THE EFFECT OF INITIAL CLASSIFICATION ON OUTLIER TESTING IN A LINEAR MODEL OF CONSTANT INTRACLASS CORRELATION

By

ABU HASSAN SHAARI MOHD NOR

Bachelor of Science
Southern Illinois University
Carbondale, Illinois
1979

Master of Science
University of Iowa
Iowa City, Iowa
1981

Submitted to the Faculty of the
Graduate College of the
Oklahoma State University
in partial fulfillment of
the requirements for
the Degree of
DOCTOR OF PHILOSOPHY
July, 1989

THE EFFECT OF INITIAL CLASSIFICATION ON OUTLIER TESTING IN A LINEAR MODEL OF CONSTANT INTRACLASS CORRELATION

Thesis Approved:

Barry Kurt Moser
Thesis Adviser

J. Keary Folks

Ronald W. McJew

Gary Stevens

Kenneth E Case

Norman N. Durham
Dean of Graduate College

ACKNOWLEDGMENTS

I wish to express my gratitute to my adviser, Dr. Barry K. Moser, for suggesting the problem and for his guidance in the preparation of this thesis.

An expression of gratitude is extended to Dr. J. Leroy Folks for serving as the Chairman on my advisory committe. Also my appreciation to Dr. Ronald W. McNew, Dr. Gary R. Stevens and Dr. Kenneth E. Case for being in my advisory committe.

I gratefully acknowledge indebtedness for the financial support from the Universiti Kebangsaan Malaysia, Bangi, Selangor, Malaysia, through the years of my study at the Oklahoma State University.

Finally, my wife and my two children deserve my deepest appreciation for their constant support, understanding and patience.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER I

INTRODUCTION

Srikantan (1961) and Ferguson (1961) were probably the first to use the mean-shift model to identify outlying observations in linear models. Since then, Gentleman and Wilks (1975a,1975b), John and Draper (1978), John (1978), Rosner (1975), Tietjen, More, and Beckman (1973) and others have addressed the problem of outlier testing in linear models. Others, such as Jain (1981b) and Balasooriya and Tse (1986) have considered comparing the powers of some outlier test procedures which have been developed both in normal samples and in linear models situations.

Box and Tiao (1968), introduced a Bayesian approach to outlier detection in linear models. Guttman (1973) and Guttman, Dutter and Freeman (1978) develop an ad hoc Bayesian approach for handling outliers in univariate and multivariate samples using the mean-shift model. Guttman and Katri (1975) extend the work of Guttman (1973) to include scale-change models. Gambino and Guttman (1984) provide a Bayesian approach to deriving the predictive distribution for future observations in the presence of outliers.

An extensive list of references on outliers can be found in a paper by Beckman and Cook (1983). The books by

Barnett and Lewis (1984) and Hawkins (1980) provide a useful survey of the literature.

All of the outlier detection procedures listed above, treat outliers in linear models with independent errors. Moser and Marco (1988), extend the literature by providing an outlier detection procedure for linear models with correlated errors.

The procedures for identifying outliers are subjective in nature (Collett, D. and Lewis, T., 1976). According to Bross (1961), it is more difficult to identify outliers in a patterned experiment as compared to an unpatterned experiment. Bross also stressed the importance of having a working definition of an outlier in a patterned experiment. Hence, the following definitions are provided to clarify the problems of identifying and testing for outliers in linear models.

1) Outlier - Any observation that has not been generated by the mechanism that generated the majority of observations in the data set. (Freeman,P.R. 1979)

2) Inlier - Any observation that has been generated by the mechanism that generated the majority of the data set.

3) Suspected Outlier - Any observation that does not fit the pattern of the data or hypothesized model and the cause of the irregularity is not clear.

4) Suspected Inlier - Any observation that appears to follow the pattern of the data or hypothesized model.

5) Classification - Partitioning of observations into

suspected outlier or suspected inlier groups so that the former can be studied in detail. An observation will be classified into the suspected outlier group if it satisfies Definition 3, otherwise it will be classified into the suspected inlier group.

6) Identification - The process of distinguishing which observations are outliers and which are inliers. The goal is to find outliers to make them available for further study.

Moser and Marco develop an outlier test procedure for a linear model of constant intraclass correlation based on the predictive density of suspected outlier observations given a set of existing inlier observations. This thesis extends their work by investigating the effect of initial classification or misclassification of outlier and inlier observations on the Probability of Correct Identification (PCI). PCI is the probability that the inliers and outliers of a data set are correctly identified.

This thesis consists of five chapters. In Chapter I, a historical background of outlier testing in linear models was presented. In Chapter II, the mathematical background of the problem is developed. A summary of the proposed numerical outputs is presented in Chapter III. Discussion of the results is presented in Chapter IV and then the thesis is briefly summarized in Chapter V.

CHAPTER II

MATHEMATICAL BACKGROUND

The main objective of this thesis is to investigate the effect of initial classification or misclassification of outliers and inliers on the PCI for a linear model of constant intraclass correlation. Since different initial classifications of observations produce different PCI values, six cases of these initial classifications are considered. In addition, the cases are compared so that the consequences of misclassifying observations can be studied in detail.

In Section 2.1, the six cases of initial classifications are defined. Then in Section 2.2, the linear model of constant intraclass correlation is stated. In Sections 2.3 and 2.4, the predictive density and the outlier test procedure for this model are described, respectively. Numerical integrations are used to calculate the PCI values. These numerical integration calculations are presented in Section 2.5.

2.1. Definitions of Cases 1-6

The following six initial classifications are considered. In each case, the observations are either

classified into the suspected outlier or suspected inlier group.

Case 1. One observation is initially classified into the suspected outlier group when all observations are inliers.

Case 2. One observation is initially classified into the suspected outlier group and it is the only outlier.

Case 3. One observation is initially classified into the suspected outlier group but actually there are two outliers in the data set.

Case 4. Two observations are initially classified into the suspected outlier group when all observations are inliers.

Case 5. Two observations are initially classified into the suspected outlier group, one of which is an outlier. All other observations are inliers.

Case 6. Two observations are initially classified into the suspected outlier group and both are outliers. All other observations are inliers.

2.2. The Model

Moser and Marco (1988) develop a procedure for testing suspected outliers when the observations conform to a linear model of constant intraclass correlation. A Bayesian approach to the problem is developed using the predictive distribution of the suspected outliers given the inliers. A test procedure based on this predictive distribution is then derived for testing the suspected outliers.

The procedure is performed as follows: first, the

observations are initially partitioned or classified into two groups; suspected inliers and suspected outliers; next, the test procedure is applied to identify which observations in the suspected outlier group are in fact outliers. Thus, as defined in Chapter I, classification is the initial partitioning of the observations into suspected inlier and suspected outlier groups. Identification, on the other hand, is the process by which Moser and Marco's test procedure distinguishes which observations in the suspected outlier group are outliers. Therefore, the final decision on which observations are inliers and which are outliers is not based on the initial classification of observations but is only made after Moser and Marco's test procedure has been performed. The objective of this thesis is to investigate the effect of different initial classifications of observations on the probability that the test procedure ultimately identifies the outliers correctly. This Probability of Correct Identification is subsequently referred to as PCI.

The following model form, as discussed by Moser and Marco is considered:

$$\underset{\sim}{Y}_i = Z_i \underset{\sim}{\theta} + T_i \underset{\sim}{\beta} + \underset{\sim}{\varepsilon}_i \qquad (2.1)$$

for the i^{th} class, $i=1,\ldots,r$. $\underset{\sim}{Y}_i$ is $(n_i+m_i)x1$ random vector of observations, $n_i \geq 1$, $m_i \geq 0$, $\Sigma\, n_i = N$, $\Sigma\, m_i = M$. There are N+M N+M total observations in the data set, N suspected inliers and M suspected outlier. Z_i is an $(n_i+m_i)xp$ matrix of

independent variables taking the form $\underset{\sim}{1}_{n_i+m_i}\underset{\sim}{z}_i$ with $\underset{\sim}{z}_i$ a (1xp) vector, and T_i is an (n_i+m_i)xq matrix of covariates. $\underset{\sim}{\theta}$, $\underset{\sim}{\beta}$ is (px1) and (qx1) vectors of unknown parameters, respectively and $\underset{\sim}{\varepsilon}_i$ is (n_i+m_i)x1 vector of random errors.

Assumptions

The following assumptions are made concerning observations from linear model (2.1).

1) Observations in different classes are independent, while observations in the same class are equicorrelated.

2) $\underset{\sim}{\varepsilon}_i \sim N(\underset{\sim}{0}, \Sigma)$, where $\Sigma=\sigma^2[(1-\rho)I_{(n_i+m_i)}+\rho J_{(n_i+m_i)}]$ with $\sigma^2>0$, and $-1/(n_i+m_i)<\rho<1$.

3) M observations are classified a priori as suspected outliers and N observations as suspected inliers.

2.3. The Predictive Density

Following Moser and Marco, each vector $\underset{\sim}{Y}_i$ is partitioned into $(\underset{\sim}{Y}_i^{(1)'}, \underset{\sim}{Y}_i^{(2)'})$ where $\underset{\sim}{Y}_i^{(1)}$ is an n_ix1 vector of suspected inliers in the i^{th} class and $\underset{\sim}{Y}_i^{(2)}$ is an m_ix1 vector of suspected outliers. Then assuming that the Nx1 vector $\underset{\sim}{Y}^{(1)}$ of suspected inliers contains only inlier observations, they have shown that under H_o: all of the suspected outliers are inliers, the predictive density of $\underset{\sim}{Y}^{(2)}|\underset{\sim}{Y}^{(1)}$, with noninformative prior

$$p(\underset{\sim}{\theta}, \underset{\sim}{\beta}, \sigma^2, \rho) \propto [\sigma^2(1-\rho)]^{-1} \qquad (2.2)$$

is an M-dimensional multivariate t distribution, with N-r-q

degrees of freedom, location vector $\underset{\sim}{\mu}$ and scale matrix $\mathfrak{D}$. The location vector $\underset{\sim}{\mu}$ and scale matrix $\mathfrak{D}$ are presented in the following forms;

$$\underset{\sim}{\mu} = \bar{\underset{\sim}{Y}}^{(1)} + (T^{(2)} - \bar{T}^{(1)})\hat{\underset{\sim}{\beta}} \tag{2.3}$$

where,

$$\hat{\underset{\sim}{\beta}} = (T^{(1)\prime} W T^{(1)})^{-1} T^{(1)\prime} W \underset{\sim}{Y}^{(1)} , \tag{2.4}$$

$$\bar{Y}^{(1)} = (\bar{Y}_1^{(1)\prime}, \ldots, \bar{Y}_r^{(1)\prime})'_{Mx1} , \quad \bar{Y}_i^{(1)} = (1/n_i) 1_{m_i} 1'_{n_i} Y_i^{(1)} .$$

$$\mathfrak{D} = \hat{s}^2 \{ I_M + \Delta_M + (T^{(2)} - \bar{T}^{(1)})(T^{(1)\prime} W T^{(1)})^{-1} (T^{(2)} - \bar{T}^{(1)})' \} \tag{2.5}$$

where,

$$\hat{s}^2 = (N-r-q)^{-1} (\underset{\sim}{Y}^{(1)} - T^{(1)} \hat{\underset{\sim}{\beta}})' W (\underset{\sim}{Y}^{(1)} - T^{(1)} \hat{\underset{\sim}{\beta}}), \tag{2.6}$$

Δ_M is an MxM block matrix whose i^{th} block is $(1/n_i) J_{m_i}$,

$T^{(1)} = (T_1^{(1)\prime}, \ldots, T_r^{(1)\prime})'_{Nxq}$, $T^{(2)} = (T_1^{(2)\prime}, \ldots, T_r^{(2)\prime})'_{Mxq}$,

$\bar{T}^{(1)} = (\bar{T}_1^{(1)\prime}, \ldots, \bar{T}_r^{(1)\prime})'_{Mxq}$, $\bar{T}_i^{(1)} = (1/n_i)\ 1_{m_i} 1'_{n_i} T_i^{(1)}$ and W is an NxN block diagonal matrix whose i^{th} block is $I_{n_i} - (1/n_i) J_{n_i}$. Hence, following Berger(1980) the predictive density of $\underset{\sim}{Y}^{(2)} | \underset{\sim}{Y}^{(1)}$ can be presented as

$$f(\underset{\sim}{Y}^{(2)}, \nu, \underset{\sim}{\mu}, \mathfrak{D})$$

$$= \begin{cases} C\left[1 + (1/\nu)(\underset{\sim}{Y}^{(2)} - \underset{\sim}{\mu})' \mathfrak{D}^{-1} (\underset{\sim}{Y}^{(2)} - \underset{\sim}{\mu})\right]^{-(\nu+M)/2}, & \underset{\sim}{Y}^{(2)} \in R^M \\ 0 \qquad \text{otherwise} \end{cases} \tag{2.7}$$

where $C=\dfrac{\Gamma[(\nu+M)/2]\,|\mathfrak{D}|^{-1/2}}{\Gamma(\nu/2)(\nu\pi)^{M/2}}$, $\nu=N-r-q>0$,

$\underset{\sim}{\mu}\varepsilon R^M$ and $\mathfrak{D}$ is an MxM positive definite symmetric matrix. When M=2, the density function in (2.7) is a bivariate t distribution.

The distribution of $\underset{\sim}{Y}^{(2)}|\underset{\sim}{Y}^{(1)}$ under H_1, assuming all elements of $\underset{\sim}{Y}^{(1)}$ are inliers is derived in a similar fashion. Assuming the perturbation in the outlier observation is caused by a shift in the mean, then under H_1 the predictive density of $\underset{\sim}{Y}^{(2)}|\underset{\sim}{Y}^{(1)}$ with prior (2.2), is an M-dimensional multivariate t distribution with N-r-q degrees of freedom, location vector

$$\underset{\sim}{\mu}=\bar{\underset{\sim}{Y}}^{(1)}+(T^{(2)}-\bar{T}^{(1)})\hat{\underset{\sim}{\beta}}+\underset{\sim}{a} \qquad (2.8)$$

where $\underset{\sim}{a}=(a_1,\ldots,a_M)'$, $\underset{\sim}{a}\varepsilon R^M$ is a vector of unknown shift parameters for the M suspected outliers, and scale matrix $\mathfrak{D}$ as given in (2.5).

Below is a summary of the predictive density derived in this section.

1) Outlier Model: The mean-shift model for the outliers can be presented in the following way.

i) $E(\underset{\sim}{Y}^{(1)})=Z^{(1)}\underset{\sim}{\theta}+T^{(1)}\underset{\sim}{\beta}$

ii) $E(\underset{\sim}{Y}^{(2)})=Z^{(2)}\underset{\sim}{\theta}+T^{(2)}\underset{\sim}{\beta}+\underset{\sim}{a}$

2) Predictive Density, assuming a mean-shift model(All

Cases except Case 3):

Under H_0: all of the suspected outliers are inliers,

$\underset{\sim}{Y}^{(2)}|\underset{\sim}{Y}^{(1)} \sim MVt(\underset{\sim}{\mu}, \mathbb{D})$

Under H_1: at least one of the suspected outliers is an outlier,

$\underset{\sim}{Y}^{(2)}|\underset{\sim}{Y}^{(1)} \sim MVt(\underset{\sim}{\mu}+\underset{\sim}{a}, \mathbb{D})$.

2.4. The Test

With the knowledge of the distribution of $\underset{\sim}{Y}^{(2)}|\underset{\sim}{Y}^{(1)}$, Moser and Marco develop a test procedure for detecting the presence of outliers in the suspected outlier group. The hypotheses of interest are

H_0: all of the suspected outliers are inliers.

H_1: at least one of the suspected outliers is an outlier.

Following Berger (1980), under H_0 the random variable $F^*=(1/M)(\underset{\sim}{Y}^{(2)}-\underset{\sim}{\mu})'\mathbb{D}^{-1}(\underset{\sim}{Y}^{(2)}-\underset{\sim}{\mu})$ has an F distribution with M and N-r-q degrees of freedom. Hence, an α level rejection rule for testing H_0 is to reject if

$$F^* > F^{\alpha}_{M,N-r-q} \quad . \tag{2.9}$$

When H_0 is rejected, a Bonferroni multiple comparison procedure is used to identify which elements of the (Mx1) vector of suspected outliers $\underset{\sim}{Y}^{(2)}$ are outliers. The following test statistic was used (Moser and Marco 1988, equation 14).

$$v_{ij}=$$

$$\frac{\xi_{ij}'(\underset{\sim}{Y}^{(2)}-\underset{\sim}{\mu})}{\hat{s}\{1+1/n_i+\xi_{ij}'(T^{(2)}-\bar{T}^{(1)})(T^{(1)'}WT^{(1)})^{-1}(T^{(2)}-\bar{T}^{(1)})'\xi_{ij}\}^{1/2}}$$

where v_{ij} has a univariate t distribution with N-r-q degrees of freedoms, ξ_{ij} is an Mx1 vector of zeros except for a one corresponding to the j^{th} suspected outlier in the i^{th} class, and all other terms are defined as before. By the Bonferroni procedure, one concludes that the j^{th} suspected outlier in the i^{th} class is an outlier if,

$$|v_{ij}| \geq t_{\alpha/2M,N-r-q} \cdot \qquad (2.10)$$

2.5. Numerical Integrations

In this section, the integrals used to calculate the Probability of Correct Identification (PCI) for Cases 1-6 are defined. In each case (except for Case 3), the integrand is the predictive density of the suspected outliers given the inlier observations. To reduce the number of parameters involved in the numerical integrations of the PCI, the random vector $\underset{\sim}{Y}^{(2)}$ is standardized.

Let $\underset{\sim}{X}=(X_1,X_2,\ldots,X_M)'$ be an Mx1 random vector with

$$X_i=\frac{(Y_i^{(2)}-\mu_i-a_i)}{(d_{ii})^{1/2}}, \quad i=1,\ldots,M,$$

where d_{ii} is the i^{th} diagonal element of $\mathfrak{D}$ given in (2.5), $Y_i^{(2)}$ is the i^{th} suspected outlier, μ_i is the i^{th} element of $\underset{\sim}{\mu}$, given in (2.3), and a_i is the i^{th} element of $\underset{\sim}{a}$ given in (2.8).

In Cases 1-3, M=1; thus $\underset{\sim}{X}=X_1=X$, where the subscript is dropped for convenience. For Cases 4-6, M=2 so that $\underset{\sim}{X}= (X_1, X_2)'$. Hence, under H_0 and assuming $\underset{\sim}{Y}^{(1)}$ is a vector of inliers, the random variable X follows a univariate t distribution with N-r-q degrees of freedom and the random vector $\underset{\sim}{X} = (X_1 , X_2)'$ follows a bivariate t distribution with N-r-q degrees of freedom, location vector $\underset{\sim}{\mu}=(0,0)'$ and scale matrix

$$D= \begin{bmatrix} 1 & \delta \\ \delta & 1 \end{bmatrix} , \qquad (2.11)$$

where

$$\delta= \frac{d_{12}}{(d_{11})^{1/2}\,(d_{22})^{1/2}} , \quad -1<\delta<1.$$

Define

$$F^*=(1/2)(X_1, X_2)D^{-1}(X_1, X_2)' . \qquad (2.12)$$

Then, following Berger under H_0, F^* has an F distribution with 2 and N-r-q degrees of freedom.

Now, the PCI for the six cases will be calculated based on the predictive density derived in this section and

test procedures in (2.9)-(2.10). For convenience, denote ν=N-r-q as the degrees of freedom.

Case 1(M=1)

One observation is initially classified into the suspected outlier group when all observations are inliers. In this case the PCI is the probability that the one observation in the suspected outlier group is identified as as an inlier. Hence,

PCI= P(accept H_0 and identify X as the inlier)

$=P(|X|<t_{\alpha/2,\nu})$

$$=\int_{-t_{\alpha/2,\nu}}^{t_{\alpha/2,\nu}} f(x,\nu)dx = 1-\alpha \text{ , for all } \nu>0, \qquad (2.13)$$

where $f(x,\nu)$ denotes the univariate t distribution with ν degrees of freedom. The limits of integration are obtained from the Student's t table with the appropriate α and ν.

Case 2 (M=1)

One observation is initially classified into the suspected outlier group and it is the only outlier. In this case the PCI is the probability the observation in the suspected outlier group is identified as an outlier. Hence,

PCI= P(reject H_0 and identify X as the outlier)

$=P(|X|>t_{\alpha/2,\nu})$

$$=\int_{-\infty}^{-t_{\alpha/2,\nu}} f(x,a^*,\nu)dx + \int_{t_{\alpha/2,\nu}}^{\infty} f(x,a^*,\nu)dx, \qquad (2.14)$$

where $f(x,a^*,\nu)$ denotes the univariate t distribution with

ν degrees of freedom, location parameter $a^*=a/d^{1/2}$. Thus t is a shifted central t distribution with a shifted location parameter, a^* (not a noncentral t distribution). As in Case 1, the limits of integration for this case are also obtained from the Student's t table.

Case 3 (M=1)

One observation is initially classified into the suspected outlier group when there are two outliers in the data set. The predictive density of $\underset{\sim}{Y}^{(2)}|\underset{\sim}{Y}^{(1)}$, as given by equation (2.7) is not applicable here since one of the observations in $\underset{\sim}{Y}^{(1)}$ is an outlier. Hence, a separate predictive density is needed for calculating the PCI of Case 3.

In this case, the effect of misclassifying one outlier observation into the suspected inlier group on the location and scale parameters of the predictive density is investigated. Then, a brief discussion on the probability of correctly identifying the one observation in the suspected outlier group is presented.

Denote $\underset{\sim}{Y}_O^{(1)}$ as the Nx1 vector of observations from the suspected inlier group where one of the observations is an outlier. Without loss of generality, the vector $\underset{\sim}{Y}_O^{(1)}$ can be presented as

$$\underset{\sim}{Y}_O^{(1)}=\underset{\sim}{Y}^{(1)}+a_2\underset{\sim}{\delta} \qquad (2.15)$$

$\underset{\sim}{Y}^{(1)}$ represents an Nx1 vector of suspected inliers, where all of its elements are inliers (this vector is the same as that given in (2.4) and (2.6)); a_2 is an unknown parameter defined such that

$$E(\underset{\sim}{Y}_O^{(1)})-E(\underset{\sim}{Y}^{(1)})=a_2\underset{\sim}{\gamma}\ , \tag{2.16}$$

where $\underset{\sim}{\gamma}$ is an Nx1 vector of 0's with a '1' corresponding to the one outlier in the suspected inlier group.

Under H_O and assuming all elements of $\underset{\sim}{Y}^{(1)}$ are inliers the predictive density of $\underset{\sim}{Y}^{(2)}|\underset{\sim}{Y}^{(1)}$ for M=1 is a univariate t distribution with ν degrees of freedom, location parameter

$$\mu=\bar{\underset{\sim}{Y}}^{(1)}+(\underset{\sim}{T}^{(2)}-\bar{\underset{\sim}{T}}^{(1)})\hat{\underset{\sim}{\beta}} \tag{2.17}$$

and scale parameter

$$d=\hat{s}^2\left\{1+1/n_i+(\underset{\sim}{T}^{(2)}-\bar{\underset{\sim}{T}}^{(1)})(T^{(1)\prime}WT^{(1)})^{-1}(\underset{\sim}{T}^{(2)}-\bar{\underset{\sim}{T}}^{(1)})'\right\}. \tag{2.18}$$

Note that the location and scale parameters are derived from (2.3) and (2.5) respectively by letting M=1. Following Berger (1980), this predictive density can be presented as

$$f(y^{(2)},\nu,\mu,d)=\frac{\Gamma[(\nu+1)/2]}{(d\nu\pi)^{1/2}\ \Gamma(\nu/2)}\left\{1+(1/\nu d)(y^{(2)}-\mu)^2\right\}^{-(\nu+1)/2} \tag{2.19}$$

where $y^{(2)}\varepsilon\ R$, $\nu>0$, $-\infty<\mu<\infty$ and $d>0$.

Using (2.15) - (2.19) and the results from Box and Tiao (1968), the predictive density of $Y^{(2)}|\underset{\sim}{Y}_O^{(1)}$ under H_O with prior (2.2) is a univariate t with ν degrees of

freedom, location and scale parameters, say μ_o and d_o, respectively. In the following paragraph, μ_o and d_o are derived as a function of μ and d, respectively for Case 3.

The location parameter μ_o is given as

$$\mu_o=\bar{Y}_o^{(1)}+(\underset{\sim}{T}^{(2)}-\bar{\underset{\sim}{T}}^{(1)})\hat{\underset{\sim}{\beta}}_o \ , \qquad (2.20)$$

where $\bar{Y}_o^{(1)}=\bar{Y}^{(1)}+\varphi(a_2/n_i)$, $i=1,\ldots,r$

φ is defined as

$$\varphi = \begin{cases} 0 & \text{if the outlier in the suspected outlier group is in a different class from the outlier in the suspected inlier group,} \\ 1 & \text{if the outlier in the suspected outlier group is in the same class as the outlier in the suspected inlier group.} \end{cases}$$

$$\hat{\underset{\sim}{\beta}}_o=(T^{(1)'}WT^{(1)})^{-1}T^{(1)'}W\underset{\sim}{Y}_o^{(1)}$$

$$=(T^{(1)'}WT^{(1)})^{-1}T^{(1)'}W(\underset{\sim}{Y}^{(1)}+a_2\underset{\sim}{\gamma})$$

$$=(T^{(1)'}WT^{(1)})^{-1}T^{(1)'}W\underset{\sim}{Y}^{(1)}+a_2(T^{(1)'}WT^{(1)})^{-1}T^{(1)'}W\underset{\sim}{\gamma}$$

$$\therefore\ \hat{\underset{\sim}{\beta}}_o=\hat{\underset{\sim}{\beta}}+a_2(T^{(1)'}WT^{(1)})^{-1}T^{(1)'}W\underset{\sim}{\gamma}.$$

Subsitute for $\bar{Y}_o^{(1)}$ and $\hat{\underset{\sim}{\beta}}_o$ into (2.20); thus μ_o can be written as

$$\mu_o=[\bar{Y}^{(1)}+\varphi(a_2/n_i)]+(\underset{\sim}{T}^{(2)}-\bar{\underset{\sim}{T}}^{(1)})\{\hat{\underset{\sim}{\beta}}+a_2(T^{(1)'}WT^{(1)})^{-1}T^{(1)'}W\underset{\sim}{\gamma}\}$$

$$=\left\{\ \bar{Y}^{(1)}+(\underset{\sim}{T}^{(2)}-\bar{\underset{\sim}{T}}^{(1)})\hat{\underset{\sim}{\beta}}\ \right\}+\varphi(a_2/n_i)+a_2(\underset{\sim}{T}^{(2)}-\bar{\underset{\sim}{T}}^{(1)})$$

$$(T^{(1)\prime}WT^{(1)})^{-1}T^{(1)\prime}W\underset{\sim}{Y}$$

$$= \mu + \varphi(a_2/n_i) + a_2(\underset{\sim}{T}^{(2)} - \bar{\underset{\sim}{T}}^{(1)})(T^{(1)\prime}WT^{(1)})^{-1}T^{(1)\prime}W\underset{\sim}{Y}.$$

Hence,

$$\mu_O = \mu + a_2\left\{\varphi/n_i + (\underset{\sim}{T}^{(2)} - \bar{\underset{\sim}{T}}^{(1)})(T^{(1)\prime}WT^{(1)})^{-1}T^{(1)\prime}W\underset{\sim}{Y}\right\} \qquad (2.21)$$

where μ is the mean of the predictive density of $Y^{(2)}|\underset{\sim}{Y}^{(1)}$ defined in (2.17).

The scale parameter d_O is given as

$$d_O = \hat{s}_O^2\left\{1 + 1/n_i + (\underset{\sim}{T}^{(2)} - \bar{\underset{\sim}{T}}^{(1)})(T^{(1)\prime}WT^{(1)})^{-1}(\underset{\sim}{T}^{(2)} - \bar{\underset{\sim}{T}}^{(1)})'\right\} \qquad (2.22)$$

where $\hat{s}_O^2 = (1/\nu)(\underset{\sim}{Y}_O^{(1)} - T^{(1)}\hat{\underset{\sim}{\beta}}_O)'W(\underset{\sim}{Y}_O^{(1)} - T^{(1)}\hat{\underset{\sim}{\beta}}_O)$.

Substitute for $\underset{\sim}{Y}_O^{(1)}$ and $\hat{\underset{\sim}{\beta}}_O$; then

$$\hat{s}_O^2 = (1/\nu)\{\underset{\sim}{Y}^{(1)} + a_2\underset{\sim}{Y} - T^{(1)}(\hat{\underset{\sim}{\beta}} + a_2(T^{(1)\prime}WT^{(1)})^{-1}T^{(1)\prime}W\underset{\sim}{Y})\}'W$$

$$\{\underset{\sim}{Y}^{(1)} + a_2\underset{\sim}{Y} - T^{(1)}(\hat{\underset{\sim}{\beta}} + a_2(T^{(1)\prime}WT^{(1)})^{-1}T^{(1)\prime}W\underset{\sim}{Y})\}$$

$$= (1/\nu)\{(\underset{\sim}{Y}^{(1)} - T^{(1)}\hat{\underset{\sim}{\beta}}) + (a_2\underset{\sim}{Y} - a_2T^{(1)}(T^{(1)\prime}WT^{(1)})^{-1}T^{(1)\prime}W\underset{\sim}{Y})\}'W$$

$$\{(\underset{\sim}{Y}^{(1)} - T^{(1)}\hat{\underset{\sim}{\beta}}) + (a_2\underset{\sim}{Y} - a_2T^{(1)}(T^{(1)\prime}WT^{(1)})^{-1}T^{(1)\prime}W\underset{\sim}{Y})\}$$

$$= (1/\nu)(\underset{\sim}{Y}^{(1)} - T^{(1)}\hat{\underset{\sim}{\beta}})'W(\underset{\sim}{Y}^{(1)} - T^{(1)}\hat{\underset{\sim}{\beta}})$$

$$+ (1/\nu)\{a_2\underset{\sim}{Y} - a_2T^{(1)}(T^{(1)\prime}WT^{(1)})^{-1}T^{(1)\prime}W\underset{\sim}{Y}\}'W$$

$$\{a_2\underset{\sim}{Y} - a_2T^{(1)}(T^{(1)\prime}WT^{(1)})^{-1}T^{(1)\prime}W\underset{\sim}{Y}\} + (2/\nu)(\underset{\sim}{Y}^{(1)} - T^{(1)}\hat{\underset{\sim}{\beta}})'$$

$$(a_2\underset{\sim}{Y} - a_2T^{(1)}(T^{(1)\prime}WT^{(1)})^{-1}T^{(1)\prime}W\underset{\sim}{Y}).$$

Thus, $\hat{s}_0^2=\hat{s}^2 + K_1+K_2$, where

$$K_1=(1/\nu)\{a_2\underset{\sim}{y}-a_2T^{(1)}(T^{(1)'}WT^{(1)})^{-1}T^{(1)'}W\underset{\sim}{y}\}'W$$

$$\{a_2\underset{\sim}{y}-a_2T^{(1)}(T^{(1)'}WT^{(1)})^{-1}T^{(1)'}W\underset{\sim}{y}\} ,$$

$$K_2=(2/\nu)(\underset{\sim}{Y}^{(1)}-T^{(1)}\hat{\underset{\sim}{\beta}})'(a_2\underset{\sim}{y}-a_2T^{(1)}(T^{(1)'}WT^{(1)})^{-1}T^{(1)'}W\underset{\sim}{y}),$$

and $K_1>0, K_2>0$ for all $a_1>0$ and $a_2>0$.

Substitute for $\hat{s}_0^2$ into (2.22); then

$$d_0=(\hat{s}^2+K_1+K_2)\left\{1+1/n_i+(\underset{\sim}{T}^{(2)}-\bar{\underset{\sim}{T}}^{(1)})(T^{(1)'}WT^{(1)})^{-1}(\underset{\sim}{T}^{(2)}-\bar{\underset{\sim}{T}}^{(1)})'\right\}$$

$$=s^2\left\{1+1/n_i+(\underset{\sim}{T}^{(2)}-\bar{\underset{\sim}{T}}^{(1)})(T^{(1)'}WT^{(1)})^{-1}(\underset{\sim}{T}^{(2)}-\bar{\underset{\sim}{T}}^{(1)})'\right\}+ K_3$$

$$\therefore\ d_0=d+K_3 \qquad (2.23)$$

where

$$K_3=(K_1+K_2)\left\{1+1/n_i+(\underset{\sim}{T}^{(2)}-\bar{\underset{\sim}{T}}^{(1)})(T^{(1)'}WT^{(1)})^{-1}(\underset{\sim}{T}^{(2)}-\bar{\underset{\sim}{T}}^{(1)})'\right\}>0$$

and d is the scale parameter of the predictive density of $Y^{(2)}|\underset{\sim}{Y}^{(1)}$ given in (2.18).

Therefore, under H_0 the predictive density of $Y^{(2)}|\underset{\sim}{Y}_0^{(1)}$, using prior (2.2) with M=1, is a univariate t distribution with ν degrees of freedom, location parameter μ_0 given in (2.21) and scale parameter d_0 given in (2.23). From (2.21), (2.23) one can observe that misclassifying one outlier into the suspected inlier group (failure to classify the outlier into the suspected outlier group) has made the location and scale parameters larger, i.e. $\mu_0>\mu$

and $d_o > d$.

A summary on the derivation of the predictive density for Case 3 is presented below:

1) Outlier Model: The mean-shift model for the outliers in both groups can be presented in the following way.

$$E(\underset{\sim}{Y}_o^{(1)}) = Z^{(1)}\underset{\sim}{\theta} + T^{(1)}\underset{\sim}{\beta} + a_2\underset{\sim}{\gamma}$$

$$E(Y^{(2)}) = Z^{(2)}\underset{\sim}{\theta} + T^{(2)}\underset{\sim}{\beta} + a_1$$

2) Predictive Density(using prior in 2.2).

Under H_o: the suspected outlier is an inlier,

$$Y^{(2)} | \underset{\sim}{Y}_o^{(1)} \sim t(\mu_o, d_o).$$

Under H_1: the suspected outlier is an outlier,

$Y^{(2)} | \underset{\sim}{Y}_o^{(1)} \sim t(\mu_o + a_1, d_o)$, where a_1 is the shift of the outlier observation in the suspected outlier group.

Now a brief discussion of the PCI for Case 3 is presented (for convenience, the discussion of the PCI is based on the predictive density of $Y^{(2)} | \underset{\sim}{Y}_o^{(1)}$, with M=1 and q=1). Recall that in Chapter I, the PCI was defined as the probability that the inliers and outliers of a data set are correctly identified. Test procedures (2.9) and (2.10) only identify outliers in the suspected outlier group. Therefore if one outlier is misclassified into the suspected inlier group, as in Case 3, the PCI is zero.

Although the PCI as defined in Chapter I is zero for

Case 3, it is still possible for the test procedures in (2.9) and (2.10) to identify correctly the one observation in the suspected outlier group. The following definitions will be useful in discussing the PCI of the one observation in the suspected outlier group for Case 3.

1) PCI* of Case 3- the probability that the outlier observation in the suspected outlier group is correctly identified for Case 3=P(reject H_0 and identify the observation in the suspected outlier group as an outlier).

2) a_1 is the shift of the outlier observation in the suspected outlier group, ($a_1 \geq 0$).

3) a_2 is the shift of the outlier observation in the suspected inlier group, ($a_2 \geq 0$).

4) Distribution 1 is the predictive distribution of $Y^{(2)} | \underset{\sim}{Y}_0^{(1)}$ under H_0 when M=1, q=1, with one observation in $\underset{\sim}{Y}_0^{(1)}$ being an outlier.

5) Distribution 2 is the predictive distribution of $Y^{(2)} | \underset{\sim}{Y}_0^{(1)}$ under H_1 when M=1, q=1, with one observation in $\underset{\sim}{Y}_0^{(1)}$ being an outlier.

6) Region A is the acceptance region for testing H_0, when one of the observations in $\underset{\sim}{Y}_0^{(1)}$ is an outlier. The area of this region under Distribution 1 is equal to $1-\alpha$.

In Figure 1.1, the predictive distributions of

$Y^{(2)}|\underset{\sim}{Y}_o^{(1)}$ with M=1, q=1 are presented. No numerical integration is done for Case 3. However, the following discussions of the PCI*, based on the distributions shown in Figure 1.1 are presented.

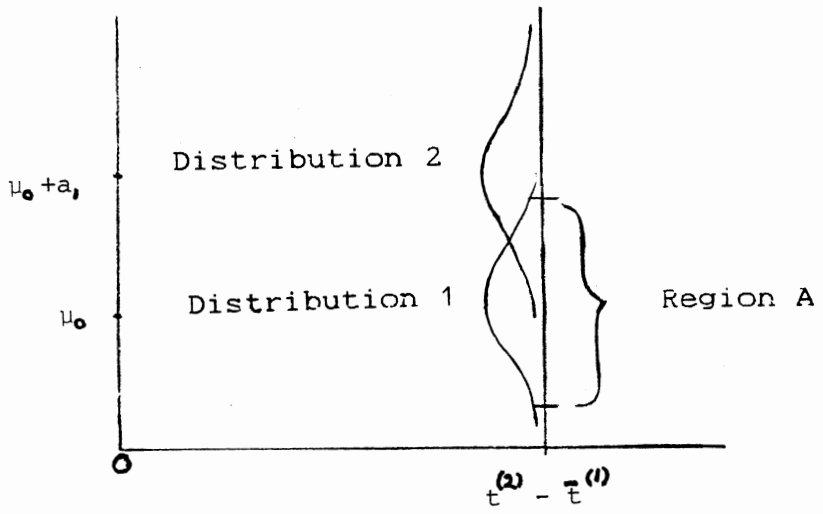

Figure 1.1. Predictive Distributions for Case 3

In Case 3, $a_1>0$ and $a_2>0$; i.e. the observation in the suspected outlier group is an outlier and one observation in the suspected inlier group is an outlier. Hence,

PCI*={area under Distribution 2, outside of Region A}.

=P(reject H_o and identify $Y^{(2)}$ as the outlier)

$$=P(|Y^{(2)}|>t^*_{\alpha/2,\nu})$$

$$=C_1\int_{-\infty}^{-t^*_{\alpha/2,\nu}}\left\{1+(1/d_o\nu)(y^{(2)}-\mu_o^*)^2\right\}^{-(\nu+1)/2}dy^{(2)}$$

$$+C_1\int_{t^*_{\alpha/2,\nu}}^{\infty}\left\{1+(1/d_o\nu)(y^{(2)}-\mu_o^*)^2\right\}^{-(\nu+1)/2}dy^{(2)}, \qquad (2.24)$$

where $C_1 = \frac{\Gamma[(\nu+1)/2]}{\Gamma(\nu/2)(\nu d_0 \pi)^{1/2}}$, $\mu_0^* = \mu_0 + a_1$.

The limits of integrations are determined from the complement of Region A under Distribution 1, i.e. the univariate t distribution with ν degrees of freedom, location parameter μ_0 and scale parameter d_0.

Although Case 3 refers to one outlier in the suspected outlier group and one outlier in the suspected inlier group (i.e. $a_1>0$, $a_2>0$), Figure 1.1 can also be used to discuss the probability of correctly identifying one inlier in the suspected outlier group when there is an outlier in the suspected inlier group. This situation corresponds to $a_1=0$, $a_2>0$ for Case 3, and the PCI is calculated as follows:

PCI={area under Distribution 1, in Region A}

=P(accept H_0 and identify $Y^{(2)}$ as an inlier)=$1-\alpha$.

In Case 2, one observation is initially classified into the suspected outlier group and it is an outlier. In Case 3, one observation is initially classified into the suspected outlier group but actually there are two outliers in the data set. To see the effect of this misclassification, a comparison between Cases 2 and 3 in terms of their predictive distributions is made. From Figures 1.2 and 1.3, it is observed that for the same shift of the outlier in the suspected outlier group, the PCI of Case 2 is greater than the PCI of Case 3, for all $\nu>0$, α and $a_2>0$. However, from (2.21) and (2.23) it can be observed that as a_2 gets

smaller and n_i gets larger, the effect of this misclassification on the PCI of Case 3 gets smaller.

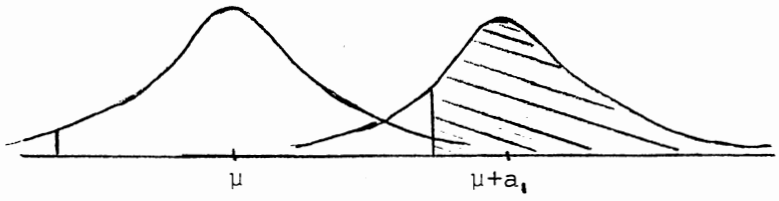

Figure 1.2. Predictive Distributions for Case 2

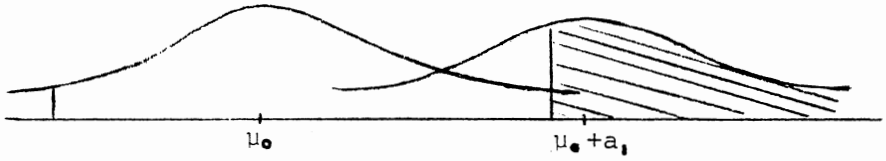

Figure 1.3. Predictive Distributions for Case 3

Case 4(M=2)

Two observations are initially classified into the suspected outlier group when all observations are inliers. In this case the PCI is the probability the two observations in the suspected outlier group are identified as inliers. Hence,

PCI=P(accept H_0 and identify both X_1, X_2 as inliers).

In this case, X_1 and X_2 will be identified as inliers when the F-test results in accepting H_0: all the suspected outliers are inliers or when the F-test results in accepting

H_1: at least one of them is an outlier, but the two t-tests which are done separately on X_1 and X_2 using the Bonferroni procedure, identify them both as inliers. The latter situation occurs when the combined effect of X_1 and X_2 leads to a large value of F^*, but each X_i (i=1,2) is not large enough to reject H_0. Hence,

$$PCI = P(F^* < F^{\alpha}_{2,\nu}) + P(F^* > F^{\alpha}_{2,\nu},\ |X_1|<t_{\alpha/4,\nu},\ |X_2|<t_{\alpha/4,\nu})$$

$$= \int\int f(\underset{\sim}{x},\nu,\underset{\sim}{a}^*,D)d\underset{\sim}{x}\ , \qquad \underset{\sim}{x}=(x_1,x_2)' \varepsilon R^2 \qquad (2.25)$$

where $f(\underset{\sim}{x},\nu,\underset{\sim}{a}^*,D)$ denotes the bivariate t distribution with ν degrees of freedom , location vector $\underset{\sim}{a}^*=(0,0)'$ and scale matrix D given in (2.11). The limits of integration are determined from the following equations:

$|X_1|<t_{\alpha/4,\nu}$ .....(i)

$|X_2|<t_{\alpha/4,\nu}$ .....(ii)

$F^* < F^{\alpha}_{2,\nu}$ .....(iii)

where F^* is defined in (2.12) and $F^{\alpha}_{2,\nu}$ is the $100(1-\alpha)$ percentile of an F distribution with 2 and ν degrees of freedom. An example of the region of integration is shown in Figure 1, for $\alpha=0.01$, $\nu=15$, and $\delta=0.50$.

Case 5(M=2)

Two observations are initially classified into the suspected outlier group,one of which is an outlier. All other observations are inliers. Without loss of generality,

let X_1 be the outlier observation and X_2 be the inlier observation in the suspected outlier group. In this case, the PCI is the probability X_1 is identified as an outlier and X_2 as an inlier. Hence,

PCI= P(reject H_0 and identify X_1 as the outlier and X_2 as the inlier)

$$=P(F^{*}>F^{\alpha}_{2,\nu}, |X_1|>t_{\alpha/4,\nu}, |X_2|<t_{\alpha/4,\nu})$$

$$=\int\int f(\underset{\sim}{x},\nu,\underset{\sim}{a}^{*},D)d\underset{\sim}{x}\ , \qquad \underset{\sim}{x}=(x_1,x_2)'\varepsilon R^2 \tag{2.26}$$

where $f(\underset{\sim}{x},\nu,\underset{\sim}{a}^{*},D)$ denotes the bivariate t distribution with ν degrees of freedom, location vector $\underset{\sim}{a}^{*}= (a_1^{*}, 0)'$, where $a_1^{*}= a_1/(d_{11})^{1/2}$ and scale matrix D as given in (2.11). The limits of integration are determined from the following equations:

$|X_1|>t_{\alpha/4,\nu}$ .....(i)

$|X_2|<t_{\alpha/4,\nu}$ .....(ii)

$F^{*} > F^{\alpha}_{2,\nu}$ .....(iii)

An example of the region of integration is shown in Figure 2, for α=0.01, ν=15 and δ=0.50.

Case 6(M=2)

Two observations are initially classified into the suspected outlier group, both are outliers. All other observations are inliers.

For this case, the PCI is the probability that both

observations in the suspected outlier group are identified as outliers. Hence,

PCI=P(reject H_0 and identify both X_1 and X_2 as outliers)

$$=P(F^{*}>F^{\alpha}_{2,\nu}, |X_1|>t_{\alpha/4,\nu} , |X_2|>t_{\alpha/4,\nu})$$

$$=\int\int f(\underset{\sim}{x},\nu,\underset{\sim}{a}^{*},D)d\underset{\sim}{x} , \qquad \underset{\sim}{x}=(x_1,x_2)'\varepsilon R^2 \tag{2.27}$$

where $f(\underset{\sim}{x},\nu,\underset{\sim}{a}^{*},D)$ denotes the bivariate t distribution with ν degrees of freedom, location vector $\underset{\sim}{a}^{*} = (a_1^{*} , a_2^{*})'$, $a_1^{*}=a_1/(d_{11})^{1/2}$, $a_2^{*}=a_2/(d_{22})^{1/2}$ and scale matrix D as given in (2.11).

The limits of integration are determined from the following equations:

$|X_1|>t_{\alpha/4,\nu}$ .....(i)

$|X_2|>t_{\alpha/4,\nu}$ .....(ii)

$F^{*} > F^{\alpha}_{2,\nu}$ .....(iii)

An example of the region of integration is shown in Figure 3, for α=0.01, ν=15 and δ=0.50.

In the next chapter, a summary of the plan for the numerical integration calculations of the PCI values will be presented.

CHAPTER III

SUMMARY OF PROPOSED NUMERICAL OUTPUTS

In Chapter II, the PCI integral formulas were derived for the six cases. From equations (2.13), (2.14), (2.24), (2.25), (2.26) and (2.27), one can see that the PCI is a function of four parameters. It is proposed to calculate the PCI for the six cases by using specific combinations of these parameters.

In this chapter, the parameters that affect the PCI are listed and defined. Then, the proposed numerical integrations for the six cases are summarized and finally, the limiting value of the PCI is derived.

3.1. List of Parameters

The PCI is a function of four parameters: ν, δ, α and a^*. The following levels of these parameters are used in calculating the PCI of the six cases:

1) $\nu=N-r-q$ is the degrees of freedom. It is a fixed constant determined by the size of the experiment and is defined for all N, r, q that fits the linear model in (2.1). Four levels of ν are considered : 5, 10, 15, 20.

2) $\delta=d_{12}/\{\sqrt{d_{11}}\sqrt{d_{22}}\}$ is the correlation between the two random variables, X_1 and X_2(for Cases 4-6). It is a known

constant, given $\underset{\sim}{y}^{(1)}$. Three levels of δ are considered: 0.25, 0.50, 0.90.

3) α is the significance level of the test. It is a fixed constant determined by the researcher. Two levels of α are considered : 0.01, 0.05.

4) $\underset{\sim}{a}^*$ is the Mx1 vector of unknown shift parameters for the outlier observations. In Cases 1-3, a^* is a scalar with M=1 while in Cases 4-6, $\underset{\sim}{a}^*$ is a vector with M=2. In Cases 1-5, the effect of $\underset{\sim}{a}^*$ is symmetric, hence only positive values are considered. Both positive and negative values of $\underset{\sim}{a}^*$ are considered for Case 6.

3.2. Summary of Numerical Integrations

In this section, the numerical integrations using those parameters listed above are summarized. Numerical outputs of the PCI values are tabulated in Appendix A, Tables I-IV. Graphical presentations of these PCI values are shown in Appendix B, Figures 4-13.

Case 1(M=1)

No numerical integration is done for this case because the PCI is constant at $1-\alpha$, for all $\nu>0$.

Case 2(M=1)

For this case the PCIs were calculated for the eight combinations of ν=5, 10, 15, 20 and α= 0.01, 0.05 with a^* values from 0.1 to 7.8 in increments of 0.1. In Table I, the PCI values are reported for all the eight combinations of ν and α and for a^*values from 1.0 to 7.8 in increments

of 0.4. The PCI values for ν=5,10,15,20, α=0.01 and a^* values from 0.1 to 7.8 in increments of 0.1 are plotted in in Figure 4. In Figures 5 and 6, the PCI values are plotted against a^*, using α=0.01, 0.05 and for values of ν=5, 20, respectively.

Case 3(M=1)

As mentioned in Chapter II, no numerical integrations are performed for this case.

Case 4(M=2)

For this case the PCIs were calculated for the twenty-four combinations of ν= 5, 10, 15, 20, α=0.01, 0.05 and δ=0.25, 0.50, 0.90. The results are not reported in any tables because the PCI is constant at $1-\alpha$, for all $\nu>0$ and δ.

Case 5(M=2)

For this case the PCIs were calculated for the twenty-four combinations of ν= 5, 10, 15, 20, α= 0.01, 0.05 and δ= 0.25, 0.50, 0.90 with a_1^* values from 0.1 to 8.2 in increments of 0.1. In Table II, the PCI values for α=0.01 are reported for all levels of ν and δ and for a_1^* values from 1.0 to 8.2 in increments of 0.4. In Table III, the PCI values for α=0.05 are reported for all levels of ν and δ and for a_1^* values from 1.0 to 8.2 in increments of 0.4. The PCI values for this case are plotted in Figures 7-10. In Figure 7, the plot of PCI against a_1^* values from 0.1 to

8.2 in increments of 0.1 for ν=5, 10, 15, 20, α=0.01 and δ=0.90 is presented. The plot of PCI against a_1^* for δ=0.25, 0.50, 0.90, at ν=20 and α=0.01 is shown in Figure 8. In Figure 9 and 10, the PCI values are plotted against a_1^* for α=0.01, 0.05, δ=0.90 and for values of ν=5 and 20, respectively.

Case 6(M=2)

In this case the PCIs were calculated for the twenty-four combinations of ν= 5, 10, 15, 20, α= 0.01, 0.05 and δ= 0.25, 0.50, 0.90 with

i) a_1^* and a_2^* values from 0.1 to 8.0 in increments of 0.1.
ii) a_1^* values from -0.1 to -8.0 in increments of -0.1 and a_2^* values from 0.1 to 8.0 in increments of 0.1. In Table IV, the PCI values are reported for α=0.01, ν=20, δ=0.90 and for:

i) a_1^* and a_2^* values from 1.0 to 8.0 in increments of 1.0.

ii) a_1^* values from -1.0 to -8.0 in increments of -1.0 and a_2^* values from 1.0 to 8.0 in increments of 1.0. For this case, the PCI values are plotted against a_1^* values from 1.0 to 7.1 in increments of 0.1, a_2^* values from 1.0 to 4.2 in increments of 0.1 and for ν=20, α=0.01 and δ=0.90. This 3-dimensional plot is presented in Figure 11.

To study the effects of misclassification of observations, the PCIs for Cases 2 and 5 are compared. In Figures 12, the plots of the PCIs for comparing Cases 2

and 5, using ν=5, α=0.01, δ=0.90 (Case 5) and a^* values from 0.1 to 8.2 in increments of 0.1 are presented. In Figure 13, the same comparison as above is made using ν=20.

The numerical integrations for Cases 4-6 were done using the IMSL (1987) subroutine *DTWODQ* on IBM and the single integral in Case 2 was evaluated using SAS (1987) Function *PROBT*.

3.2. Limiting Values of PCI

In this section the limiting values of the PCI for the six cases are calculated. In order to determine the precisions of the IMSL and SAS subroutines, the PCI values from the numerical integrations are compared to the limiting values of the PCI as $a_i^* \longrightarrow \infty$, i=1,2.

For purpose of discussion, let $\nu>0$, $\delta>0$, $a_i^*>0$, and $C=\frac{\Gamma\{(\nu+M)/2\}(\nu\pi)^{-M/2}|D|^{-1/2}}{\Gamma(\nu/2)}$ where M=1,2. Then the limiting value of the PCI is calculated as follows:

Case 1(M=1)

No limiting PCI values are appropriate here since a^*=0 and the PCI is constant at 1-α.

Case 2(M=1)

The limiting value of the PCI as $a^* \longrightarrow \infty$ is calculated using (2.14) in the following way:

$$\lim_{a^* \to \infty} \text{PCI} = \lim_{a^* \to \infty} C\int_{-\infty}^{-t_{\alpha/2,\nu}} \left\{1+(1/\nu)(x-a^*)\right\}^{-(\nu+1)/2} dx$$

$$+\lim_{a^* \to \infty} C\int_{t_{\alpha/2,\nu}}^{\infty} \left\{1+(1/\nu)(x-a^*)\right\}^{-(\nu+1)/2} dx \ , \qquad -\infty<x<\infty.$$

Let $x^* = x-a^*$, then for $-\infty<x^*<\infty$,

$$\lim_{a^* \to \infty} \text{PCI} = \lim_{a^* \to \infty} C\int_{-\infty}^{-t_{\alpha/2,\nu}-a^*} \left\{1+(1/\nu)x^{*2}\right\}^{-(\nu+1)/2} dx^*$$

$$+ \lim_{a^* \to \infty} C\int_{t_{\alpha/2,\nu}-a^*}^{\infty} \left\{1+(1/\nu)x^{*2}\right\}^{-(\nu+1)/2} dx^* \qquad (3.1)$$

$$= 0 + C\int_{-\infty}^{\infty} \left\{1+(1/\nu)x^{*2}\right\}^{-(\nu+1)/2} dx^*$$

$= 1$, for all $\nu>0$ and α.

$\therefore \lim_{a^* \to \infty} \text{PCI} = 1$, for all $\nu>0$ and α.

Hence, if one outlier exists in the data set and it is initially classified into the suspected outlier group, then it will be correctly identified as an outlier wp 1 as $a^* \longrightarrow \infty$, for all α and $\nu>0$. From Table I, one can see that this result is consistent with the result from the numerical integration.

The limiting value of the PCI as $a^* \longrightarrow 0$, can be calculated for Case 2 using (3.1). Hence, from (3.1),

$$\lim_{a^* \longrightarrow 0} PCI = C\int_{-\infty}^{-t_{\alpha/2,\nu}} \left\{1+(1/\nu)x^{*2}\right\}^{-(\nu+1)/2}dx^*$$

$$+C\int_{t_{\alpha/2,\nu}}^{\infty} \left\{1+(1/\nu)x^{*2}\right\}^{-(\nu+1)/2}dx^*$$

$$=2C\int_{t_{\alpha/2,\nu}}^{\infty} \left\{1+(1/\nu)x^{*2}\right\}^{-(\nu+1)/2}dx^*$$

$=2(\alpha/2)=\alpha$, for all $\nu>0$.

Therefore in this case, as a^* approaches 0, the PCI approaches α.

Case 3(M=1)

Refer to the discussion of this case in Chapter II.For this case, the limiting value of the PCI^* as $a_1 \longrightarrow \infty$, $a_2>0$ fixed, will be calculated. Let $C_1 = \dfrac{\Gamma[(\nu+1)/2]}{\Gamma(\nu/2)(d_o\nu\pi)^{1/2}}$ be the constant in the p.d.f in (2.19). Thus, from (2.24) the limiting value is calculated as follows:

$$\lim_{a_1 \longrightarrow \infty} PCI = \lim C_1\int_{-\infty}^{-t^*_{\alpha/2,\nu}} \left\{1+(1/d_o\nu)(y^{(2)}-\mu_o^*)^2\right\}^{-(\nu+1)/2}dy^{(2)}$$

$$+ \lim C_1\int_{t^*_{\alpha/2,\nu}}^{\infty} \left\{1+(1/d_o\nu)(y^{(2)}-\mu_o^*)^2\right\}^{-(\nu+1)/2}dy^{(2)},$$

where $\mu_o^*=\mu_o+a_1$. Let $w=y^{(2)}-\mu_o$ then,

$$\lim PCI^{*}=\lim_{a_1\longrightarrow\infty} C_1\int_{-\infty}^{-t^{*}_{\alpha/2,\nu}} \left\{1+(1/d_0\nu)(w-a_1)^2\right\}^{-(\nu+1)/2}dw$$

$$+\lim_{a_1\longrightarrow\infty} C_1\int_{t^{*}_{\alpha/2,\nu}}^{\infty} \left\{1+(1/d_0\nu)(w-a_1)^2\right\}^{-(\nu+1)/2}dw$$

Let $w^{*}=w-a_1$ then,

$$\lim PCI=\lim_{a_1\longrightarrow\infty} C_1\int_{-\infty}^{-t^{*}_{\alpha/2,\nu-a_1}} \left\{1+(1/d_0\nu)w^{*2}\right\}^{-(\nu+1)/2}dw^{*}$$

$$+\lim_{a_1\longrightarrow\infty} C_1\int_{t^{*}_{\alpha/2,\nu-a_1}}^{\infty} \left\{1+(1/d_0\nu)w^{*2}\right\}^{-(\nu+1)/2}dw^{*} \qquad (3.2)$$

$=0+1=1$, for all $\nu>0$, α and any a_2.

Therefore, if the one outlier is initially classified into the suspected outlier group, then it will be correctly identified as an outlier wp 1, as the shift of the outlier in the suspected outlier group approaches ∞. This is true for all $\nu>0$, and regardless of the shift of the outlier in the suspected inlier group. The limiting value of PCI^{*} as $a_1\longrightarrow 0$, $a_2>0$ can be calculated using (3.2). Hence,

$$\lim_{a_1\longrightarrow 0} PCI^{*} = C_1\int_{-\infty}^{-t^{*}_{\alpha/2,\nu}} \left\{1+(1/d_0\nu)w^{*2}\right\}^{-(\nu+1)/2}dw^{*}$$

$$+ C_1\int_{t^*_{\alpha/2,\nu}}^{\infty} \left\{1+(1/d_0\nu)w^{*2}\right\}^{-(\nu+1)/2}dw^*$$

$$=2C_1\int_{t^*_{\alpha/2,\nu}}^{\infty} \left\{1+(1/d_0\nu)w^{*2}\right\}^{-(\nu+1)/2}dw^*$$

$=2(\alpha/2)=\alpha$, for all $\nu>0$ and $a_2>0$.

Therefore, the PCI* approaches α as the shift of the outlier in the suspected outlier group approaches 0, regardless of the shift of the outlier in the suspected inlier group.

Case 4(M=2)

No limiting PCI values are appropriate here since $a_1^*=a_2^*=0$ and the PCI is constant at $1-\alpha$.

Case 5(M=2)

Using (2.25), the limiting value of the PCI as $a_1^*\longrightarrow\infty$, when $a_2^*=0$ is calculated as follows

$$\lim_{a_1^*\longrightarrow\infty} PCI = \lim C\int_{-t_{\alpha/2,\nu}}^{t_{\alpha/2,\nu}}\int_{-\infty}^{-t_{\alpha/4,\nu}} \left\{1+1/[\nu(1-\delta^2)]\right.$$

$$\left.\{(x_1-a_1^*)^2-2\delta(x_1-a_1^*)x_2+x_2^2\}\right\}^{-(\nu+2)/2}dx_1dx_2$$

$$+\lim C\int_{-t_{\alpha/4,\nu}}^{t_{\alpha/4,\nu}}\int_{t_{\alpha/4,\nu}}^{\infty} \left\{1+1/[\nu(1-\delta^2)]\right.$$

$$\left.\{(x_1-a_1^*)^2-2\delta(x_1-a_1^*)x_2+x_2^2\}\right\}^{-(\nu+2)/2}\, dx_1dx_2$$

Let $x_1^* = x_1 - a_1^*$, then

$$\lim_{a_1^* \to \infty} PCI = \lim_{a_1^* \to \infty} C \int_{-t_{\alpha/4,\nu}}^{t_{\alpha/4,\nu}} \int_{-\infty}^{-t_{\alpha/4,\nu} - a_1^*} \left\{1+1/[\nu(1-\delta^2)]\right.$$

$$\left.\{x_1^{*2} - 2\delta x_1^* x_2 + x_2^2\}\right\}^{-(\nu+2)/2} dx_1^* dx_2$$

$$+\lim_{a_1^* \to \infty} C \int_{-t_{\alpha/4,\nu}}^{t_{\alpha/4,\nu}} \int_{t_{\alpha/4,\nu} - a_1^*}^{\infty} \left\{1+1/[\nu(1-\delta^2)]\right.$$

$$\left.\{x_1^{*2} - 2\delta x_1^* x_2 + x_2^2\}\right\}^{-(\nu+2)/2} dx_1^* dx_2. \qquad (3.3)$$

$$= 0 + C \int_{-t_{\alpha/4,\nu}}^{t_{\alpha/4,\nu}} \left\{1+(1/\nu)x_2^2\right\}^{-(\nu+1)/2} dx_2 = 1-\alpha/2 \text{ , for all } \nu>0.$$

$\therefore \lim_{a_1^* \to \infty} PCI = 1-\alpha/2$, for all $\nu>0$.

In this case the limiting value is independent of ν and δ but is dependent on α. Thus for all δ and $\nu>0$, the limiting PCI at $\alpha=0.01$ is greater than at $\alpha=0.05$, (see Figure 10).

The limiting value of the PCI, as $a_1^* \to \infty$ is less than 1 as a consequence of misclassifying one inlier observation into the suspected outlier group. From Tables II and III, one can see that the limiting values of the PCIs for both $\alpha=0.01$ and $\alpha=0.05$ are consistent with the results from the numerical integrations.

The limiting value of the PCI as $a_1^* \longrightarrow 0$ when $a_2^*=0$, can be calculated for Case 5 using (3.3) . An upper bound on the limiting PCI is obtained for Case 5, since the exact limiting value is difficult to calculate. Thus,

$$\lim_{a_1^* \longrightarrow 0} PCI = 2C \int_{-t_{\alpha/4,\nu}}^{t_{\alpha/4,\nu}} \int_{t_{\alpha/4,\nu}}^{\infty} \{1+1/[\nu(1-\delta^2)]$$

$$(x_1^{*2}-2\delta x_1^* x_2+x_2^2)\}^{-(\nu+2)/2} dx_1^* dx_2$$

$$< 2C \int_{-\infty}^{\infty} \int_{t_{\alpha/4,\nu}}^{\infty} \{1+1/[\nu(1-\delta^2)](x_1^{*2}-2\delta x_1^* x_2+x_2^2)\}^{-(\nu+2)/2} dx_1^* dx_2 .$$

Reverse the order of integration to obtain

$$= 2C \int_{t_{\alpha/4,\nu}}^{\infty} \int_{-\infty}^{\infty} \{1+(1/[\nu(1-\delta^2)](x_1^{*2}-2\delta x_1^* x_2+x_2^2)\}^{-(\nu+2)/2} dx_2 dx_1^*$$

$$= 2C \int_{t_{\alpha/4,\nu}}^{\infty} \{1+(1/\nu)x_1^{*2}\}^{-(\nu+1)/2} dx_1^* = 2(\alpha/4) = \alpha/2 .$$

Therefore, $\lim_{a_1^* \longrightarrow 0} PCI < \alpha/2$, for all $\nu>0$, δ and $a_2^*=0$.

Hence, as $a_1^* \longrightarrow 0$, $a_2^*=0$, the limiting PCI for Case 5 approaches an upper bound of $\alpha/2$.

Case 6(M=2)

For both $a_1^*>0$ and $a_2^*>0$, the limiting value of the PCI

was calculated as both a_1^* and a_2^* approach ∞. Denote the p.d.f. of a bivariate t distribution with location vector $\underset{\sim}{a}^*=(a_1^*, a_2^*)'$ and scale matrix $D=\begin{bmatrix} 1 & \delta \\ \delta & 1 \end{bmatrix}$, as

$$f(\underset{\sim}{x}, \nu, \underset{\sim}{a}^*, D) =$$

$$C\left\{1+\{1/\nu(1-\delta^2)\}(x_1-a_1^*)^2-2\delta(x_1-a_1^*)(x_2-a_2^*)+(x_2-a_2^*)^2\right\}^{-(\nu+2)/2}$$

$(x_1, x_2) \in R^2$, $a_1^*>0$ and $a_2^*>0$ and $-1<\delta<1$. Then, using (2.26) the limiting value for this case is calculated as follows:

$$\lim_{\substack{a_i^* \longrightarrow \infty \\ \{i=1,2\}}} PCI = \lim_{a_2^* \longrightarrow \infty} \int_{t_{\alpha/4,\nu}}^{\infty} \lim_{a_1^* \longrightarrow \infty} \int_{t_{\alpha/4,\nu}}^{\infty} f(x_1, x_2, \nu, \underset{\sim}{a}^*, D)\,dx_1\,dx_2$$

$$+ \lim_{a_2^* \longrightarrow \infty} \int_{t_{\alpha/4,\nu}}^{\infty} \lim_{a_1^* \longrightarrow \infty} \int_{-\infty}^{-t_{\alpha/4,\nu}} f(x_1, x_2, \nu, \underset{\sim}{a}^*, D)\,dx_1\,dx_2$$

$$+ \lim_{a_2^* \longrightarrow \infty} \int_{-\infty}^{-t_{\alpha/4,\nu}} \lim_{a_1^* \longrightarrow \infty} \int_{-\infty}^{-t_{\alpha/4,\nu}} f(x_1, x_2, \nu, \underset{\sim}{a}^*, D)\,dx_1\,dx_2$$

$$+ \lim_{a_2^* \longrightarrow \infty} \int_{-\infty}^{-t_{\alpha/4,\nu}} \lim_{a_1^* \longrightarrow \infty} \int_{t_{\alpha/4,\nu}}^{\infty} f(x_1, x_2, \nu, \underset{\sim}{a}^*, D)\,dx_1\,dx_2$$

Let $x_1^*=x_1-a_1^*$ and $x_2^*=x_2-a^*$, then

$$=\lim_{a_2^*\longrightarrow\infty}\int_{t_{\alpha/4,\nu}-a_2^*}^{\infty}\ \lim_{a_1^*\longrightarrow\infty}\int_{t_{\alpha/4,\nu}-a_1^*}^{\infty} f(x_1^*,x_2^*,\nu,D)dx_1^*dx_2^*$$

$$+\lim_{a_2^*\longrightarrow\infty}\int_{t_{\alpha/4,\nu}-a_2^*}^{\infty}\ \lim_{a_1^*\longrightarrow\infty}\int_{-\infty}^{-t_{\alpha/4,\nu}-a_1^*} f(x_1^*,x_2^*,\nu,D)dx_1^*dx_2^*$$

$$+\lim_{a_2^*\longrightarrow\infty}\int_{-\infty}^{-t_{\alpha/4,\nu}-a_2^*}\ \lim_{a_1^*\longrightarrow\infty}\int_{-\infty}^{-t_{\alpha/4,\nu}-a_1^*} f(x_1^*,x_2^*,\nu,D)dx_1^*dx_2^*$$

$$+\lim_{a_2^*\longrightarrow\infty}\int_{-\infty}^{-t_{\alpha/4,\nu}-a_2^*}\ \lim_{a_1^*\longrightarrow\infty}\int_{t_{\alpha/4,\nu}-a_1^*}^{\infty} f(x_1^*,x_2^*,\nu,D)dx_1^*dx_2^* \quad (3.4)$$

$\therefore \lim_{a_i^*\longrightarrow\infty} PCI = 1+0+0+0=1$, for all δ, α and $\nu>0$.
$\{i=1,2\}$

Hence, as $a_1^*\longrightarrow\infty$ and $a_2^*\longrightarrow\infty$, both outliers will be correctly identified wp 1, regardless of ν, δ and α.

The limiting value of the PCI as $a_1^*\longrightarrow 0$ and $a_2^*\longrightarrow 0$ can be calculated for Case 6 using (3.4). An upper bound on the limiting PCI value is obtained for this case since an exact limiting PCI value is difficult to calculate. Thus,

$$\lim_{\substack{a_i^* \longrightarrow 0 \\ \{i=1,2\}}} PCI = \int_{t_{\alpha/4,\nu}}^{\infty} \int_{t_{\alpha/4,\nu}}^{\infty} f(x_1^*, x_2^*, \nu, D)\,dx_1^* dx_2^*$$

$$+ \int_{t_{\alpha/4,\nu}}^{\infty} \int_{-\infty}^{-t_{\alpha/4,\nu}} f(x_1^*, x_2^*, \nu, D)\,dx_1^* dx_2^*$$

$$+ \int_{-\infty}^{-t_{\alpha/4,\nu}} \int_{-\infty}^{-t_{\alpha/4,\nu}} f(x_1^*, x_2^*, \nu, D)\,dx_1^* dx_2^*$$

$$+ \int_{-\infty}^{-t_{\alpha/4,\nu}} \int_{t_{\alpha/4,\nu}}^{\infty} f(x_1^*, x_1^*, \nu, D)\,dx_1^* dx_2^*$$

$$< 2 \int_{t_{\alpha/4,\nu}}^{\infty} \int_{-\infty}^{\infty} f(x_1^*, x_2^*, \nu)\,dx_1^* dx_2^*$$

$$= 2 \int_{t_{\alpha/4,\nu}}^{\infty} f(x_2^*, \nu)\,dx_2^* = 2(\alpha/4) = \alpha/2, \text{ for all } \nu>0 \text{ and } \delta.$$

Therefore, $\lim_{\substack{a_i^* \longrightarrow 0 \\ \{i=1,2\}}} PCI < \alpha/2$, for all $\nu>0$ and δ.

Hence, as $a_i^* \longrightarrow 0$ $\{i=1,2\}$ the PCI approaches an upper bound of $\alpha/2$.

CHAPTER IV

DISCUSSION OF RESULTS

In this chapter, calculated PCI values are presented. These PCI values were generated using the IMSL subroutine *DTWODQ* and SAS Function *PROBT*. The outputs from these subroutines are provided in Tables I-IV and Figures 4-13, in Appendix A and B, respectively. These Tables and Figures are interpreted for each case below in Section 4.1. Then in Section 4.2, the PCIs for Cases 2 and 5 are compared.

4.1. Results for Cases 1-6

In this section, the results on the PCI for the six cases of initial classifications are reported. A brief discussion on the PCI for each case is also presented.

Case 1(M=1)

The PCI is constant at $1-\alpha$, for all $\nu>0$. However, the larger the significance level of the test used the lower will be the PCI.

Case 2(M=1)

From Table I, it is observed that the PCI increases as a^* increases. For fixed α and a^*, increasing ν increases the PCI (see Figure 4). From Figures 5 and 6, it can be seen that as ν increases the influence of α on the PCI

decreases. For all ν and α, the PCI equals to its limiting value at $a^{*}=7.4$ (see Table I).

Case 3(M=1)

No numerical integration calculations were done for Case 3. The main result for this case is the derivation of the predictive density of $Y^{(2)}|\underset{\sim}{Y}_{O}^{(1)}$. From (2.21) and (2.23) the following conclusions are made.

i) The predictive density has a different location and scale parameter. As a consequence of this misclassification the location parameter $\mu_{O}>\mu$ and the scale parameter $d_{O}>d$.

ii) The PCI depends on ν, α and both the shifts of the outlier observations in the suspected outlier and suspected inlier groups.

iii) The PCI also depends on whether the outlier in the suspected inlier group is in the same or different class as the outlier in the suspected outlier group.

Case 4(M=2)

The results from the numerical integration of this case are not presented because the PCI is constant at $1-\alpha$, for all $\nu>0$ and δ. In the following paragraph, the effect of using a Bonferroni procedure(test procedure 2.10) on the PCI of Cases 1 and 4 is briefly discussed.

Recall that in Case 1, one inlier is misclassified into the suspected outlier group. In Case 4,two inliers are misclassified into the suspected outlier group. However, for both cases the PCI is $1-\alpha$. This is due to the fact

that the Bonferroni procedure uses a $t_{\alpha/2M,\nu}$ rejection rule. The $\alpha/2M$ alpha level produces a constant PCI value of $1-\alpha$, for Cases 1 and 4.

Case 5(M=2)

From Tables II, III and Figures 7-10, the following conclusions are made.

i) For all $\nu>0$, δ and α, the PCI increases as a_1^* increases.

ii) For fixed δ, a_1^* and α, the PCI increases with increasing ν. From Figure 7 for fixed a_1^*, α and δ, the influence of ν on the PCI decreases as ν increases.

iii) In general, for fixed ν and α, the PCI increases with increasing value of δ. From Figure 8, it is apparent however that the effect of δ on the PCI is small.

iv) For fixed ν, δ, and $0< a_1^*< 5$, the PCI at $\alpha=0.05$ is larger than the PCI at $\alpha=0.01$, but for $a_1^*>5$ the reverse is true (see Figure 9). This effect occurs because for $a_1^*>5$, the PCI for both values of α approach its limiting value. Since the limiting value of the PCI for Case 5 is equal to $1-\alpha/2$, the result follows.

v) From Figure 9 and 10 for fixed δ, decreasing the value of ν will make the difference between the PCI at $\alpha=0.01$ and $\alpha=0.05$ larger.

Case 6(M=2)

From Table IV and the 3-d plot of Figure 11, it can be observed that:

i) the PCI increases as $a_1^*>0$ and $a_2^*>0$ increases, for all ν,

δ and α.

ii) The PCI increases faster when both a_1^* and a_2^* increase as opposed to just a_1^* or a_2^* increasing and the other one held fixed.

iii) At ν=20, δ=0.90 and α=0.01, the PCI for $a_1^*>0$, $a_2^*>0$ is always larger than the PCI for $a_1^*<0$, $a_2^*>0$ (see Table IV).

For fixed ν and α, the PCI increases as δ increases when $a_1^*>0$, $a_2^*>0$, while the PCI decreases as δ increases when $a_1^*<0$, $a_2^*>0$. No 3-dimensional plot for different values of δ is presented since it requires a lot of computer time.

4.2. Comparisons of Cases

It is of interest to examine the consequences of misclassifying one extra inlier into the suspected outlier group. In Case 2, the observation in the suspected outlier group is an outlier, hence there is no initial misclassification. In Case 5, one of the two observations in the suspected outlier group is an inlier, hence there is an initial misclassification. Comparing Cases 2 and 5 will address this problem of interest.

From Figures 12 and 13, misclassifying one inlier into the suspected outlier group reduces the PCI. However, this initial misclassification has little effect on the PCI when ν is large. For instance, at δ=0.90, α=0.01, a_1^*=5.0 and ν=5, the difference between the PCI of Cases 2 and 5 is 0.227, while this difference at ν=20 is 0.021. Hence,

given the choice of classifying an observation into the suspected outlier or suspected inlier group, it is safer to classify the observation into the suspected outlier group. This is due the fact that when ν is large ($\nu>20$), misclassifying an extra inlier into the suspected outlier group has little effect on the PCI. This choice is further supported by the fact that, if an outlier is not initially classified into the suspected outlier group, then one loses the chance of identifying it. It is also noted that δ and α have little effect on this comparison.

CHAPTER V

SUMMARY, CONCLUSIONS, AND RECOMMENDATIONS

This study is devoted to the investigation of the effect of initial classification or misclassification of outliers and inliers on the PCI for a linear model of constant intraclass correlation. The following six cases of initial classifications are considered:

Case 1(M=1)

One observation is initially classified into the suspected outlier group when all observations are inliers.

Case 2(M=1)

One observation is initially classified into the suspected outlier group and it is the only outlier.

Case 3(M=1)

One observation is initially classified into the suspected outlier group but actually there are two outliers in the data set.

Case 4(M=2)

Two observations are initially classified into the suspected outlier group when all observations are inliers.

Case 5(M=2)

Two observations are initially classified into the suspected outlier group, one of which is an outlier. All

other observations are inliers.

Case 6(M=2)

Two observations are initially classified into the suspected outlier group and both are outliers. All other observations are inliers.

For all the cases except Case 3, the PCI is calculated based on the predictive density of the suspected outlier given a set of inliers. For Case 3, this calculation is based on the predictive density of the suspected outlier given a set inliers with one outlier in it.

From this study, it can be concluded that misclassifying an extra inlier into the suspected outlier group has the effect of decreasing the PCI. However, this initial misclassification has little effect on the PCI if ν is large ($\nu>20$). Hence for large enough ν ($\nu>20$), given the choice of classifying an observation into the suspected outlier or suspected inlier group, it is safer to initially classify the observation into the suspected outlier group. This is due to the fact that misclassifying an extra inlier observation into the suspected outlier group has little effect on the PCI, when ν is large. Thus, the test procedures in (2.9) and (2.10) are not affected very much by initial misclassification, when ν is large.

The effect of misclassifying an extra outlier into the suspected inlier group is studied through the examination of Case 3. As a consequence of this misclassification, the location and scale parameters of the predictive density

have increased. Furthermore, the probability of correctly identify the outlier in the suspected outlier group is smaller than the probability when there is no outlier in the suspected inlier group (Case 2). Further investigation of this case is recommended, so that the effect of misclassifying one outlier observation into the suspected inlier group on the PCI can be studied in detail.

The Bonferroni procedure offers protection against declaring too many observations to be outliers. When there is no outlier in the data set, the Bonferroni procedure in (2.10) ensures that the PCI is constant at $1-\alpha$, even if the number of misclassifications increases. From the numerical integrations, it is noticed that the four parameters play an important role in determining the PCI values. However, the PCI for Case 5 is insensitive to changes in δ (the correlation between the two observations in the suspected outlier group).

Lastly, we recommend the investigations of the following topics for future studies:

1) Extend this study to include the investigations of PCIs when there are more than two outliers in the data set.
2) Use a different prior for the derivation of the predictive density of the suspected outliers given a set of inliers, perhaps an informative one.
3) The use of scale-change model for the M suspected outliers.

BIBLIOGRAPHY

Balasooriya, U. and Tse, Y.K. (1986). Outlier Detection in Linear Model: A Comparative study in simple linear regression. Communication in Statistics, Theory and Method, 15(12), 3589-3597.

Barnett, V. and Lewis, T.(1984) . Outliers in Statistical Data, New York, John Wiley.(2nd Edition).

Beckman, R.J, and Cook, R.D. (1983). "Outlier.....s," Technometrics, 25, 119-149 (Discussion on pp 150-163).

Berger, J.O. (1980). Statisical Decision Theory, Springer-Verlag New York Heidelberg Berlin.

Box, G.E.P. and Tiao, G.C. (1968). A Bayesian approach to some outlier problems, Biometrika 55, 119-129.

Box, G.E.P. and Tiao, G.C. (1973). Bayesian Inference in Statistical Analysis. Addison-Wesley-London.

Bross, I.D.J. (1961). Outliers in patterned experiments: A Strategic Appraisal, Technometrics, 3, 91-102.

Collett, D. and Lewis, T. (1976). The Subjective Nature of Outlier Rejection Procedures. Applied Statistics, 25 No. 3, 228-237.

Freeman, P.R. (1979). On the number of outlier in data from a linear model (with Discussion). Pages 349-365 of Bernardo, J.M., Degroot, M.H., Lindley, D.V., and Smith, A.F.M. (Eds.). (1979). Bayesian Statistics. University Press, Valencia, Spain.

Gambino, J.,and Guttman, I. (1984). A Bayesian Approach to Prediction in the Presence of spurious observations for several models. Communication in Statistics A. Theory and Methods. 13(7), 791-812.

Gentleman, J.F.,and Wilk, M.B. (1975a). Detecting Outlier:

II Supplementing the direct analysis of residuals. Biometrics, 31, 387-410.

Gentleman, J.F., and Wilk, M.B. (1975b). Detecting outliers in a two-way table. I. Statistical behaviour of residuals. Technometrics, 17, 1-14.

Guttman, I. (1973). Care and Handling of Univariate or Multivariate Outliers in Detecting Spuriosity-A Bayesian approach Technometrics, 15, 723-738.

Guttman, I., Dutter, R., and Freeman, P.R. (1978). Care and Handling of Univariate Outliers in the General Linear Model to detect spuriosity-A Bayesian approach, Technometrics, 20, 187-193.

Guttman, I., and Katri, C.G. (1975). A Bayesian approach to some problems involving the detection of spuriosity. 111-145 of Gupta, R.P. (Ed.) (1975). Applied Statistics, North-Holland, Amsterdam.

Graybill, F.A. (1976). Theory and Application of the Linear Model, Duxbury Press.

Hawkins, D.M. (1980). Identification of Outliers. London, Chapman and Hall.

IMSL (1987). International Mathematical and Statistical Library, Edition 9.2. IMSL, INC. of Houston, Texas.

Jain, R.B. (1981b). Detecting Outliers:Power and some other considerations, Communication in Statistics A, Theory and Methods 10, 2299-2314.

John, J.A. (1978). Outliers in factorial experiments. Applied Statistics, 27, 111-119.

John, J.A., and Draper N.R. (1978). On Testing for Two Outliers or One Outlier in Two-Way Tables, Technometrics, 20, 69-78.

Moser, B.K. (1986). Outlier Testing in Unbalanced Linear Models with Constant Intraclass Correlation, submitted to Biometrics.

Moser, B.K. and Marco, V.R.(1988). Bayesian Outlier Testing using the Predictive Distribution For a Linear Model of Constant Intraclass form, Communication in Statistics A, Theory and Methods 17, No 3, 849-861.

Moser, B.K. and McCabe, G.P. (1987). Closed Form Estimators for an Analysis of Covariance Model with Constant Intraclass Correlation and Unbalanced Data, submitted to Biometrics.

Raiffa, H., and Schlaifer, R. (1961). Applied Statistical Decision Theory. Boston, Massachusetts: Harvard University Press.

Rosner, B. (1975). On the detection of many outliers. Technometrics, 17, 221-227.

SAS User's Guide: Basics. Version 5 Edition. Sas Institute Inc. Box 8000, Cary, N.C. 27511-8000.

Srikantan, K.S. (1961). Testing for the single Outlier in a Regression Model, Sankhya, A, 23, 251-260.

Tietjen, G.L., More, R.H. and Beckman, R.J. (1973). Testing for a Single Outlier in Linear Regression. Technometrics, 15, 717-721.

Zellner, A. (1971). An Introduction to Bayesian Inference in Econometrics. New York, John Wiley and Sons, Inc.

APPENDIXES

APPENDIX A

TABLES

TABLE I

PCI FOR CASE 2: PCI=PCI*0.001

	v=N-r-q							
	5		10		15		20	
	Alpha		Alpha		Alpha		Alpha	
	.01	.05	.01	.05	.01	.05	.01	.05
* a								
1.0	17	97	29	128	36	141	40	148
1.4	25	152	54	216	72	240	82	251
1.8	39	241	101	340	135	373	154	389
2.2	64	365	178	490	233	527	263	545
2.6	106	513	291	642	367	677	404	694
3.0	175	658	435	771	521	801	561	814
3.4	278	779	589	866	671	888	707	898
3.8	413	864	729	927	796	942	825	949
4.2	564	918	837	962	885	972	905	976
4.6	703	951	909	980	940	987	953	990
5.0	811	971	951	990	971	994	978	996
5.4	885	982	975	995	987	997	991	998
5.8	931	989	987	997	994	999	996	999
6.2	959	993	994	999	997	999	998	999
6.6	975	995	997	999	999	999	999	999
7.0	984	997	998	999	999	999	999	999
7.4	990	998	999	999	999	999	999	999
7.8	994	998	999	999	999	999	999	999

TABLE II

PCI FOR CASE 5: ALPHA=0.01, PCI=PCI*0.001

	v=N-r-q											
	5			10			15			20		
	Delta			Delta			Delta			Delta		
	.25	.50	.90	.25	.50	.90	.25	.50	.90	.25	.50	.90
* a1												
1.0	5	5	5	10	10	10	14	14	15	16	16	18
1.4	8	8	8	20	20	24	30	30	37	36	37	44
1.8	12	12	13	40	42	50	61	65	77	75	79	92
2.2	19	20	22	76	82	96	118	127	146	143	153	172
2.6	31	33	38	138	153	173	209	228	251	248	267	288
3.0	51	57	66	236	263	285	337	367	389	388	415	435
3.4	86	99	112	370	407	428	491	526	544	545	576	591
3.8	144	169	186	525	567	582	646	680	692	695	723	733
4.2	235	277	294	676	713	723	777	804	811	816	837	842
4.6	362	418	433	798	826	832	872	890	894	898	912	915
5.0	513	572	584	884	901	904	932	942	944	948	955	957
5.4	662	712	719	936	946	948	965	970	972	974	977	979
5.8	783	818	822	965	970	972	981	984	985	986	988	989
6.2	867	888	891	980	983	984	989	990	992	989	991	993
6.6	919	932	934	988	989	991	992	993	994	993	994	995
7.0	951	958	959	990	992	993	994	995	995	995	995	995
7.4	969	973	974	993	994	995	995	995	995	995	995	995
7.8	979	981	983	994	995	995	995	995	995	995	995	995
8.2	985	987	988	995	995	995	995	995	995	995	995	995

TABLE III

PCI FOR CASE 5: ALPHA=0.05, PCI=PCI*0.001

	v=N-r-q											
	5			10			15			20		
	Delta			Delta			Delta			Delta		
	.25	.50	.90	.25	.50	.90	.25	.50	.90	.25	.50	.90
* a1												
1.0	50	50	50	50	50	50	60	60	60	70	70	70
1.4	56	56	56	100	101	110	121	122	133	132	133	145
1.8	95	97	103	178	184	200	213	220	238	231	238	257
2.2	159	166	178	292	305	325	341	355	375	365	379	400
2.6	256	272	287	436	457	476	492	512	531	519	539	556
3.0	388	412	427	589	613	628	643	665	679	668	688	702
3.4	538	565	577	726	747	758	770	788	799	790	807	817
3.8	679	703	712	828	844	853	861	874	882	874	887	895
4.2	790	807	814	895	905	913	916	925	933	925	933	941
4.6	865	876	882	934	940	947	947	952	959	953	957	963
5.0	911	918	924	942	958	964	962	965	970	966	968	972
5.4	938	943	949	955	967	972	970	971	974	971	972	975
5.8	953	956	962	966	971	974	973	973	975	974	974	975
6.2	962	964	969	970	973	975	974	975	975	974	975	975
6.6	967	968	972	973	974	975	975	975	975	975	975	975
7.0	970	971	973	974	975	975	975	975	975	975	975	975
7.4	972	972	974	975	975	975	975	975	975	975	975	975
7.8	973	974	975	975	975	975	975	975	975	975	975	975
8.2	974	975	975	975	975	975	975	975	975	975	975	975

TABLE IV

PCI FOR CASE 6: v=20, DELTA=0.90, AND ALPHA=0.01

a2* / a1*	1	2	3	4	5	6	7	8
-1	0.0000	0.0000	0.0000	0.0938	0.0991	0.0992	0.0992	0.0992
-2		0.0000	0.0002	0.1073	0.1153	0.1154	0.1154	0.1154
-3			0.0259	0.1201	0.1275	0.1276	0.1276	0.1276
-4				0.1383	0.1409	0.1409	0.1409	0.1409
-5					0.1435	0.1436	0.1436	0.1436
-6						0.1436	0.1436	0.1436
-7							0.1436	0.1436
-8								0.1436
1	0.0130	0.0211	0.0215	0.0835	0.0921	0.1175	0.1252	0.1253
2		0.0928	0.1288	0.1325	0.1368	0.1435	0.1442	0.1442
3			0.3664	0.4358	0.4363	0.4363	0.4364	0.4364
4				0.6058	0.6529	0.6536	0.6536	0.6536
5					0.8015	0.8159	0.8161	0.8161
6						0.8491	0.8513	0.8514
7							0.8556	0.8563
8								0.8563

APPENDIX B

FIGURES

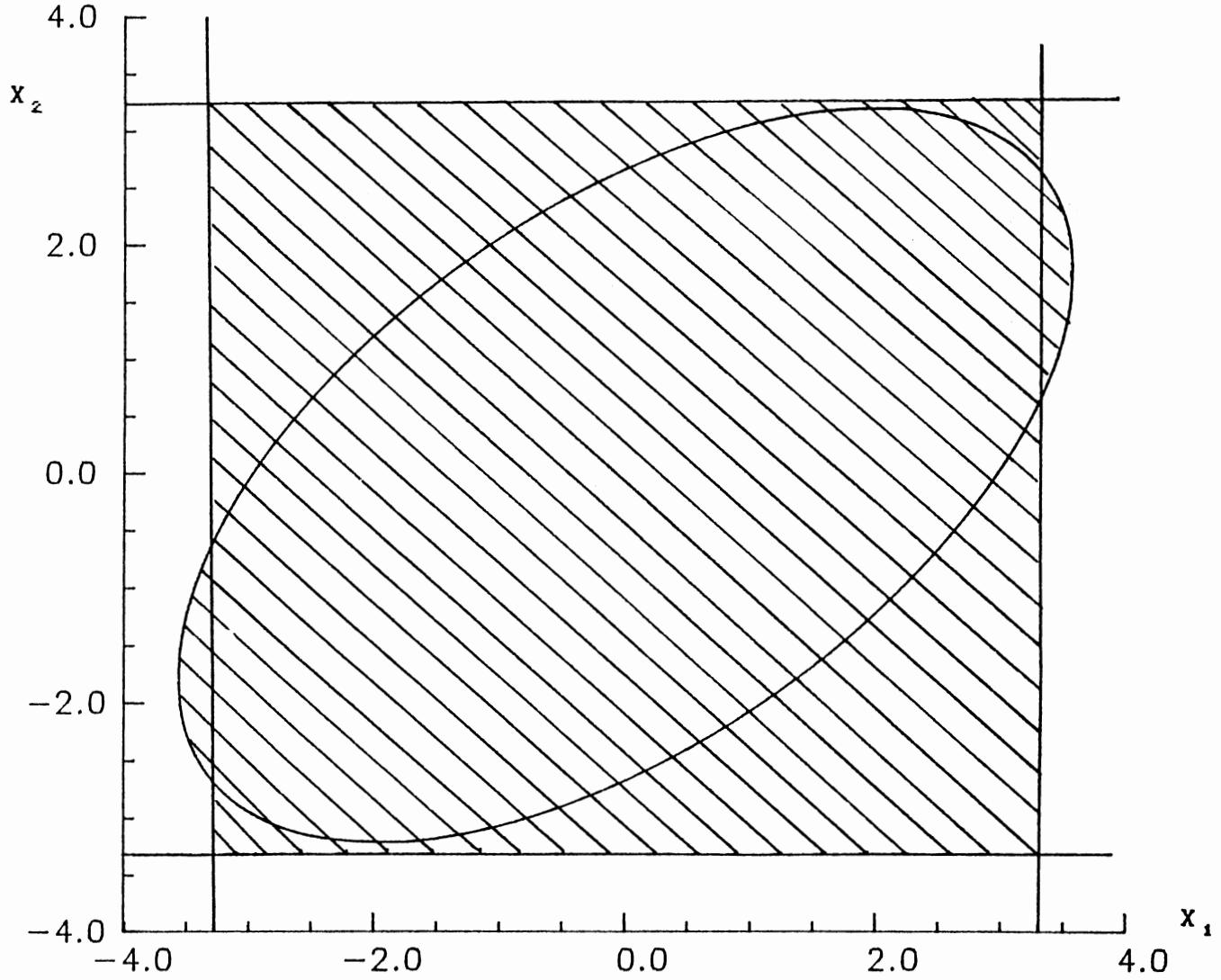

Figure 1. The Region of Integration for Case 4, when ν=15, δ=0.50 and α=0.01

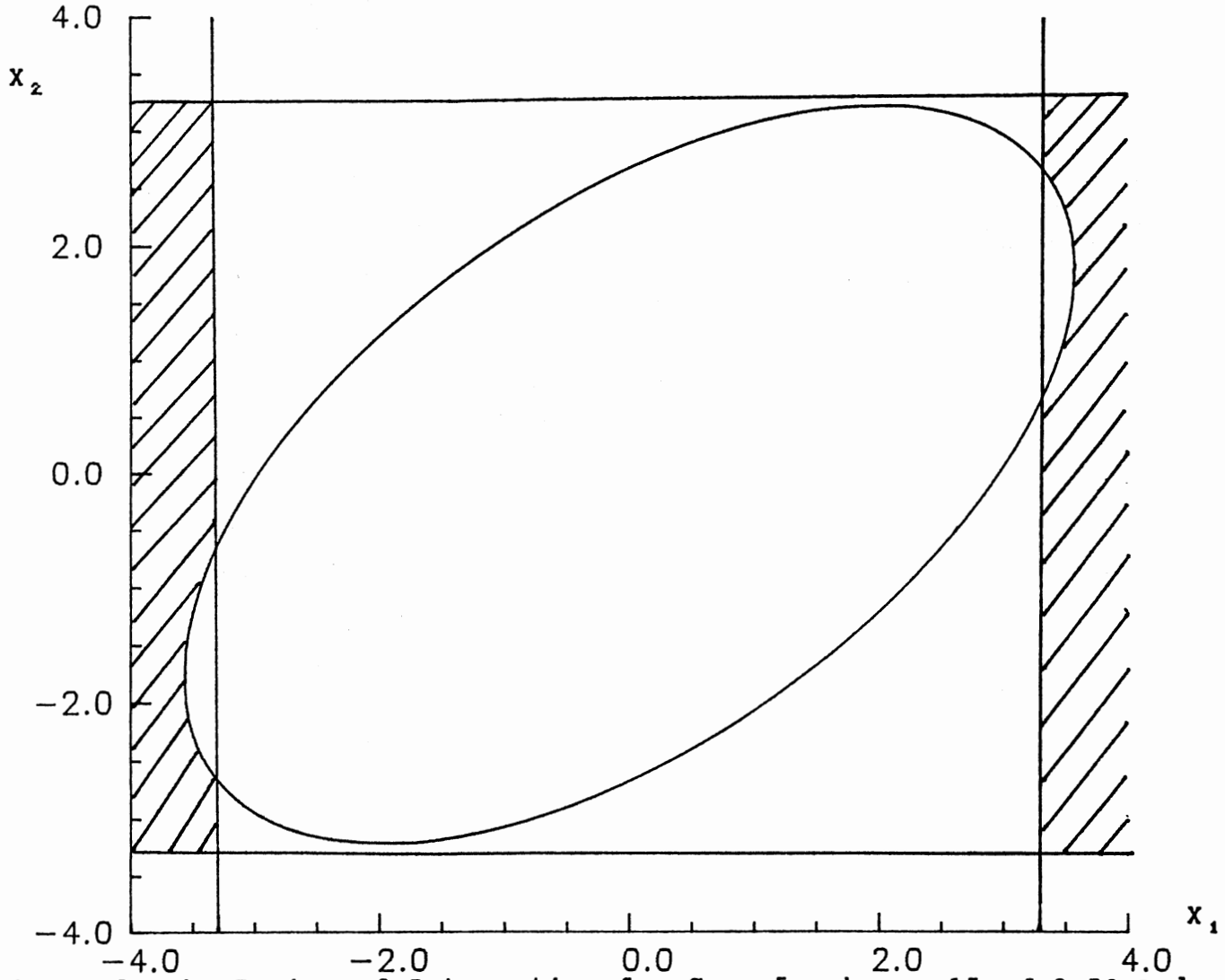

Figure 2. The Region of Integration for Case 5, when ν=15, δ=0.50 and α=0.01

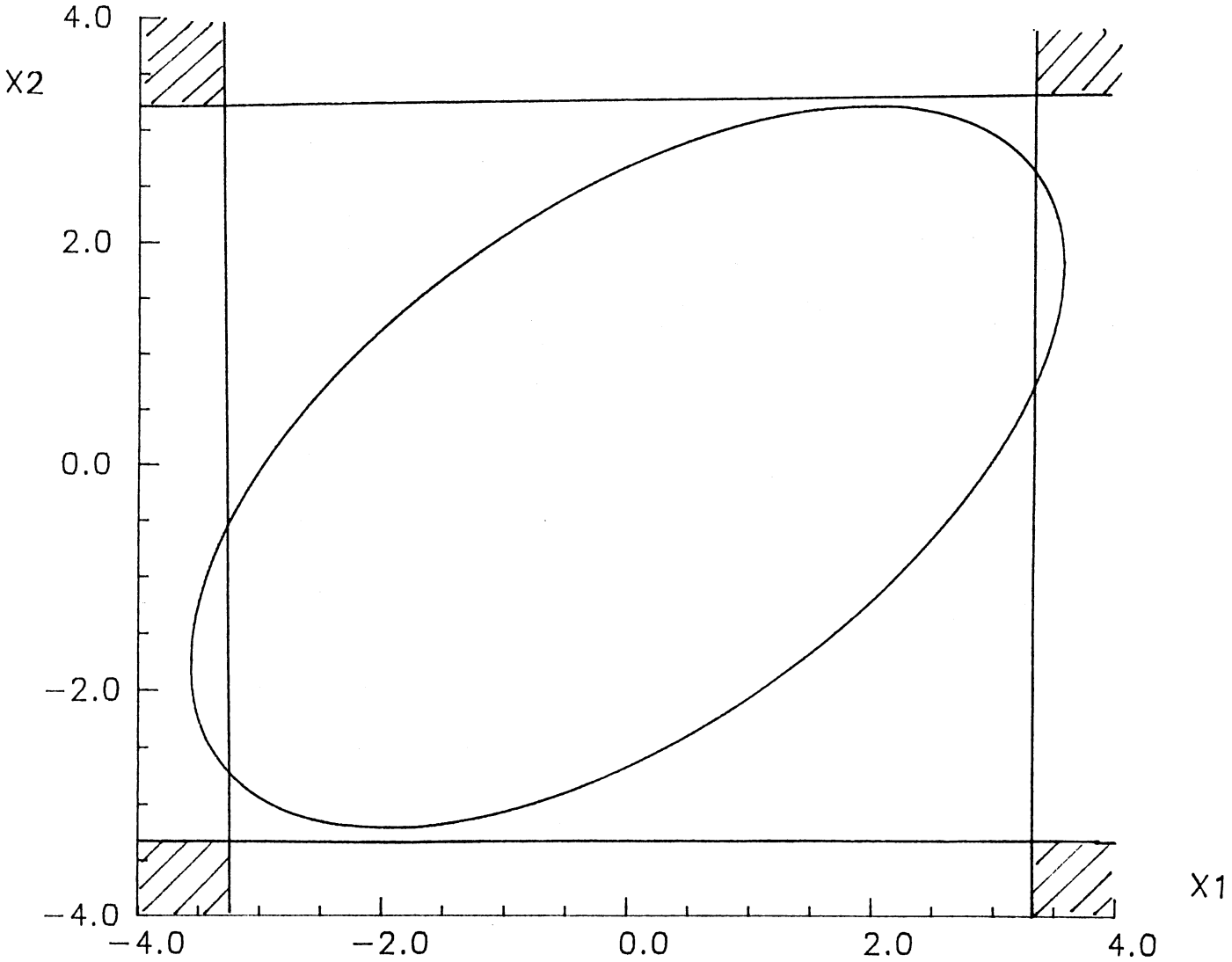

Figure 3. The Region of Integration for Case 6, when ν=15, δ=0.50 and α=0.01

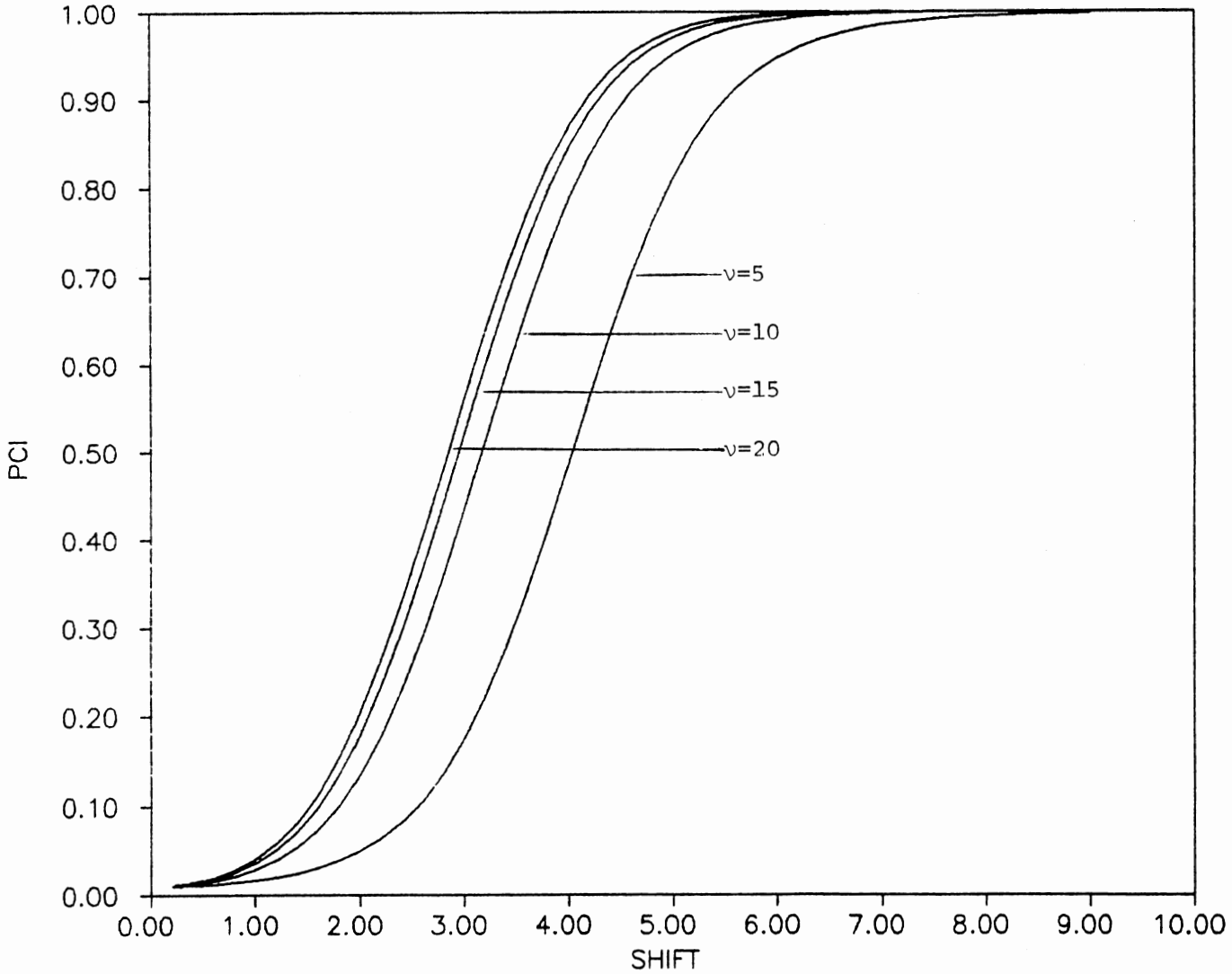

Figure 4. The Effect of ν=5, 10, 15, 20 on the PCI of Case 2, when α=0.01

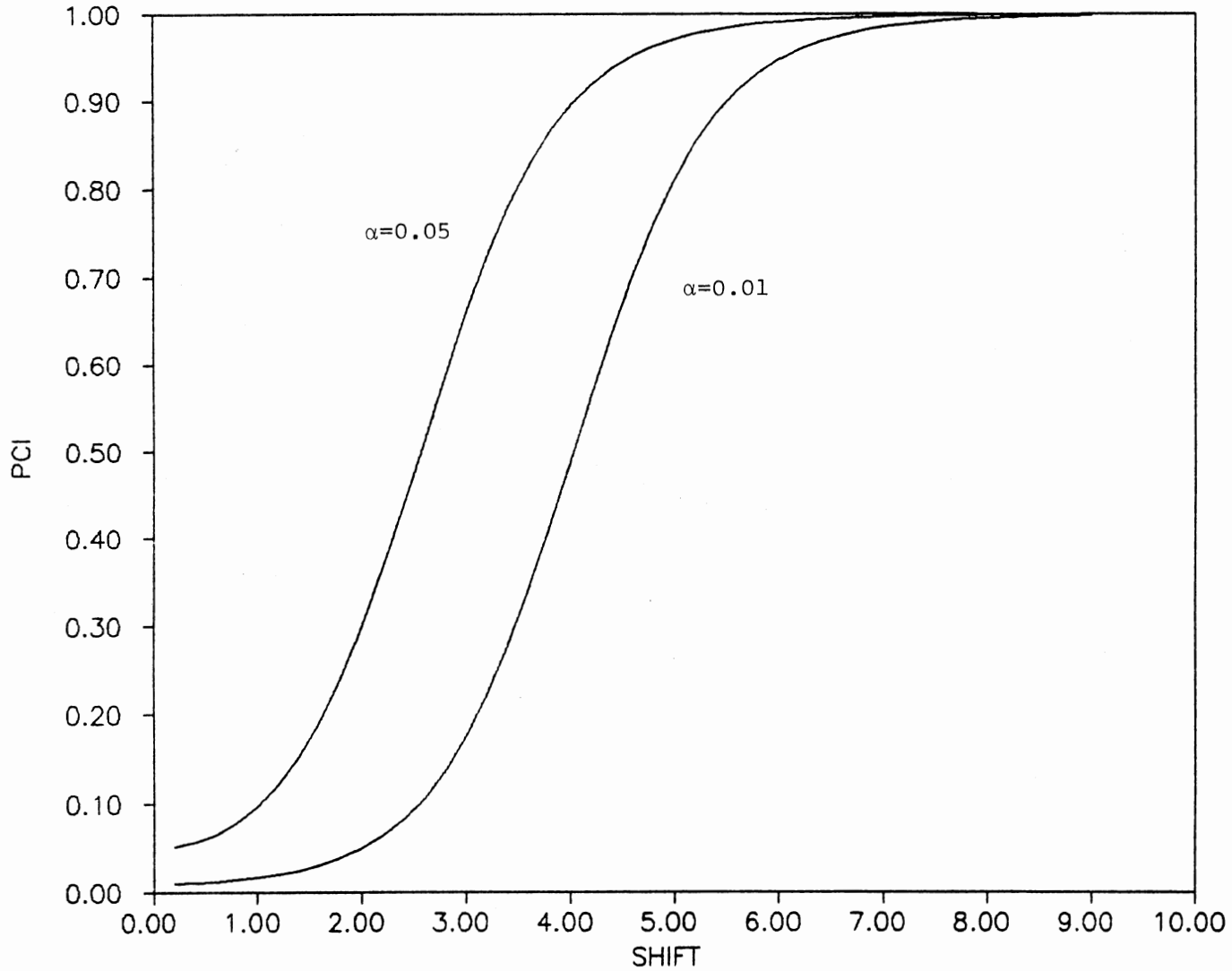

Figure 5. The Effect of α=0.01, 0.05 on the PCI of Case 2, when ν=5

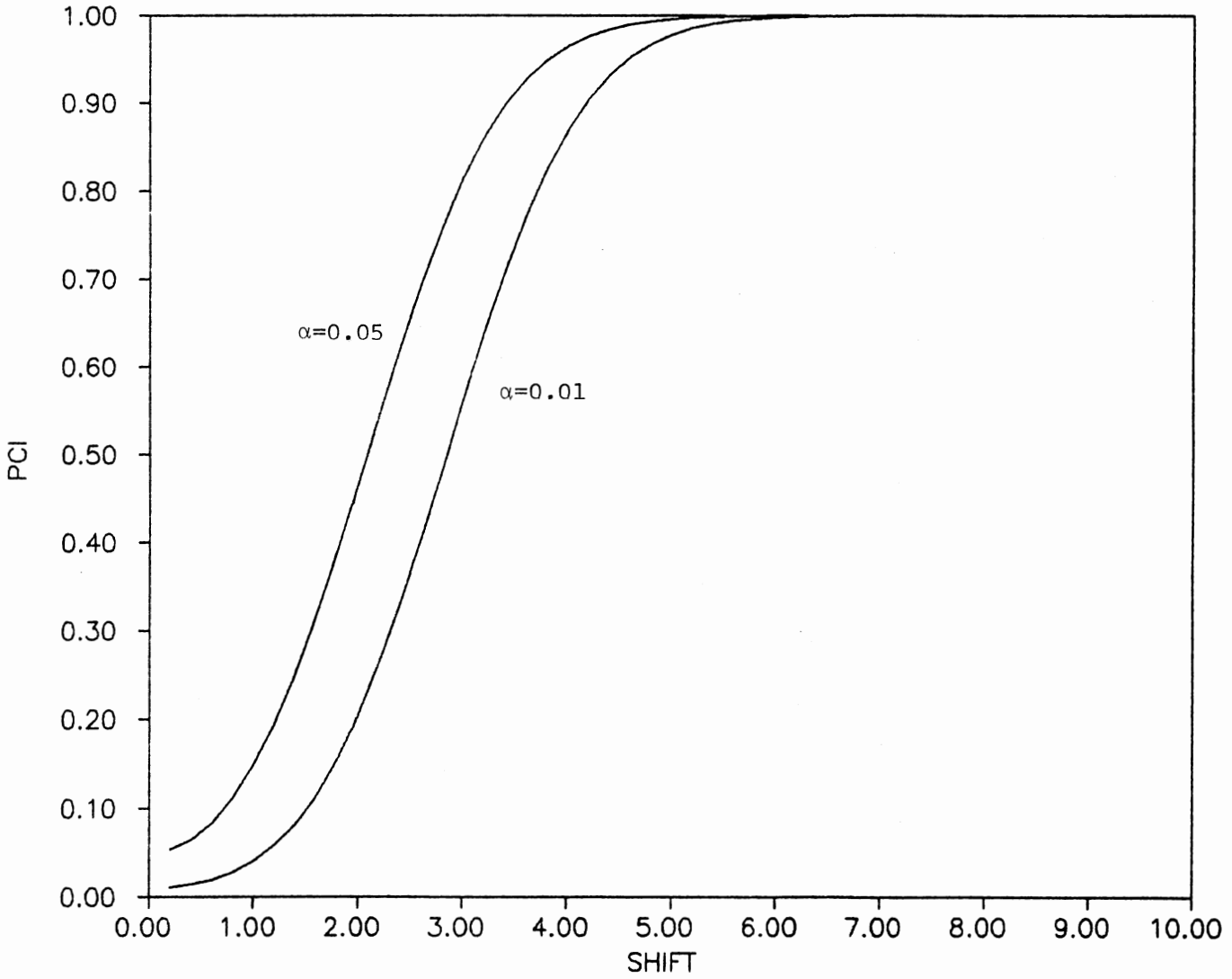

Figure 6. The Effect of α=0.01, 0.05 on the PCI of Case 2, when ν=20

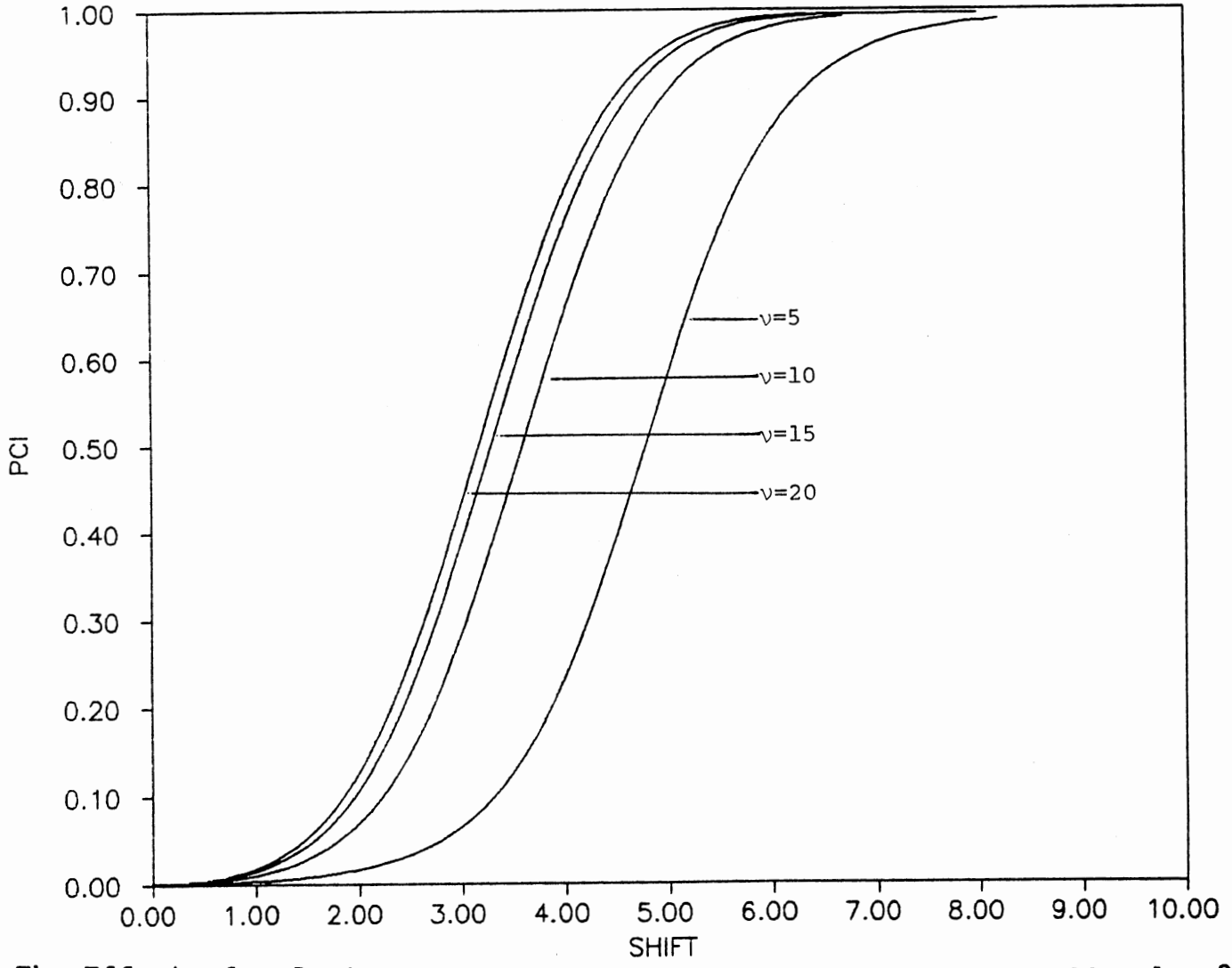

Figure 7. The Effect of ν=5, 10, 15, 20 on the PCI of Case 5, when δ=0.90 and α=0.01

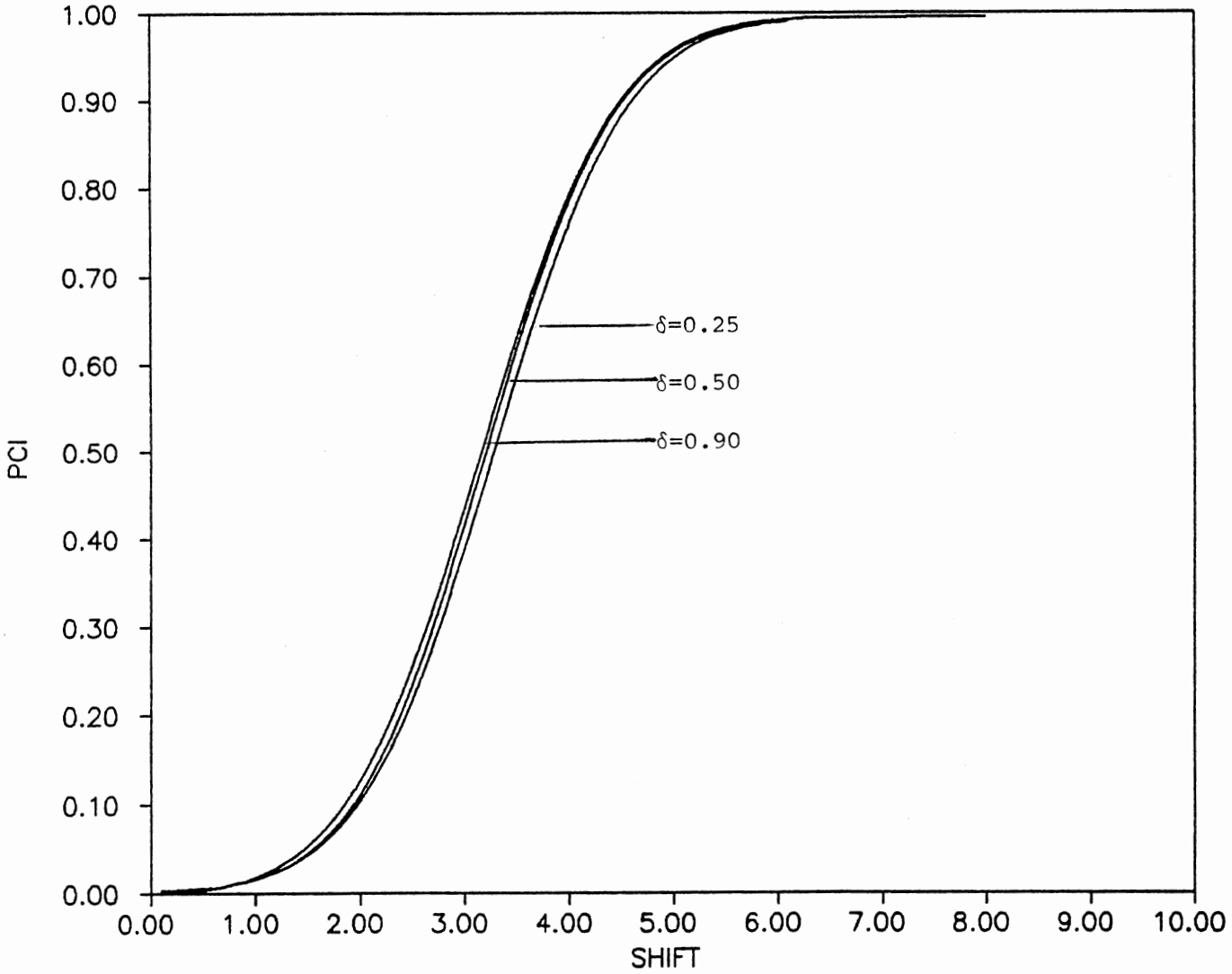

Figure 8. The Effect of δ=0.25, 0.50, 0.90 on the PCI of Case 5, when ν=20 and α=0.01

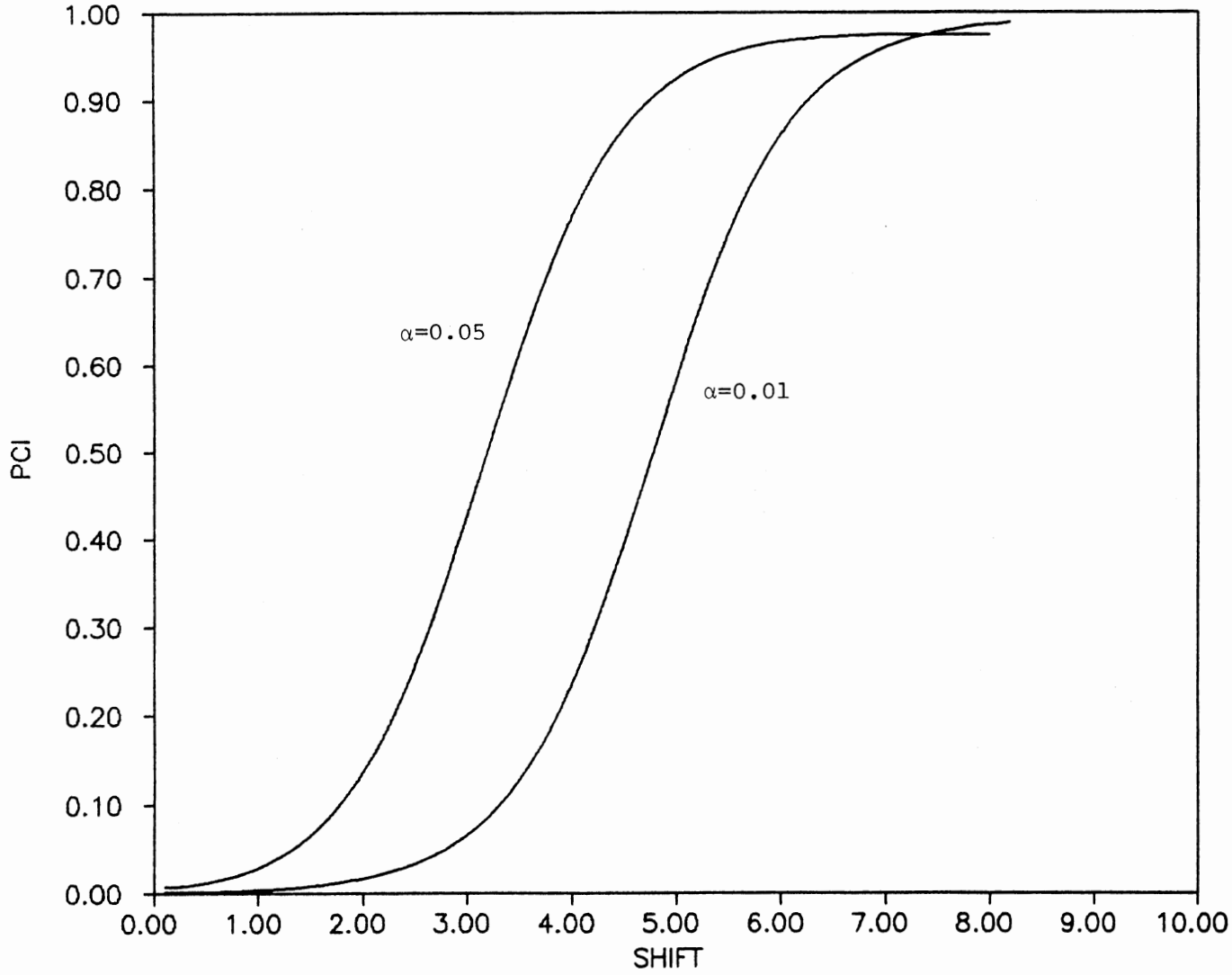

Figure 9. The Effect of α=0.01, 0.05 on the PCI of Case 5, when δ=0.90 and ν=5

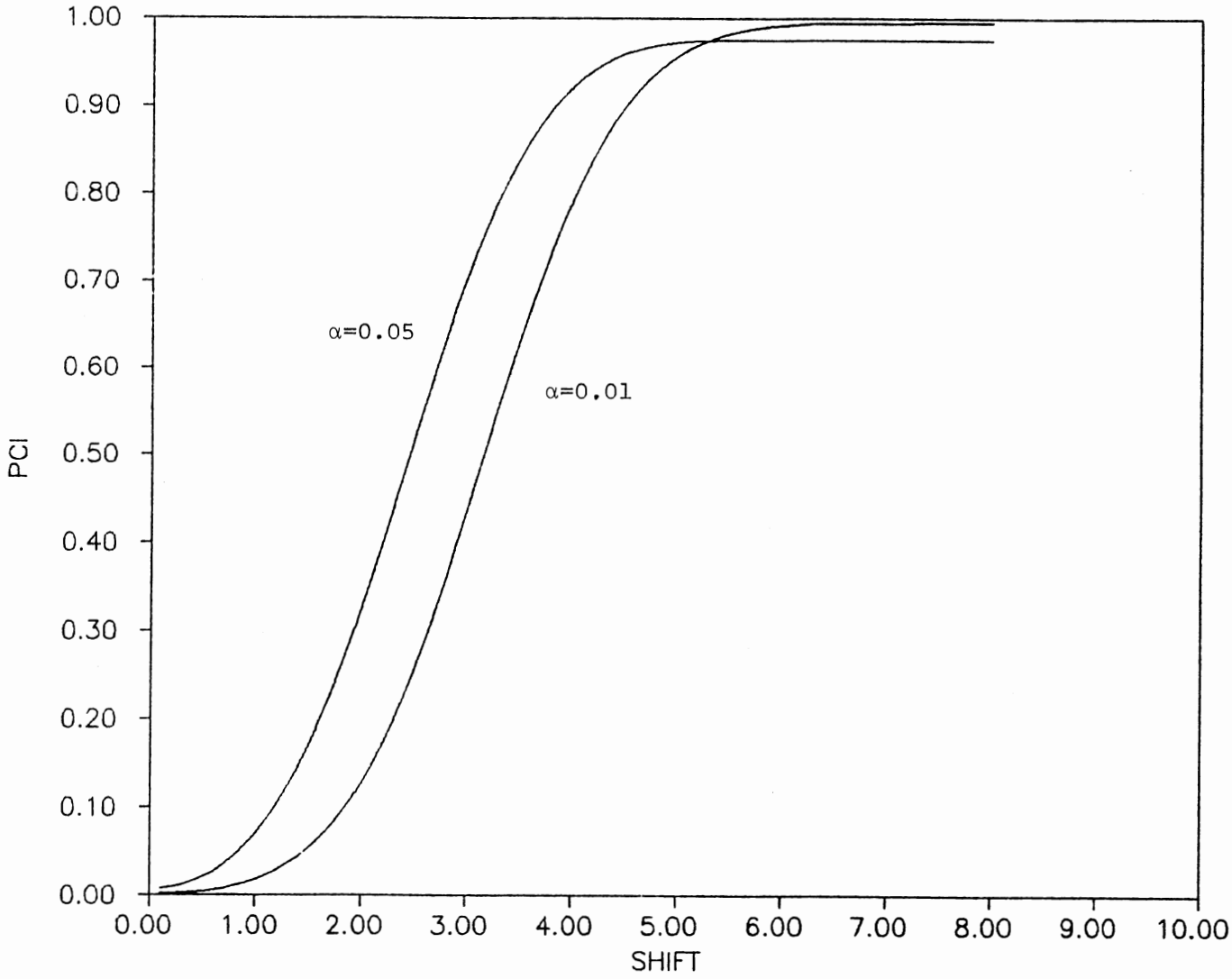

Figure 10. The Effect of α=0.01, 0.05 on the PCI of Case 5, when δ=0.09 and ν=20

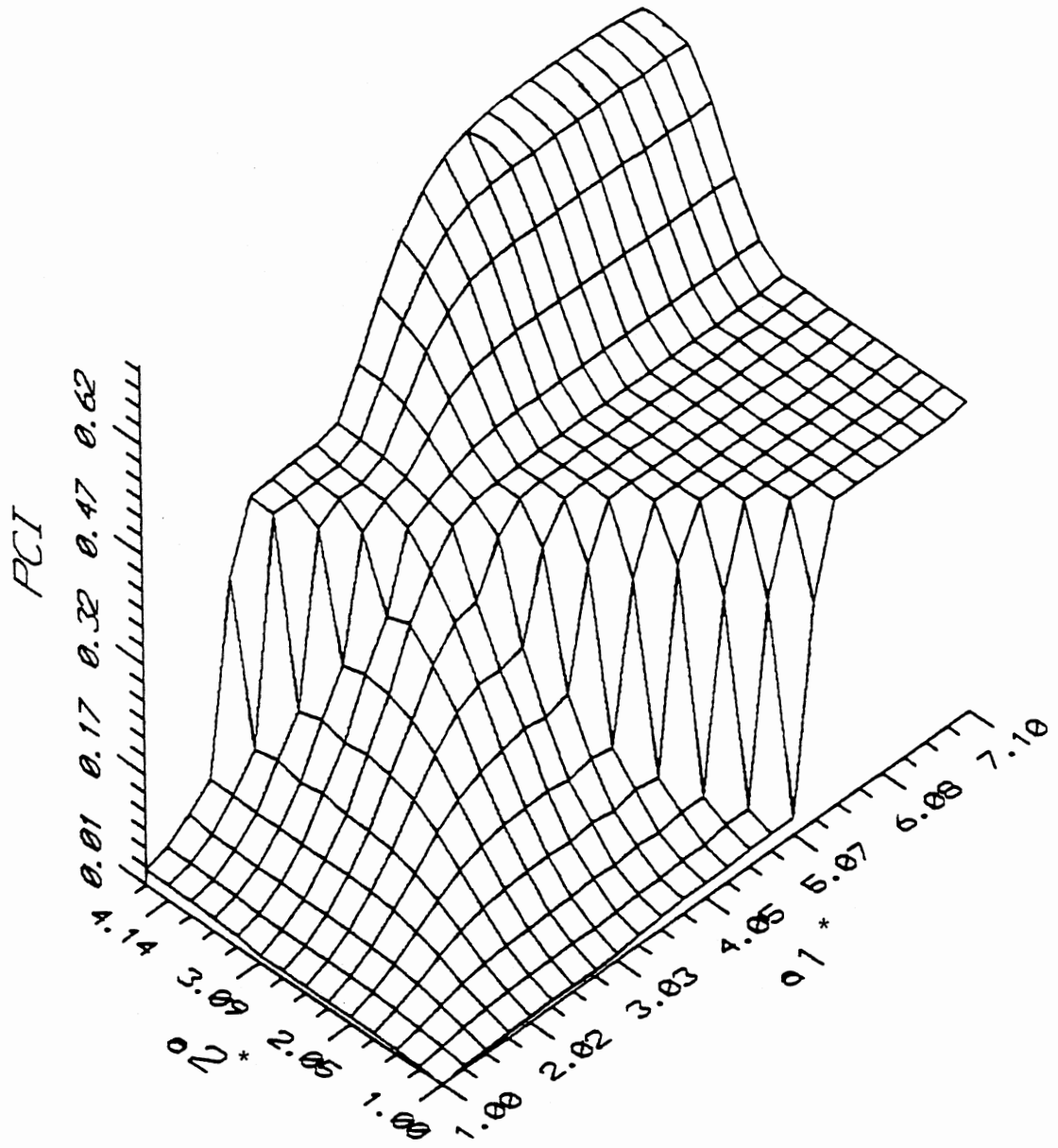

Figure 11. The Effect of $a_1^*>0$ and $a_2^*>0$ on the PCI of Case 6, when $\nu=20$, $\delta=0.90$ and $\alpha=0.01$

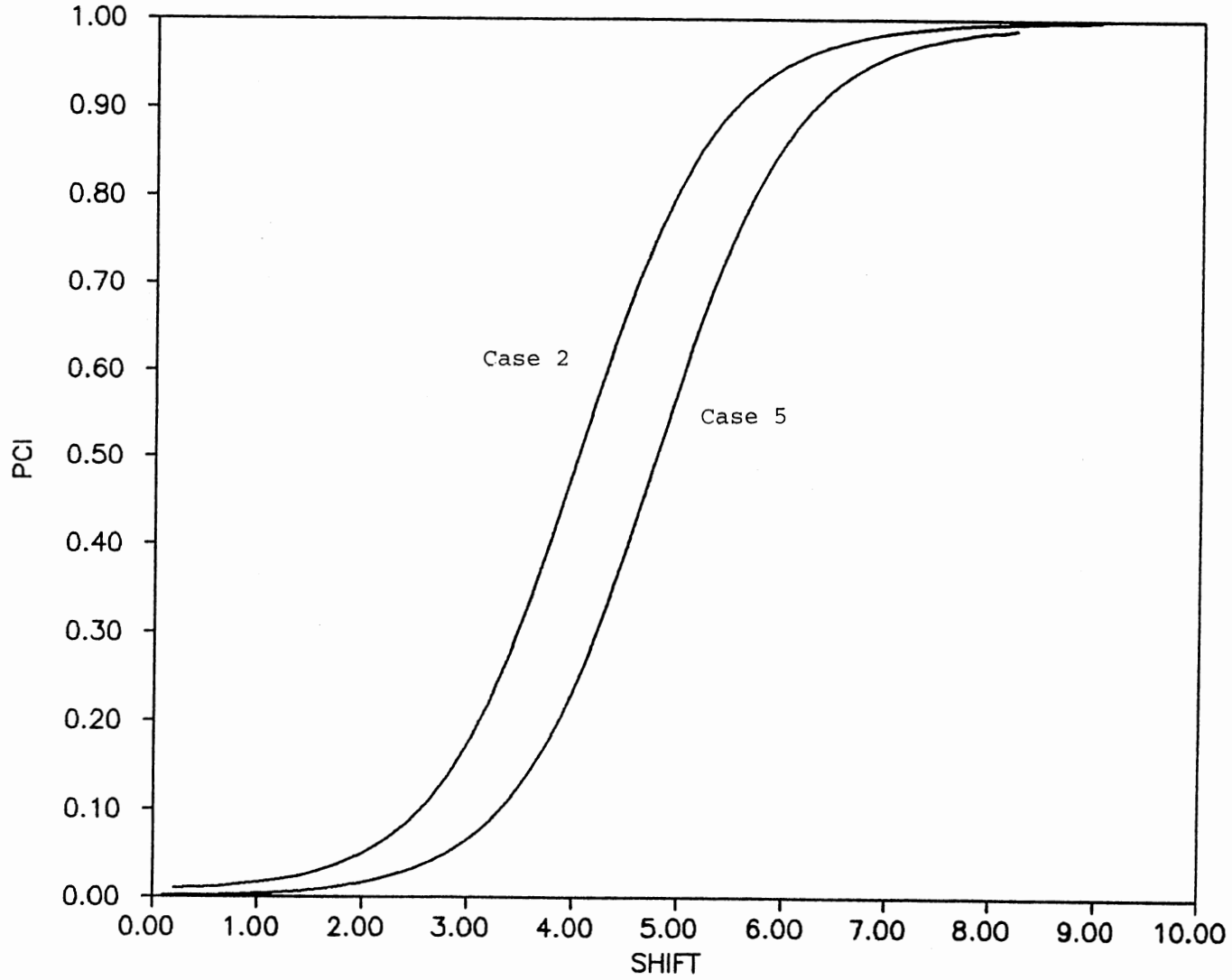

Figure 12. Comparison of Cases 2 and 5, when ν=5, δ=0.90 and α=0.01

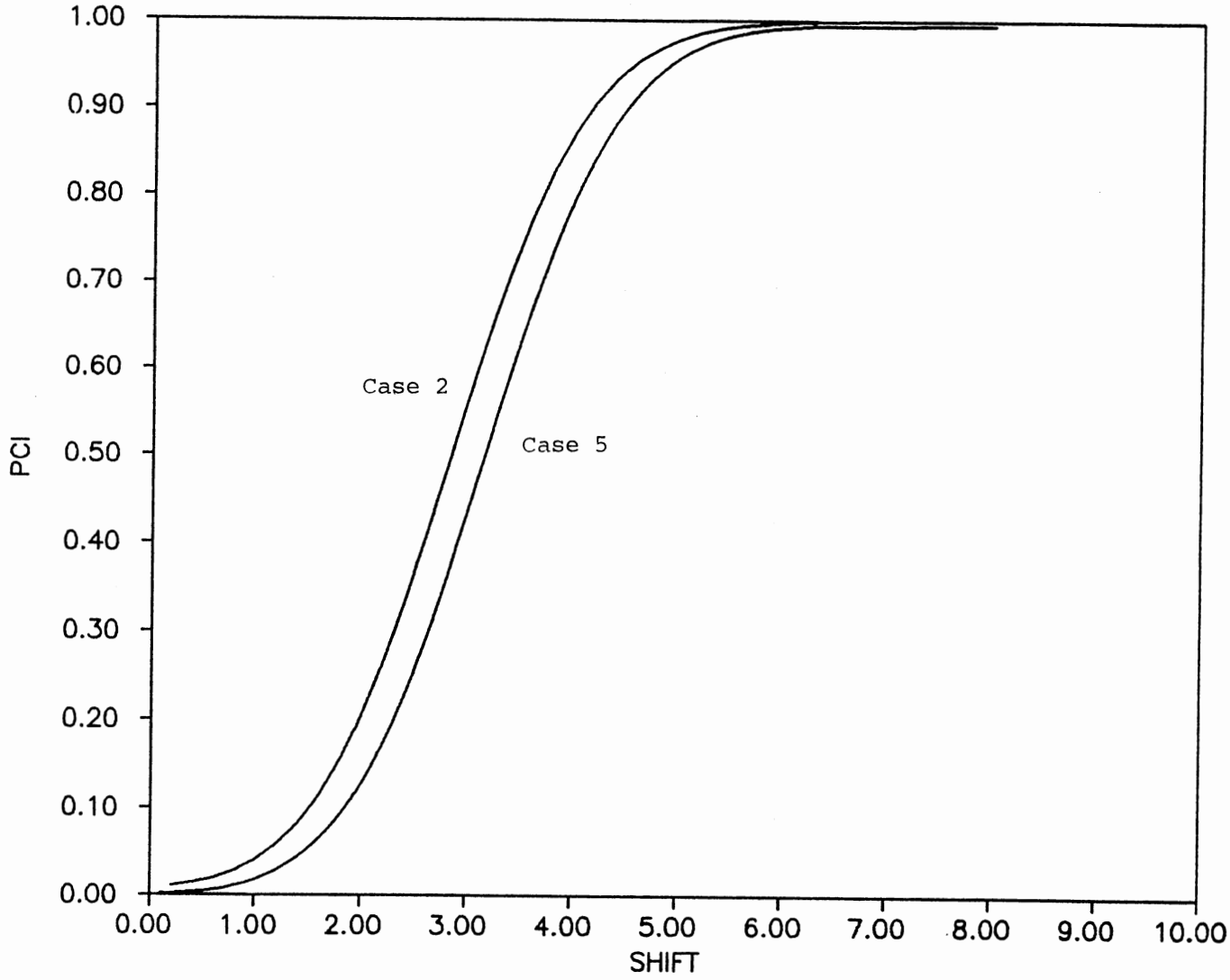

Figure 13. Comparison of Cases 2 and 5, when ν=20, δ=0.90 and α=0.01

VITA

Abu Hassan Shaari Mohd Nor

Candidate for the Degree of

Doctor of Philosophy

Thesis: THE EFFECT OF INITIAL CLASSIFICATION ON OUTLIER TESTING IN A LINEAR MODEL OF CONSTANT INTRACLASS CORRELATION

Major Field: Statistics

Biographical:

Personal Data: Born in Kg. Chuah, Seremban, N.Sembilan, Malaysia, March 12, 1957.

Education: Received Bachelor of Science degree in Mathematics from Southern Illinois University, Carbondale, in 1979; Master of Science degree in Statistics from the University of Iowa, Iowa City, in 1981; completed requirements for the Doctor of Philosophy degree at Oklahoma State University in July, 1989.

Professional Experience: Lecturer at the Universiti Kebangsaan Malaysia, 1981-1985.