)d\varepsilon & -\int_{-\infty}^{+\infty} \sum_{k=1}^{N} \frac{\partial \psi_2}{\partial \lambda} f(\varepsilon)d\varepsilon \end{bmatrix} \begin{bmatrix} \frac{\partial \theta}{\partial h} \\ \frac{\partial \lambda}{\partial h} \end{bmatrix} = \begin{bmatrix} \sum_{k=1}^{N} \phi(k)^T \psi_1(z) \\ \sum_{k=1}^{N} \psi_2(z) \end{bmatrix}$$

Note that $\frac{\partial \psi_i}{\partial \theta} = -\frac{\partial \psi_i}{\partial \varepsilon}\phi(k)$ with $i = 1, 2$ as $\varepsilon(k) = y(k) - \phi(k)^T\theta$. From [2],

$$\mathrm{IF}(z) = \left. \begin{bmatrix} \frac{\partial \theta}{\partial h} \\ \frac{\partial \lambda}{\partial h} \end{bmatrix} \right|_{h=0, g=\delta(z)}$$

Equation (E.2) then becomes

$$\mathrm{IF}(z) = \begin{bmatrix} \int_{-\infty}^{+\infty} \sum_{k=1}^{N} \phi(k)^T \frac{\partial \psi_1(\varepsilon)}{\partial \varepsilon} f(\varepsilon)\phi(k)d\varepsilon & -\int_{-\infty}^{+\infty} \sum_{k=1}^{N} \phi(k)^T \frac{\partial \psi_1(\varepsilon)}{\partial \lambda} f(\varepsilon)d\varepsilon \\ \int_{-\infty}^{+\infty} \sum_{k=1}^{N} \frac{\partial \psi_2(\varepsilon)}{\partial \varepsilon} f(\varepsilon)\phi(k)d\varepsilon & -\int_{-\infty}^{+\infty} \sum_{k=1}^{N} \frac{\partial \psi_2(\varepsilon)}{\partial \lambda} f(\varepsilon)d\varepsilon \end{bmatrix}^{-1} \times \begin{bmatrix} \sum_{k=1}^{N} \phi(k)^T \psi_1(z) \\ \sum_{k=1}^{N} \psi_2(z) \end{bmatrix}$$

This gives Equation (5.17).

Appendix F

Derivation of Equation (5.23)

Take expectation of the first Equation of (5.22) with assumption of unit batch size

$$\int_{-\infty}^{\infty} \frac{1}{f(\varepsilon)} \frac{\partial f(\varepsilon)}{\partial s} 2\Sigma^{-1} \varepsilon f(\varepsilon) d\varepsilon - A^T \gamma^T = 0 \tag{F.1}$$

Sub $f = (1-h)f + hg$ into (F.1) and diff w.r.t h

$$\int_{-\infty}^{\infty} 2\Sigma^{-1} \frac{\partial \psi(\varepsilon)}{\partial h} \Big((1-h) + hg \Big) d\varepsilon + \int_{-\infty}^{\infty} 2\Sigma^{-1} \psi(\varepsilon)(-f+g) d\varepsilon = 0 \tag{F.2}$$

where $\psi(\varepsilon) = \frac{1}{f(\varepsilon)} \frac{\partial f(\varepsilon)}{\partial s} \varepsilon$. Note that

$$\frac{\partial \psi}{\partial h} = \frac{\partial \psi}{\partial \varepsilon} \frac{\partial \varepsilon}{\partial x} \frac{\partial x}{\partial h} = -\frac{\partial \psi}{\partial \varepsilon} \frac{\partial x}{\partial h} \tag{F.3}$$

Sub (F.3) into (F.2) and let $h = 0$

$$\int_{-\infty}^{\infty} 2\Sigma^{-1} \left(-\frac{\partial \psi}{\partial \varepsilon} \frac{\partial x}{\partial h} \right) f(\varepsilon) d\varepsilon + \int_{-\infty}^{\infty} 2\Sigma^{-1} \psi(\varepsilon)(-f+g) d\varepsilon = 0 \tag{F.4}$$

Add $A^T \gamma^T$ to both sides of (F.4)

$$\int_{-\infty}^{\infty} 2\Sigma^{-1} \left(-\frac{\partial \psi}{\partial \varepsilon} \frac{\partial x}{\partial h} \right) f(\varepsilon) d\varepsilon + \int_{-\infty}^{\infty} 2\Sigma^{-1} \psi(\varepsilon)(-f+g) d\varepsilon + A^T \gamma^T = A^T \gamma^T \tag{F.5}$$

Using (F.1), (F.5) becomes

$$\int_{-\infty}^{\infty} 2\Sigma^{-1}\left(-\frac{\partial\psi}{\partial\varepsilon}\frac{\partial x}{\partial h}\right) f(\varepsilon)d\varepsilon + \int_{-\infty}^{\infty} 2\Sigma^{-1}\psi(\varepsilon)gd\varepsilon = A^T\gamma^T$$

At $g = \delta(z)$

$$\int_{-\infty}^{\infty} 2\Sigma^{-1}\frac{\partial\psi}{\partial\varepsilon}f(\varepsilon)d\varepsilon\frac{\partial x}{\partial h}(z) + A^T\gamma^T = 2\Sigma^{-1}\psi(z) \tag{F.6}$$

Make use of the constraint in (5.20)

$$A\frac{\partial x}{\partial h}(z) = 0 \tag{F.7}$$

Define

$$\mathrm{IF}(z) = \frac{\partial x}{\partial h}(z)$$

(F.6) and (F.7) can be rewritten in matrix form

$$\begin{bmatrix} M & A^T \\ A & 0 \end{bmatrix}\begin{bmatrix} \mathrm{IF}(z) \\ \gamma \end{bmatrix} = \begin{bmatrix} 2\Sigma^{-1}\psi(z) \\ 0 \end{bmatrix} \tag{F.8}$$

where

$$M = \int_{-\infty}^{\infty} 2\Sigma^{-1}\frac{\partial\psi}{\partial\varepsilon}f(\varepsilon)d\varepsilon$$

Solving (F.8) gives

$$\mathrm{IF}(z) = \left(M^{-1} - M^{-1}A\left(AM^{-1}A^T\right)AM^{-1}\right)2\Sigma^{-1}\psi(z) \tag{F.9}$$

Equation (F.9) is the Equation (5.23).