1999

Process monitoring and fault classification for an air handling unit

Kyung-Jin Jang
Iowa State University

INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps. Each original is also photographed in one exposure and is included in reduced form at the back of the book.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6" x 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.

UMI

A Bell & Howell Information Company
300 North Zeeb Road, Ann Arbor MI 48106-1346 USA
313/761-4700 800/521-0600

Process monitoring and fault classfication for an air handling unit

by

Kyung-Jin Jang

A dissertation submitted to the graduate faculty

in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Major: Mechanical Engineering

Major Professor: Ron M. Nelson

Iowa State University

Ames, Iowa

1999

Graduate College
Iowa State University

This is to certify that the Doctoral dissertation of

Kyung-Jin Jang

has met the dissertation requirements of Iowa State University

Signature was redacted for privacy.

Committee Member

Signature was redacted for privacy.

Committee Member

Signature was redacted for privacy.

Committee Member

Signature was redacted for privacy.

Committee Member

Signature was redacted for privacy.

Major Professor

Signature was redacted for privacy.

For the Major Program

Signature was redacted for privacy.

For the Graduate College

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

NOMENCLATURE

$\mathbf{a}$	PCA Eigenvectors
D^2	Sample squared distance between two means
$\mathbf{E}$	Within sums of squared matrix
$\mathbf{H}$	Between sums of squared matrix
k	Number of groups
n	Number of observations
p	Number of variables
$\mathbf{S}_j$	Sample covariance matrix from group j
$\mathbf{S}_{pl}$	Pooled sample covariance matrix
$\mathbf{U}_\mathrm{i}$	i^{th} Principal component vectors
$\bar{y}$	Observation mean vectors
$\mathbf{y}$	n by p observation matrix
$\mathbf{y}_\mathrm{i}$	Observation vectors from i^{th} group
$\dot{Q}_w$	Volume flow rate of hot water, l/s
$\dot{Q}_{air}$	Volume flow rate of air, m^3/h
ΔP	Pressure rise across AHU, inches of water
Pow	Fan power, W
V	Volume, m^3
$\dot{W}_{fan}$	Fan power, W
$\dot{m}_w$	Mass flow rate of water, kg/s
$\dot{m}_a$	Mass flow rate of air, kg/s
T_{win}	Hot water inlet temperature, ° C
T_{wout}	Hot water outlet temperature, ° C
T_{ain}	Air inlet temperature, ° C
T_{aout}	Air outlet temperature, ° C
C_{p_w}	Specific heat of water, kJ/kg· K

$C_{p_{air}}$ Specific heat of air, kJ/kg· K

Greek Symbols

β Logistic coefficient vector

χ chi-square probability distribution

λ Vector of Eigenvalues

μ Population mean

ρ Density, kg/m^3

σ Standard deviation

Subscripts

α Level of significance

ν degrees of freedom

1 INTRODUCTION

Overview

Concerns for reliability and fault tolerance have challenged many scientific fields of study. Because there are vast numbers of different systems with their own system behaviors, one must pay close attention to the requirements of each individual system. These systems consist of multiple components interacting with each other in complex ways that can only be seen as an overall system performance. The prediction of faults for these systems is a challenge. With respect to this challenge a quote from E.F. Schumacher is found in the preface of the G.E. Box (1978, pg. vi).

> When the Lord created the world and people to live in it-an enterprise which, according to modern science, took a very long time-I could well imagine that He reasoned with Himself as follows: "If I make everything predictable, these human beings, whom I have endowed with pretty good brains, will undoubtedly learn to predict everything, and they will thereupon have no motive to do anything at all, because they will recognize that the future is totally determined and cannot be influenced by any human action. On the other hand, if I make everything unpredictable, they will gradually discover that there is no rational basis for any decision whatsoever and, as in the first case, they will thereupon have no motive to do anything at all. Neither scheme would make sense. I must therefore create a mixture of the two. Let some things be predictable and let others be unpredictable. They will then, amongst many other things, have the very important task of finding out which is which." (From *Small is Beautiful*)

The complex automatic systems in modern commerce and industry can consist of hundreds of inter-dependent working parts that are individually subjected to malfunction or failure. Total failure of these systems can present unacceptable economic loss or hazards to personnel. Therefore, provision for the required schedule of operation of the entire system should be implemented by the following scheme:

- A maintenance plan which will replace worn parts before they malfunction or fail.
- A monitoring plan which detects and identifies a fault as it occurs.

The above scheme of operation can contribute to sustainable "reliability" of the product and its overall service so that the system continues to operate satisfactorily. The major concern is the monitoring function including the detection, identification, and prediction of faults during real time operation of a dynamic system.

The main interest in the automation of technical processes in recent years can be observed in instrumentation, feedforward and feedback control, alarm monitoring and protection, etc. Good progress can also be seen in the technology and performance of modern measurement and control systems.

The improvements in process control are, on one-hand, based on the development of the components for sensors, transducers, control systems, and actuators. On the other hand, the understanding and modeling of process dynamics together with applied control theory can enhance our understanding of faults and their identification. The main idea is to detect process changes and faults during normal operation and to take actions to avoid damage to the process or injury to human operators.

The basis of this study relies on the fact that any change of the system induces a change in the behavior of the system. For example, in a building air handling unit, a sudden change in pressure rise and/or fan speed indicates the possibility of a fault that could result in the deterioration of system performance. A good diagnostic system should integrate all the information sources, process dynamics and control theory.

This research experimentally investigates system fault detection and looks into the relationship of the physical components of an air handling unit (AHU) in a heating and ventilation and air conditioning (HVAC) system. The study uses a systematic approach in applying multivariate methods for identification of system faults. Furthermore, this study examines the minimal set of the measured variables needed to develop fault detection model that produce a minimal number of false alarms.

Problem statement

An important objective of detecting faults in an AHU is to keep the number of false alarms to a minimum. When a fault does occur, it is beneficial to know what kind of fault it is and where the fault took place. The fault detection model should distinguish among different types of faults and normal operation with the minimal set of information. This study investigates categorizing and identifying faults and normal operation of an AHU with the fewest physical variables.

Although there are many methods for detecting faults, stochastic models provide information that is needed to identify the performance of a given system of interest. In this research, various multivariate statistics are used to classify faults with measured information from the physical variables of mechanical HVAC components. The methods of analysis used in this research include principal component analysis, discriminant and classification analysis, and logistic regression analysis.

One aspect of this study is to see whether there are noticeable differences in the detection of faults with a reduced set of variables. A reduced number of variables results in minimal sensor information, simplified monitoring schemes, and lower data acquisition costs.

Fault categories

This study investigates three types of faults in a Constant Air Volume (CAV) Air Handling Unit (AHU): RPM faults, valve faults, and coil faults. These faults can be detected by the use of the first law of thermodynamics and statistical classification. It is assumed that no faults occur simultaneously, since the probability of this occurrence is very small.

RPM faults are due to slippage of the fan belt. This results in the reduced CFM and possible changes in temperature measures. Features to look for are changes of the RPM measures leading to related heat transfer changes between the air and the hot water.

Coil faults are due to debris accumulated over the face of the heat exchanger. The main outcomes are decreased pressure rise across the AHU and decreased heat transfer rate between the air and the hot water. The level of the blockage may cause a temperature difference of the

inlet and outlet of AHU.

Valve faults are detected through changes in the mass flow rate of the water and by possible changes in heat transfer between the air and hot water.

Appropriate variables to monitor the performance of CAV systems can be tested by factorial experiments. For an example, any value of power that must be compared with a prediction to determine whether that power is acceptable or represents a fault condition, can be correlated with several explanatory variables such as temperature change, ΔT, mass flow rates, $\dot{m}$, or pressure rise, ΔP. Fan power imparted to the air stream, $\dot{W}_t$, depends on airflow through the fan, total pressure rise across the fan and an efficiency. The power measured at the fan shaft, $\dot{W}_s$, equation 1.1, is the mechanical power scaled by the efficiency of the fan.

$$\dot{W}_s = \frac{\dot{m}\Delta P}{\rho_{air}\eta_{fan}} \tag{1.1}$$

This research is conducted on the fan, valve and hot water heat exchanger coil in an AHU for normal and fault conditions during the heating modes of HVAC operation. The investigation involves experimental study to allow exploration of the measured variables to understand the principles involved in the detection of faults in an AHU. This study involves experimental design and statistical analysis.

2 LITERATURE REVIEW

The performance of HVAC components may change seasonally as the load and the weather change. Self-learning procedures may be needed to modify component models according to the changed state. In recent years, several approaches have been explored making use of steady state input-output relationships, energy balances, knowledge of required equipment sequencing, control logic, and dynamic system models.

This section summarizes the studies that have been done on the prediction of HVAC faults and diagnostics. The traditional approaches for process monitoring and fault detection compare measured rates of change to prescribed limits. A signal is generated if an abnormal situation occurs. In recent years, control researchers have developed a wide variety of new diagnostic techniques such as model-based methods including parameter estimation, residual evaluation and statistically based techniques including χ^2 tests, and artificial intelligence techniques such as neural networks and expert systems. Many fault detection techniques are based on the evaluation of model residuals, which are the differences between actual measurements and predictions from a process model. In general, large residuals are indicative of behavior that may be due to faults or unusual disturbances. A variety of statistical tests are used to determine if the residuals are statistically significant. The model predictions are generated with a variety of process models, such as steady-state or dynamic processes, physical or empirical approaches.

Lee, W.Y. *et al.* (1997) presented an application of a two-stage artificial neural network (ANN) for fault diagnosis and sensor recovery methods in a simulated air handling unit. The first stage ANN is trained to identify the subsystem faults and the second stage ANN is trained to diagnose the specific cause of a fault at the system level. The input variables used to train the neural networks were residuals between the normal operation model and

the "non-normal" temperature sensor values. Lee, W.Y. *et al.* (1996) developed a scheme for detecting faults using residual and parameter identification methods of autoregressive moving average with exogenous input (ARMAX) and autoregressive with exogenous input (ARX) model for single-input/single-output (SISO) and multi-input/single-output (MISO) structures. The model parameters are estimated using the Kalman-filter recursive identification method. Experimental data were generated from a laboratory variable air volume air handling unit operated with and without faults. In the study, eight complete faults of equipment and sensors were tested under constant load conditions and for short periods. Faults were detected when residuals and identification parameters changed significantly and thresholds were exceeded. Lee, W.Y. *et al.* (1996) applied the previous fault detection method to train a neural network for various faults conditions and successfully identified each fault.

Fasolo *et al.* (1995) utilized a controller performance index developed by Desborough *et al.* (1992, 1993) for an online monitoring fault detection technique for a hot water heat exchanger in an air duct. The controller performance index fault detection model is based on a standard time-series model of the process and a stochastic disturbance:

$$Y(i) - \mu = \frac{\omega(B)B^b}{\delta(B)}u(i) + \frac{\theta(B)}{\phi(B)\Delta^d}a(i) \tag{2.1}$$

where $Y(i)$ is the i^{th} output variable or the controlled variable, μ is the population mean, and $w(B)$, $\delta(B)$, $\phi(B)$, and $\theta(B)$ are the autoregressive integrated moving average coefficients. The input variable $u(i)$, or the manipulated variable, is expressed as a deviation from the reference value required to keep Y at μ. The stochastic disturbance, $a(i)$, assumed to be independent and identically distributed (iid) with zero mean and some variance. Six fault conditions were observed and the results were compared to the standard statistical quality control charts. The simulation study resulted in the detection of the faults in the feedback control loop operation and efficient computation in obtaining the results.

Haves *et al.* (1996) utilized a radial basis function network to generate data for the complete operating range of a system and used this for the estimation of the parameters of the first principles model of 2 faults in the cooling coil of an air handling unit. In their study, tracking

of the degradation faults were observed by changes in the weights of the radial basis function. Dexter *et al.* (1996) utilized a fuzzy fault diagnostic method for terminal boxes and heating coils. In this method, no training is required from the actual plant and it is suitable for real time implementation in packaged digital controllers or in energy management and control systems. Yoshida *et al.* (1996) investigated a mathematical system dynamics model using the autoregressive exogenous (ARX) model and the extended Kalman filter. In their study, sudden faults were generated using the HVACSIM+ program. Li *et al.* (1996) applied neural networks for developing a fault diagnosis method. In their study, simulations of a building were generated using commercial simulation software. Six different faults were selected for the boiler. Stylianou *et al.* (1996) utilized a combination of thermodynamic modeling, pattern recognition, and expert knowledge to diagnose the faults of a reciprocating chiller. Peitsman *et al.* (1996) implemented ARX and neural network approaches to identify the faults in a simulated reciprocating chiller. Tsutsui *et al.* (1996) studied the suitability of topological case-based modeling (TCBM) for district heating and cooling systems.

Tugnait (1992) applied cumulant statistics to noise prone system signals. In his study of parameter estimation and system identification for stochastic linear systems, inputs were assumed to be non-Gaussian while the noises were assumed Gaussian. Tugnait considered that the use of higher order cumulant statistics can yield consistent parameter estimates. The performance criterion based on the fourth cumulant of a generalized error signal was proposed to estimate the system parameters. In the study, Tugnait concluded that without knowing the noise statistics, the estimation yielded biased parameter estimates.

In the identification of the signal in space modeling of a microwave landing system (MLS), signal error source identification was conducted by Kelly (1992) by using the ARMA model approach to detect invariant signal structure in the random error components. The purpose was to make an intelligent estimate of the unexplained data or residuals. In particular the focal point was in the filter transfer function computed from the flight test. A MATLAB identification tool box was employed for model identification.

The integration of system identification and robust control design uncertainties was studied

by Karlov *et al.* (1994). The purpose was to identify a control design model that results in the robust performance of a closed loop system. The essential element of this scheme is an ability to evaluate the residual uncertainties in the parameters and determine the cost in terms of robust control performance. Karlov *et al.* view present identification theory and practice effort to be directed toward the estimation of uncertain parameters rather than evaluation of their accuracy characteristics (bounds). In it, the Empirical Transfer Function Estimation (ETFE), nonlinear Least Squares and Maximum Likelihood Methods, Prediction Error Methods, Eigensystem Realization Algorithm, and Q-Markov allow the construction of models from measurement data that describe model structure and its parameters involving nonlinear operations which hinder analytical error analysis. In the study, an extended Kalman filter is applied in the identification of model parameters and Petersen-Hollot's bounds for the robust control algorithm.

Haberl and Claridge (1987) established relationships, based on historic data, between environmental, building load, and occupancy variables (input) and fuel consumption measured at the site (output). These relationships were derived from manually recorded daily energy consumption data and from load recorders and electronic sensors. Initial data were reviewed for the catastrophic faults (undetected weekend equipment operation), which when corrected established a period of normal operation. Subsequent input-output correlations were compared with the reference period of normal operation to detect problems, using a rule-based system in some applications.

Anderson et al. (1989) combined a statistical preprocessor and a rule based system to diagnose HVAC system faults, using data automatically recorded by a data logger. Statistical analysis consisted of two steps: first, redundancy checks for sensor failures, and second, a comparison of measurement against predictor established by correlating historical data. Differences between measurement and prediction were flagged if in excess of a variation based on the standard deviation of the fit between historical data and the model used to define the predictors. The data fits were made via singular value decomposition (Press 1988) to pinpoint and remove singularities associated with correlated independent variables.

One possible means to handle the anomalous data identification procedures is to compute

measurement error estimates and use them as random variables of concern and make decisions on the basis of a hypothesis testing that utilize the statistical properties of the data. This hypothesis testing identification was investigated by Mili *et al.* (1984). One of the approaches used is the methodology of non-quadratic criteria. This state estimation allows online monitoring of power systems to identify bad measurements and clear the final data base from the induced errors. The hypothesis testing identification (HTI) is based on the computation of measurement error estimates and on the definition of the decision rules. This method uses the weighted least squares (WLS) estimates based on the quadratic criterion to estimate the static state. The bad data are detected under hypothesis testing. Although this method was considered as a new concept and highly encouraged in 1984, much improvement can be expected by the Bayesian approach.

Norford *et al.* (1987, 1990) took advantage of an extensive set of sensors attached to a data logger to establish relationships between chiller efficiency (chiller electrical use) and chiller load and between fan power and airflow. Duct temperature and flow data were also used to establish energy balances that could reveal excessive or deficient outdoor air intake into the building. Control logic faults such as chilled water pumps in prolonged operation with chillers off, were detected. Fault detection was made via visual inspection of graphs or with a rule based system.

In a study by Isermann (1989), a fault and the time of its occurrence is recognized through the determination of the changes in the process coefficients with respect to the normal operation reference. His methods curtail theoretical process modeling and estimates of the continuous time models, and Bayes decision and classification. His findings in the case study of heat transfer in the steam coil were that due to significant differences, despite simplification and assumptions for theoretical models, in faulty state and normal state that detailed theoretical modeling was not necessary and the faults were detected through patterns of changes of the study. Furthermore, he has shown that two measured signals are sufficient to monitor three process model parameters and to detect four artificially generated faults.

Baruch (1984) and Berman *et al.* (1983) developed a system identification procedure to

estimate the parameters of a structural system from dynamic test results. They determined the parameters by identifying a set of minimum changes in the original stiffness or mass matrices so that the analytical mode shapes agree with test measurements. A probabilistic scheme for system identification for nonlinear systems using a recursive filtering algorithm was proposed by Yun *et al.* (1980). Collins *et al.* (1974) developed a statistical identification method that uses measurements of mode shapes and natural frequencies to estimate the parameters of a linear dynamic system. A probabilistic system identification scheme was investigated by Gangadharan *et al.* (1991). Here, a weighted regression analysis is applied to experimental measurements or results from a detailed finite element analysis to estimate the parameters of a welded joint in a car body. The aforementioned studies have provided the basis for this research on classification and prediction of HVAC faults and diagnostics using an experimental approach for a building air handling unit.

3 METHODS OF ANALYSIS

This chapter covers AHU process dynamics, multivariate methods of principal component analysis, discriminant and classification analysis, and logistic regression used for the research. A variety of statistical and linear model techniques have been utilized to develop a fault detection model. With the application of the process model procedure and probability density function, a more descriptive representation of the variation of fault detection and classification are investigated. One particular link that arose from this study is the similarity of the equations for the AHU process model and logistic model.

Process dynamics in an air handling unit (AHU)

The AHU system consists of the fan and heating water heat exchanger shown schematically in Figure 3.1. The air outlet temperature, T_{aout}, is considered as the output variable, and the flow rate of hot water, Q_w, air flow rate, Q_{air}, inlet temperature of the hot water, T_{win}, outlet temperature of the hot water, T_{wout}, and inlet air temperature, T_{ain} are considered as the input variables.

The first law of thermodynamics is applied to the open control volume system to obtain the dynamic process model equation 3.1. Equation 3.1 assumes negligible kinetic and potential energy terms, constant properties of the fluids, and adiabatic air handling unit housing. Moreover it is assumed that most of the air in the AHU is at the outlet air temperature.

$$\rho V C_p \frac{dT_{aout}}{dt} = \dot{W}_{fan} + \dot{m}_w C_{p_w}(T_{win} - T_{wout}) + \dot{m}_a C_{p_{air}}(T_{ain} - T_{aout}) \tag{3.1}$$

where

- ρ = Density of air, kg/m^3

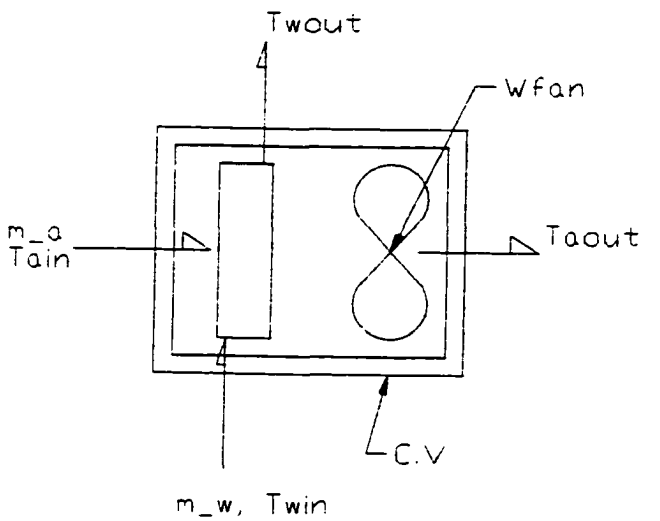


Figure 3.1 Control volume of the AHU

- V = Volume of the air in the AHU, m^3
- $\dot{W}_{fan}$ = Fan power, W
- $\dot{m}_w$ = Mass flow rate of water, kg/s
- $\dot{m}_a$ = Mass flow rate of air, kg/s
- T_{win}, T_{wout} = Hot water temperature, $°C$
- T_{ain}, T_{aout} = Air temperature, $°C$
- $C_{p_w}, C_{p_{air}}$ = Specific heat of water and air, $kJ/kg \cdot K$

The system has 5 input variables, namely the $\dot{W}_{fan}$, $\dot{m}_w$, $\dot{(m)}_a$, T_{win}, and T_{ain}, affecting the process output. If we assume that the air flow rate is maintained constant and the inlet water temperature is fixed, the system has three input variables. Although the number of input variables have been reduced, nonlinearity exists in the product terms between $\dot{m}_w$ and T_{aout}. Since classical linear process control theory has been developed for linear process systems, the model equation needs to be linearized. A linear approximation of a nonlinear steady state model is most accurate near the point of linearization and the same is true for dynamic

process models, Seborg (1989). Although wide changes in operating conditions for a nonlinear process cannot be approximated satisfactorily by linear expressions, HVAC processes remain in the vicinity of a particular operating state. Hence the linearized model may be sufficiently accurate.

In analyzing process dynamics, the model is made as general as possible. If the process model was linear from the beginning, one way to eliminate the explicit dependence of the model on the original steady state condition is to subtract the steady state relation from the differential equation model. The nonlinear unsteady state equation can be linearized by using a Taylor series expansion and truncating the higher order terms. The reference point for linearization is the normal steady state operating points. Hence if the unsteady state equation is of the form $dy/dt = f(y, x)$, estimated at the reference point $(\bar{y}, \bar{x})$, linearization is accomplished by the following procedure. Here, a bar over the variables indicates steady state variables.

$$f(y,x) \approx f(\bar{y},\bar{x}) + \left(\frac{\partial f}{\partial y}\right)_{\bar{y},\bar{x}} (y-\bar{y}) + \left(\frac{\partial f}{\partial x}\right)_{\bar{y},\bar{x}} (x-\bar{x}) \tag{3.2}$$

From the Taylor series expansion, deviation variables are formed naturally and because the steady state condition corresponds to $f(\bar{y}, \bar{x}) = 0$, the linearized differential equation in terms of $y^\star$ and $x^\star$ is developed in equation 3.3. The deviation variable, $T^\star \equiv T - \overline{T}$, is described as the measured deviation from the original steady state and is sometimes referred to as a perturbation variable.

$$\frac{dy^\star}{dt} = \left(\frac{\partial f}{\partial y}\right)_{\bar{y},\bar{x}} y^\star + \left(\frac{\partial f}{\partial x}\right)_{\bar{y},\bar{x}} x^\star \tag{3.3}$$

Thus with the steady state values $(\overline{T}_{aout}, \overline{\dot{W}}_{fan}, \overline{T}_{ain}, \overline{\dot{m}}_w)$, and by assuming the mass flow of the air is supplied with near constant value (CAV), the following linear process model is obtained.

$$\rho V C_p \frac{d\overline{T}^\star_{a\,out}}{dt} = \dot{W}^\star_{fan} + C_{p_w}(\overline{T}_{win} - \overline{T}_{wout})\dot{m}^\star_w - \overline{\dot{m}}_w C_{p_w} T^\star{}_{wout} + \dot{m}_a C_{p_{air}}(T^\star_{a\,in} - T^\star_{a\,out})$$

Simplifying further and bringing the $T^\star_{a\,out}$ to left side of the equation and dividing by the

constant term $\dot{m}_a C_{p_{air}}$ the following is obtained.

$$\frac{\rho V C_p}{\dot{m}_a C_{p_{air}}} \frac{dT^*_{a\,out}}{dt} + T^*_{a\,out} = \frac{1}{\dot{m}_a C_{p_{air}}} \dot{W}^*_{fan} + \frac{C_{p_w}(\overline{T}_{win} - \overline{T}_{wout})}{\dot{m}_a C_{p_{air}}} \dot{m}^*_w - \frac{\overline{\dot{m}}_w C_{p_w}}{\dot{m}_a C_{p_{air}}} T^*_{wout} + T^*_{a\,in} \tag{3.4}$$

The above equation is general in that it applies to any specified operating point. Next, define variables τ, K_1, K_2, K_3.

$$\tau = \frac{\rho V}{\dot{m}_a} \tag{3.5}$$

$$K_1 = \frac{1}{\dot{m}_a C_{p_{air}}} \tag{3.6}$$

$$K_2 = \frac{C_{p_w}(\overline{T}_{win} - \overline{T}_{wout})}{\dot{m}_a C_{p_{air}}} \tag{3.7}$$

$$K_3 = \frac{\overline{\dot{m}}_w C_{p_w}}{\dot{m}_a C_{p_{air}}} \tag{3.8}$$

Then, substitute the constants and take the Laplace transform of equation 3.4 with the initial steady state condition $T^*(0) = 0$ to obtain:

$$\tau s T^*_{a\,out}(s) + T^*(s)_{a_{out}} = K_1 \dot{W}^*(s)_{fan} + K_2 \dot{M}^*(s)_w - K_3 T^*(s)_{w_{out}} + T^*(s)_{a_{in}} \tag{3.9}$$

where $\dot{M}^*(s) = \mathcal{L}[\dot{m}^*(t)]$, etc . Collecting the terms and dividing through $(\tau s + 1)$ results in the recognizable form of process model with transfer functions.

$$T^*(s)_{a_{out}} = G_1 \dot{W}^*(s)_{fan} + G_2 \dot{M}^*(s)_w - G_3 T^*(s)_{w_{out}} + G_4 T^*(s)_{a_{in}} \tag{3.10}$$

$$G_1 = \frac{T^*(s)_{a_{out}}}{\dot{W}^*(s)_{fan}} = \frac{K_1}{\tau s + 1} \tag{3.11}$$

$$G_2 = \frac{T^*(s)_{a_{out}}}{\dot{M}^*(s)_w} = \frac{K_2}{\tau s + 1} \tag{3.12}$$

$$G_3 = \frac{T^*(s)_{a_{out}}}{T^*(s)_{w_{out}}} = \frac{K_3}{\tau s + 1} \tag{3.13}$$

$$G_4 = \frac{T^*(s)_{a_{out}}}{T^*(s)_{a_{in}}} = \frac{1}{\tau s + 1} \tag{3.14}$$

The transfer functions, G_1, G_2, G_3, G_4 have the same first order dynamics but different gains, K_1, K_2, K_3. Ordinarily, the transfer functions relate changes in process output to changes in process input and they contain information about the steady state and dynamic relationship between input and output, namely the process gain and the process time constant, τ. Since

the transfer function is defined only for the linear equations, a nonlinear process model must be linearized.

Process fault diagnosis based on dynamic model

According to Isermann (1989), the process model enables the estimation of process state variables and parameters influenced by faults. This study also adopts his methods based on process parameters and links them to the multivariate method, namely logistic regression and discriminant and classification. Parameter estimation, feature extraction, fault decision and classification using the dynamic model are outlined in the following paragraphs.

The mathematical process model is of the form: $\vec{Y} = f[\vec{U}, \vec{N}, \vec{X}, \vec{\Theta}]$. $\vec{U}(t)$ and $\vec{Y}(t)$ are measurable input and output variables. $\vec{N}(t), \vec{X}(t)$, and $\vec{\Theta}$ are disturbance variables, process state variables, and constant or slowly time-varying process parameters, respectively.

A process fault generally causes changes in process parameters and process state variables depending on the type of faults. Therefore, the output is also changed according to the process dynamic and static characteristics.

Fault detection methods may be classified according to the use of the following quantities.

- Measurable signals, $\vec{U}(t), \vec{Y}(t)$.
- State variables (mostly unmeasurable), $\vec{X}(t)$.
- Process parameters (mostly unmeasurable), $\vec{\Theta}$.
- Characteristic quantities, $\eta = f[\vec{U}, \vec{Y}, \vec{\Theta}]$.

Isermann's method of detecting faults are as follows:

- Range or limit checking $\vec{Y}_{min} < \vec{Y}(t) < \vec{Y}_{max}$
- Analyze output signal, Low and High frequencies of $\vec{Y}(t)$. The analysis of higher frequency components by autocorrelation or spectral analysis can give additional information concerning the inner state of a process.
- Calculate η i.e efficiencies, consumption rate, wear rate, etc changes. $\Delta\eta$ may give overall information on internal changes. But in most cases a detailed fault diagnosis is not possible.

- If a process model is known, one can try to estimate the usually unmeasurable process state variables $\vec{X}(t)$ or process parameters $\vec{\Theta}$ based on the measurable inputs $\vec{U}(t)$ and outputs $\vec{Y}(t)$ by using state variable or process parameter estimation methods and *to detect changes* $\Delta\vec{X}(t)$ or $\Delta\vec{\Theta}(t)$. This is also known as analytical redundancy.

Application of the process model based fault detection is applicable to preventive maintenance, maintenance on request (instead of fixed schedules), remote diagnosis of processes, and automatic inspection of products without disassembling in manufacturing.

A generalized scheme of technical fault diagnosis is shown in Figure 3.2. For the parameter estimation and theoretical modeling flow chart see Figure 3.3. Isermann's contribution concentrates on model-based fault detection methods based on parameter estimation. Each phase is described in the following paragraphs. The procedure holds well for the fault diagnosis methods based on signal models and process models in non-parametric or parametric form. In the case of parametric process models the reduced information may consist of process model parameters or process model state variables.

Data processing. The measured signals are processed by methods of filtering and estimation such that the information reduction becomes suitable for fault detection and diagnosis. The reduced information, for example, exists in filtered signal components, correlation functions, or in parameter or state-variable estimates (if process models are applied).

Fault detection. With the reduced process information, features are extracted allowing the detection of faults in the process. Changes of these features are then determined with reference to the normal process. These changes are subsequently used to recognize the event of a fault and the time of its occurrence. For this task statistical decision methods may be used. This is the issue for our research using the discrimination between fault classes or groups.

Fault diagnosis. After a fault event is detected, the features and their changes are submitted to a classification procedure with the aim to determine the fault type, fault location, fault size and the cause of the fault. The above procedure is linked to the discrimination and classification in the multivariate analysis. It is noted by Isermann that fault decision and classification are so combined that their clear separation may not be possible. Nevertheless, there is clear connotation in the classification analysis and it is the subject of this study.

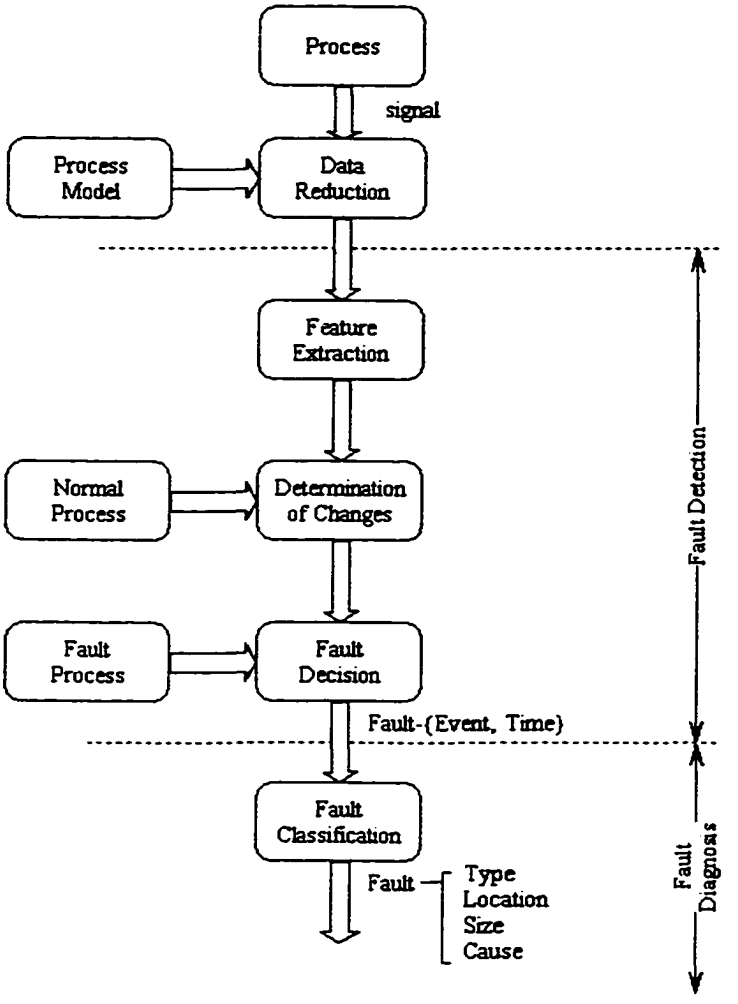


Figure 3.2 Generalized technical fault diagnosis flow chart, Isermann (1984)

The fundamental idea is that many process faults appear as changes of process coefficients, $\vec{p}$, like resistances, capacitances, mass, stiffness, etc. These process coefficients are contained in the parameters $\vec{\Theta}$ of a process model. Process model parameters are understood as constants or time dependent coefficients in the process which appear in the mathematical description of the relationship between the input and output signals, the process model. A distinction can be made between *static process models* in the form of polynomial equation, $Y(U) = \beta_0 + \beta_1 U + \beta_2 U^2 + \ldots$ and *dynamic process models* in the form of lumped parameters expressed in differential equations, $a_0 y(t) + a_1 \dot{y}(t) + \ldots + a_n y^n(t) = b_0 u(t) + b_1 \dot{u}(t) + \ldots + b_m u^m(t)$. For

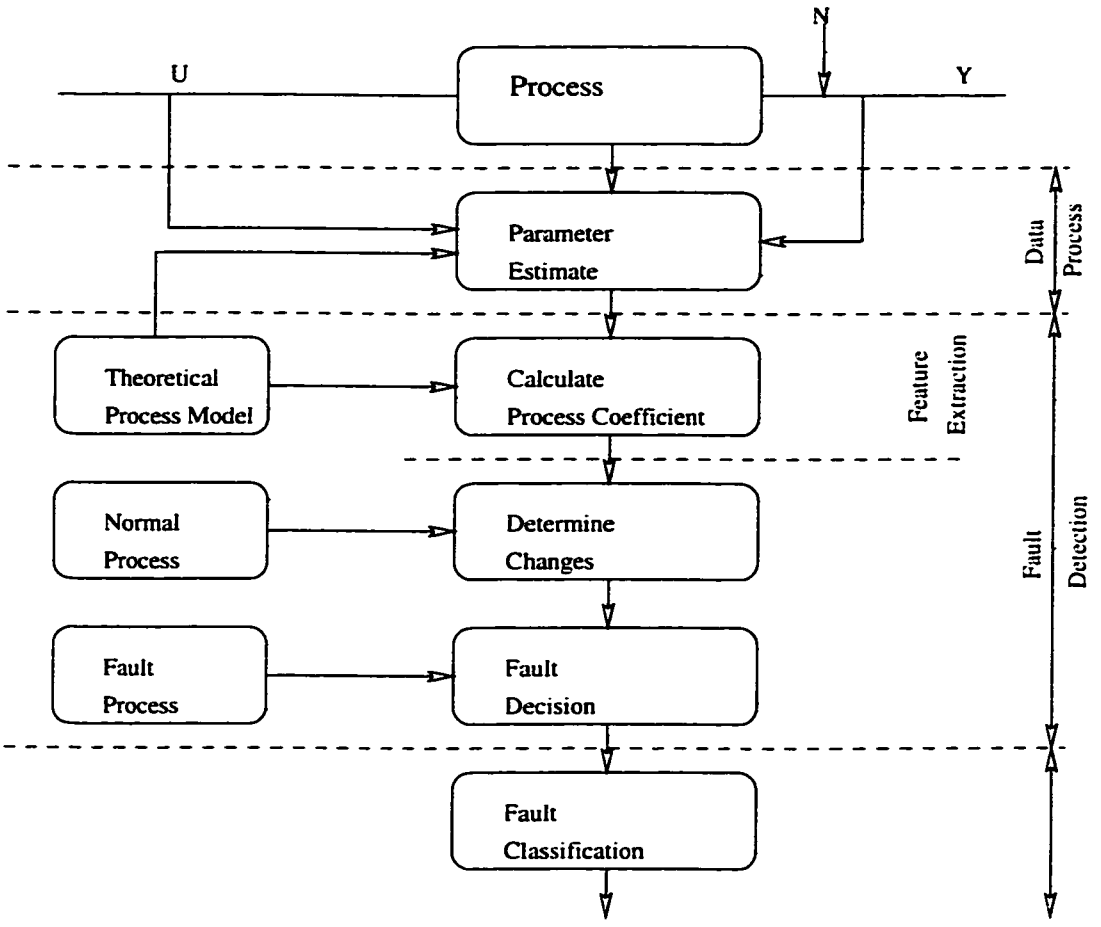


Figure 3.3 Technical fault diagnosis flow chart for parameter estimate model, Isermann(1984)

the simple case, the equations are linearized about one operating point. The process model parameters are given by, $\vec{\Theta^T} = [\beta_0\beta_1\beta_2\ldots]$ for the linear case and $\vec{\Theta^T} = [a_0a_1a_2\ldots a_{n-1} : b_0b_1b_2\ldots b_m]$ for the differential model. The estimation of the fault parameters in forms of logistic coefficients are estimated in logistic regression analysis which is discussed in a later section.

Multivariate methods

The most frequently used statistical method to analyze data is the method of analysis of variance. In the analysis of variance, or ANOVA, the effects of each independent variable are examined separately. For two or more dependent variables with one or more independent

variables, multivariate analysis of variance, or MANOVA, provides examination of not just the effects of each independent variable but also includes effects of combinations or interactions among independent variables. MANOVA allows a simultaneous test across all dependent variables. That is, MANOVA finds a linear combination of the dependent measures that maximize separations among groups.

In many applied disciplines, especially in the field of the engineering, researchers often measure several variables on each subject or experimental unit. Although the variable-by-variable approach may be productive in some cases, in most instances, the system is complex, and the variables are intertwined in such a manner that when analyzed separately, they yield little information about the system. All of the variables must be examined simultaneously in order to determine the key features of the system under study. This can be done with multivariate analysis, regardless of how many variables there are or how they are intercorrelated. The multivariate approach allows exploration of the joint performance of the variables and determination of the effect of each variable in the presence of the others. Because multivariate analysis provides both descriptive and inferential procedures, we can search for patterns in the data or test hypotheses about patterns of a priori interest. Multivariate descriptive technique allows researchers to look into the complex interaction of variables on the surface and extract the salient information about the system under study. Multivariate inferential procedures include hypothesis tests that process any number of variables without inflating the *Type I* error rate and allow for any intercorrelations the variable have, (Rencher 1995). The *Type I* error is committed when the significance test decision results in favor of the alternate hypothesis, H_a, when the null hypothesis is true. Although there are many multivariate descriptive and inferential software packages, *MATLAB* programs are written for the computational convenience. Each of the codes are attached in the Appendix C.

Principal component analysis (PCA)

Principal component analysis examines the variance structure in the data. In particular, PCA is the procedure used to change or transform correlated variables into uncorrelated vari-

ables. One purpose is to see if the first few principal components account for most of the variation in the data. If they do, then these few principal component variables can be used to describe the data without loss of information. One advantage of having the PCA is to reduce the dimensionality of the data, that is PCA removes the linear dependencies in the independent variables to adequately summarize the data. Also when the data are standardized (using the correlation matrix instead of the variance-covariance matrix) the unit measurement differences are eliminated between variables. The last few principal components are useful for detecting non-linear relationships between variables. This can be valuable information when several variables are measured on each of the units in a study describing the relationship among the variables.

A practical application is in a situation where independent variables outnumber the available number of observations that make a test ineffective. Another situation is where highly correlated independent variables produce unstable estimates. In these cases, the independent variables can be reduced to a smaller number of principal components, uncorrelated new variable, that will produce better test or stable estimates of the regression coefficients, Rencher (1995).

PCA also identifies essential features like the amount of variability or scatter in the relationship and units having measures unusually different from the others. This can also be achieved by means of simple linear regression and correlation coefficients. When correlation coefficients are calculated among variables it is an empirical fact that the coefficients differ from zero and often the difference from zero can not be explained by the variability caused by measurement/sampling/experimental errors. In other words there is a tendency for the average value of one variable to increase (or decrease) as the average value of the second variable increases. Sometimes this relationship is well known and expected. For other variables the existence of this relationship may be unexpected. It is useful to look into the reason of the causality of the correlation. One reason for the variables to be correlated maybe that one change in a certain variable causes the others to change. Another reason may be that an unknown variable exists and it causes the changes in known variables. This is the case when the

correlation between say Y_1 and Y_2 is not a measure of a causal relationship but a measure of how the variables are responding to a mutual cause. Certainly some of the variables measured in the flow loop may exhibit these effects and the utilization of PCA will enhance the model building by reducing variables without loss of information.

Because principal components analysis deals with a single sample of observations with no structure in the observations or among the variables within an observation vector, it is ideal for processing fault group data. First principal component has the largest variance and last principal component has the smallest variance. Because the variance of z_1 is λ, the largest eigenvalue, and the variance of z_p is λ_p, the smallest eigenvalue, we can speak of the proportion of variance explained by the first k components. Thus we try to represent the p-dimensional points $(y_{i1}, y_{i2}, \ldots, y_{ip})$ with a few principal components $(z_{i1}, z_{i2}, \ldots, z_{ip})$ that account for a large proportion of the total variance. If one of the variances is larger than the other variances, then that term in the principal component will have a large coefficient and all other coefficients will be smaller. When a ratio of the proportion of variance is used for discriminant functions and canonical variates, it is frequently referred to as percent of variance. In the case of discriminant functions and canonical variates, the eigenvalues are not variances as they are in principal components.

Selection of principal variables

In principal components, there are no dependent variables, as in regression, or no groupings among the observations, as in discriminant analysis. With no external influence, the objective of the selection variables is to find the subsets that best capture the internal variation of the variables. Jolliffe (1972, 1973) discussed selection methods and referred to the process as *discarding variables.*

The method is based on multiple correlation, clustering of variables, and principal components. One of the correlation methods proceeds in a stepwise fashion, deleting at each step the variable with the largest multiple correlation with the other variables. The clustering methods partition the variables into groups or clusters and select a variable from each cluster.

The principal component method associates a variable with each of the first few components and retains these few components. Another approach is to associate a variable with each of the last remaining principal components and delete them. To associate a variable with a principal component, choose the variable, which has not been selected previously, corresponding to the largest coefficient in the component. This research utilizes Jolliffe's PCA variable selection of principal component for the AHU fault group variable selection.

Logistic regression

Logistic regression is frequently used as a statistical classification method. Logistic regression is used when predictors, or independent variables, are qualitative or quantitative and the criterion variable, or dependent variable, is dichotomous, that is the variable takes on two values (fault = 1, no fault = 0) only. The method identifies the relative importance of variables and can be utilized to develop a classification model for prediction solely on the basis of the independent variables. In logistic regression, the relationship between the predictor and the predicted values is assumed to be nonlinear and when plotted, the curve takes the "S" shape, or sigmoidal function. Moreover, the curve never falls below 0 or reaches above 1; even for extreme values of the predictor. Thus, the predicted values obtained using the logistic model can be interpreted as probabilities. For example, if the dependent variable is coded as 0 and 1, the logistic regression analysis predicts a probability value that an observation belongs to a group.

As in linear regression analysis, several conditions must be met for the method to be valid. The following conditions are needed for logistic regression to be valid:

- It is assumed that the random variable of interest, i.e. occurrence of fault, is a dichotomous variable taking the value 1 with probability P_1 and the value 0 with probability $P_0 = 1 - P_1$.
- The outcomes must be statistically independent. In other words, a single case can be represented in the data set only once. If they are not independent then the result of the test may be inaccurate.

- The categories must be mutually exclusive. Therefore, a case cannot be in more than one outcome category at a time. This is analogous to either fault has occurred or not occurred.

To test a hypothesis for the logistic regression coefficients, larger samples are required than for linear regression analysis. This is because standard errors for maximum likelihood coefficients are large sample estimates. For small samples, the test result may be inaccurate. For most applications, a minimum of 50 cases per predictor variable is sufficient (Aldrich & Nelson, 1984).

While the coefficients of the linear regression parameters are selected according to the least squares criterion, the coefficients of the logistic parameters estimates are chosen according to the maximum likelihood criterion. Let $x_{i1}, x_{i2}, \ldots, x_{ip}$ be the values of the p number of regressor variables for the i^{th} observation, and let y_i be the i^{th} response. The logistic regression model assumes that y_i are independent and $y_i \sim \text{binomial}(M_i, \pi_i)$, $i = 1, \ldots, n$ and

$$\begin{aligned} logit(\pi_i) &= \log \frac{\pi_i}{1-\pi_i} \\ &= \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}, \quad i = 1, \ldots, n \end{aligned} \tag{3.15}$$

where $M_i \geq 1$ are fixed integers for trials conducted and

$$logit(\pi_i) = \log \frac{\pi_i}{1-\pi_i} = \beta_0 + \beta_1 x_i \quad \text{for } i = 1, \ldots, n \tag{3.16}$$

Here, $\beta_i \in \Re$ are parameters, and x_i's are known values of a regressor variable. The function $\log \frac{\pi}{1-\pi}$ is called the logit of π and maps the unit interval on the real line. Another notion for the same function is *log-odds* for the ratio $\pi/(1-\pi)$ when π is the probability of an event. The logit of the success probability π is a linear function of the regressor variable. If the logit of π is zero, then $\pi = 1/2$. Probabilities larger and smaller than $1/2$ correspond to positive and negative values on the logit scale. The regression problem is to estimate the conditional mean of y given x, (Hosmer 1989).

$$E[Y_i|x_i] = M_i \pi_i = M_i \cdot \frac{\exp(\alpha + \beta x_i)}{1 + \exp(\alpha + \beta x_i)} \tag{3.17}$$

The maximum likelihood estimators of the parameters α and β are given next.

The likelihood function for observed counts $y_1, \ldots, y_n$ is given by equation 3.18.

$$\mathbf{L}(\alpha, \beta; y_1, \ldots, y_n) = \prod \binom{M_i}{y_i} \pi_i^{y_i} (1 - \pi_i)^{M_i - y_i} \tag{3.18}$$

The success probabilities π_i are functions of α and β, according to the equation 3.19.

$$\pi_i = \frac{exp(\alpha + \beta x_i)}{1 + exp(\alpha + \beta x_i)}, \quad i = 1, \ldots, n \tag{3.19}$$

The log-likelihood function is given by equation 3.20.

$$\ell(\alpha, \beta) = C + \sum_{i=1}^{n} [y_i \log \pi_i + (M_i - y_i) \log(1 - \pi_i)] \tag{3.20}$$

where C is a constant that does not depend on the unknown parameters. Substitution of equation 3.16 and rearrangement results in equation 3.21.

$$\ell(\alpha, \beta) = C + \sum_{i=1}^{n} [y_i(\alpha + \beta x_i) + M_i \log(1 - \pi_i)] \tag{3.21}$$

Derivatives of equation 3.19 with respect to α and β are shown in equations 3.22 and 3.23, respectively.

$$\frac{\partial \pi_i}{\partial \alpha} = \pi_i(1 - \pi_i) \tag{3.22}$$

$$\frac{\partial \pi_i}{\partial \beta} = x_i \pi_i (1 - \pi_i) \tag{3.23}$$

Two equations, 3.24 and 3.25, result and need to be solved simultaneously to find the likelihood estimates.

$$\begin{aligned} S_1 = \frac{\partial \ell(\alpha, \beta)}{\partial \alpha} &= \sum_{i=1}^{n} \left[y_i + M_i \frac{\partial \log(1 - \pi_i)}{\partial \pi_i} \cdot \frac{\partial \pi_i}{\partial \alpha} \right] \\ &= \sum_{i=1}^{n} (y_i - M_i \pi_i) \end{aligned} \tag{3.24}$$

$$\begin{aligned} S_2 = \frac{\partial \ell(\alpha, \beta)}{\partial \beta} &= \sum_{i=1}^{n} \left[x_i y_i + M_i \frac{\partial \log(1 - \pi_i)}{\partial \pi_i} \cdot \frac{\partial \pi_i}{\partial \beta} \right] \\ &= \sum_{i=1}^{n} x_i (y_i - M_i \pi_i) \end{aligned} \tag{3.25}$$

For numerical maximization and to find the asymptotic distribution of the maximum likelihood estimators, information function, equation 3.26 is needed.

$$\begin{aligned} \mathbf{I}(\alpha,\beta) &= -\begin{pmatrix} \frac{\partial S1}{\partial \alpha} & \frac{\partial S1}{\partial \beta} \\ \frac{\partial S1}{\partial \alpha} & \frac{\partial S1}{\partial \beta} \end{pmatrix} \\ &= \begin{pmatrix} \sum_{i=1}^{n} M_i \pi_i (1-\pi_i) & \sum_{i=1}^{n} x_i M_i \pi_i (1-\pi_i) \\ \sum_{i=1}^{n} x_i M_i \pi_i (1-\pi_i) & \sum_{i=1}^{n} x_i^2 M_i \pi_i (1-\pi_i) \end{pmatrix} \end{aligned} \tag{3.26}$$

For numerical procedure, Newton-Raphson algorithm is used. The algorithm requires first and second derivatives of the function to be maximized, namely equations 3.24, 3.25, and 3.26. The algorithm steps are as follows. At the current value (α, β) of the arguments of the log-likelihood function $\ell(\alpha, \beta)$, it computes a second order Taylor series approximation. This leads to new values of α and β, which in turn a new Taylor series approximation is calculated, and step loops over until some specified tolerance is met.

$$\begin{pmatrix} \alpha^{j+1} \\ \beta^{j+1} \end{pmatrix} = \begin{pmatrix} \alpha^{j} \\ \beta^{j} \end{pmatrix} + [\mathbf{I}(\alpha^j, \beta^j)]^{-1} \begin{pmatrix} S_1(\alpha^j, \beta^j) \\ S_2(\alpha^j, \beta^j) \end{pmatrix} \tag{3.27}$$

Starting with arbitrary values (α^0, β^0), the repeated evaluations of equation 3.27 produce parameter estimates that converges to the maximum likelihood estimate $(\hat{\alpha}, \hat{\beta})$ if the maximum exists.

As always in estimation problems, the distribution of the estimated parameters are of importance. For large sample, the joint distribution of $\hat{\alpha}$ and $\hat{\beta}$ is approximately bivariate normal with means α and β, and with covariance matrix Σ (Flury 1997). Equation 3.28 defines the estimated covariance matrix of $\hat{\alpha}$ and $\hat{\beta}$.

$$\begin{aligned} \hat{\Sigma} &= [\mathbf{I}(\hat{\alpha}, \hat{\beta})]^{-1} = \begin{pmatrix} \hat{\sigma}_{11} & \hat{\sigma}_{12} \\ \hat{\sigma}_{21} & \hat{\sigma}_{22} \end{pmatrix} \\ &= \begin{pmatrix} \sum_{i=1}^{n} M_i \hat{\pi}_i (1-\hat{\pi}_i) & \sum_{i=1}^{n} x_i M_i \hat{\pi}_i (1-\hat{\pi}_i) \\ \sum_{i=1}^{n} x_i M_i \hat{\pi}_i (1-\hat{\pi}_i) & \sum_{i=1}^{n} x_i^2 M_i \hat{\pi}_i (1-\hat{\pi}_i) \end{pmatrix}^{-1} \end{aligned} \tag{3.28}$$

The square roots of the diagonal elements of $\hat{\Sigma}$ is referred to as standard errors of the parameter estimates.

$$\text{se}(\hat{\alpha}) = \sqrt{\hat{\sigma}_{11}} \quad \text{and} \quad \text{se}(\hat{\beta}) = \sqrt{\hat{\sigma}_{22}} \tag{3.29}$$

The estimated success probabilities and the estimated expected frequencies are defined by equations 3.30 and 3.31 respectively.

$$\hat{\pi}_i = g(x_i; \hat{\alpha}, \hat{\beta}) = \frac{\exp(\hat{\alpha} + \hat{\beta}x_i)}{1 + \exp(\hat{\alpha} + \hat{\beta}x_i)} \quad i = 1, \ldots, n \tag{3.30}$$

$$\hat{y}_i = M_i \hat{\pi}_i \tag{3.31}$$

A formal assessment of the fit of a logistic curve can be made using the notion of *deviance*. For given maximum likelihood estimates, $\hat{\beta}$, log-likelihood function is evaluated at its maximum and writing the maximum as a function of the estimated probabilities $\hat{\pi}$ becomes equation 3.32.

$$\ell(\hat{\pi}) = C + \sum_{i=1}^{n} [y_i \log \hat{\pi}_i + (M_i - y_i) \log(1 - \hat{\pi}_i)] \tag{3.32}$$

It is then compared to the value of the log-likelihood function for a model with success probabilities of $\tilde{\pi}_i = y_i/M_i$. $\tilde{\pi}_i$ is known as the saturated model or maximal model. The maximal model is given by equation 3.33.

$$\ell(\tilde{\pi}) = C + \sum_{i=1}^{n} \left[y_i \log \frac{y_i}{M_i} + (M_i - y_i) \log \frac{M_i - y_i}{M_i} \right] \tag{3.33}$$

The *deviance* is then defined as twice the difference of equation 3.33 and equation 3.32. Since $\ell(\tilde{\pi}) \geq \ell(\hat{\pi})$, the deviance is always nonnegative and can be viewed as a measure of fit of model. Asymptotic theory shows that if all M_i are large and the specified logistic regression model is correct, the distribution of deviance is approximately χ^2 with $(n - p - 1)$ degrees of freedom (Flury 1997). In short, if deviance is large, there is indication of the poor fit of the model. The deviance can also be used to compare different models. There is always advantage for choosing parsimonious models since there are fewer parameters to be specified and generally do well on future predictions.

Discrimination and classification

This analysis uses the term *group* to represent either a population or a sample from the population. For consistency with the multivariate statistics, (Johnson 1992), the term *discriminant analysis* is referred only in connection with the group separation and the term *classification*

analysis is associated with the prediction or allocation of the observations to one of the group. The distinction between the two terms is described as follows:

- Description of group separation, (or the descriptive aspect of *discriminant analysis* to separate groups), refers to linear functions or *discriminant functions* of the variables used to distinguish the differences between two or more groups. The goal of discriminant analysis includes identifying the relative contribution of the p variables for separation of the groups and finding the optimal plane on which the points can be projected to best illustrate the configuration on the groups.
- Prediction or allocation of the observed group, or *classification analysis*, refers to linear or quadratic *classification functions* of the variables used to assign an individual sampling unit to one of the groups. The measured values for an individual or object are evaluated by the classification functions to see to which group the individual most likely belongs.

In engineering and computer science, classification is usually referred to as *pattern recognition*. Some have used the term *cluster analysis* to refer to classification analysis. In classification, a sampling unit is assigned to a group on the basis of the vector of p measured values, $\mathbf{y}$. To classify the unit, a previously available sample of observation vectors is needed. A simple approach is to compare $\mathbf{y}$ with the mean vectors $\bar{\mathbf{y}}_1, \bar{\mathbf{y}}_2, \ldots, \bar{\mathbf{y}}_k$ of the k group samples and assign the unit to the group with the closest mean.

Standard distance and the linear discriminant function

This section describes the measure of distance between two multivariate distributions, called the multivariate standard distance. Standard distance is closely related to the linear discriminant function, which can be thought of as a linear combination of the variables that best separates the two distributions as much as possible.

Since the first discriminant function maximally separates the groups, the examination of the discriminant function coefficients reveals the contribution of each variable in separating the groups. Hence, if the discriminant function $z = a_1y_1 + a_2y_2 + \ldots + a_ky_k$ and a_2 is larger than the other a_i's, then the contribution of the y_2 weighs heavily towards the separation of the groups. Also because of the explicit unit differences in the variables that are measured within each group, a method of standardization is employed to adjust for differences in the scale among

the variables. The relative size of a_i shows the contribution of y_i in the presence of the other variables, in a manner analogous to a standardized regression coefficient or "beta weight." The individual F tests on $y_1, y_2, \ldots, y_p$ ignore the presence of other variables and thus do not take into account the correlation of each variable with the others. Because the primary interest is in the collective behavior of the variables, it would seem that the discriminant function coefficients provide more relevant information than the tests on individual variables.

Hubert (1975) compared the standardized coefficients to some correlations that can be shown to be related to individual variable tests. In a limited simulation, the discriminant coefficients were found to be more valid than the univariate tests in identifying those variables that contribute least to separation of groups.

In regression analysis, linear regression functions are defined in linear combinations of p regressors $X_1, \ldots, X_p$ such that the dependent variable Y is optimally approximated using the method of least squares. While regression utilizes the minimization of residual variance, linear discriminant function uses maximization of the distance between two groups. The following equation defines the standard distance between two numbers.

$$\Delta_Y(y_1, y_2) = \frac{|y_1 - y_2|}{\sigma} \tag{3.34}$$

The standard distance becomes a unit of Euclidean distance if σ is one. For the multivariate case, let $y_{11}, y_{12}, \ldots, y_{1N_1}$ denote the observed data vectors from group 1, and $y_{21}, y_{22}, \ldots, y_{2N_2}$ the data vectors from group 2. These $N_1 + N_2$ observations constitute the training samples. Equation 3.35 is the sample mean vectors, and equation 3.36 the usual sample covariance matrices.

$$\bar{\mathbf{y}}_j = \frac{1}{N_j} \sum_{i=1}^{N_j} \mathbf{y}_{ji}, \quad j = 1, 2 \tag{3.35}$$

$$\mathbf{S}_j = \frac{1}{N_j - 1} \sum_{i=1}^{N_j} (\mathbf{y}_{ji} - \bar{\mathbf{y}}_j)(\mathbf{y}_{ji} - \bar{\mathbf{y}}_j)', \quad j = 1, 2 \tag{3.36}$$

where the prime denotes the operation of *transposing* a column to a row. The only difficulty is how to accommodate the sample covariance matrices, $\mathbf{S}_1$ and $\mathbf{S}_2$, rather than a single common

covariance matrix. This difficulty is overcome by using a weighted average of $\mathbf{S}_1$ and $\mathbf{S}_2$, the *pooled sample covariance matrix.* The pooled sample covariance matrix is defined in equation 3.37.

$$\mathbf{S}_{pl} = \frac{1}{N_1 + N_2 - 2}[(N_1 - 1)\mathbf{S}_1 + (N_2 - 1)\mathbf{S}_2] \tag{3.37}$$

The use of this pooled estimate is justified through the unbiasedness estimator of the common covariance matrix of two populations, irrespective of the exact distribution. The sample multivariate standard distance between $\bar{\mathbf{y}}_1$ and $\bar{\mathbf{y}}_2$ is given by equation 3.38.

$$D(\bar{\mathbf{y}}_1, \bar{\mathbf{y}}_2) = \max_{a \in \Re^p a \neq 0} \frac{|\mathbf{a}'(\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_2)|}{(\mathbf{a}'\mathbf{S}_{\mathbf{pl}}\mathbf{a})^{1/2}} \tag{3.38}$$

The term $\mathbf{a}'\mathbf{S}_{\mathbf{pl}}\mathbf{a}$ is the pooled-sample variance of the linear combination $z = \mathbf{a}'\mathbf{y}$. Here, $\mathbf{a} = (a_1, \ldots, a_p)' \in \Re^P$ denote the vector of coefficients of a linear combination. The equation defines maximum univariate standard distance over all linear combinations, provided the maximum exists. Any linear combination for which the maximum is attained is called a *linear discriminant function* for the given samples. With the assumption that the pooled sample covariance matrix, $\mathbf{S}_{pl}$, is nonsingular, the multivariate standard distance between $\bar{\mathbf{y}}_1$ and $\bar{\mathbf{y}}_2$ is given by equation 3.39 and the vector of coefficients of the linear discriminant function is given by equation 3.40.

$$D(\bar{\mathbf{y}}_1, \bar{\mathbf{y}}_2) = \sqrt{(\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_2)'\mathbf{S}_{pl}^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)} \tag{3.39}$$

$$\mathbf{b} = \mathbf{S}_{pl}^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2) \tag{3.40}$$

Note, that if $N_1 + N_2 < p + 2$, then $\mathbf{S}_{pl}$ is always singular. This restricts the use of linear discriminant analysis to samples large enough for the pooled covariance matrix to be positive definite. This is the only assumption on the current descriptive approach and does not assume that the data constitute random samples from a particular family of distributions or that the covariance matrices in the populations are identical.

Discriminant analysis for several groups

In this section, group membership of 2 and more, $k \geq 2$, is discussed. Here k is the number of groups. Analogous to the discriminant analysis for two groups, discriminant analysis for several groups can serve any one of the following goals.

- Examine group separation in a two-dimensional plot. More than two groups requires more than one discriminant function to describe group separation. If the points in the p-dimensional space are projected into a 2-dimensional space represented by the first two discriminant functions, then the groups are best separated.
- Find a subset of the original variables that separates the groups almost as well as the original set.
- Rank the variables in terms of their relative contributions to group separation. The standardized discriminant function coefficients provide valid comparison of the variables.
- Interpret the new dimensions represented by the discriminant functions.
- Follow up to fixed effects MANOVA.

The linear functions contributing to the description of group separation are often referred to as canonical variates or discriminant coordinates. As with discriminant analysis for two groups, the objective is finding linear combinations of variables that best separate groups of multivariate observations. For two groups and more, description of group separation requires more than one discriminant function. As a second objective it is of our interest to find the subset of the original variables that separates the groups almost as well as the original set.

For k groups with n_i observations in the i^{th} group, each observation vector y_{ij} is transformed to obtain $z_{ij} = \mathbf{a}'\mathbf{y}_{ij}, \quad i = 1, 2, \ldots, k; \; j = 1, 2, \ldots, n_i$, and find the means $\bar{z}_i = \mathbf{a}'\bar{\mathbf{y}}_i$. The problem is to find the vector $\mathbf{a}$ that maximally separates $\bar{z}_1, \bar{z}_2, \ldots, \bar{z}_k$.

Selection of minimal variables

One objective of this research is to determine how many and what variables make significant contribution to detection of faults. There are several methods by which this is done. Tests of significance can help determine the needed information.

Tests for additional information during group separation can show presence of the redundancy of the variables in terms of separating the groups. The hypothesis test on the subset of variables may show the contributional significance of the information already available in preset vectors for separating the groups. This is in accordance with full and reduced model tests in regression analysis, (Rencher 1995).

Let $\mathbf{y}$ be a $p \times 1$ vector of measurements and $\mathbf{x}$ be a $q \times 1$ vector measuring other variables in addition to $\mathbf{y}$. The test of $H_o : \mu_1 = \mu_2 = \ldots = \mu_k$ yields whether $\mathbf{x}$ makes a significant contribution. In other words, the question is whether the separation of groups achieved by $\mathbf{x}$ be predicted from the separation yielded by $\mathbf{y}$ alone? In a sense we wish to know if the variables in $\mathbf{x}$ can be deleted because they do not contribute to rejecting the null hypothesis, H_o. The one-way MANOVA procedure is conducted for this test. Assume that there are k group samples of n observations,

$$\begin{pmatrix} \mathbf{y}_{ij} \\ \mathbf{x}_{ij} \end{pmatrix} \qquad \begin{matrix} i = 1, 2, \ldots, k \\ j = 1, 2, \ldots, n \end{matrix} \tag{3.41}$$

calculate the "within", $\mathbf{E}$, and the "between", $\mathbf{H}$, sums of squared matrices. They are defined in equation 3.42 and 3.43.

$$\mathbf{H} = \sum_{i=1}^{k} n_i (\bar{\mathbf{y}}_{i.} - \bar{\mathbf{y}}_{..})(\bar{\mathbf{y}}_{i.} - \bar{\mathbf{y}}_{..})' \tag{3.42}$$

$$\mathbf{E} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} n_i (\mathbf{y}_{ij} - \bar{\mathbf{y}}_{i.})(\mathbf{y}_{ij} - \bar{\mathbf{y}}_{i.})' \tag{3.43}$$

where $\bar{\mathbf{y}}_{i.}$ is the mean of the i^{th} group, and $\bar{\mathbf{y}}_{..}$ is the total sample mean.

$$E = \begin{pmatrix} \mathbf{E}_{yy} & \mathbf{E}_{yx} \\ \mathbf{E}_{xy} & \mathbf{E}_{xx} \end{pmatrix} \tag{3.44}$$

$$H = \begin{pmatrix} \mathbf{H}_{yy} & \mathbf{H}_{yx} \\ \mathbf{H}_{xy} & \mathbf{H}_{xx} \end{pmatrix} \tag{3.45}$$

where $\mathbf{E}$ and $\mathbf{H}$ are $(p+q) \times (p+q)$ and $\mathbf{E}_{yy}$ and $\mathbf{H}_{yy}$ are $(p \times p)$ etc. Then

$$\Lambda(y, x) = \frac{|\mathbf{E}|}{|\mathbf{E} + \mathbf{H}|} \tag{3.46}$$

is distributed as $\Lambda_{(p+q),(\nu H),(\nu E)}$ and tests the significance of group separation using the full set. In the balanced one-way model, the degrees of freedom are $\nu_H = k-1$, and $\nu_E = k(n-1)$. To test group separation using the reduced set, then use equation 3.47

$$\Lambda(y) = \frac{|\mathbf{E_{yy}}|}{|\mathbf{E_{yy}} + \mathbf{H_{yy}}|} \tag{3.47}$$

Reduced set is distributed as $\Lambda_{(p),(\nu H),(\nu E)}$. Equation 3.48 is calculated to test the hypothesis that the extra variables in $\mathbf{x}$ do not contribute anything significant to separating the groups beyond the existing information already available in $\mathbf{y}$.

$$\Lambda(\mathbf{x}|\mathbf{y}) = \frac{\Lambda(\mathbf{y},\mathbf{x})}{\Lambda(\mathbf{y})} \tag{3.48}$$

This has distribution of $\Lambda_{(p),(\nu H),(\nu E-p)}$. The error degree of freedom of $\Lambda(\mathbf{x}|\mathbf{y})$ is $(\nu E - p)$ because it has been adjusted for the p y's. Thus to test for the contribution of additional variables to separation of groups, take the ratio of Wilks' Λ for the full set of variables to the reduced set of Wilks' Λ. If the addition of $\mathbf{x}$ reduces the ratio sufficiently, then $\Lambda(\mathbf{x}|\mathbf{y})$ will be small enough to reject the hypothesis testing and conclude addition of the variables does not necessarily improve the group separation.

For checking for the effect of a single variable, then $q = 1$ and equation 3.48 becomes

$$\Lambda(x|y_1,\ldots,y_p) = \frac{\Lambda(y_1,\ldots,y_p,x)}{\Lambda(y_1,\ldots,y_p)} \tag{3.49}$$

which is distributed as $\Lambda_{1,(\nu H),(\nu E-p)}$.

Classification analysis

In classification, an unknown group membership sampling unit is assigned to a group on the basis of the vector of p measured values, $\mathbf{y}$, associated with the unit. To classify the unit, the analysis needs previously obtained samples of observation vectors from each group. One approach to classification is to compare the measured values $\mathbf{y}$ with the mean vectors $\bar{\mathbf{y}}_1$, $\bar{\mathbf{y}}_2,\ldots,\bar{\mathbf{y}}_k$ of the k group samples and assign the unit to the closest to one of the group.

In the case of two populations, a classification procedure by Fisher (1938) can be utilized. Fisher's idea was to transform the multivariate observation into univariate observations such

that the transformed observations have clear distinction between the groups. For Fisher's procedure, the principal assumption is that the two populations have the same covariance matrix because the method applies pooled estimate of the common covariance matrix. The assumption of normality is not required. The first step is to obtain the sample means, $\bar{\mathbf{y}}_1 = \sum_{j=1}^{n_1} \mathbf{y_{1j}}/\mathbf{n_1}$ and $\bar{\mathbf{y}}_2 = \sum_{j=1}^{n_2} \mathbf{y_{2j}}/\mathbf{n_2}$, and the pooled variance, $\mathbf{S}_{pl}$ from both sample groups.

$$\mathbf{S}_{pl} = \frac{(n_1 - 1)\mathbf{S_1} + (n_2 - 1)\mathbf{S_2}}{n_1 + n_2 - 2} \tag{3.50}$$

where n_1, n_2, S_1, and S_2 are number of sample observations and sample covariance matrices from group 1 and group 2 respectively. For reasons of unbiasedness, the pooled within group covariance matrix for multiple groups is given by equation 3.51

$$\mathbf{S}_{pl} = \frac{1}{n_1 + n_2 + \ldots + n_k - k} \sum_{j=1}^{k} (n_j - 1)\mathbf{S}_j \tag{3.51}$$

A simple classification can be based on the discriminant function z, equation 3.53.

$$z_o = \mathbf{a}'\mathbf{y} = (\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_2)'\mathbf{S}_{pl}^{-1}\mathbf{y} \tag{3.52}$$

$$z = \mathbf{a}'\mathbf{y} = \frac{1}{2}(\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_2)'\mathbf{S}_{pl}^{-1}(\bar{\mathbf{y}}_1 + \bar{\mathbf{y}}_2) \tag{3.53}$$

where $\mathbf{y}$ is the vector of the new measurements and z_o is the associated discriminant function. Here, $\mathbf{a} = (a_1, \ldots, a_p)' \in \Re^P$ denote the vector of coefficients of a linear combination. z is the prior linear discriminant function estimated from the two population samples. Allocate new observation into group 1 if the $z_o - z \geq 0$ and into group 2 if $z_o - z < 0$ The linear discriminant functions was developed under the assumption that the two populations, whatever their form, have a common covariance matrix. Another classification function, w, expressed in equation 3.54 is the result of simplified form of $z_o - z$.

$$w = \mathbf{a}'\mathbf{y} = \frac{1}{2}(\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_2)'\mathbf{S}_{pl}^{-1}(\mathbf{x} - \frac{1}{2}\bar{\mathbf{y}}_1 + \bar{\mathbf{y}}_2) \tag{3.54}$$

The allocation of the group membership is if $w \geq 0$ $\mathbf{y}$ is assigned to group 1 and if $w < 0$ then $\mathbf{y}$ is assigned to group 2. It is noted that, provided the two normal populations have the same covariance matrix, Fisher's classification rule is equivalent to the minimum expected cost of

misclassification (ECM) rule with equal prior probabilities and equal costs of misclassification, Johnson (1992).

For two populations, the maximum relative separation that can be achieved through linear combination of the multivariate observations is the sample squared distance between the two means, D^2, given in equation 3.38. This is convenient because D^2 can be used to test whether the population means are significantly different. Consequently, a test for the difference of the mean vectors can be interpreted as a test for the separation that can be achieved.

Classification with several groups

The theory of optimal classification allows multiple group classification. The minimum expected cost of misclassification, ECM, is used in the study. Let $f_i(\mathbf{x})$ be the density (assume multivariate normal density for most part) associated with population g_i, and π_i denote the prior probability of population g_i, $i = 1, \ldots, k$. Let $c(j|i)$ be the cost of allocating an item to g_j when it belongs to g_i. For $j = i$, $c(i|i) = 0$. Let R_j be the set of $\mathbf{x}$'s classified as g_j and Let

$$\Pr(j|i) = \Pr(\text{classify item as } g_j|g_i) = \int_{R_k} f_i(\mathbf{x})d\mathbf{x} \tag{3.55}$$

for $j, i = 1, \ldots, k$, with $\Pr(i|i) = 1 - \sum_{k=1_{k\neq i}}^{k} \Pr(k|i)$. The conditional expected cost of misclassifying an $\mathbf{x}$ from g_1 into g_2, or $g_3, \ldots, g_k$ is defined in equation 3.56.

$$\begin{aligned} ECM(1) &= \Pr(2|1)c(2|1) + \Pr(3|1)c(3|1) + \ldots + \Pr(g|1)c(g|1) \\ &= \sum_{j=2}^{k} \Pr(j|1)c(j|1) \end{aligned} \tag{3.56}$$

This conditional expected cost occurs with prior probability π_1, the probability of g_1. In a similar manner, the conditional expected costs of misclassification, $ECM(2), \ldots, ECM(k)$ can be obtained. Multiplying each conditional ECM by its prior probability and summing gives overall ECM, equation 3.57.

$$\begin{aligned} ECM &= \pi_1 ECM(1) + \pi_2 ECM(2) + \ldots + \pi_k ECM(k) \\ &= \sum_{i=1}^{k} \pi_i \left(\sum_{j=1\ j\neq i}^{k} \Pr(j|i)c(j|i) \right) \end{aligned} \tag{3.57}$$

Table 3.1 Minimum classification rule with equal misclassification costs

Allocate $\mathbf{x}$ to g_j	If $\pi_j f_j(\mathbf{x}) > \pi_i f_i(\mathbf{x})$ for all $i \neq k$	
Allocate $\mathbf{x}$ to g_j	If $\ln \pi_j f_j(\mathbf{x}) > \ln \pi_i f_i(\mathbf{x})$ for all $i \neq k$	
Allocate $\mathbf{x}$ to g_j	If d_j^Q, equation 3.61 $= \max_i(d_i^Q)$ $i = 1, \ldots, k$	Unequal Σ_i
Allocate $\mathbf{x}$ to g_j	If d_j^L, equation 3.62 $= \max_i(d_i^L)$ $i = 1, \ldots, k$	Equal Σ_i
Allocate $\mathbf{x}$ to g_j	largest of $-1/2D_i^2(\mathbf{x}) + \ln \pi_i$	Equal Σ_i

Determining an optimal classification procedure amounts to choosing the mutually exclusive and exhaustive classification regions R_1, $R_2, \ldots,$ R_k such that ECM is minimum. The classification regions that minimize the ECM are defined by allocating $\mathbf{x}$ to that population g_j, $j = 1, \ldots, k$ for smallest of

$$\sum_{i=1 \ i \neq j}^{k} \pi_i f_i \mathbf{x} c(j|i) \tag{3.58}$$

.

Suppose all the misclassification costs are equal, then the ECM rule is the minimum total probability of misclassification rule. So, allocate $\mathbf{x}$ to the population g_j, $j = 1, \ldots, k$ when the $\sum_{i=1 \ i \neq j}^{k} \pi_i f_i \mathbf{x}$ is smallest. When the misclassification costs are the same, the minimum expected cost of misclassification has the following rule which is identical to the one that maximizes the posterior probability, $\Pr(g_i|\mathbf{x})$, equation 3.59.

$$P(g_j|\mathbf{x}) = \frac{\pi_j f_j(\mathbf{x})}{\sum_{i=1}^{k} \pi_i f_i(\mathbf{x})} \quad \text{for } j = 1, 2, \ldots, k \tag{3.59}$$

For special case of multivariate normal densities, equation 3.60, with mean vectors μ_i and covariance matrices Σ_i along with equal misclassification costs, the classification rule for allocating $\mathbf{x}$ to the group g_j is that if $\ln \pi_j f_j(\mathbf{x}) = \max_i(\ln \pi_i f_i)$ then allocate $\mathbf{x}$ to π_j. This is a form of quadratic equation. Equation 3.61 defines the quadratic discrimination score for the i^{th} population.

$$f_i(\mathbf{x}) = \frac{1}{(2\pi)^{p/2}|\Sigma_i|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x} - \mu_i)'\Sigma^{-1}(\mathbf{x} - \mu_i)\right] \quad i = 1, 2, \ldots, k \tag{3.60}$$

$$d_i^Q(\mathbf{x}) = -\frac{1}{2}\ln|\Sigma_i| - \frac{1}{2}(\mathbf{x} - \mu_i)'\Sigma^{-1}(\mathbf{x} - \mu_i) + \ln \pi_i \tag{3.61}$$

The quadratic score is composed of contributions from the generalized variance $|\Sigma_i|$, the prior probability π_i, and the squared distance from $\mathbf{x}$ to the population mean μ_i. For equal population covariance matrices, Σ, linear discriminant score has the following form.

$$d_i^L(\mathbf{x}) = \mu_i \Sigma_i^{-1} \mathbf{x} - \frac{1}{2} \mu_i \Sigma_i^{-1} \mathbf{x} + \ln \pi_i \tag{3.62}$$

The estimate of the linear discriminant score is based on the pooled estimate of Σ, equation 3.51. Equation 3.62 is a convenient linear function of $\mathbf{x}$. A similar form of classifier, equation 3.63, for the equal covariance case can be seen from the previous discussion of discrimination analysis, equation 3.39.

$$D_i^2(\mathbf{x}) = (\mathbf{x} - \bar{\mathbf{x}}_i)' \mathbf{S}_{pl}^{-1} (\mathbf{x} - \bar{\mathbf{x}}_i) \tag{3.63}$$

This can be interpreted in terms of the squared distances from $\mathbf{x}$ to the sample mean vector $\bar{\mathbf{x}}_i$. The allocation rule is in Table 3.1. This rule assigns $\mathbf{x}$ to the closest population and the distance measure is penalized by $\ln \pi_i$. If the prior probabilities are unknown, the usual procedure is to use the observation frequency proportion.

Linear vs quadratic

Linear and quadratic classification rules of the general multivariate normal distributions have been discussed in this section. The optimal classification rule can be formulated in terms of the posterior probabilities or in terms of the classification functions. Many studies comparing expected actual error rates of linear and quadratic classification rules have reached the conclusion that linear discrimination performs better than quadratic discrimination as long as the differences in variability are only moderate, and the sample sizes are relatively small. This is due to the fact that the quadratic rule estimates more parameters, i.e. two covariance matrices instead of one, which gives a better fit to the observed data but less stable estimates. Unfortunately, the theoretical calculation of expected actual error rates is rather complicated and can be done only by simulation.

4 EXPERIMENTAL SET-UP AND OPERATION

An important quality characteristic of system faults is the deviation from "known" operating conditions. AHU operation will change as building conditions change. Some of these changes are not necessarily due to fault conditions. Therefore, one strategy of conducting the experimental runs is to simulate the normal operation and induce an abrupt fault. Because these abrupt changes bring added source of variability in the true nature of the faults, most of the fault conditions considered in the study are simulated faults.

However, the faults introduced in this study may exist in real operations. One of the faults is reduced water flow through the heat exchanger which may be due to clogging or sticking control valves. A second fault is building up of debris in the airside of the heat exchanger surfaces after long term exposure to circulating particles. A third fault is the degradation of the fan performance from prolonged operations. The flow loop used for this research is equipped to handle these type of fault conditions.

Air handling unit

This study was carried out with the HVAC air flow test loop, shown in Figure 4.1, in room 2103 of the H. M. Black Engineering building at Iowa State University. The loop has been in service for over 10 years for teaching and research purposes. The main components of the loop consist of a 2000 cubic feet per minute (cfm) fan unit (centrifugal fan comprised of conventional forward curved and operates at the low pressure class for the blower and coil section), a full set of heat exchanger coils (HX) (hot water HX, refrigerant HX, chilled water HX, and steam HX), pneumatic dampers, air-to-air heat exchanger, variable air volume (VAV) boxes, and a pneumatic control system.

The instrumentation of the loop allows measurements for flow rates, temperatures, pressures, and dew point temperatures of the fluid medium. Additional instrumentation for the fan rotational speed, power to the fan motor, and static pressure across the air handling unit were introduced for this study. All measurements were directed to the central processing computer for automatic data acquisition.

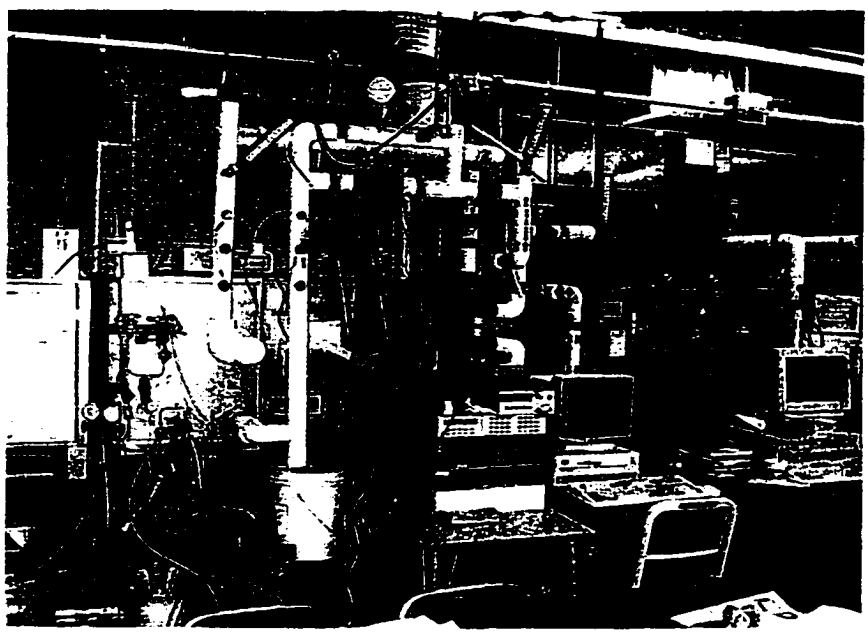


Figure 4.1 Charles L. Schwab HVAC test loop

The air handling unit (AHU) consists of a filter, refrigerant coil, hot water coil, humidifier, and variable speed fan as shown in Figure 4.2. The filter construction is of spun glass medium with spun nylon backing which is tolerant of 100 percent saturated air. For this study, the main components of interest are the hot water coil and the fan. The hot water for the hot water coil is produced from processed steam from the Iowa State University power plant and provides heating of the air. A pneumatically controlled valve allows the rate of the steam flow to control the inlet water temperature. The frequency controlled fan provides air flow rates up to 2000 cfm at a fan speed of 1750 revolution per minute (rpm). The various damper settings allow control of the recirculated and outdoor air mixing rates. The flow measuring stations

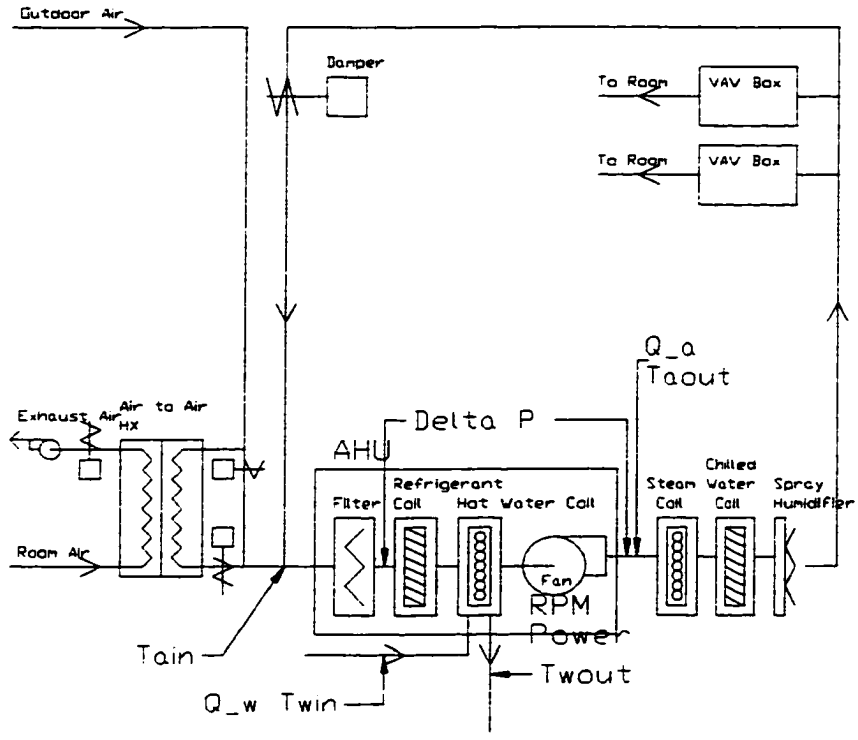


Figure 4.2 Test loop sensor location

provide measurement of the volume flow rate of the air. The specification of the hot water coil and the fan are summarized in the Table 4.1.

Heating mode operation was used for this study. The chilled water coil, in the load section of the loop, enabled the dehumidification and cooling of the conditioned supply air to simulate the loads that may occur in a building during winter time operation. Heating was provided by the hot water coil without steam humidification. A mixed air controller allowed variations in the return and outdoor air mixtures. The entire system was operated as a fixed fan speed or constant air volume (CAV) system.

Table 4.1 Specification of fan and hot water coil

Component	Specification	
Hot Water Coil	Face Area	0.362 m^2
	Face Velocity	2.6 m/s
	Total Btuh	11,941 W
	Water Flow Rate	25 l/min
	Temperatures	Entering Air 21.1°C
		Leaving Air 31.4°C
		Entering Water 71°C
		Leaving Water 65.6°C
Fan	Type	FC Frequency controlled
	Maximum Speed	1400 rpm
	Flow Rate	3398 m^3/h

Data acquisition

The various sensors consisting of thermocouples for the temperature measurements, pressure transducers for the static pressure measurement and volume flow rate of the air, rotary flow meter for the hot water flow rate had already been installed and did not require special attention. On the other hand, a magnetic pick up coil for measuring rotation of the fan, a 3 phase power transducer for the motor power, and static pressure probes for measuring pressure rise across the AHU were added. The sensor measurements and signals were processed through the National Instrument GPIB-34A data acquisition board, Hewlett Packard (HP) 5316B Universal Counter, HP3488A Switch Control Unit, HP3455A Digital Voltmeter, and HP3495A Scanner. The main processor utilized was an IBM PS/2 model 50Z consisting of Intel 286 central processing unit. Figure 4.3 shows the connection between the data acquisition system and the sensors. A QBasic program, MAIN.BAS, was written to initiate the A/D conversions and store the sensor measurements in ASCII files.

Experimental procedure

In general terms, the purpose of the study and the scope of the problem to be addressed are to detect faults in the performance of the fan, blockage of the coils, and sticking valves

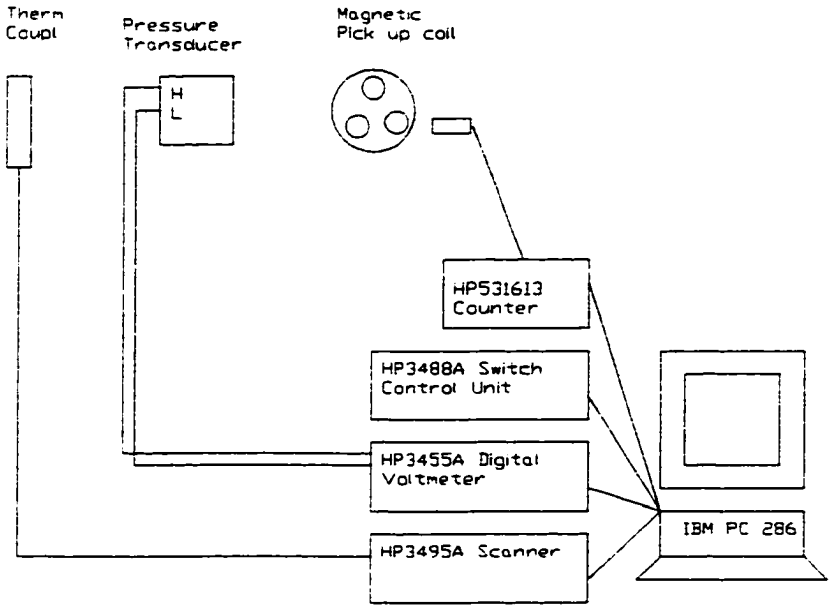


Figure 4.3 Sensor and data acquisition system

with the fewest measurable variables. The study encompasses nine effective variables used to identify the faults without misclassifying normal operation or keeping the misclassification at low rate. The nine effective variables used are:

- T_{ain} Inlet air temperature, (°C)
- T_{aout} Outlet air temperature, (°C)
- T_{win} Inlet water temperature, (°C)
- T_{wout} Outlet water temperature, (°C)
- ΔP Pressure rise across AHU, (in w.c.)
- $\dot{Q}_w$ Volume flow rate of hot water, (l/s)
- $\dot{Q}_{air}$ Volume flow rate of Air, (ft^3/min)
- RPM Rotational Fan speed, (rev/min)
- Pow Power measure to motor, (W)

These variables were selected and arranged by their measurability. This research looks at the effects of these variables on fault detection and diagnostics.

Because these variables take on values for heating operation in an experimental environment, different selection of the controlling devices are needed. The test flow loop consists of three main setting that change the operation of the experiments, namely the damper settings for mixing the outside and recirculated air, automatic/manual control setting of the hot water valve, and improvision of the debris to simulate the accumulation on the face of the hot water heat exchanger coil.

In order to use these different settings, three levels for each of the factor combinations are required. An appropriate experimental design for this situation is a 3^4 factorial design. This design uses every combination of the three levels for each of the three factors. The root number 3 indicates the number of levels used for each factor, and the exponent 4 indicates the total number of factors. Table 4.2 lists the factors and their levels.

Table 4.2 Factor level combination

Factor	Level 1	Level 2	Level 3
RPM	70	80	90
Damper	closed	1/2 open	open
Valve	closed	1/2 open	open
Coil	1/4 screen	1/2 screen	open

The experimental design for this study is listed in Table 4. In it the actual design was performed in random order to minimize the impact of any potential biases. With each setting of the process conditions, the process was allowed to reach steady state, and for each test run, fault values were assigned to each event. In Chapter 5 and 6, an appropriate analysis is performed to produce a model that allows discrimination and classification or prediction of the faults occurring during system operation given a specific combination of the process conditions.

In some cases, a specific model should explain the behavior of the data. The next section provides an experimental plan to estimate the proposed model as efficiently as possible with the resources available. The data collected from these experiments help to determine the

adequacy of the proposed model. If the results are satisfactory with the model, then the result can be used to predict the behavior of the response over the ranges of the factors studied in the experiment. Here, each treatment combination is applied to the operation of the air handling system. This study seeks to determine the basic relationships among the factors and the system faults.

The contribution of this study is to identify the minimal set of variables required to detect the faults in the AHU. It is hypothesized that with fewer variables, the prediction of a fault is more robust but looses its accuracy. Here, *robust* refers to the deterioration in error rates caused by using a classification procedure (prediction) with data that do not conform to the assumption on which the procedure was based. Also, cost is reduced by measuring fewer variables. Because it is desired to not have too many false alarms, there must be some minimal set of variables that balances the dimensional aspect with the accuracy of the prediction.

Physical preparation

In preparing the laboratory experiment, data acquisition and sensors were added to the existing setup. The fan power measurement was added, but due to the difficulties in obtaining the true calibration for the transducer, a representable index of power measure was used in this study. The difficulty arose due to the 3 phase power connection and the variable frequency drive for the fan. More temperature sensors were added to the already existing thermocouples. In addition, a magnetic pickup sprocket disk with metal nuts was installed on the shaft of the fan to measure the rotational speed of the fan. A signal pick up coumter was used to measure the fan speed in revolution per minute. Static pressure probes were placed on either end of the air handling unit.

Sensor calibration

When using data, one is quickly faced with the fact that variation is omnipresent. Some of that variation comes about because the objects studied are never exactly alike. Some has its origin in the fact that measurement processes also have their own inherent variability. The

variability or error that is inevitable in measuring systems can be thought of as having both internal and external components. Internal components of variability comes from the sensors and external components comes from the operator.

Initially, all the sensors were calibrated and tested for their precision and accuracy. A working definition of precision is when a measurement produces small variation in repeated measurement of the same object. Precision is the internal consistency of a measurement system. Precision maybe improved with consistency. Although precision is important, for many purpose it alone is not adequate. Several dry runs of the experiments were conducted and any sensor faults were corrected. The faults were mainly due to miscalculation of the intercept of the calibration curve. The Table 4.3 summarizes the calibration result of the sensors.

Table 4.3 Sensor precision and standard deviation

Sensor	Precision	Standard Error
Copper Constantine Type T Thermocouples	$0.05°C$	$0.21°C$
Pressure Transducer	1.548e-3$inW.C.$	1.626e-3$inW.C.$
Water Flow rate	$0.155kg/min$	$0.879kg/min$
Air Velocity	$0.02m/s$	$0.131m/s$

A measurement system is called accurate or sometimes referred as unbiased if on average it produces the true or correct value of a quantity being measured. Accuracy, by statistical standard, is the agreement of a measuring system with some external standard of measurement. Poorly calibrated measuring devices may be sufficient for comparing local conditions. This was the case with the fan power measure. The main problem was inadequate calibration equipment available for the 3 phase power transducer, but the consistency in the measurement was observed. But if one is to establish the values of quantities in any absolute (rather than relative) sense or to expect local values to have any meaning at other places and other times, it is important to calibrate measurement systems against a constant standard.

The possibility of bias or inaccuracy in measuring systems has at least two important implications for planning the studies. First, measurement system accuracy is time dependent. Hence periodic recalibration is needed. In this study, pressure transducer needed to be zeroed

before each test run. Instrument drift can ruin a well planned study. Second, if possible, a single system should be used to measure response. This was not a problem with this research since only one system was in operation.

Process dynamics in AHU

As a preliminary evaluation of the data, the fan pressure rise and capacity measurements were collected and compared to the manufacturer's fan performance graph. The result indicated near perfect match. The collection of the data began with the control sequence and the data were examined through different variable settings. Figures 4.4 through 4.7 show the response of the system with different control settings.

For the input setting of hot water, Figure 4.4, response of the inlet and outlet air temperature is provided in Figure 4.5. The settings for the damper control was set at 10 psi, 1/2 return and 1/2 outdoor air. The tests indicated the proper operation of the sequencing of the control settings and performance of the test operation in their range of experimental settings. Most of the transients lasted for about 20 minutes. Therefore, the data taking can begin 20 minutes after each chang in control settings.

Data collection

Data collection is considered an important activity and careful consideration was exercised. Hence, during the data collection each of the measured data were ensured so that the data on relevant variables was collected with known and adequate quality. Table 4 lists the order of the experimental run for the 3^4 factorial experiment.

The first column indicates the variable combinations identification. The second column indicates randomized experimental run. The randomization order was obtained from the SAS program. The randomization guarantees inferential validity in the face of unspecified disturbances. Fisher (1936) argued that with a randomized experiment it is possible to conduct a significance test without making any assumptions about the distribution. Another reason for randomization was to avoid the biases of the single experimental unit for which there may be

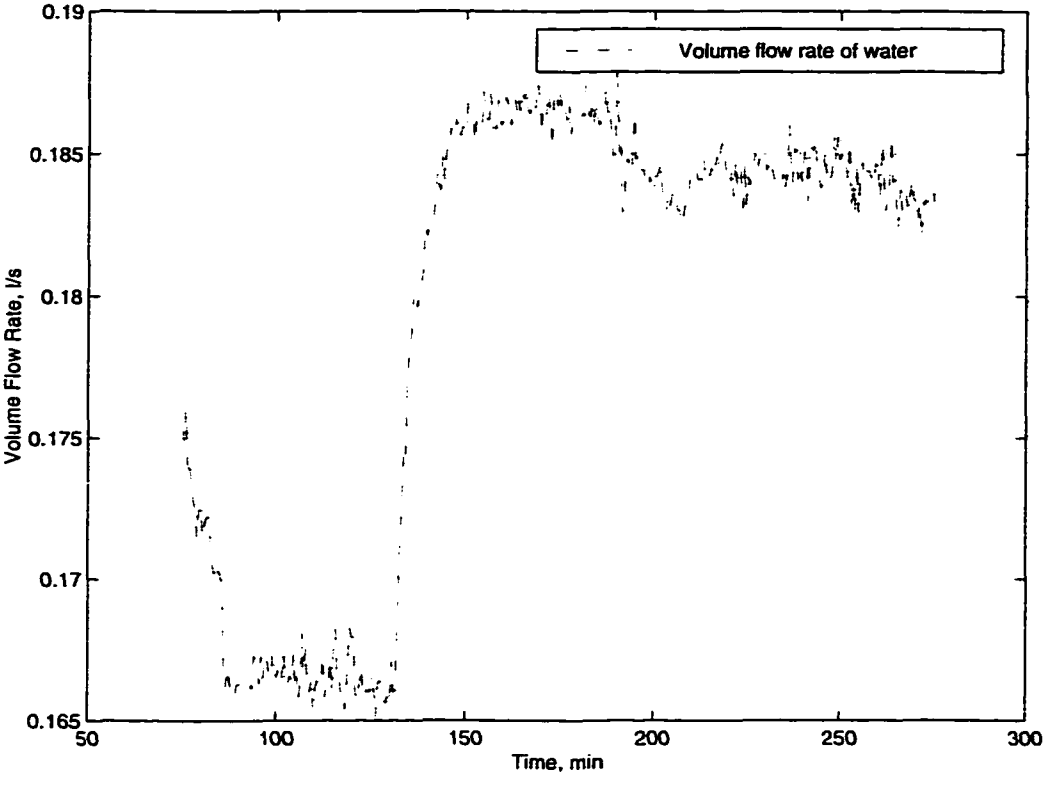


Figure 4.4 Step input of water flow rate

residual effect on the gradual changes of the system for continuous data taking.

It was later discovered that more data points of the normal operation were needed at different settings. This was necessary to regard the possible prior information required to perform classification of the fault groups. It was assumed for the research that the occurrence of the faults are less probable than the normal operation. Hence, during the classification process, more weight is given for normal operation than fault operation. Because of the limited control settings of the apparatus, the damper settings were refined to include 7 different manual pneumatic settings of 10, 7, 13, 8, 12, 9, and 11 *psi*. A similar approach was employed during the data taking months of April 1998 through September 1998. In each experimental run, the variable air volume box was opened to allow one time passage through the air handling unit without recirculation.

The data points were taken at an interval of 1 minute for a period of 20 minutes per run.

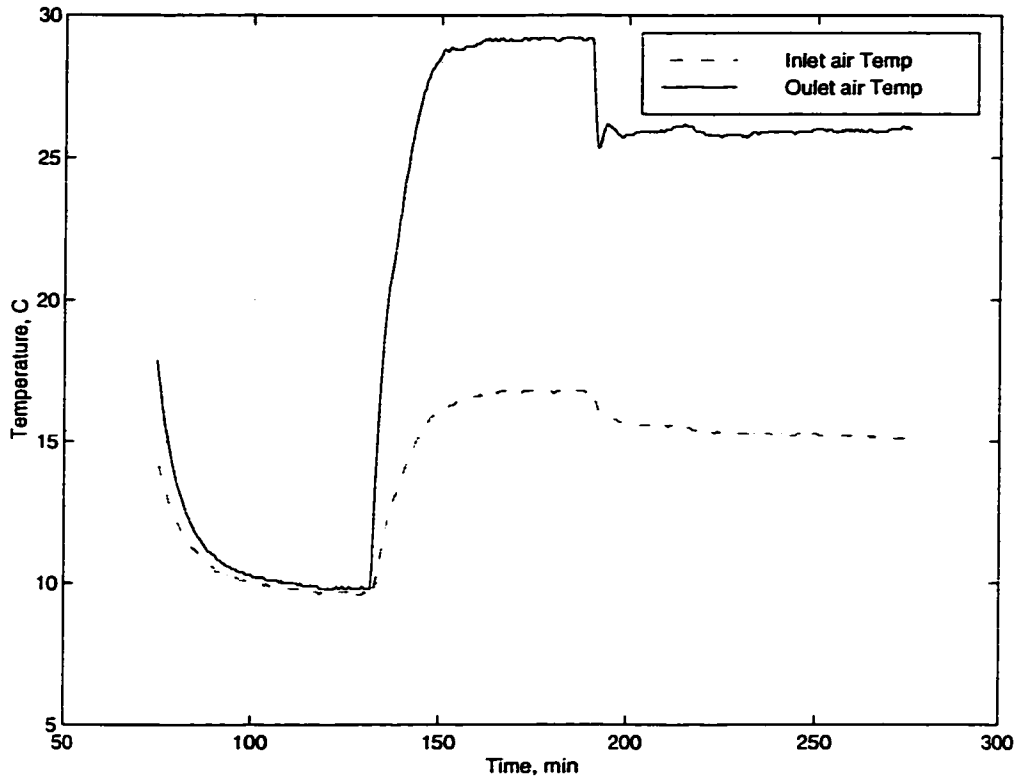


Figure 4.5 Response of outlet air temperature for step input of inlet water temperature

Steady state was observed for each run. However, some recent runs included transient points to aid the detection of normal operation during transient modes. Although multiple simultaneous fault groups can be investigated with the data gathered, the current study dealt only with the single fault groups.

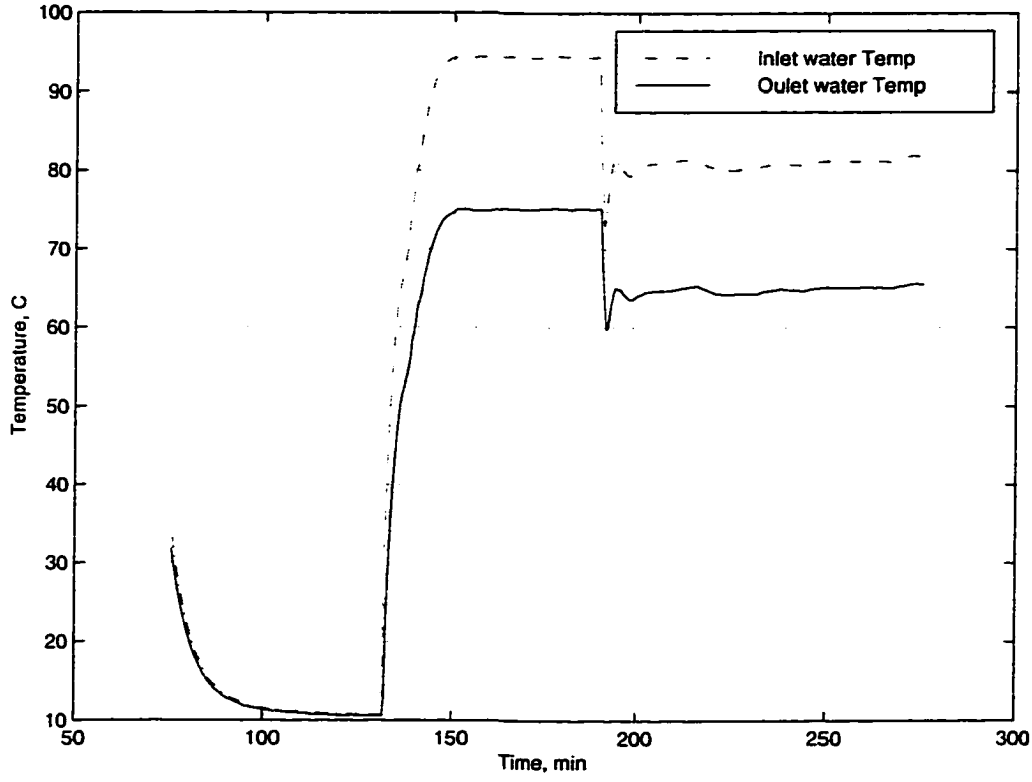


Figure 4.6 Response of outlet water temperature for step input of inlet water temperature

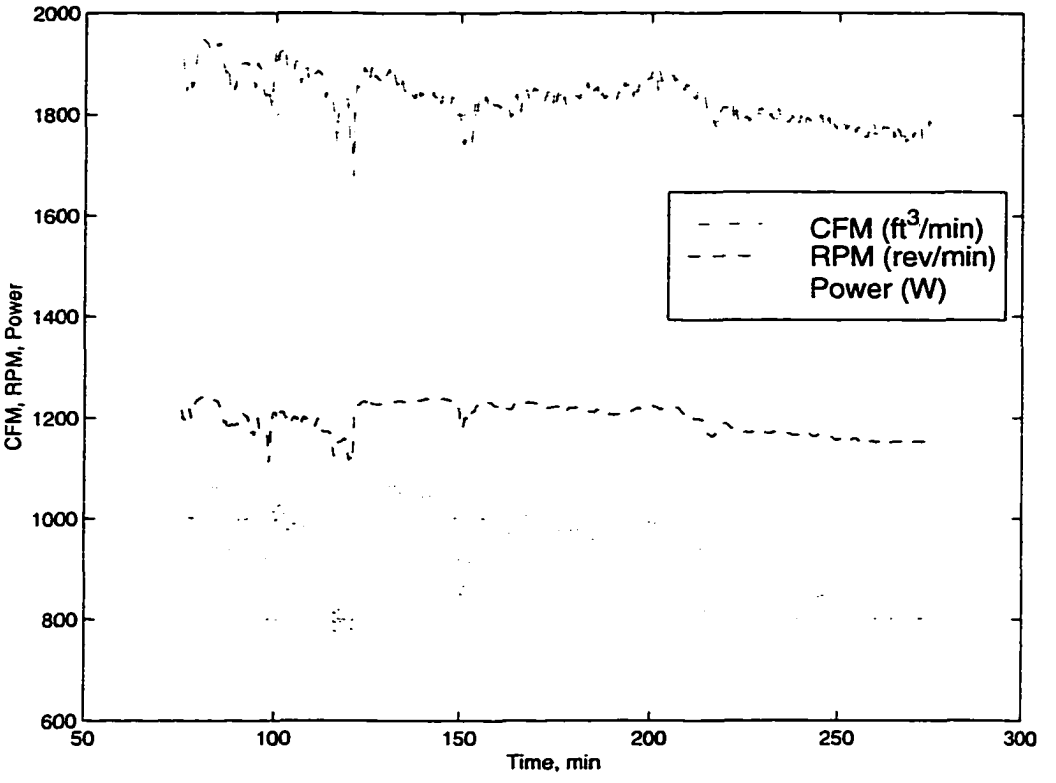


Figure 4.7 Response of air flow rate, fan speed, and fan power for step input of water flow rate

Table 4.4 Test order of 3^4 factorial experiment

Combination	Test Order	FT ID	RPM	Damper	Valve	Coil	Faults*
1	40	1111	90	10psi	Open	Open	n
2	18	1112	90	10psi	Open	Half	c
3	6	1113	90	10psi	Open	1/4 Block	c
4	8	1121	90	10psi	Half	Open	v
5	24	1122	90	10psi	Half	Half	vc
6	26	1123	90	10psi	Half	1/4 Block	vc
7	50	1131	90	10psi	1/4Open	Open	v
8	54	1132	90	10psi	1/4Open	Half	vc
9	31	1133	90	10psi	1/4Open	1/4 Block	vc
10	51	1211	90	13psi	Open	Open	n
11	44	1212	90	13psi	Open	Half	c
12	67	1213	90	13psi	Open	1/4 Block	c
13	5	1221	90	13psi	Half	Open	v
14	41	1222	90	13psi	Half	Half	vc
15	25	1223	90	13psi	Half	1/4 Block	vc
16	2	1231	90	13psi	1/4Open	Open	v
17	7	1232	90	13psi	1/4Open	Half	vc
18	68	1233	90	13psi	1/4Open	1/4 Block	vc
19	1	1311	90	7psi	Open	Open	n
20	48	1312	90	7psi	Open	Half	c
21	57	1313	90	7psi	Open	1/4 Block	c
22	11	1321	90	7psi	Half	Open	v
23	33	1322	90	7psi	Half	Half	vc
24	79	1323	90	7psi	Half	1/4 Block	vc
25	61	1331	90	7psi	1/4Open	Open	v
26	62	1332	90	7psi	1/4Open	Half	vc
27	47	1333	90	7psi	1/4Open	1/4 Block	vc
28	3	2111	80	10psi	Open	Open	r
29	35	2112	80	10psi	Open	Half	rc
30	76	2113	80	10psi	Open	1/4 Block	rc
31	14	2121	80	10psi	Half	Open	rv
32	15	2122	80	10psi	Half	Half	rvc
33	81	2123	80	10psi	Half	1/4 Block	rvc
34	63	2131	80	10psi	1/4Open	Open	rv
35	46	2132	80	10psi	1/4Open	Half	rvc
36	65	2133	80	10psi	1/4Open	1/4 Block	rvc
37	34	2211	80	13psi	Open	Open	r
38	49	2212	80	13psi	Open	Half	rc
39	4	2213	80	13psi	Open	1/4 Block	rc
40	38	2221	80	13psi	Half	Open	rv
41	45	2222	80	13psi	Half	Half	rvc
42	71	2223	80	13psi	Half	1/4 Block	rvc
43	42	2231	80	13psi	1/4Open	Open	rv
44	74	2232	80	13psi	1/4Open	Half	rvc
45	17	2233	80	13psi	1/4Open	1/4 Block	rvc
46	12	2311	80	7psi	Open	Open	r
47	32	2312	80	7psi	Open	Half	rc
48	64	2313	80	7psi	Open	1/4 Block	rc

Table 4.4 (Continued)

49	9	2321	80	7psi	Half	Open	rv
50	37	2322	80	7psi	Half	Half	rvc
51	30	2323	80	7psi	Half	1/4 Block	rvc
52	59	2331	80	7psi	1/4Open	Open	rv
53	13	2332	80	7psi	1/4Open	Half	rvc
54	19	2333	80	7psi	1/4Open	1/4 Block	rvc
55	75	3111	70	10psi	Open	Open	r
56	73	3112	70	10psi	Open	Half	rc
57	28	3113	70	10psi	Open	1/4 Block	rc
58	52	3121	70	10psi	Half	Open	rv
59	10	3122	70	10psi	Half	Half	rvc
60	36	3123	70	10psi	Half	1/4 Block	rvc
61	60	3131	70	10psi	1/4Open	Open	rv
62	23	3132	70	10psi	1/4Open	Half	rvc
63	43	3133	70	10psi	1/4Open	1/4 Block	rvc
64	53	3211	70	13psi	Open	Open	r
65	22	3212	70	13psi	Open	Half	rc
66	20	3213	70	13psi	Open	1/4 Block	rc
67	66	3221	70	13psi	Half	Open	rv
68	16	3222	70	13psi	Half	Half	rvc
69	70	3223	70	13psi	Half	1/4 Block	rvc
70	69	3231	70	13psi	1/4Open	Open	rv
71	78	3232	70	13psi	1/4Open	Half	rvc
72	39	3233	70	13psi	1/4Open	1/4 Block	rvc
73	72	3311	70	7psi	Open	Open	r
74	27	3312	70	7psi	Open	Half	rc
75	55	3313	70	7psi	Open	1/4 Block	rc
76	21	3321	70	7psi	Half	Open	rv
77	56	3322	70	7psi	Half	Half	rvc
78	29	3323	70	7psi	Half	1/4 Block	rvc
79	77	3331	70	7psi	1/4Open	Open	rv
80	58	3332	70	7psi	1/4Open	Half	rvc
81	80	3333	70	7psi	1/4Open	1/4 Block	rvc

* n: normal operation, c: coil fault, r: fan fault, v: valve fault

5 DATA ANALYSIS

This chapter describes the procedures used to organize the experimental data and perform an initial assessment of the data collected. As a preliminary determination of the fault groups, the data are first put into two group categories and the classification of the fault groups are performed.

Organization of data and simple descriptive statistics

As graphs and tabular arrangements of the observations help in the data analysis, summary numbers portraying certain features of the data are also needed to describe the data. In this research, the notation y_{ij} indicates a particular value of the i^{th} variable that is observed on the j^{th} item or observation. The notation of n, p, and k are assigned to the number of observations, the number of variables, and the number of groups. All of the observations on all of the variables are contained in a rectangular array of n rows and p columns and is given a notation $\mathbf{y}$. The n measurements on p variables are displayed as follows.

$$\mathbf{y} = \begin{pmatrix} y_{11} & y_{12} & \cdots & y_{1j} & \cdots & y_{1p} \\ y_{21} & y_{22} & \cdots & y_{2j} & \cdots & y_{2p} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ y_{i1} & y_{i2} & \cdots & y_{ij} & \cdots & y_{ip} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ y_{n1} & y_{n2} & \cdots & y_{nj} & \cdots & y_{np} \end{pmatrix}$$

In a simple descriptive summary, the measure of location (sample mean), spread (variance), and linear association (correlation) of the observations are calculated by the following equations.

$$\bar{\mathbf{y}} = \frac{\mathbf{y}'\mathbf{j}}{n}$$

where $\bar{\mathbf{y}}$ represents sample mean vector and $'$ denotes transpose of the observation matrix, $\mathbf{y}$. A vector of 1's is denoted by $\mathbf{j}$.

$$\mathbf{j} = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}$$

$$\mathbf{S} = (s_{ij}) = \begin{pmatrix} s_{11} & s_{12} & \cdots & s_{1p} \\ s_{21} & s_{22} & \cdots & s_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ s_{p1} & s_{p2} & \cdots & s_{pp} \end{pmatrix}$$

$\mathbf{S}$ is sample variance covariance matrix and the elements consists of variance, s_{ii}, of the i^{th} variable, and covariance, s_{ij}, of the i^{th} and j^{th} variables. Equations 5.1 and 5.2 define sample variance and covariance elements.

$$s_{ii} = s_i^2 = \frac{1}{n-1}\sum_{m=1}^{n}(y_{mi} - \bar{y}_i)^2 \tag{5.1}$$

$$s_{ij} = \frac{1}{n-1}\sum_{m=1}^{n}(y_{mi} - \bar{y}_i)(y_{mj} - \bar{y}_j) \tag{5.2}$$

The sample correlation between the i^{th} and j^{th} variables and the sample correlation matrix are given in equation 5.3 and 5.4. The sample correlation matrix is analogous to the covariance matrix with correlations in place of covariances. The second row, for example, contains the

correlation of y_2 with each of the y's. This matrix is symmetric, since $r_{ij} = r_{ji}$.

$$r_{ij} = \frac{s_{ij}}{\sqrt{s_{ii}s_{jj}}} \tag{5.3}$$

$$\mathbf{R} = r_{ij} = \begin{pmatrix} 1 & r_{12} & \dots & r_{1p} \\ r_{21} & 1 & \dots & r_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ r_{p1} & r_{p2} & \dots & 1 \end{pmatrix} \tag{5.4}$$

The correlation matrix can be obtained from the covariance matrix and vice versa.

$$\mathbf{D}_s = diag(\sqrt{s_{11}}, \sqrt{s_{11}}, \dots, \sqrt{s_{pp}}) \tag{5.5}$$

$$= diag(s_1, s_2, \dots, s_p) \tag{5.6}$$

$$= \begin{pmatrix} s_1 & 0 & \dots & 0 \\ 0 & s_2 & \dots & 0 \\ \vdots & \vdots & \dots & \dots \\ 0 & 0 & \dots & s_p \end{pmatrix} \tag{5.7}$$

then

$$\mathbf{R} = \mathbf{D}_s^{-1}\mathbf{S}\mathbf{D}_s^{-1} \tag{5.8}$$

and

$$\mathbf{S} = \mathbf{D}_s\mathbf{R}\mathbf{D}_s \tag{5.9}$$

The sample correlation coefficient, r_{ij} is a standardized version of the sample covariance, where the product of the square roots of the sample variances provides the standardization. Also, the sample correlation coefficient can be viewed as a sample covariance as seen from above equation. In a univariate analysis, when the original values of y_{ij} and y_{kj} are replaced by *standardized* values according to the equations 5.11, the values are commensurable because both sets are centered at zero and expressed in standard deviation units.

$$x_{ij} = \frac{(y_{ij} - \bar{y}_i)}{\sqrt{s_{ii}}} \tag{5.10}$$

$$x_{kj} = \frac{(y_{kj} - \bar{y}_k)}{\sqrt{s_{kk}}} \tag{5.11}$$

Although the signs of the sample correlation and the sample covariance are the same, the correlation is easier to interpret. The sample correlation coefficient is just the sample covariance of the standardized observations. Although the sample correlation coefficients and the variances do not convey all there is to know about the association between two variables, because of the existence of nonlinearity, they provide measures of linear association along a line. They tend to be very sensitive to outliers and can indicate association when little exists.

Initial data assessment

The scatter plot matrix, Figure 5.1, of the variables shows strong correlation between inlet and outlet water temperatures. Some relationship is found between air flow rate (CFM) and pressure rise (Dp) and also with air flow rate and the fan power (Pow). Another association is detected between hot water flow rate (Mwater) and hot water outlet temperatures (Twout). The plot also shows patterns suggesting that there are two or more separate clumps of observations. In the plot there does not seem to be any unusual observation with regard to the outliers. As seen in the patterns of the plot, the real physical indications are from the different damper settings, valve operations, and fan speed controls. Because the observations are entirely from normal operation, these different patterns do not in any way suggest faulty behavior. Figure 5.2 show scatter plot for both normal and fault operation modes. The plot shows some outliers and clumps of distinct central locations. At a glance it is difficult to tell if the outliers are due to fault mode or not. However, the plot seem to show more variation in group clusters than as seen in Figure 5.1.

Table 5.1 shows sample mean and standard deviation summaries of normal operation, fan fault, valve fault, and coil fault. Table 5.2 shows sample univariate 95% confidence interval for normal operation. When the means of the fault conditions are compared to the normal operation means, almost all the fault condition means fall within the 95% interval for normal operation. This suggests that data taken for normal operation overlap the fault conditions and univariate analysis may not be able to separate the fault conditions from that of the normal modes. This observation hints at the importance of the interaction among the variables.

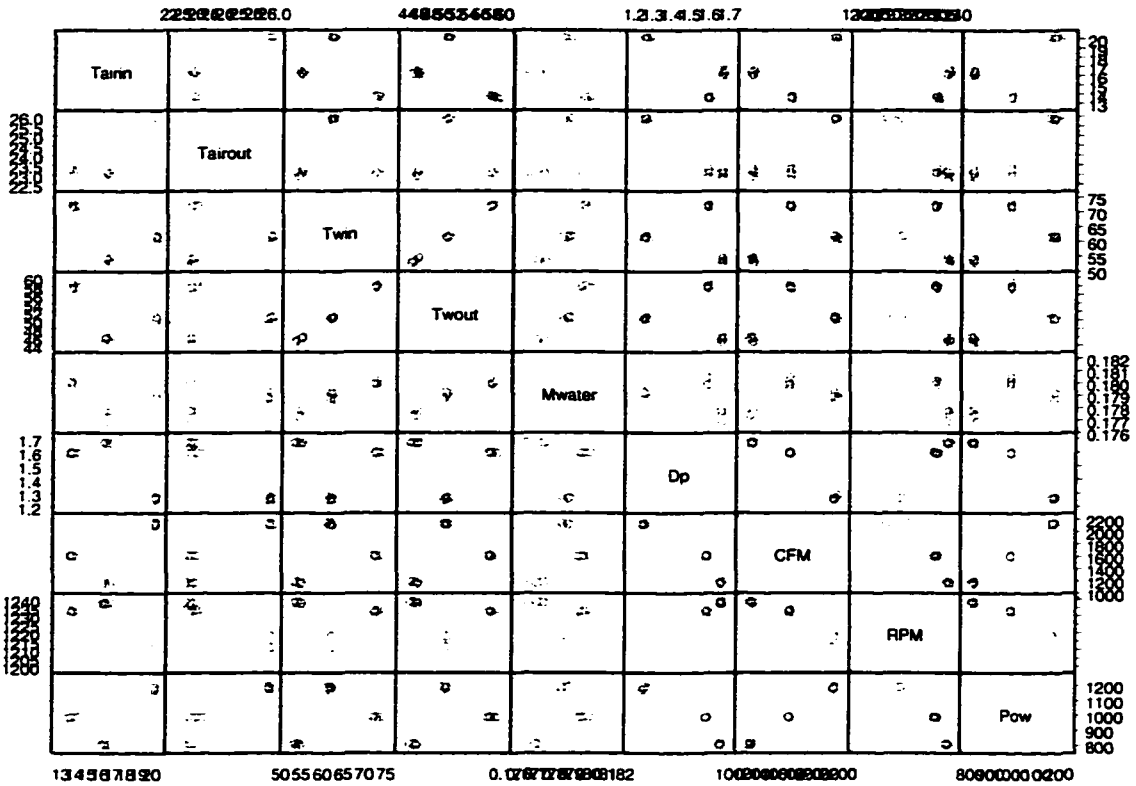


Figure 5.1 Normal operation scatter plot of AHU variables

Moreover, the study of the relationship among the variables leads to multivariate analysis. Tables 5.3 through 5.10 show variance and correlation matrices for each mode of operation.

Assessing Multivariate Normality

Investigating multivariate normality is not as straightforward as assessing univariate normality. Consequently, the state of the art is not as well developed (Rencher 1995). Numerous procedures have been proposed for checking for multivariate normality. The procedure used in this research is based on the standardized distance, equation 5.12, from each data point to its mean.

$$D_i^2 = (\mathbf{y_i} - \bar{\mathbf{y}})'\mathbf{S}^{-1}(\mathbf{y_i} - \bar{\mathbf{y}}) \tag{5.12}$$

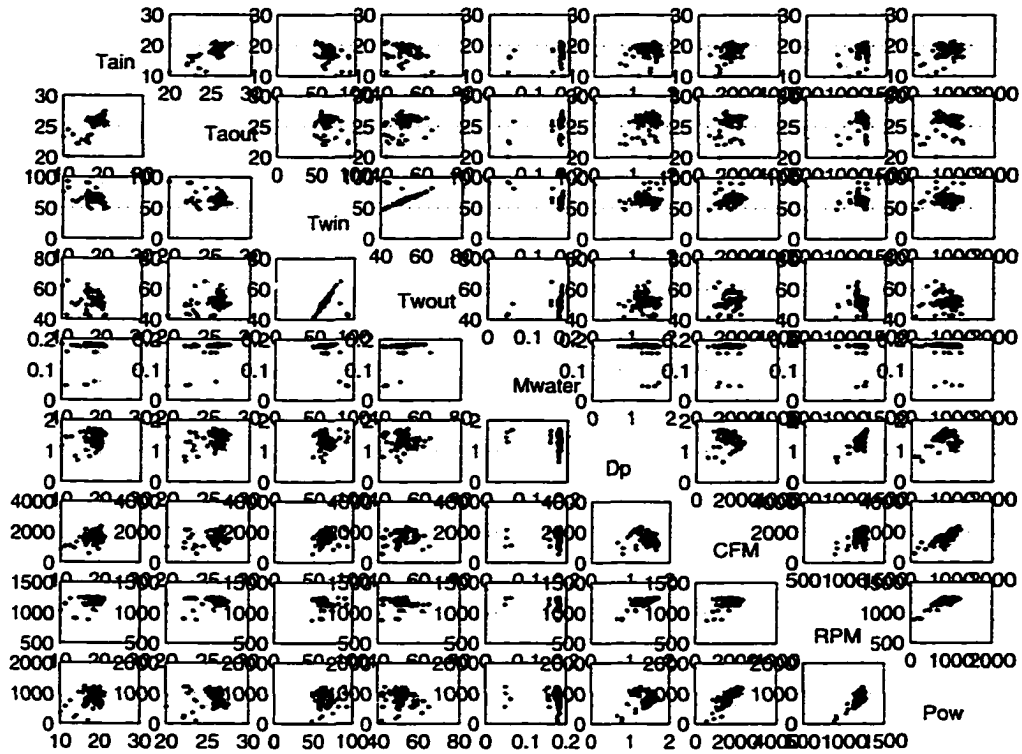


Figure 5.2 Normal and fault operation scatter plot of AHU variables

Gnanadesikan and Kettenring (1972) showed that if the $\mathbf{y_i}$'s are multivariate normal, then

$$u_i = \frac{n * D_i^2}{(n-1)^2} \tag{5.13}$$

has a beta distribution, which is related to the F distribution. To obtain the Quantile,Q-Q, plot, the values $u_1, u_2, \ldots, u_n$ are ranked to give $u_{(1)} \leq u_{(2)} \leq \ldots \leq u_{(n)}$, and are plotted against the quantiles v_i, equations 5.14 to 5.16.

$$\alpha = \frac{p-2}{2p} \tag{5.14}$$

$$\beta = \frac{n-p-2}{2(n-p-1)} \tag{5.15}$$

$$v_i = \frac{i-\alpha}{n-\alpha-\beta+1} \tag{5.16}$$

A nonlinear pattern in the plot would indicate a departure from normality.

Table 5.1 Sample mean and standard deviation vectors for each mode of operation

Fault type	Tair in	Tair out	Tw in	Tw out	Q water	ΔP	CFM	RPM	Power
Normal	15.97	23.98	67.39	49.71	0.1486	1.2675	1364.3	1142.7	733.7
Fan	15.49	23.78	60.91	50.74	0.1788	0.9164	1097.5	968.9	339.6
Valve	17.16	24.37	75.38	49.54	0.1131	1.4903	1599.8	1222.4	967.2
Coil	16.41	24.27	58.40	48.23	0.1776	1.3231	1484.2	1229.6	927.4
Normal	1.43	0.74	5.81	4.12	0.0026	0.1736	299.2	44.8	182.6
Fan	3.37	1.64	8.90	6.41	0.0025	0.1974	381.1	82.2	204.4
Valve	3.16	2.11	13.31	7.36	0.0566	0.1621	461.7	44.2	252.2
Coil	2.10	1.11	3.78	2.76	0.0016	0.2194	370.7	5.6	158.0

Table 5.2 A 95% confidence interval for the mean of AHU variables

Variable	Low limit	Mean	High limit
Tair in	14.636	15.965	17.294
Tair out	23.203	23.976	24.749
Tw in	61.099	67.391	73.683
Tw out	46.977	49.705	52.433
Q water	0.125	0.149	0.172
ΔP	1.124	1.268	1.411
CFM	1161.307	1364.298	1567.288
RPM	1082.880	1142.719	1202.558
Power	576.284	733.696	891.106

The normal operation of the AHU data as illustrated in figure 5.3 suggests that there are a few outliers and the assumption of normality may not be appropriate. In the event that normality is not a viable assumption, one alternative is to ignore the findings of a normality check and proceed as if the data were normally distributed. A second alternative is to make nonnormal data look more like normal by considering transformations of the data.

Performing a test for an outlier based on the distances D_i^2 in a graphical procedure for checking multivariate normality allows detection of the outliers of the data set. After the outliers were removed, Figure 5.4 suggest that an assumption of multivariate normality maybe applicable.

A power transformation is applicable to positive variables such as the physical variables obtained from this research. The transformed observation provide cosmetic effects in the sense that it is only the appearance of the data themselves that influence the transformation. Figure

Table 5.3 Sample covariance for normal operation

	Tair in	Tair out	Tw in	Tw out	Q water	ΔP	CFM	RPM	Power
Tair in	1.9433	0.7129	-4.6105	-3.2065	-0.0002	-0.0683	117.2552	2.5105	57.0918
Tair out	0.7129	0.7053	0.0673	0.4949	0.0010	-0.0187	-10.8800	-13.5213	-51.8523
Tw in	-4.6105	0.0673	28.9359	20.4545	0.0043	0.0519	123.8647	-58.1740	-120.6620
Tw out	-3.2065	0.4949	20.4545	15.4216	0.0040	0.0799	-56.9579	-52.1381	-191.8197
Q water	-0.0002	0.0010	0.0043	0.0040	0.0000	-0.0002	-0.0260	-0.0622	-0.1959
ΔP	-0.0683	-0.0187	0.0519	0.0799	-0.0002	0.0358	-30.9607	5.3509	-4.3284
CFM	117.2552	-10.8800	123.8647	-56.9579	-0.0260	-30.9607	71829.3436	394.4017	34718.6036
RPM	2.5105	-13.5213	-58.1740	-52.1381	-0.0622	5.3509	394.4017	1939.1981	4694.9343
Power	57.0918	-51.8523	-120.6620	-191.8197	-0.1959	-4.3284	34718.6036	4694.9343	29860.1487

Table 5.4 Sample correlation matrix for normal operation

	Tair in	Tair out	Tw in	Tw out	Q water	ΔP	CFM	RPM	Power
Tair in	1.0000	0.6089	-0.6148	-0.5857	-0.0621	-0.2591	0.3138	0.0409	0.2370
Tair out	0.6089	1.0000	0.0149	0.1501	0.4447	-0.1179	-0.0483	-0.3656	-0.3573
Tw in	-0.6148	0.0149	1.0000	0.9683	0.3075	0.0510	0.0859	-0.2456	-0.1298
Tw out	-0.5857	0.1501	0.9683	1.0000	0.3886	0.1075	-0.0541	-0.3015	-0.2827
Q water	-0.0621	0.4447	0.3075	0.3886	1.0000	-0.3078	-0.0372	-0.5411	-0.4345
ΔP	-0.2591	-0.1179	0.0510	0.1075	-0.3078	1.0000	-0.6107	0.6424	-0.1324
CFM	0.3138	-0.0483	0.0859	-0.0541	-0.0372	-0.6107	1.0000	0.0334	0.7497
RPM	0.0409	-0.3656	-0.2456	-0.3015	-0.5411	0.6424	0.0334	1.0000	0.6170
Power	0.2370	-0.3573	-0.1298	-0.2827	-0.4345	-0.1324	0.7497	0.6170	1.0000

Table 5.5 Sample covariance for fan fault

	Tair in	Tair out	Tw in	Tw out	Q water	*DelP*	CFM	RPM	Power
Tair in	9.0275	3.6347	-12.3764	-8.8043	-0.0018	0.0272	375.0875	56.2559	248.4918
Tair out	3.6347	2.4473	0.3543	0.5878	0.0011	0.1741	219.6168	91.4648	229.4533
Tw in	-12.3764	0.3543	59.3315	42.3505	0.0106	0.6955	996.5260	362.3459	741.9189
Tw out	-8.8043	0.5878	42.3505	30.6256	0.0085	0.5754	588.7789	273.8291	522.2024
Q water	-0.0018	0.0011	0.0106	0.0085	0.0000	0.0003	-0.1248	0.1117	0.1296
ΔP	0.0272	0.1741	0.6955	0.5754	0.0003	0.0330	-1.6936	11.7303	18.1755
CFM	375.0875	219.6168	996.5260	588.7789	-0.1248	-1.6936	118724.7866	12596.1835	52217.1228
RPM	56.2559	91.4648	362.3459	273.8291	0.1117	11.7303	12596.1835	5718.4126	12679.7993
Power	248.4918	229.4533	741.9189	522.2024	0.1296	18.1755	52217.1228	12679.7993	35379.7514

Table 5.6 Sample correlation matrix for fan fault

	Tair in	Tair out	Tw in	Tw out	Q water	ΔP	CFM	RPM	Power
Tair in	1.0000	0.7733	-0.5348	-0.5295	-0.2638	0.0498	0.3623	0.2476	0.4397
Tair out	0.7733	1.0000	0.0294	0.0679	0.3075	0.6127	0.4074	0.7732	0.7798
Tw in	-0.5348	0.0294	1.0000	0.9935	0.6021	0.4971	0.3755	0.6221	0.5121
Tw out	-0.5295	0.0679	0.9935	1.0000	0.6707	0.5724	0.3088	0.6543	0.5017
Q water	-0.2638	0.3075	0.6021	0.6707	1.0000	0.8256	-0.1585	0.6465	0.3017
ΔP	0.0498	0.6127	0.4971	0.5724	0.8256	1.0000	-0.0271	0.8539	0.5319
CFM	0.3623	0.4074	0.3755	0.3088	-0.1585	-0.0271	1.0000	0.4834	0.8057
RPM	0.2476	0.7732	0.6221	0.6543	0.6465	0.8539	0.4834	1.0000	0.8915
Power	0.4397	0.7798	0.5121	0.5017	0.3017	0.5319	0.8057	0.8915	1.0000

Table 5.7 Sample covariance matrix for valve fault

	Tair in	Tair out	Tw in	Tw out	Q water	ΔP	CFM	RPM	Power
Tair in	4.4855	2.0280	-17.1629	-2.3056	0.0342	0.0577	149.9536	45.0913	169.1933
Tair out	2.0280	2.4590	-17.0966	5.3429	0.0627	0.0909	-282.8069	6.4888	-97.1598
Tw in	-17.1629	-17.0966	166.0968	-32.6804	-0.6113	-0.4221	1492.3958	-32.2022	468.9523
Tw out	-2.3056	5.3429	-32.6804	32.5149	0.2146	-0.0739	-858.5299	-97.9546	-551.3378
Q water	0.0342	0.0627	-0.6113	0.2146	0.0027	0.0000	-5.4997	-0.3900	-2.7964
ΔP	0.0577	0.0909	-0.4221	-0.0739	0.0000	0.0268	-57.6371	1.4969	-21.4538
CFM	149.9536	-282.8069	1492.3958	-858.5299	-5.4997	-57.6371	185864.1807	2774.6018	84569.2474
RPM	45.0913	6.4888	-32.2022	-97.9546	-0.3900	1.4969	2774.6018	820.0150	2869.5436
Power	169.1933	-97.1598	468.9523	-551.3378	-2.7964	-21.4538	84569.2474	2869.5436	41935.8196

Table 5.8 Sample correlation matrix for valve fault

	Tair in	Tair out	Tw in	Tw out	Q water	ΔP	CFM	RPM	Power
Tair in	1.0000	0.6106	-0.6288	-0.1909	0.3081	0.1663	0.1642	0.7435	0.3901
Tair out	0.6106	1.0000	-0.8460	0.5975	0.7633	0.3542	-0.4183	0.1445	-0.3026
Tw in	-0.6288	-0.8460	1.0000	-0.4447	-0.9060	-0.2000	0.2686	-0.0873	0.1777
Tw out	-0.1909	0.5975	-0.4447	1.0000	0.7190	-0.0791	-0.3492	-0.5999	-0.4722
Q water	0.3081	0.7633	-0.9060	0.7190	1.0000	-0.0004	-0.2437	-0.2601	-0.2608
ΔP	0.1663	0.3542	-0.2000	-0.0791	-0.0004	1.0000	-0.8166	0.3193	-0.6399
CFM	0.1642	-0.4183	0.2686	-0.3492	-0.2437	-0.8166	1.0000	0.2247	0.9579
RPM	0.7435	0.1445	-0.0873	-0.5999	-0.2601	0.3193	0.2247	1.0000	0.4893
Power	0.3901	-0.3026	0.1777	-0.4722	-0.2608	-0.6399	0.9579	0.4893	1.0000

Table 5.9 Sample covariance matrix for coil fault

	Tair in	Tair out	Tw in	Tw out	Q water	ΔP	CFM	RPM	Power
Tair in	3.9049	1.3192	-0.3791	-0.3076	-0.0002	-0.0708	379.5364	3.6226	173.4438
Tair out	1.3192	1.1203	1.9594	1.7226	0.0011	0.0242	72.2033	0.9432	32.4369
Tw in	-0.3791	1.9594	13.1226	9.2524	0.0035	-0.1710	416.0351	-3.6101	170.2743
Tw out	-0.3076	1.7226	9.2524	7.0441	0.0031	-0.0300	156.9815	-2.2375	61.3388
Q water	-0.0002	0.0011	0.0035	0.0031	0.0000	0.0001	-0.1332	-0.0043	-0.0687
ΔP	-0.0708	0.0242	-0.1710	-0.0300	0.0001	0.0420	-58.2241	0.9453	-23.5859
CFM	379.5364	72.2033	416.0351	156.9815	-0.1332	-58.2241	116461.2284	-404.0861	49850.7488
RPM	3.6226	0.9432	-3.6101	-2.2375	-0.0043	0.9453	-404.0861	146.0310	97.8694
Power	173.4438	32.4369	170.2743	61.3388	-0.0687	-23.5859	49850.7488	97.8694	22132.6687

Table 5.10 Sample correlation matrix for coil fault

	Tair in	Tair out	Tw in	Tw out	Q water	ΔP	CFM	RPM	Power
Tair in	1.0000	0.6307	-0.0530	-0.0586	-0.0631	-0.1748	0.5628	0.1517	0.5900
Tair out	0.6307	1.0000	0.5110	0.6132	0.6339	0.1114	0.1999	0.0737	0.2060
Tw in	-0.0530	0.5110	1.0000	0.9623	0.5735	-0.2303	0.3365	-0.0825	0.3160
Tw out	-0.0586	0.6132	0.9623	1.0000	0.6977	-0.0551	0.1733	-0.0698	0.1553
Q water	-0.0631	0.6339	0.5735	0.6977	1.0000	0.2325	-0.2322	-0.2128	-0.2748
ΔP	-0.1748	0.1114	-0.2303	-0.0551	0.2325	1.0000	-0.8321	0.3815	-0.7732
CFM	0.5628	0.1999	0.3365	0.1733	-0.2322	-0.8321	1.0000	-0.0980	0.9819
RPM	0.1517	0.0737	-0.0825	-0.0698	-0.2128	0.3815	-0.0980	1.0000	0.0544
Power	0.5900	0.2060	0.3160	0.1553	-0.2748	-0.7732	0.9819	0.0544	1.0000

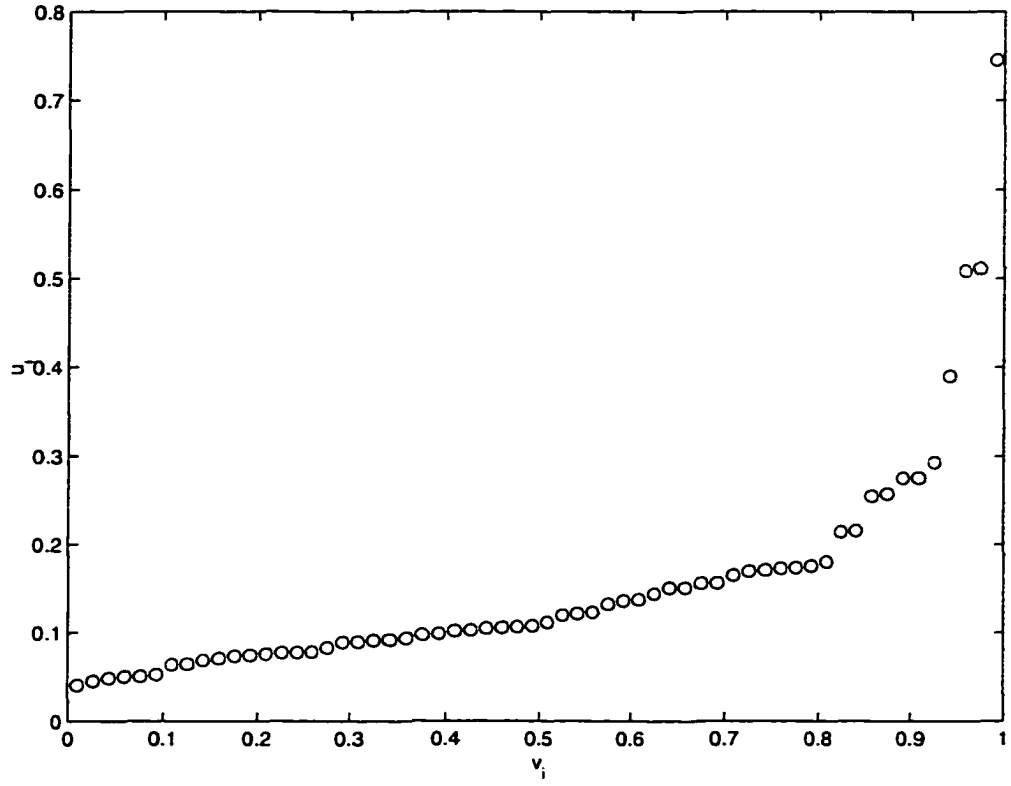


Figure 5.3 $Q-Q$ plot of u_i and v_i for normal operation data with outliers

5.5 suggests that the power transformation did in fact remove or relax some non-linearity in the data set.

Box and Cox (1964) considered the power transformations by the following equation for $\lambda \neq 0$.

$$y_t^{(\lambda)} = \frac{y^\lambda - 1}{\lambda}$$

where $y_t^{(\lambda)}$ is the new transformed observation for the old values of y, and λ is the degree of power to be used for transformation. It is recognized that the transformation improves the approximation to normality, (Johnson 1992). However, there is no guarantee that even the best choice of λ will produce a transformed set of values that adequately conform to a normal distribution. Therefore, a provisional assumption of normality is applied for the analysis.

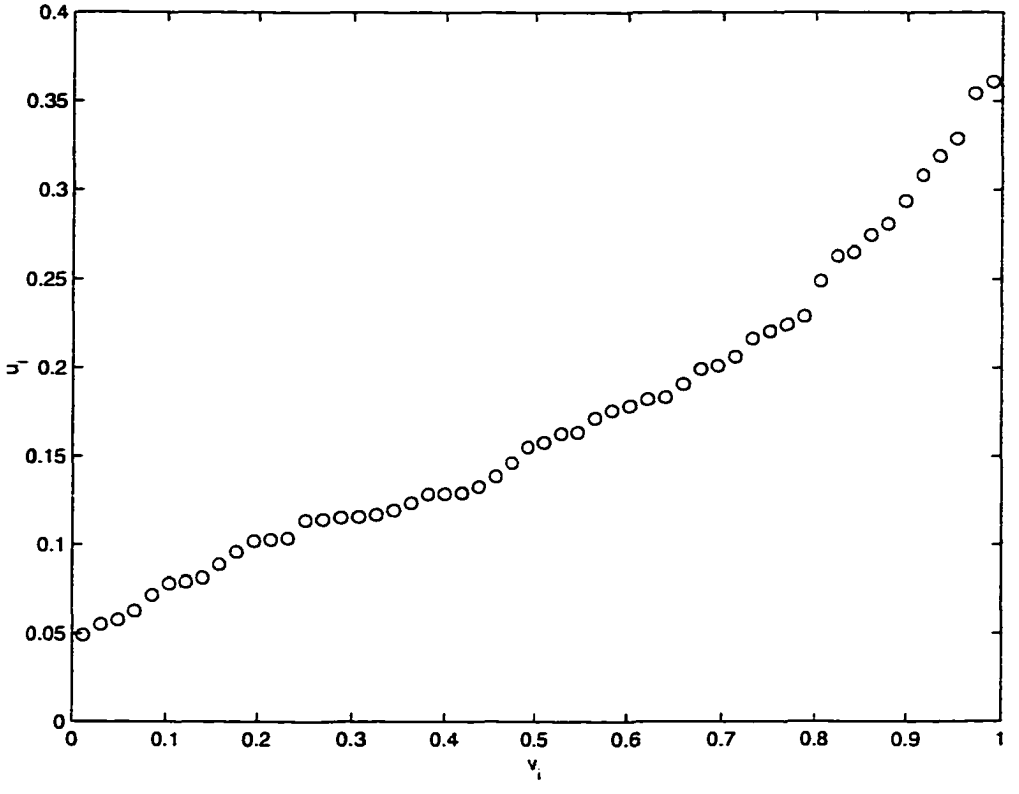


Figure 5.4 $Q - Q$ plot of u_i and v_i for normal operation data without outliers

Multivariate analysis of variance, one-way model

In this section, k sample groups are tested to compare the group means to see if they are sufficiently different from each other. MANOVA one-way model is utilized for the test. Table 5.11 shows estimated eigenvalues.

Calculated Λ statistic resulted in a value of 0.0287 and Wilks Lambda critical value, $\Lambda_{\alpha,p,\nu H,\nu E}$, of 0.97 is entered from the Table A.9 in Rencher (1995) for α of 0.05, p number of variables 9, ν_H, degrees of freedom for hypothesis 3, and ν_E, degrees of freedom for error of 1478. Wilks' Test Statistic is defined as follows.

$$\Lambda = \frac{|\mathbf{E}|}{|\mathbf{E}+\mathbf{H}|}$$

If $\Lambda \leq \Lambda_{\alpha,p,\nu H,\nu E}$, then reject H_o where

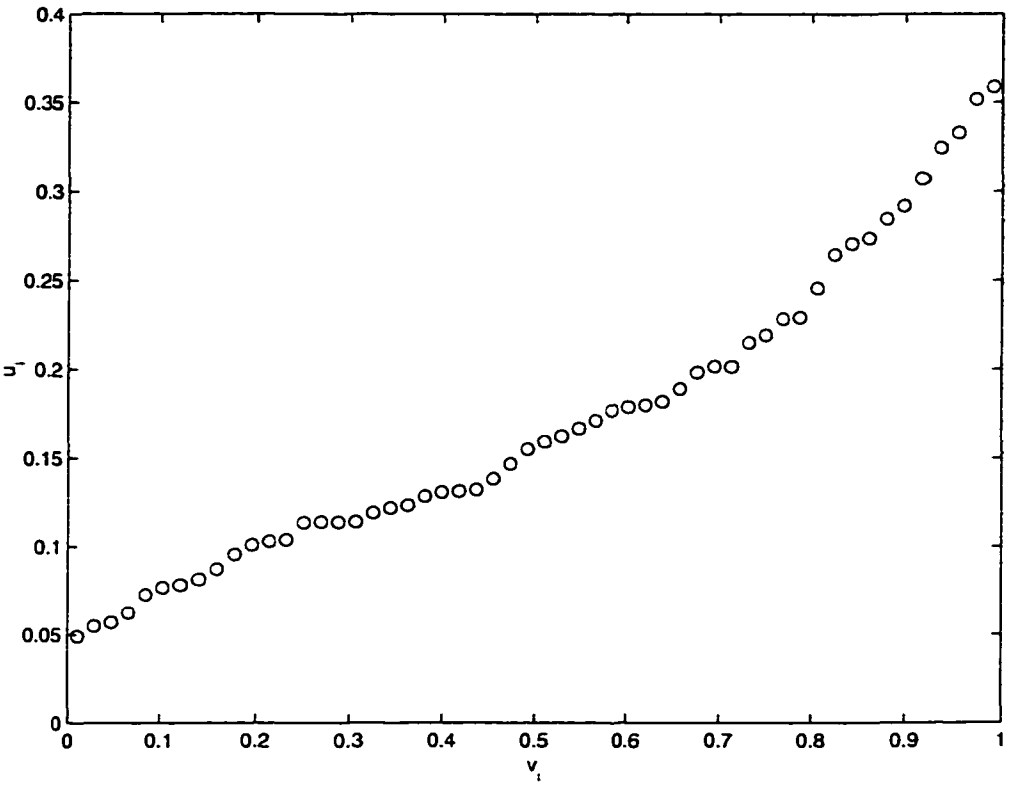


Figure 5.5 $Q - Q$ plot of u_i and v_i for transformed normal operation data with $\lambda = 0.9$

- Λ = Wilks' Λ
- α = level of significance, 0.05
- p = number of variables (dimension)
- νH = degrees of freedom for hypothesis
- νE = degrees of freedom for error

Therefore, by Wilks' test statistic, the conclusion is to Reject the null hypothesis, $Ho : \mu_1 = \mu_2 = \ldots = \mu_p$ and an indication of the pattern of the mean vectors is given by the eigenvalues of $E^{-1}H$, Table 5.11.

If there is one large eigenvalue and the others are small, the mean vectors lie close to a line in space. If there are two large eigenvalues, the mean vectors lie mostly in two dimensions, and so on. Here, there are three large eigenvalues and thus the mean vectors lie mostly in three

Table 5.11 Eigenvalues of means and associated proportions

	Eigen value	Proportion
PC 1	2.9549	0.4204
PC 2	2.6805	0.8017
PC 3	1.3940	1.0000
PC 4	0.0000	1.0000
PC 5	0.0000	1.0000
PC 6	0.0000	1.0000
PC 7	0.0000	1.0000
PC 8	0.0000	1.0000
PC 9	0.0000	1.0000

dimensions. Because the Roy test uses only the largest eigenvalue of $\mathbf{E}^{-1}\mathbf{H}$ it is more powerful than others if the mean vectors are collinear. However, Wilks' Λ has played the dominant role in significance tests in MANOVA because it was the first to be derived and has well known χ^2 and F approximations, Rencher (1995).

Discriminant Analysis

As a precursor to the more general study on fault classification in an AHU, this section covers discrimination between normal operation and fan faults. Discriminant analysis can be described as a means of group separation. The term *group* represents either a population or a sample from the population. The major objective in separation of groups involves the description of group separation, where linear functions (discriminant functions) of the variables are used to describe or clarify the differences between two or more groups.

Using linear discriminant function

The procedure for determining the discriminant function between two group membership is as follows:

- Calculate covariance matrices for the groups
- Calculate pooled covariance matrix
- Calculate the inverse of the pooled covariance matrix

- Calculate the vector of mean differences
- Calculate the coefficients of the linear discriminant function
- Calculate the bivariate standard distance

After calculating the bivariate standard distance, we know the optimal linear combination of the data vectors, $\mathbf{x}$, and that the means of the groups, are D standard distance apart. A question of how many variables should be included or excluded remains to be answered. In some instances adding more variables can degrade the ability of the discriminant function to classify future observations correctly (Flury 1997). The problem is to observe the subsets of variables to be used for discrimination and decide, for each variable, whether or not the increase in standard distance justifies the addition of the variable to the analysis. A not so appealing but acceptable method is to look at the summary of an 'all subsets linear discriminant analysis'. The standard distance plays a role that is similar to the role of the coefficient of determination (R^2) in multiple regression. Furthermore, it gives an overall assessment of the success of the analysis (Flury 1997).

When it comes to assessing the importance of regressors or subsets of regressors, one notion is the redundancy of the variables and under what circumstances does a seemingly worthless variable contribute to multivariate discrimination (associated testing theory). In the classification of the RPM faults against normal operation, univariate sample statistics for all 9 variables are given in Table 5.12.

Table 5.13 lists the standard distance between the two groups by using pooled covariance, normalized covariance matrix, and 2^{nd} group covariance. The classification cutoff values came near the estimation by pooled covariance. This estimates the fan fault cutoff point to discriminate against the normal operating condition.

From the Table 5.14, the linear discriminant function, V, can be written as follows: $V = -1.871 * x_1 + 3.814 * x_2 - 0.255 * x_3 - 0.467 * x_4 + 491.37 * x_5 + 16.599 * x_5 + 0.033 * x_6 + 0.147 * x_7 - 0.036 * x_8$. The center of the discriminant function and the standard deviation is listed in Table 5.15. The entire subset of the linear discriminant analysis needs to be performed to see

Table 5.12 Univariate summary statistics in RPM and normal discrimination

Normal mean	RPM mean	Normal S	RPM S	Standard Distance
18.500	15.282	1.6928	3.1125	1.7547
25.990	23.715	0.8808	1.5603	2.4019
64.627	61.139	5.9736	7.8316	0.5691
53.379	50.941	4.4029	5.5814	0.5419
0.181	0.179	0.0027	0.0024	0.6966
0.139	0.917	0.1798	0.1814	2.6379
1766.700	1085.000	273.9100	351.5500	2.4320
1186.000	968.120	42.1570	75.2840	4.8002
893.220	335.740	176.2300	188.3600	3.1470

Table 5.13 Multivariate discriminant analysis summary statistics for RPM vs. normal operation

Multivariate standard distance using pooled covariance	6.64
Multivariate standard distance using normal covariance matrix	8.41
Multivariate standard distance using RPM covariance matrix	10.77

the pattern of redundancy. The result should reveal the effects from the multivariate standard distance. If the standard distance does not change much, then there is evidence of redundancy.

Using the normal theory classification rule, it is reasonable to assume prior probabilities by the relative frequencies. Table 5.16 lists the prior probabilities for normal and fan faults. Normal theory classification rule places the observations into group 1 if

$$\mathbf{b}'\mathbf{Y} - 1/2\mathbf{b}(\bar{\mathbf{y}}_1 + \bar{\mathbf{y}}_2) + \log(\pi_1/\pi_2) > 0 \tag{5.17}$$

Table 5.14 Coefficients of the linear discriminant function

Coefficient by Pooled	Coefficient by Normal	Coefficient by RPM
-1.871	-0.945	-43.409
3.814	5.919	50.887
-0.255	0.665	-2.646
-0.467	-0.728	-12.502
491.370	227.771	4934.844
16.599	-24.703	718.542
0.033	0.031	0.183
0.147	0.416	-2.994
-0.036	-0.073	0.640

Table 5.15 Discrimination function means and standard deviations

Group	V	standard deviation
Normal	336.356	5.553
RPM	292.232	14.599

Table 5.16 Two group faults and prior probabilities

Group	Sample size	Prior probabilities
Normal	1514	0.92656
Fan fault	120	0.07344

and classifies them into Group 2 if

$$\mathbf{b}'\mathbf{Y} - 1/2\mathbf{b}(\bar{\mathbf{y}}_1 + \bar{\mathbf{y}}_2) + \log(\pi_1/\pi_2) < 0 \tag{5.18}$$

The half distance of the group means was 84.75 and the Normal Theory Classification cutoff point was estimated to be 82.44. From the Figures 5.6 and 5.7 we can see that, although the two linear combinations appear to be quite different, they are highly correlated and yield about the same group separation. The plot suggests that a single linear combination would be adequate. Normal error rates for classification were 0.357%. Misclassification of 78 out of 1334 observation resulted in a Plug-in-Rate of 5.84% while leave-one-out methods gave misclassification of 80 out of 1334 with leave-one-out error rate of 5.99%. Figure 5.8 shows normal theory classification based on the distribution of the discriminant function based on 2nd group data. It is clearly demarcated by the z score of 0 where the separation takes place. As the posterior probability increase and decrease the group identity thus is assigned to 2 when the posterior probability is greater than 0.5 and group 1 if the posterior probability is less than 0.5 accordingly.

The results obtained for groups of 4 modes of operation are discussed in the next chapter. The goals of discriminant analysis include identifying the relative contribution of the p variables to separation of the groups and finding the optimal plane on which the points can be projected to best illustrate the configuration of the groups

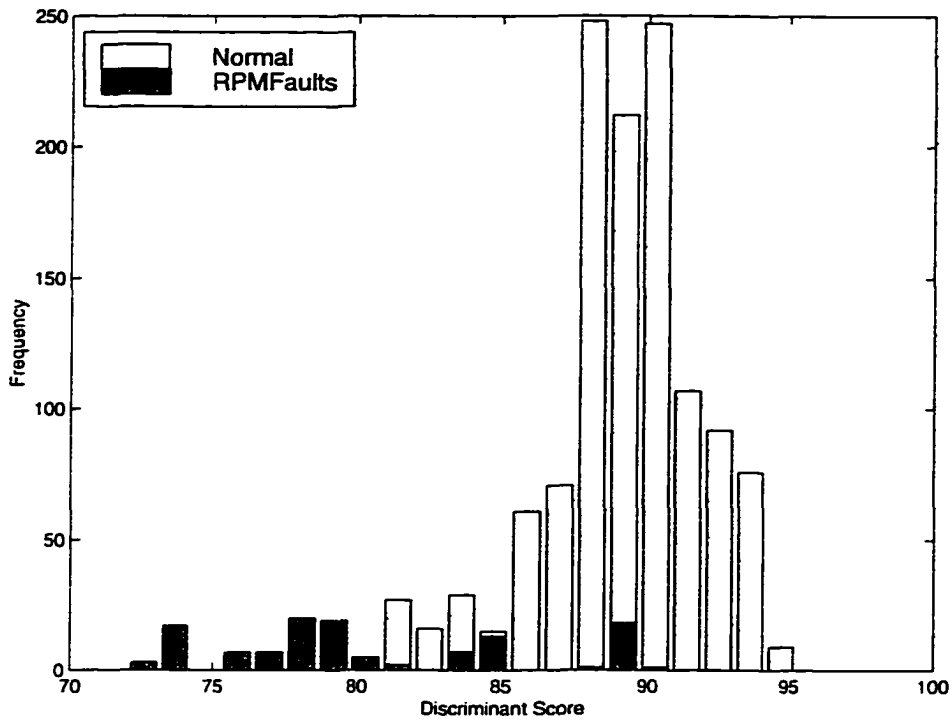


Figure 5.6 Histogram of the discriminant function

In classification, prediction or allocation are utilized where linear or quadratic functions (classification functions) of the variables are employed to assign an individual sampling unit to one of the groups. The measured values (in the observation vector) for an individual or object are evaluated by the classification functions to see to which group the individual most likely belongs.

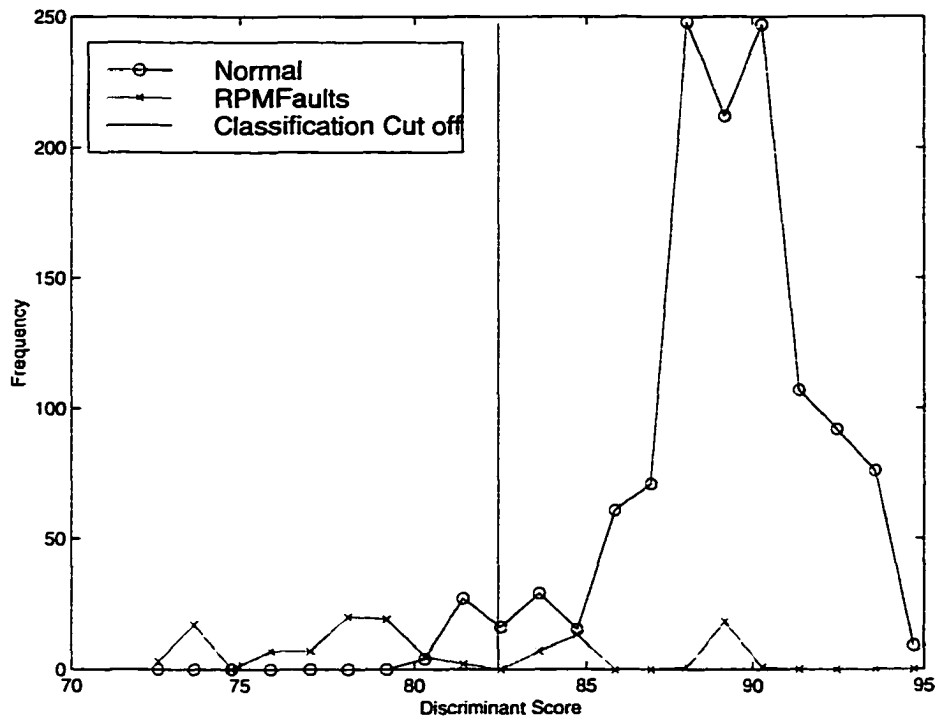


Figure 5.7 Distribution of the linear discriminant function for fault groups

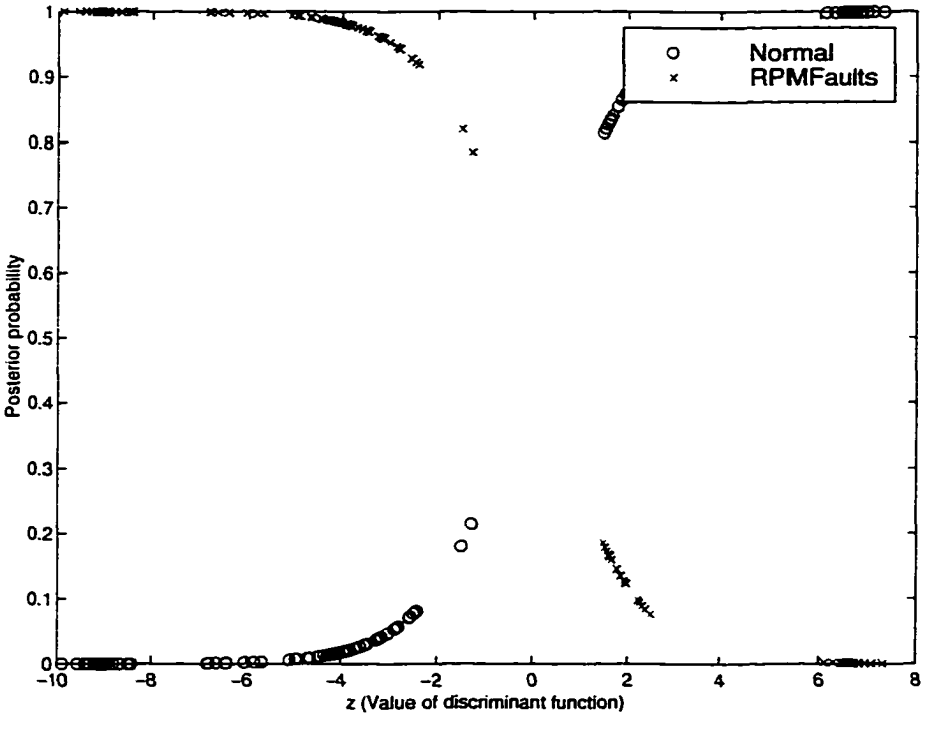


Figure 5.8 Normal theory classification based on the distribution of the discriminant function

6 RESULTS AND DISCUSSION

This chapter is organized into three sections: principal component of the AHU; discrimination and classification of the fault group associated with the AHU operation; and logistic regression of the fault classification. The first section provides description of the AHU data variance covariance structure using a few linear combinations of the original variables for data reduction and interpretation. Following this discussion, the results of a discrimination and classification of the AHU faults are presented and discussed. Together with the results obtained from the logistic regression analysis, the final section combines PCA and discriminant analysis and discusses the discoveries in detection and classification of the faults in AHU. This study attempts to provide solutions to the detection of the faults through the use of minimal information obtained from experimental data.

Principal component analysis

One of the objectives of this study was to determine the three distinct faults present in the operation of the AHU. Although occurrence of the faults maybe difficult to identify using typical procedures, principal component analysis can be used to search for groupings among the AHU data collected.

PCA identifies the smallest number of factors that together account for all of the total variance in the correlation matrix of the original variables. If the variables are highly correlated, the essential dimensionality is much smaller than the p number of variables. That is, if the first few eigenvalues are large, the proportion of the variance will be close to 1 for a small value of group number, k. On the other hand, if the correlations among the variables are all small, the dimensionality is close to the number of p variables, and the eigenvalues will be nearly equal.

In this case, no useful reduction in dimensionality is achieved because the principal components essentially duplicate the variables. Tables 6.1 and 6.2 list covariance and correlation matrices for the AHU data. The details of the PCA result are discussed next.

Determining the number of principal components

Several different types of stopping rules have been developed. When conducting PCA, researchers specify, sometimes, a priori, that successive factors will be extracted until some absolute percentage of the total variance has been explained. This rule is known as percentage of variance criterion. Another stopping rule is known as the a priori criterion. Kaiser's (1960) stopping rule extract only eigenvectors with eigenvalues of at least 1. Cattell (1966) proposed a graphical procedure, known as the scree test. To conduct the scree test, the eigenvalues are plotted on the y axis in order of magnitude. The eigenvalues in the steep descent are retained, and the eigenvalues in the gradual descent including the eigenvalues occurring in the transition from steep to gradual descent are dropped.

Summarizing research concerning the accuracy of the stopping procedures, Stevens (1986) concluded that Kaiser's stopping rule should be used for applications in which there are fewer than 30 variables. Otherwise, the scree test should be used in applications for which there are at least 200 observations and the commonalities are reasonably large. This research used Cattell's scree test due to availability of the large sample observations.

Figure 6.1 shows the scree plot of the AHU data. The scree plot exhibits a semi-ideal pattern. The first two eigenvalues form a steep curve followed by a bend and another set of steep curve then a straight line trend. The recommendation is to retain 5 principal components extracted from the AHU data points. In other instances, the turning point between the steep curve and the straight line may not be as distinct as this or there may be many bends. The data points used were from all operation modes. The total number of the AHU data included 403 data points and 9 variables.

To account for 92% of the variance, Table 6.3 indicates that 4 components should be retained. This percentage of the variance is high enough for most descriptive purposes. Rencher

Table 6.1 Covariance matrix for AHU variables

Tain (C)	Taout (C)	Twin (C)	Twout (C)	Qwater (l/s)	ΔP (inH2O)	CFM (ft^3/min)	RPM	Pow (W)
4.5208E+00	2.0409E+00	-8.5455E+00	-2.5857E+00	1.5811E-02	8.4324E-02	3.7246E+02	5.6979E+01	2.2950E+02
2.0409E+00	1.6851E+00	-2.5000E+00	2.1661E+00	1.7699E-02	5.7080E-02	1.5497E+02	1.8533E+01	4.9954E+01
-8.5455E+00	-2.5000E+00	6.2675E+01	1.8249E+01	-1.1957E-01	-6.3982E-02	4.0263E+02	-3.2719E+01	4.9076E+01
-2.5857E+00	2.1661E+00	1.8249E+01	2.1583E+01	4.3356E-02	-1.0678E-01	3.0002E+02	-6.3184E+01	-1.0616E+02
1.5811E-02	1.7699E-02	-1.1957E-01	4.3356E-02	6.4788E-04	-7.2546E-04	1.0852E+00	-2.2903E-01	-3.4670E-01
8.4324E-02	5.7080E-02	-6.3982E-02	-1.0678E-01	-7.2546E-04	4.4805E-02	-1.5237E+01	1.1216E+01	1.1439E+01
3.7246E+02	1.5497E+02	4.0263E+02	3.0002E+02	1.0852E+00	-1.5237E+01	1.3101E+05	8.3296E+03	6.4576E+04
5.6979E+01	1.8533E+01	-3.2719E+01	-6.3184E+01	-2.2903E-01	1.1216E+01	8.3296E+03	4.9693E+03	1.2382E+04
2.2950E+02	4.9954E+01	4.9076E+01	-1.0616E+02	-3.4670E-01	1.1439E+01	6.4576E+04	1.2382E+04	5.1872E+04

Table 6.2 Correlation matrix for AHU variables

Tain (C)	Taout (C)	Twin (C)	Twout (C)	Qwater (l/s)	ΔP (inH2O)	CFM (ft^3/min)	RPM	Pow (W)
1.0000	0.7394	-0.5077	-0.2618	0.2922	0.1874	0.4840	0.3802	0.4739
0.7394	1.0000	-0.2433	0.3592	0.5357	0.2077	0.3298	0.2025	0.1690
-0.5077	-0.2433	1.0000	0.4962	-0.5934	-0.0382	0.1405	-0.0586	0.0272
-0.2618	0.3592	0.4962	1.0000	0.3666	-0.1086	0.1784	-0.1929	-0.1003
0.2922	0.5357	-0.5934	0.3666	1.0000	-0.1346	0.1178	-0.1276	-0.0598
0.1874	0.2077	-0.0382	-0.1086	-0.1346	1.0000	-0.1989	0.7517	0.2373
0.4840	0.3298	0.1405	0.1784	0.1178	-0.1989	1.0000	0.3265	0.7834
0.3802	0.2025	-0.0586	-0.1929	-0.1276	0.7517	0.3265	1.0000	0.7712
0.4739	0.1690	0.0272	-0.1003	-0.0598	0.2373	0.7834	0.7712	1.0000

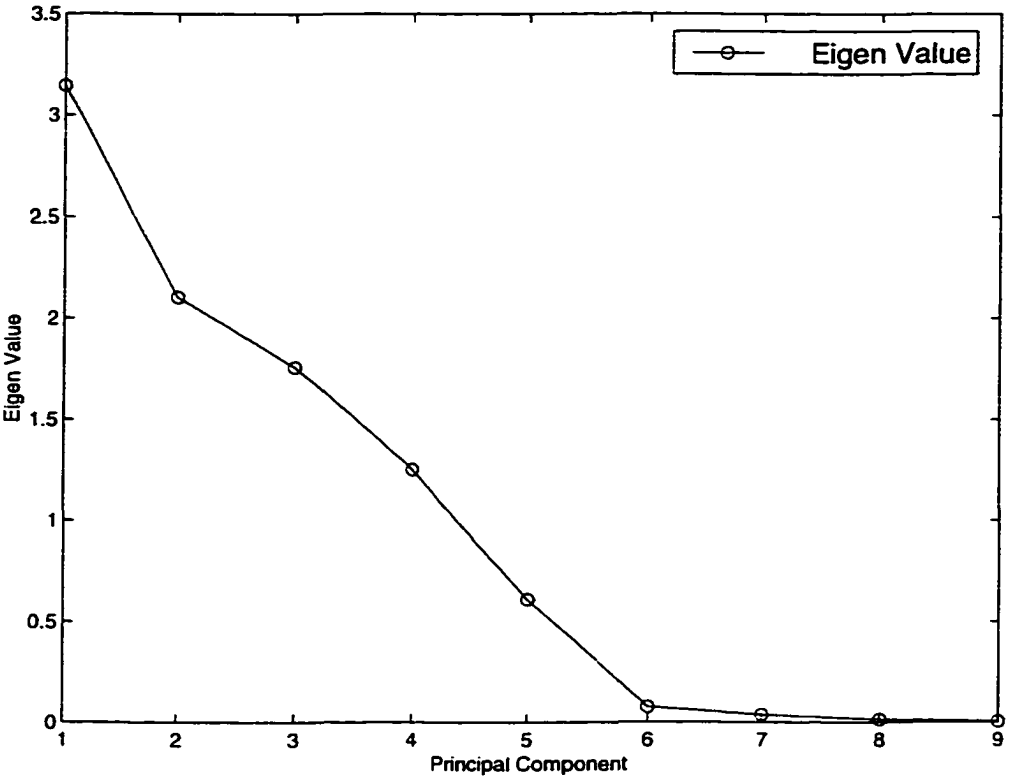


Figure 6.1 Scree graph for eigenvalues of AHU data

Table 6.3 PCA summary

Component Identity	Eigenvalue	Proportion of Variance	Cumulative Proportion	Standard Error
PC 1 (9)	3.142771	0.349197	0.349197	0.43793
PC 2 (8)	2.103960	0.233773	0.582970	0.29318
PC 3 (7)	1.758355	0.195373	0.778343	0.24502
PC 4 (6)	1.251093	0.139010	0.917353	0.17434
PC 5 (5)	0.610136	0.067793	0.985146	0.08502
PC 6 (4)	0.076006	0.008445	0.993591	0.01059
PC 7 (3)	0.038298	0.004255	0.997847	0.00534
PC 8 (1)	0.014020	0.001558	0.999404	0.00195
PC 9 (2)	0.005360	0.000596	1.000000	0.00075

Table 6.4 PCA test statistics

Eigenvalue	Components	Test Statistics u	df	$\chi^2_{0.05,df}$
3.14277	9	1291.09	44	60.46
2.10396	8	1158.66	35	49.11
1.75836	7	1048.22	27	40.11
1.25109	6	883.34	20	31.41
0.61014	5	618.68	14	23.68
0.07601	4	170.91	9	16.92
0.03830	3	88.73	5	11.07
0.01402	2	21.86	2	5.99
0.00536	1	0.00	0	0.00

(1995) found that 82% of the variance is high enough for most descriptive purposes. If on the other hand we kept 5 components, then 98% of the variance would be accounted for with the penalty of an added dimension. Kaiser's stopping rule recommends 4 components. Hence, 4 components agrees approximately with the scree plot decision. The error estimates for the eigenvalues seem stable in that the values are much smaller than the estimates of the eigenvalues.

The test of significance result is listed in Table 6.4. The test for the components assumed multivariate normality, which is not required for the estimation of the principal components. To test the significance of the larger components, the hypothesis test is formed that the last k population eigenvalues are small and equal, $H_{o_k} : \gamma_{p-k+1} = \gamma_{p-k+2} = \ldots = \gamma_p$, where $\gamma_1, \gamma_2, \ldots, \gamma_p$ denote the population eigenvalues, namely the eigenvalues of the variance covariance matrix, Σ. The implication is that the first sample components capture all the essential dimensions, while the last components reflect noise. If H_o is true, the last k sample eigenvalues will tend to have the pattern shown by the straight line in the scree graph.

The test indicates that only the last population eigenvalues are equal and we should retain the first eight. This differs from the results of scree, a priori criterion, and percentage of the total variance decision schemes. Rencher (1995) pointed out that the test of significance has a major disadvantage in that the method tends to retain more principal components than are useful. Since the first three component selection criteria are in close agreement, four components should be retained for this study.

Tables 6.5 and 6.6 lists eigenvectors and error estimates for the normal operation. Tables 6.7, and 6.8 lists eigenvectors and error estimates for the AHU data. For all the eigenvectors, all the standard errors are small and they seem stable. Since the data points are standardized, all coefficients must be between -1 and 1; therefore, a standard error of 0.5 or more automatically means that the corresponding coefficients could be practically anything. This means that it is impossible to attach any interpretation to a principal component whose components have large standard errors. But, since the standard estimates of the coefficients are small, we may be able to attach interpretation to the principal components obtained for the AHU data.

Interpretation of principal components

In the covariance or correlation matrices, Tables 6.1 and 6.2, a distinguishing pattern may be recognized from which the structure of the principal components can be deduced. The variables CFM, RPM, and Pow all had the largest variances amongst the AHU fault variables. Since, the static pressure will vary directly as the square of the CFM, the RPM varies directly with the CFM, the static pressure varies as the square of the RPM, and the fan power varies as the cube of the RPM, the relative positions among the ΔP, CFM, RPM, and Pow showed high correlations. These variable should account for most of the first principal component.

The variables between inlet and outlet air temperature show high positive correlations mainly due to the heating of the air. Another groupings of the high correlation occurred between Qwater and Twin, and between air side temperatures. Because the increase in the water flow rate, while holding everything else constant, would reduce the inlet and outlet water temperature differences, the inlet hot water temperatures and water flow rate show inverse correlations.

Hence there were total of three groupings with high correlation that can be visually detected in the correlation matrix. A case in which a component will duplicate a variable occurs when the variable is uncorrelated with other variables. The variable, Twout, seem to show this behavior if the correlation between the Twin is considered small. If the correlations are all small, the principal components will largely duplicate the variables. When all elements

Table 6.5 PCA eigenvectors and error estimates for normal operation

Variable	PC 1	PC 2	PC 3	PC 4	PC 5	PC 6	PC 7	PC 8	PC 9
Tain	1.3708E-01	5.0052E-01	3.7204E-01	-6.1187E-01	-1.0440E-01	3.0203E-01	5.2586E-02	9.8998E-02	3.2126E-01
Taout	9.0050E-03	6.3526E-02	7.7850E-02	-4.6929E-01	1.4895E-01	-1.6372E-01	-1.6126E-01	-9.9117E-02	-8.2751E-01
Twin	7.2331E-02	-5.4691E-01	-2.3166E-01	-2.3358E-01	-6.2410E-03	7.5013E-01	-1.2393E-01	7.1979E-02	-6.4526E-02
Twout	2.0120E-02	-4.5212E-01	-1.8237E-01	-5.5752E-01	1.6988E-01	-4.9809E-01	1.0720E-01	-1.1503E-01	3.8654E-01
Qwater	-1.3261E-02	-6.9400E-02	-2.1831E-02	-3.3942E-02	3.4933E-02	-1.5282E-01	8.8503E-02	9.7652E-01	-8.5967E-02
ΔP	-2.1620E-01	3.6553E-01	-7.8577E-01	-1.7146E-01	-3.5909E-01	1.7313E-02	2.0060E-01	-8.1408E-03	-5.7155E-02
CFM	5.8045E-01	-1.9477E-01	9.3277E-02	1.0089E-02	-7.5665E-01	-1.8211E-01	-1.3717E-02	-1.2652E-02	-1.0156E-01
RPM	3.1808E-02	1.3690E-01	-1.9133E-01	-1.5408E-02	-1.7480E-02	-1.1466E-01	-9.4038E-01	9.0300E-02	1.9358E-01
Pow	7.6853E-01	2.1636E-01	-3.2464E-01	9.5409E-02	4.8480E-01	3.5009E-02	1.0861E-01	1.1894E-04	-5.1754E-04

Table 6.6 Standard error on eigenvectors for normal operation

Variable	PC 1	PC 2	PC 3	PC 4	PC 5	PC 6	PC 7	PC 8	PC 9
Tain	1.5742E-02	2.5041E-02	3.3911E-02	1.4340E-02	5.2479E-02	1.5779E-02	2.9811E-02	2.3457E-02	9.6316E-03
Taout	3.4104E-03	7.7667E-03	1.0187E-02	1.4248E-02	4.0356E-02	2.4426E-02	3.0893E-02	5.1207E-02	9.0294E-03
Twin	1.5965E-02	1.5389E-02	3.5233E-02	1.7217E-02	3.0795E-02	1.2161E-02	6.3725E-02	3.0143E-02	1.6472E-02
Twout	1.3441E-02	1.3630E-02	3.0764E-02	1.8817E-02	4.9422E-02	1.7935E-02	4.5418E-02	3.1317E-02	1.3302E-02
Qwater	2.2367E-03	2.7094E-03	5.5577E-03	1.0656E-02	1.4548E-02	3.7670E-02	7.3866E-02	1.0152E-02	5.8133E-02
ΔP	1.6843E-02	5.0274E-02	2.4154E-02	3.3918E-02	1.8529E-02	2.1121E-02	8.8477E-03	1.6393E-02	7.2804E-03
CFM	6.4083E-03	1.8593E-02	1.8739E-02	6.3470E-02	6.4931E-03	2.3332E-02	2.1995E-02	1.3850E-02	7.6491E-03
RPM	5.1888E-03	1.2497E-02	9.6998E-03	1.4282E-02	1.9743E-02	7.9603E-02	1.3262E-02	7.1060E-02	2.9673E-02
Pow	8.4067E-03	3.0145E-02	2.0022E-02	4.1281E-02	1.0036E-02	1.7738E-02	1.0527E-02	1.0658E-02	5.5600E-03

Table 6.7 PCA eigenvectors

Variable	PC 1	PC 2	PC 3	PC 4	PC 5	PC 6	PC 7	PC 8	PC 9
Tain	0.48154	-0.16873	-0.13899	-0.14929	-0.47944	-0.08860	0.43223	-0.05975	-0.51956
Taout	0.37652	-0.36125	0.12438	0.29572	-0.46381	0.29810	-0.32695	-0.05381	0.45939
Twin	-0.20395	0.34081	0.55382	0.15268	-0.28689	-0.06190	0.03996	-0.64229	-0.10934
Twout	-0.05288	-0.23658	0.59805	0.42217	0.15278	0.11089	0.26868	0.48691	-0.24584
Qwater	0.17521	-0.57726	-0.00265	0.12487	0.53789	-0.16906	0.01670	-0.54448	-0.07633
ΔP	0.23508	0.30906	-0.23615	0.63316	0.01342	-0.56690	0.15710	0.05943	0.20689
CFM	0.36063	0.03036	0.44748	-0.40565	0.04921	-0.55437	-0.39528	0.19460	0.01914
RPM	0.41630	0.39065	-0.06875	0.22590	0.26412	0.38031	-0.43950	-0.03452	-0.45779
Pow	0.43890	0.29433	0.20025	-0.23802	0.29612	0.29197	0.50806	-0.06972	0.43366

Table 6.8 Standard error on eigenvectors

	PC 1	PC 2	PC 3	PC 4	PC 5	PC 6	PC 7	PC 8	PC 9
Tain	0.05629	0.14736	0.13701	0.09644	0.03941	0.06627	0.02979	0.06610	0.02025
Taout	0.09896	0.13198	0.22856	0.10660	0.05544	0.05281	0.04900	0.05757	0.01798
Twin	0.12650	0.30986	0.19676	0.17753	0.06097	0.03863	0.06250	0.01401	0.06356
Twout	0.12383	0.33785	0.17869	0.17934	0.08057	0.04963	0.05113	0.03629	0.04971
Qwater	0.14440	0.06342	0.32354	0.13156	0.04844	0.03753	0.05892	0.01512	0.05424
ΔP	0.10772	0.18512	0.25239	0.09382	0.09431	0.02962	0.08161	0.03947	0.01887
CFM	0.08671	0.27183	0.13332	0.13544	0.07440	0.05870	0.08098	0.04747	0.03024
RPM	0.10010	0.11836	0.23589	0.09502	0.05190	0.06523	0.05818	0.06463	0.02227
Pow	0.08480	0.16150	0.19220	0.10078	0.05209	0.07462	0.04701	0.06623	0.02461

of the first principal component eigenvectors are positive, the first component is a weighed average of the variables. It is sometimes referred to as a measure of size. Likewise, the positive and negative coefficients in subsequent components may be regarded as defining shape. And since the principal component coefficients have differing signs for the AHU data, the implication may be that of the shape or grouping indicators for the faults in the AHU. The first principal component is a (roughly) weighted sum, or index of the AHU components. The second component represents a contrast between the fan performance (ΔP, CFM, RPM, Pow) and the heat exchanger performance (Tain, Taout Twout, and Qwater). The third principal component represents a contrast between the (Taout, Twin, Twout, CFM, and Pow) and (Tain, Qwater, ΔP, RPM). The corresponding principal component equations are given by equation 6.1. The coefficients are taken from Tables 6.7.

$$\begin{aligned}
PC1 &= 0.48Tain + 0.38Taout - 0.2Twin - 0.05Twout + 0.18Qwater + 0.24\Delta P \\
&\quad + 0.36CFM + 0.42RPM + 0.44Pow \\
PC2 &= -0.17Tain - 0.36Taout - 0.34Twin - 0.24Twout - 0.58Qwater + 0.31\Delta P \\
&\quad + 0.03CFM + 0.39RPM + 0.29Pow \\
PC3 &= -0.14Tain + 0.12Taout + 0.55Twin + 0.60Twout - 0.003Qwater - 0.24\Delta P \\
&\quad + 0.45CFM - 0.07RPM + 0.20Pow \\
PC4 &= -0.15Tain + 0.30Taout + 0.15Twin + 0.42Twout + 0.12Qwater + 0.63\Delta P \\
&\quad - 0.41CFM + 0.23RPM - 0.24Pow
\end{aligned} \tag{6.1}$$

Since the principal components represent a rotation of axes, the components $U_i = a_i'y$ and $U_j = a_j'y$ are orthogonal for $i \neq j$, that is $a_i'a_j' = 0$. This orthogonality is also confirmed by the fact that a_i' and a_j' are eigenvectors of the symmetric matrix S. Principal components have a secondary property of being uncorrelated in the sample. That is, the covariance of z_i and z_j is zero. This property for the AHU data is shown in Table 6.9.

Discriminant functions and canonical variates, on the other hand, have the weaker property of being uncorrelated but not the stronger property of orthogonality. Thus in a plot of the first two discriminant functions or canonical variates on perpendicular coordinate axes, there is some distortion of their true relationship because the actual angle between their axes is not

Table 6.9 PCA covariance structure for AHU

Component	PC 1	PC 2	PC 3	PC 4	PC 5	PC 6	PC 7	PC 8	PC 9
PC 1	0.0140	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
PC 2	0.0000	0.0054	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
PC 3	0.0000	0.0000	0.0383	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
PC 4	0.0000	0.0000	0.0000	0.0760	0.0000	0.0000	0.0000	0.0000	0.0000
PC 5	0.0000	0.0000	0.0000	0.0000	0.6101	0.0000	0.0000	0.0000	0.0000
PC 6	0.0000	0.0000	0.0000	0.0000	0.0000	1.2511	0.0000	0.0000	0.0000
PC 7	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	1.7584	0.0000	0.0000
PC 8	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	2.1039	0.0000
PC 9	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	3.1428

90°. PCs are not scale invariant because if the scale on one or more of the y's change, the shape of the data group will be changing and will require different components to represent the new points.

Discussion

Essentially, PCA is a one-sample technique applied to data with no groupings among the observations and no partitioning of the variables into subsets. PCA are concerned only with the core structure of a single sample of observations on p variables. None of the variables is designated as dependent, and no grouping of observation is assumed. PCA searches for a dimension along which the observation is maximally separated or spread out. PCA also provides some interpretation and gives some useful information to be used as an input to another analysis.

Let U_1, U_2 and U_3 denote the values of the first, second, and third sample principal components of AHU for the nine variables. Figure 6.2 shows a scatter plot of U_1 vs. U_2 using the normal operation data set. This graph is to be viewed like a residual plot in regression. If the one dimensional principal component approximation is self consistent, then $E[U_2|U_1] = 0$. Figure 6.2 supports this assumption of self consistency because at for the whole range of U_1, all the points have scattered about the zero mean of U_2. Likewise, Figures 6.3 and 6.4 all have data points scattered about the zero mean of U_3.

The orthogonalities are observed by the PCA property that they are centered at the origin

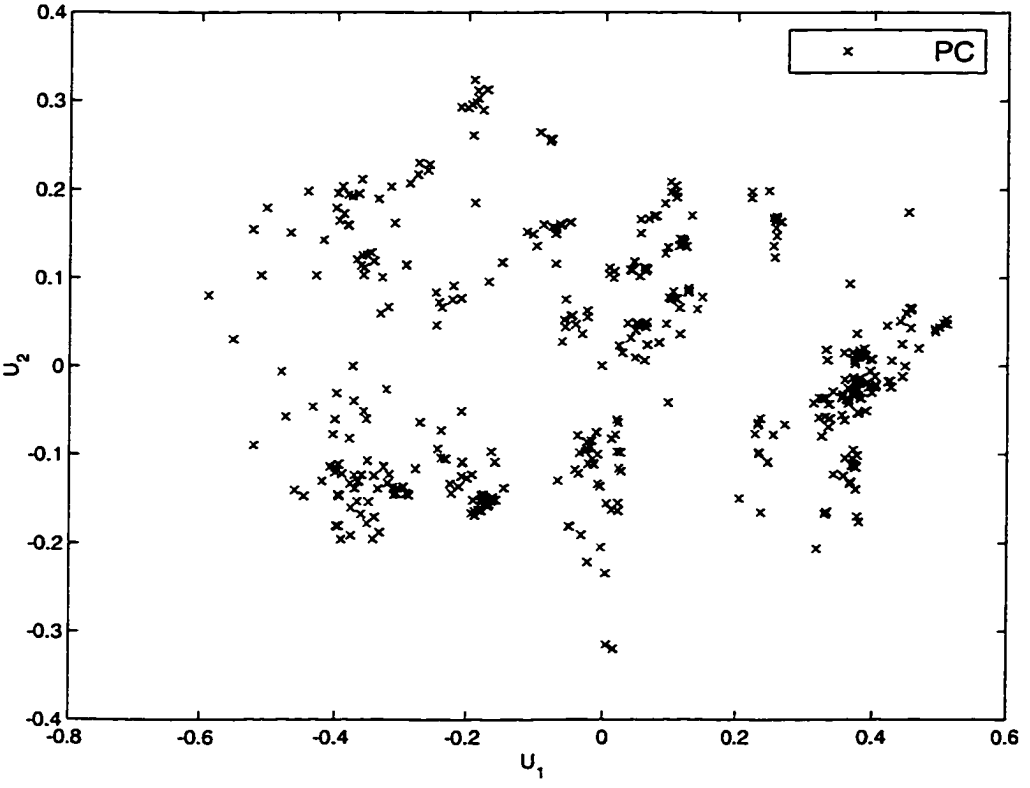


Figure 6.2 Scatter plot of principal components U_1 vs U_2 for AHU

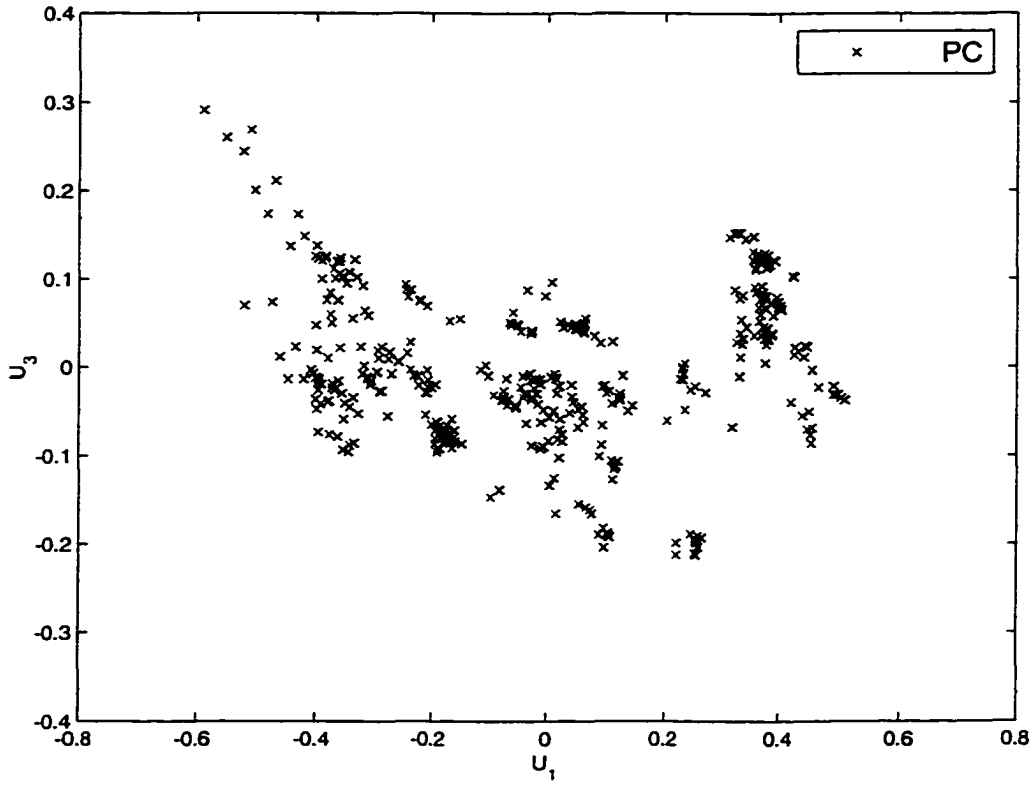


Figure 6.3 Scatter plot of principal components U_1 vs U_3 for AHU

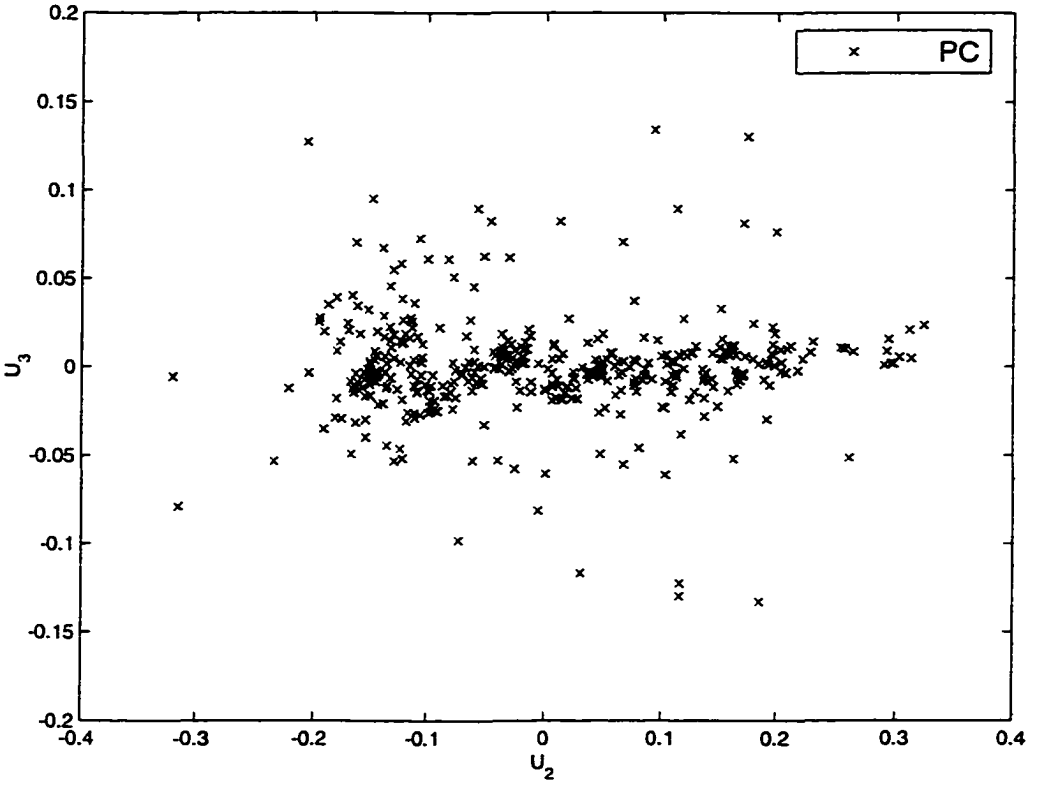


Figure 6.4 Scatter plot of principal components U_2 vs U_3 for AHU

and uncorrelated. The entries of the first principal component eigenvectors provide estimates of constants relating to normal operation. In similar plots, Figures 6.5, 6.6, and 6.7 show first three principal component combination of AHU data including fault groups. The components were extracted from the correlation matrix. All three scatter plots show the self consistency in that all the data points are clustered around the origin and scattered about each fault group centers. Unlike the normal operation plots, the plots now show distinct departure of the group clusters into the fault groups. From Figure 6.5, the first principal components indicate clear group separation of RPM fault and normal operation is observed. From Figure 6.7, the second principal components, fault group separation between valve and the normal operation is observed. In Figure 6.6, possible distinction can be observed between coil fault and the normal operation using the third principal component. Since the clusters form virtually on top of each other, the coil fault and normal modes of operation may be separated by the quadratic discriminant rule. Next section describes the functional form of the group separation.

Discriminant Analysis

In this section, discriminant analysis uses linear functions of the variables to describe the differences between two or more groups. Discriminant functions are linear combinations of variables that best separate groups. The objective of this study is to see the link between normal operation and each fault mode of operation. Nine measurements from the air handling unit were made for each test combination. There were 1150 observations for normal operation, 102 observations for the fan fault, 116 observations for the valve fault, and 114 observations for the coil fault group.

A stepwise discrimination procedure was used for selecting the subset of quantitative variables to produce a good discrimination model. The set of variables that make up each class is assumed to be multivariate normal with a common covariance matrix. Variables are chosen to add or remove from the model according to the squared partial correlation for predicting the variable under consideration from the group variables. The stepwise discrimination process was performed using the statistical analysis software (SAS) program with several different levels

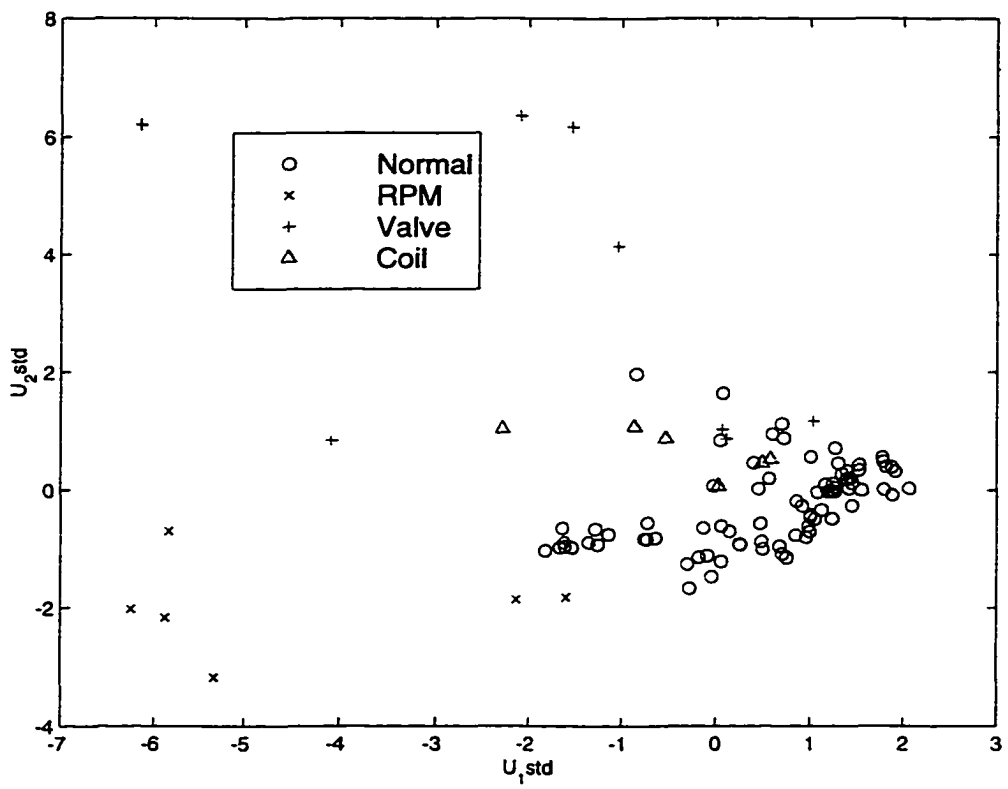


Figure 6.5 Scatter plot of principal components U_1 vs U_2 for AHU including fault groups

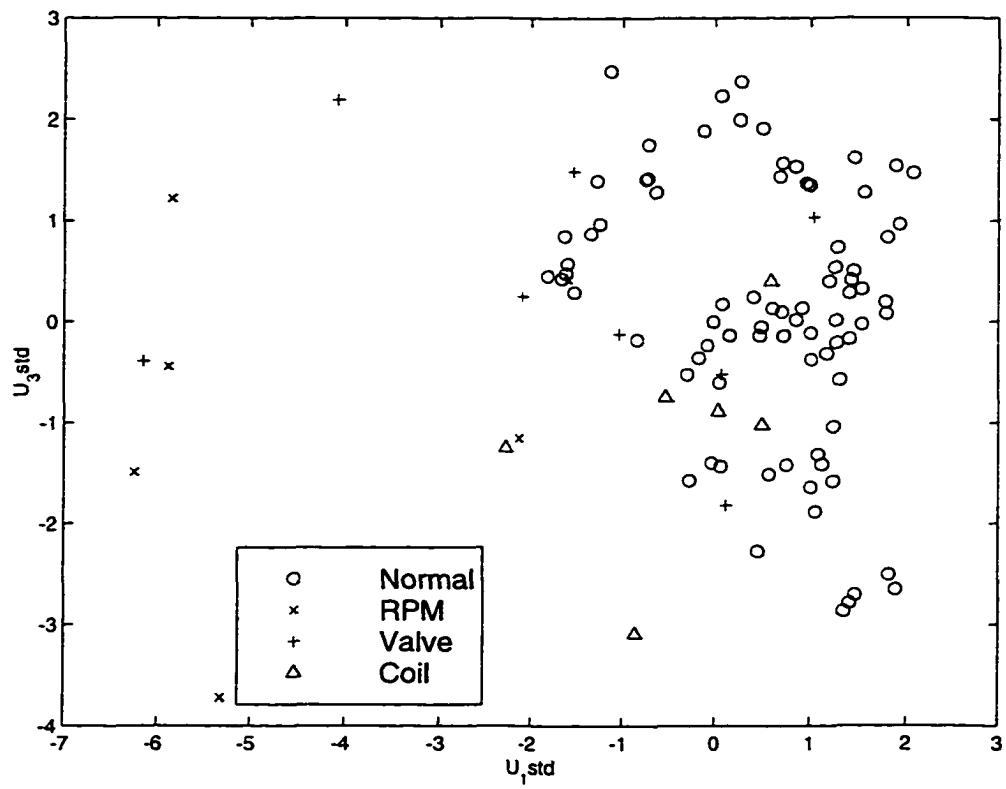


Figure 6.6 Scatter plot of principal components U_1 vs U_3 for AHU including fault groups

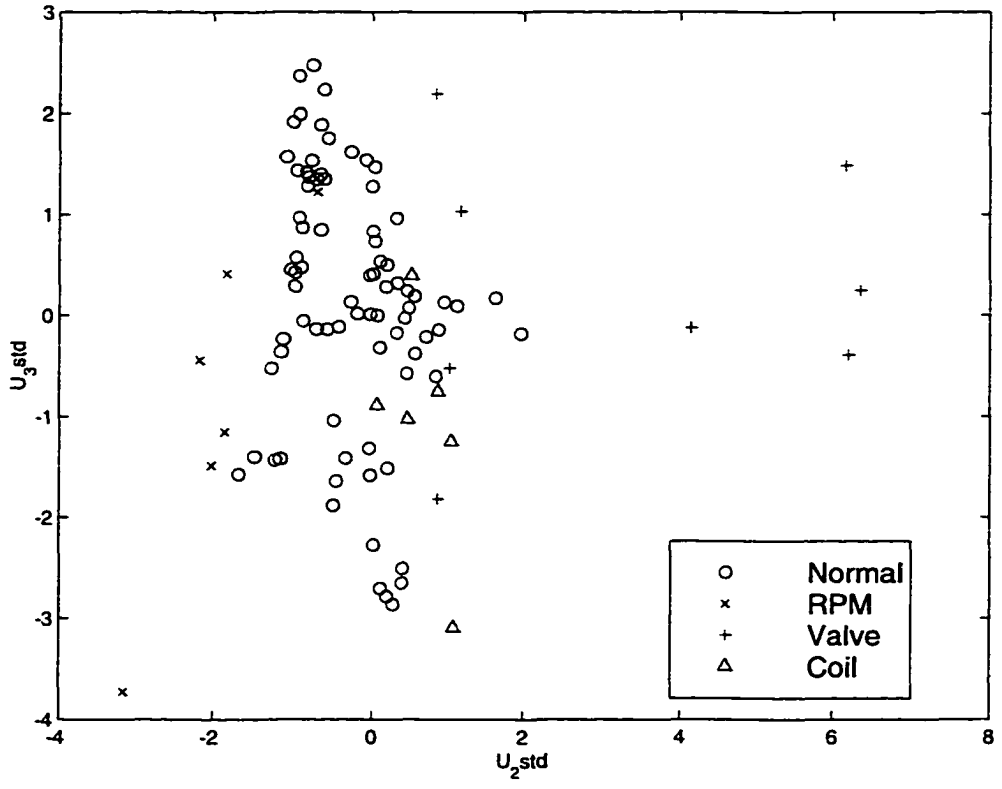


Figure 6.7 Scatter plot of principal components U_2 vs U_3 for AHU including fault groups

Table 6.10 Stepwise discriminant selection summary

Step	Variable Entered	Removed	Number In	Partial R^2	F Statistic	Prob > F
1	Qwater		1	0.6511	61.582	0.0001
2	RPM		2	0.5777	44.681	0.0001
3	Taout		3	0.3012	13.938	0.0001
4	Twout		4	0.1354	5.013	0.0028
5	Twin		5	0.3713	18.704	0.0001
6	Tain		6	0.1518	5.607	0.0014
7	ΔP		7	0.4169	22.163	0.0001

Step	Variable Entered	Removed	Number In	Wilks' Lambda	Prob < Lambda	Average Squared Canonical Correlation	Prob > ASCC
1	Qwater		1	0.3489	0.0001	0.2170	0.0001
2	RPM		2	0.1474	0.0001	0.4095	0.0001
3	Taout		3	0.1030	0.0001	0.4968	0.0001
4	Twout		4	0.0890	0.0001	0.5146	0.0001
5	Twin		5	0.0560	0.0001	0.5596	0.0001
6	Tain		6	0.0475	0.0001	0.5951	0.0001
7	ΔP		7	0.0277	0.0001	0.6845	0.0001

of cutoff values. For a squared partial correlation cutoff value of 0.15, three most significant variables (hot water flow rate, RPM, and outlet air temperature) for the model were retained. A squared partial correlation cutoff value of 0.1 retained 7 most significant variables (hot water flow rate, RPM, outlet air temperature, outlet water temperature, inlet water temperature, inlet air temperature, and pressure rise across AHU). Any value smaller than 0.01 retained all variables with the exception of the outlet air temperature. The resulting stepwise variable selection level of $R^2 = 0.1$ is listed in Table 6.10. To see if the minimal sensor measures can discriminate the fault groups, stepwise variable selection level of $R^2 = 0.1$ was performed using the variables, Tain, Taout, Twin, Twout, and ΔP. The resulting selection of the variables is listed in Table 6.11.

The discriminant function coefficients for the 7 variables are listed in Table 6.12, and the

Table 6.11 Stepwise discriminant selection summary for reduced set

Step	Variable Entered	Removed	Number In	Partial R^2	F Statistic	Prob > F
1	Taout		1	0.4439	26.347	0.0001
2	ΔP		2	0.2936	13.579	0.0001
3	Twin		3	0.3251	15.572	0.0001
4	Tain		4	0.2384	10.019	0.0001
5	Twout		5	0.1693	6.455	0.0005
6		Taout	4	0.0197	0.637	0.5929

Step	Variable Entered	Removed	Number In	Wilks' Lambda	Prob < Lambda	Average Squared Canonical Correlation	Prob > ASCC
1	Taout		1	0.5561	0.0001	0.1480	0.0001
2	ΔP		2	0.3928	0.0001	0.2394	0.0001
3	Twin		3	0.2651	0.0001	0.3312	0.0001
4	Tain		4	0.2019	0.0001	0.3729	0.0001
5	Twout		5	0.1677	0.0001	0.4061	0.0001
6		Taout	4	0.1711	0.0001	0.4024	0.0001

Table 6.12 Discriminant function coefficients for 7 variables

	Disc 1	Disc 2	Disc 3
Tain	0.538971602	-1.642389602	3.196556208
Taout	0.167080562	1.959361185	-1.979086109
Twin	4.039812379	1.045401565	0.564486492
Twout	-3.430863625	-1.528763705	1.700456965
Qwater	5.353430243	0.662158944	-0.414712388
ΔP	0.671296200	-0.750393696	1.725011343
RPM	-0.600076586	2.247460913	-1.684691200

first discriminant function is given by equation 6.2.

$$Z_1 = 0.53Tain + 0.17Taout + 4.04Twin - 3.43Twout + 5.35Qwater + 0.67\Delta P - 0.60RPM \tag{6.2}$$

The scatter plot of the resulting discriminant scores are shown in Figures 6.8, 6.9, and 6.10. Figure 6.8 shows a scatterplot of discriminant function Z_1 vs Z_2; the good separation of the normal and valve is clearly visible. Along the Z_2 there is some indication of the group separation among normal, fan, valve, and coil. The indication of the separation is more pronounced in Figure 6.9. Figure 6.9 shows clear separation of the groups between fan vs normal and between coil vs normal.

Table 6.13 lists the discriminant function coefficients for the 5 variables selected. The first discriminant function is given by equation 6.3. Figures 6.11, 6.12, and 6.13 show scatterplot of discriminant functions Z_1, Z_2, and Z_3 obtained from the reduced set of variables.

$$Z_1 = -1.49Tain - 0.37Taout + 1.05Twin - 0.41Twout + 0.96\Delta P \tag{6.3}$$

Using the result from the selection of the variables, The variable Taout was removed and the resulting discriminant function coefficients is listed in Table 6.14. and the first discriminant function is given by equation 6.4. Figures 6.14, 6.15, and 6.16 are similar to that of the figures using 5 variables. Thus, this agrees with the stepwise discriminant process. Hence, the redundant variable, Taout, is eliminated from the study. Next section discusses the classification of the AHU faults.

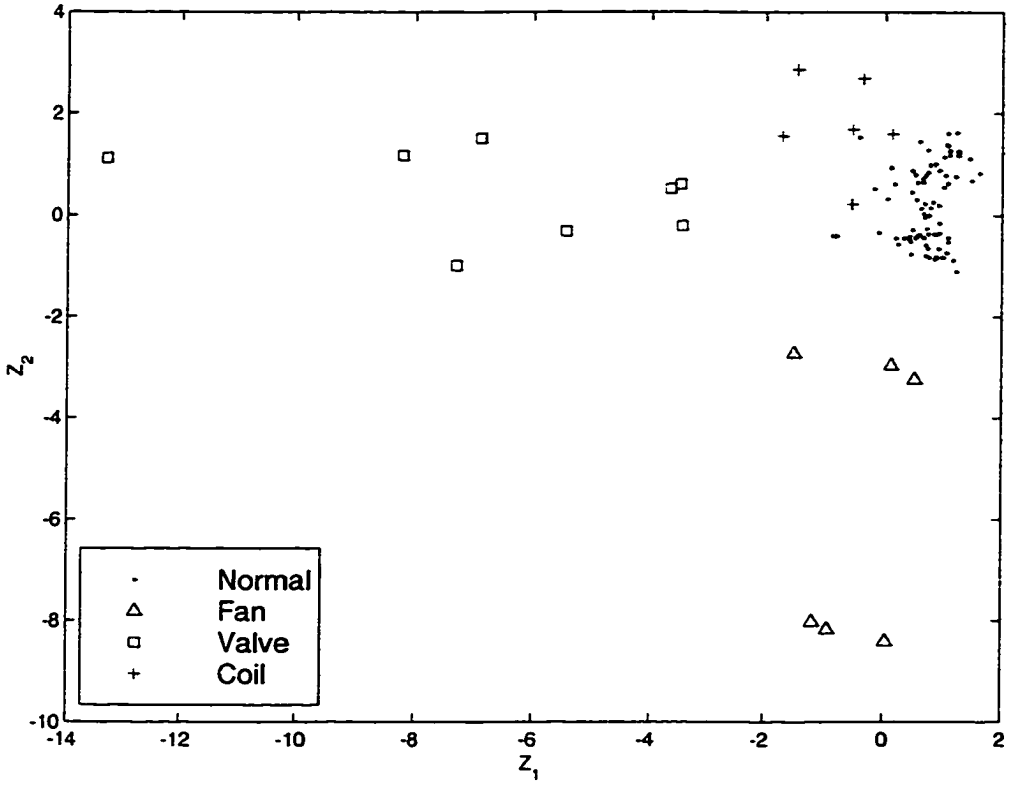


Figure 6.8 Scatter plot of the two discriminant functions Z_1 vs Z_2 for AHU train data

Table 6.13 Discriminant function coefficients for 5 variables

	Disc 1	Disc 2	Disc 3
Tain	1.491681905	0.818515226	1.120102395
Taout	-0.365701855	-0.357422160	-0.777460598
Twin	1.046217277	-1.256971023	0.530154690
Twout	0.405983470	1.382058786	0.740727778
ΔP	0.964467424	-0.055537747	-0.664834141

Table 6.14 Discriminant function coefficients for 4 variables

	Disc 1	Disc 2	Disc 3
Tain	1.173965903	0.561658558	0.527414857
Twin	1.136473488	-1.170703809	0.663211290
Twout	0.126402862	1.148543199	0.253742855
ΔP	0.923978552	-0.067856718	-0.770237153

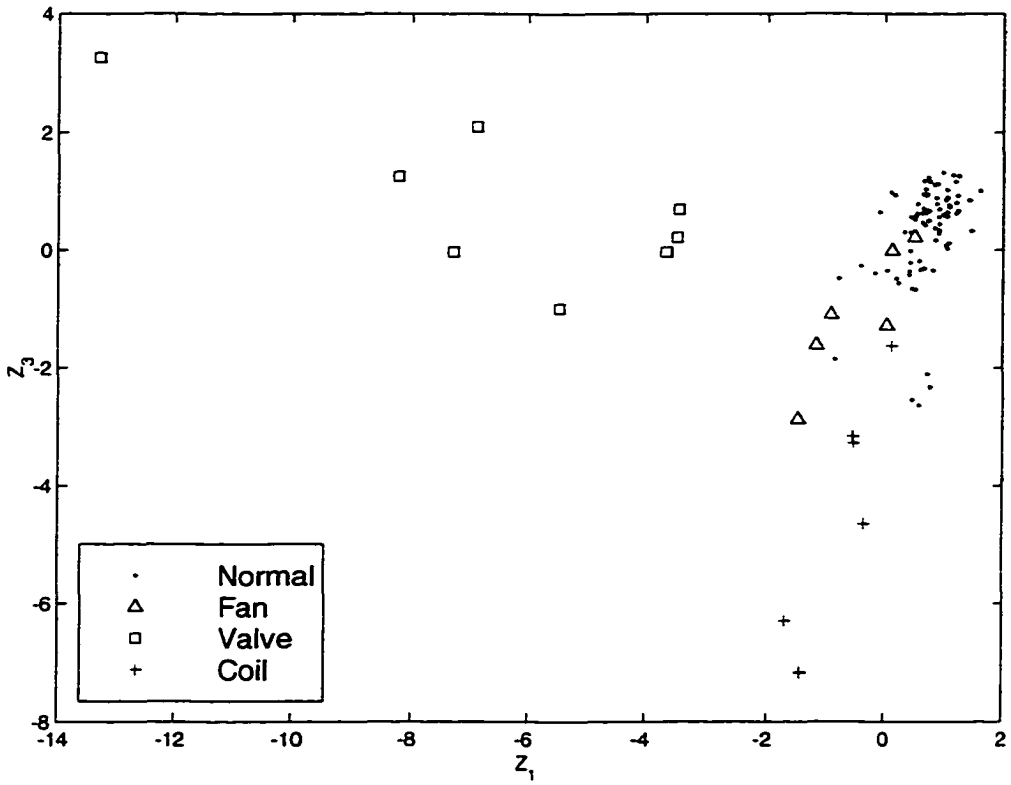


Figure 6.9 Scatter plot of the two discriminant functions Z_1 vs Z_3 for AHU train data

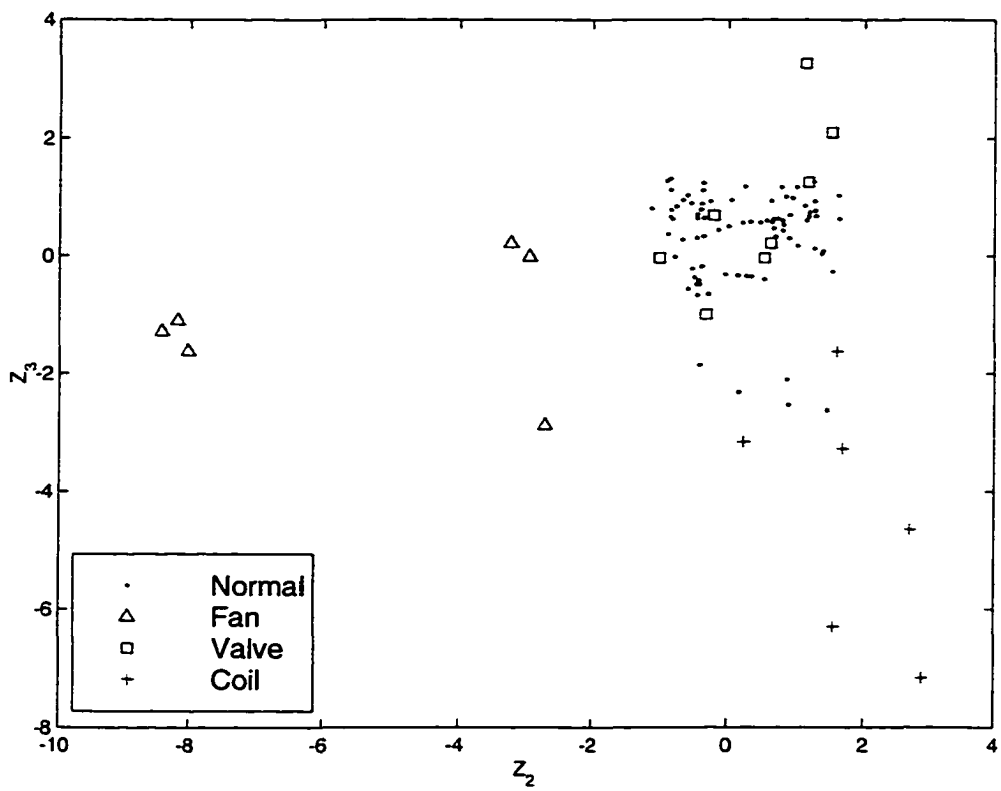


Figure 6.10 Scatter plot of the two discriminant functions Z_2 vs Z_3 for AHU train data

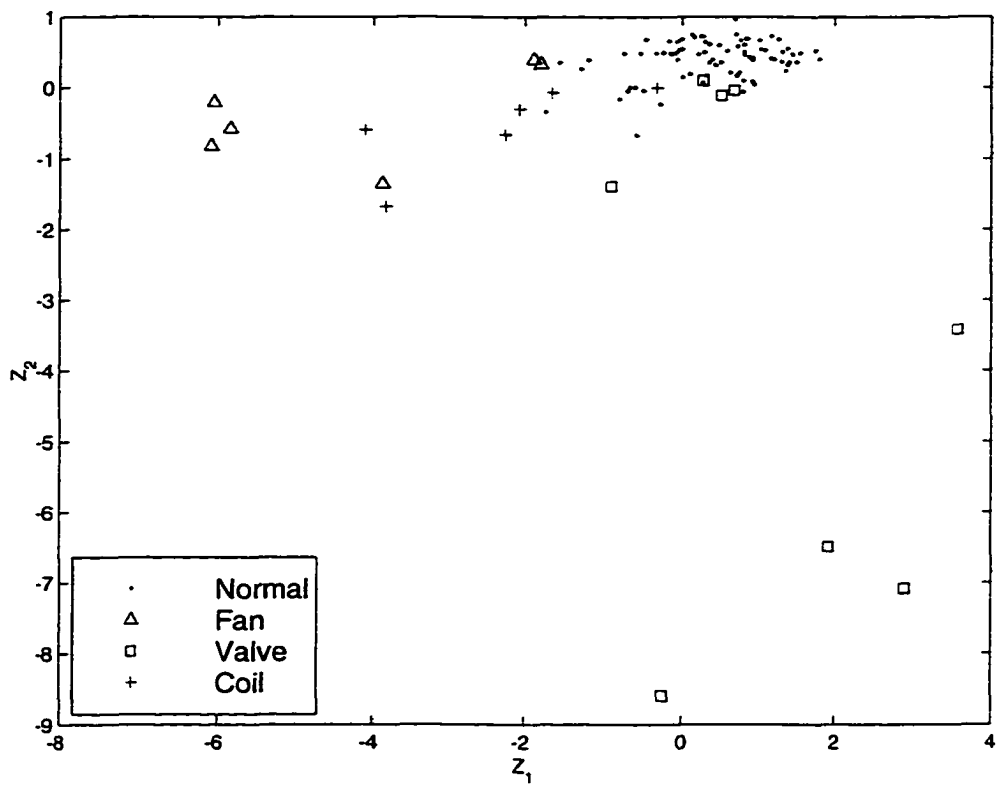


Figure 6.11 Scatter plot of the two discriminant functions Z_1 vs Z_2 for AHU reduced set

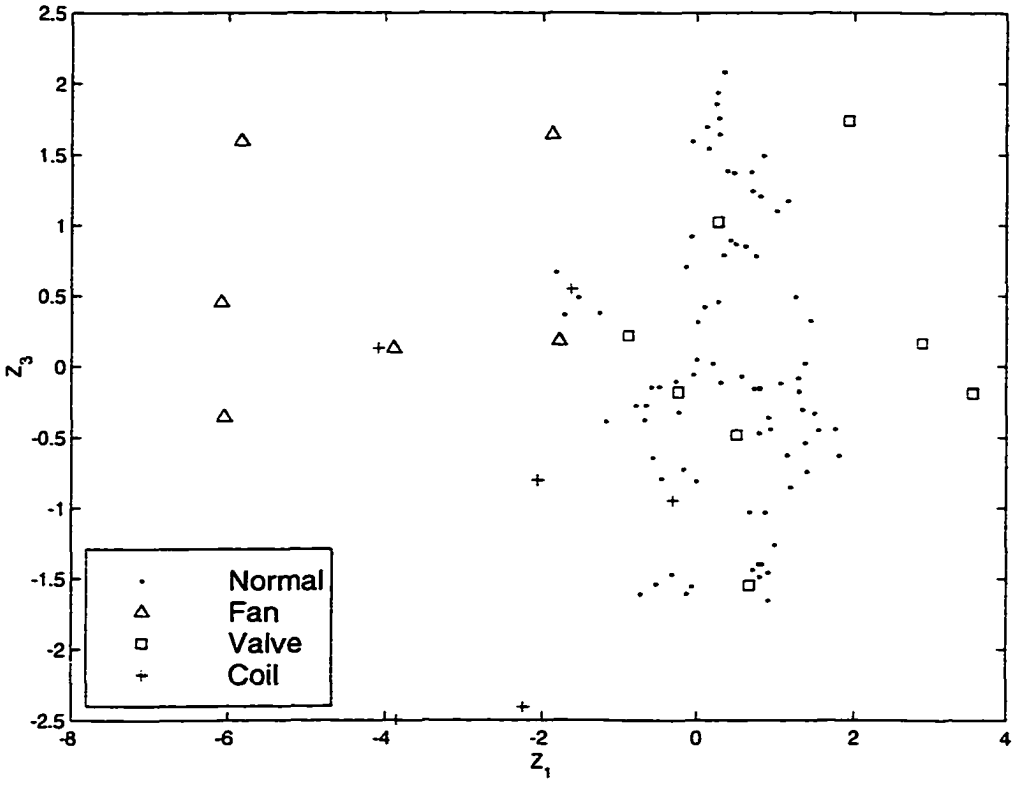


Figure 6.12 Scatter plot of the two discriminant functions Z_1 vs Z_3 for AHU reduced set

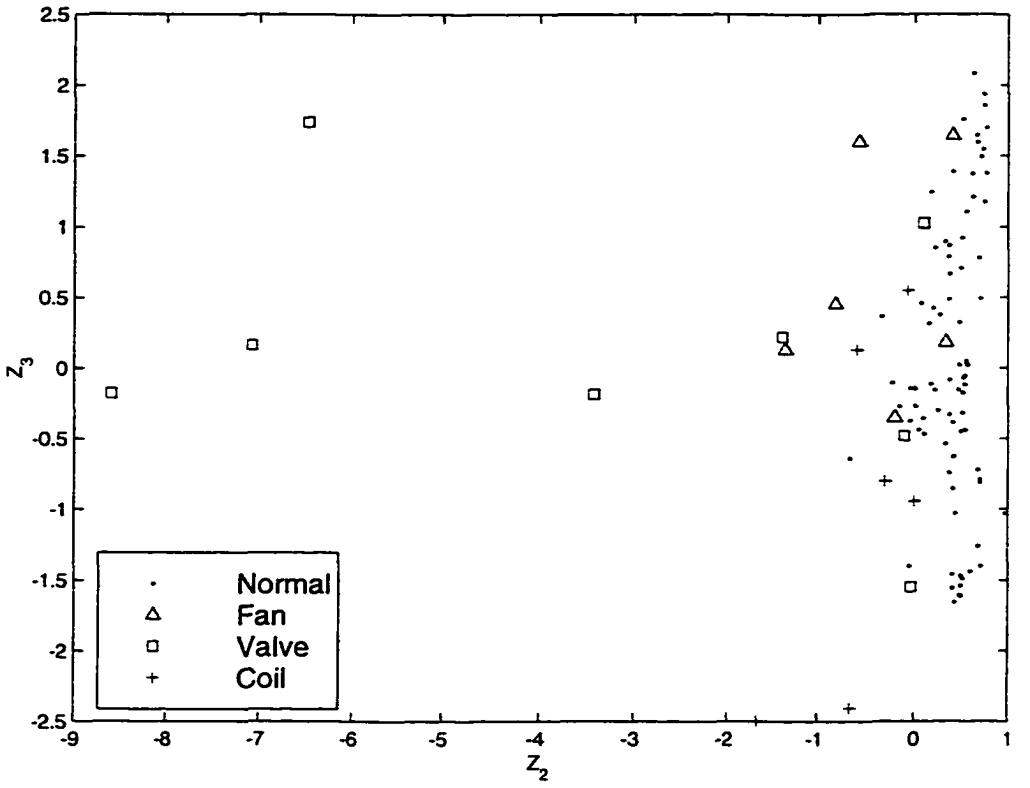


Figure 6.13 Scatter plot of the two discriminant functions Z_2 vs Z_3 for AHU reduced set

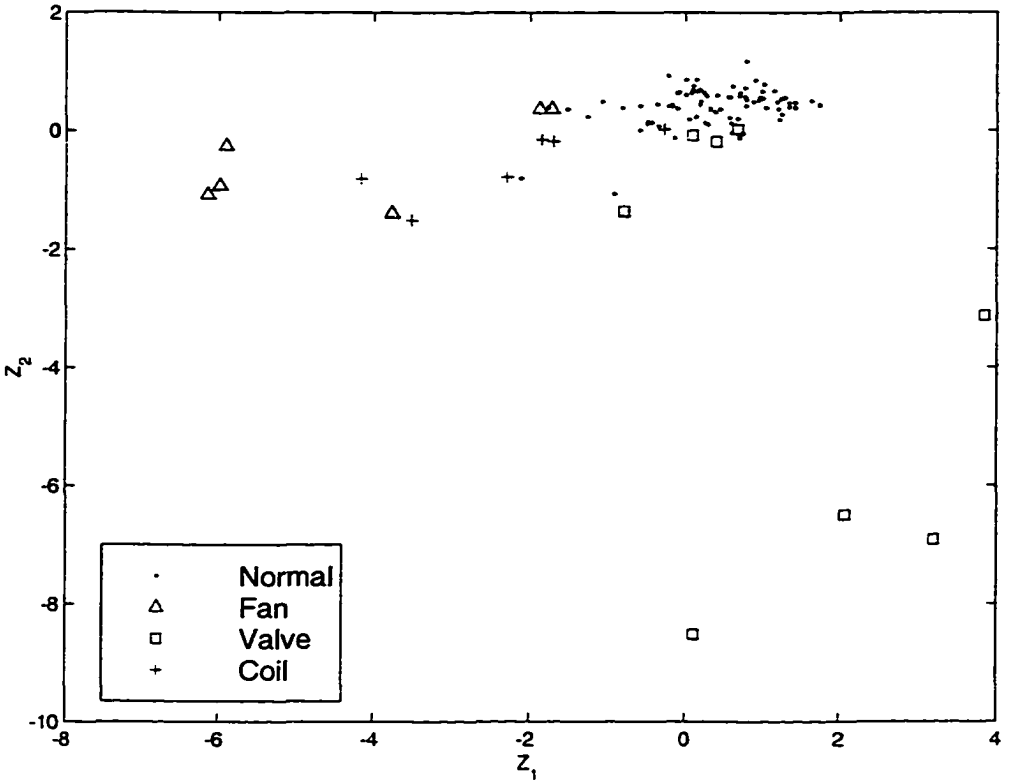


Figure 6.14 Scatter plot of the two discriminant functions Z_1 vs Z_2 for AHU train data

$$Z_1 = 1.17Tain + 1.14Twin + 0.13Twout + 0.92\Delta P \tag{6.4}$$

Classification Analysis

Classification procedures based on normal populations have been used in this research due to their simplicity and efficiency over other population models. A good classification procedure should result in low error rates. As one of the objectives for the research, the chances or probabilities of misclassification should be made small. To satisfy this objective, a rule that minimizes the chances of making mistakes is implemented. An assumption is made that normal operation has a higher tendency to occur than any of the faults. Furthermore, this research assumes all faults have an equal chances of occurring. If these assumptions hold true and

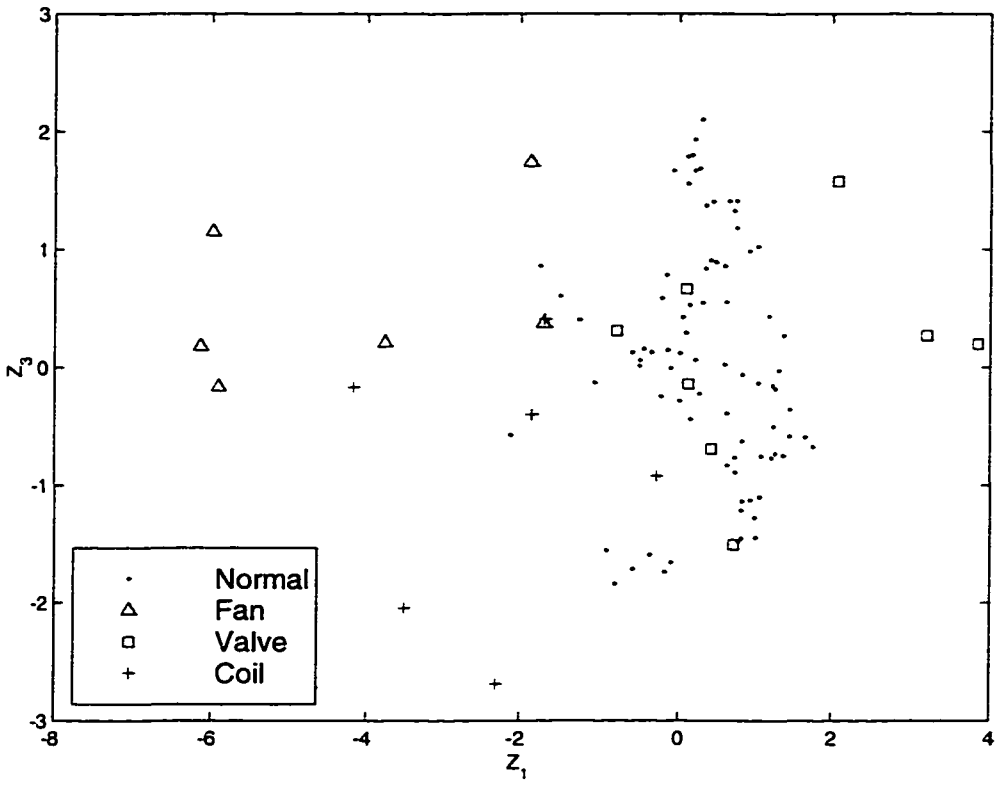


Figure 6.15 Scatter plot of the two discriminant functions Z_1 vs Z_3 for AHU train data

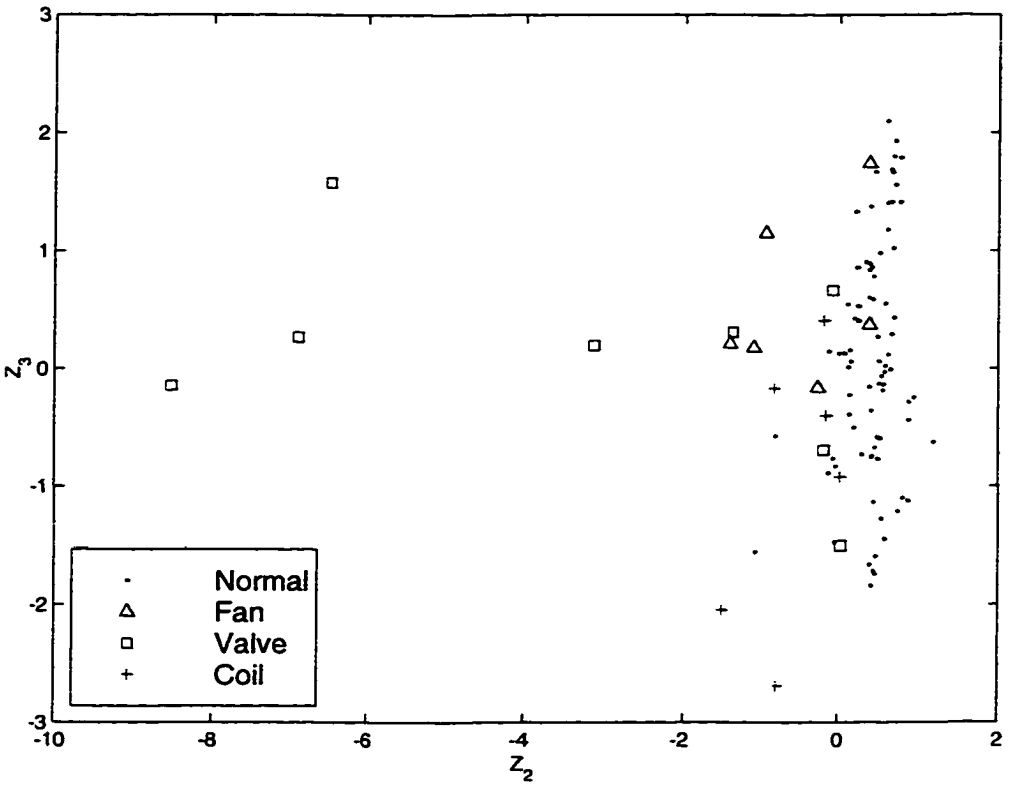


Figure 6.16 Scatter plot of the two discriminant functions Z_2 vs Z_3 for AHU train data

Table 6.15 Classification summary using linear classification rule for full set of variables

	Predicted				
Observed group	Normal	Fan	Valve	Coil	Total count
Normal	83	0	0	0	83
Fan	1	5	0	0	6
Valve	0	0	8	0	8
Coil	1	0	0	5	6
Total	85	5	8	5	103
Priors	0.8058	0.0583	0.0777	0.0583	
percent classified					
Normal	100.00	0.00	0.00	0.00	
Fan	16.67	83.33	0.00	0.00	
Valve	0.00	0.00	100.00	0.00	
Coil	16.67	0.00	0.00	83.33	

because there is an evidence that multivariate normality seem to hold, discussed in chapter 5, the prior probability could be estimated by the observations obtained from the experiments.

Table 6.15 lists the classification results using the full set of variables. Full set of 9 variables Number of Observations Classified into Y:

The classification error rate estimates for the full model resulted with 0% for the normal operation, 17% for the fan fault, 0% for the valve fault, and 17% for the Coil faults. For verification purposes, a set of test data apart from the data used for analysis was applied to the full model.

The test data set includes 144 normal observations, 23 fan fault, 33 valve faults, 12 coil faults. Figure 6.17 shows classification results using the linear discriminant function. Table 6.16 lists classification results from the corresponding discriminant function on the test set. While the error rate is zero for normal operation, fan, and coil faults, the error rate for classifying into valve fault was 39%. Using the quadratic classification, all the observation was classified into normal group. Overall correct classification resulted with 94%.

The classification summary for the reduced set of variables using Tain, Twin, Twout, and ΔP, is listed in Table 6.17. The error rate for the classification using linear discriminant

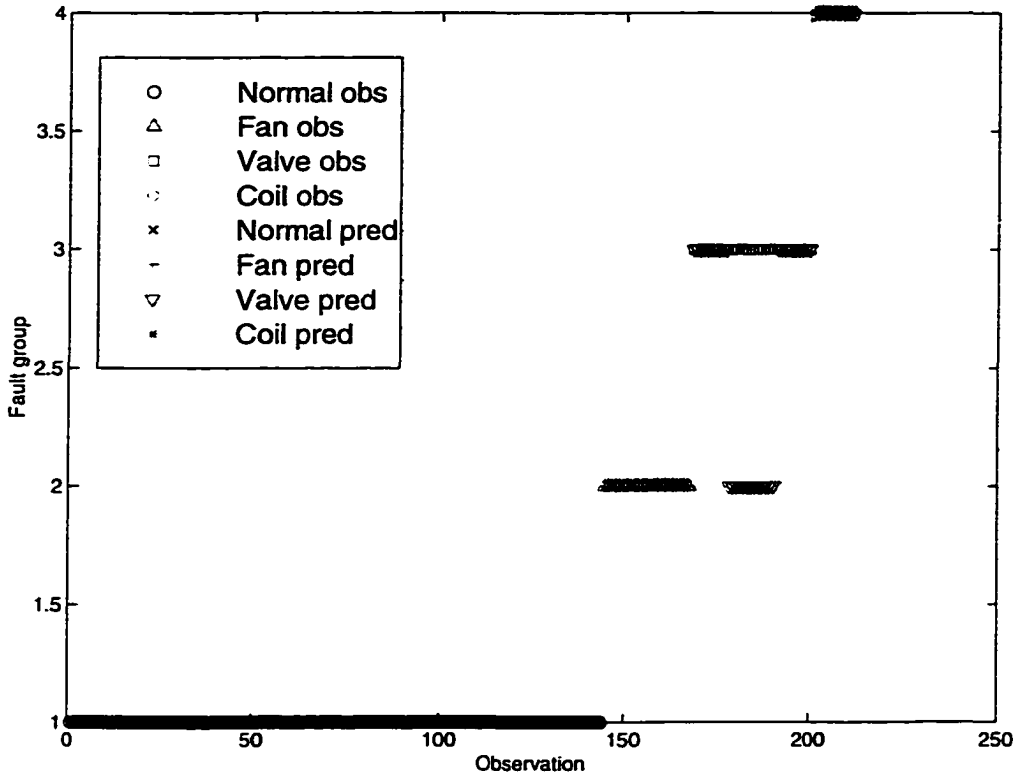


Figure 6.17 Classification using linear discriminant function on test data using full set of variables

Table 6.16 Classification summary using linear classification rule for full set of variables on the test data

	Predicted				
Observed group	Normal	Fan	Valve	Coil	Total count
Normal	144	0	0	0	144
Fan	0	23	0	0	23
Valve	0	13	20	0	33
Coil	0	0	0	12	12
Total	144	36	20	12	212
Priors	0.8058	0.0583	0.0777	0.0583	
percent classified					
Normal	100.00	0.00	0.00	0.00	
Fan	0.00	100.00	0.00	0.00	
Valve	0.00	39.39	61.61	0.00	
Coil	0.00	0.00	0.00	100.00	

Table 6.17 Classification summary using linear classification rule for reduced set of 4 variables on the training set

Observed group	Predicted Normal	Fan	Valve	Coil	Total count
Normal	82	0	0	1	83
Fan	2	4	0	0	6
Valve	4	0	4	0	8
Coil	3	1	0	2	6
Total	91	5	4	3	103
Priors	0.8058	0.0583	0.0777	0.0583	
percent classified					
Normal	98.80	0.00	0.00	1.20	
Fan	33.33	66.67	0.00	0.00	
Valve	50.00	0.00	50.00	0.00	
Coil	50.00	16.67	0.00	33.33	

function has increased to 1% for normal operation, 33% for the fan fault, 50% for the valve fault, and 67% for the coil faults. Overall correct classification resulted in 89%.

Figure 6.18 shows classification results of a reduced number of variables (from 9 to 4) using the test data set. Although the number of variables was reduced more than 1/2 the total number of variables, the classification error rate did not increase too drastically. Overall correct classification was at 83%. Table 6.18 lists the classification summary of test result using linear classification rule using 4 variables. There were no misclassification on the normal and fan faults. However, 70% and 100% classification error rate resulted for the valve and coil faults, respectively. Not too surprisingly, the quadratic discriminant rule showed better overall classification rate than that of linear rule. On the training data, the overall correct classification rate was 96%. The resulting error rates were 2%, 0%, 37%, and 16% for normal, fan, valve, and coil respectively. Figure 6.19 shows result of the quadratic classification on the test data. Table 6.19 lists the classification summary using the quadratic rule. The resulting error rates were 4%, 0%, 67%, and 0% for normal, fan, valve, and coil respectively.

Inclusion of the outlet air temperature to the reduced 4 variables did not show much improvement with the overall correct classification rate of 83% using linear classification function

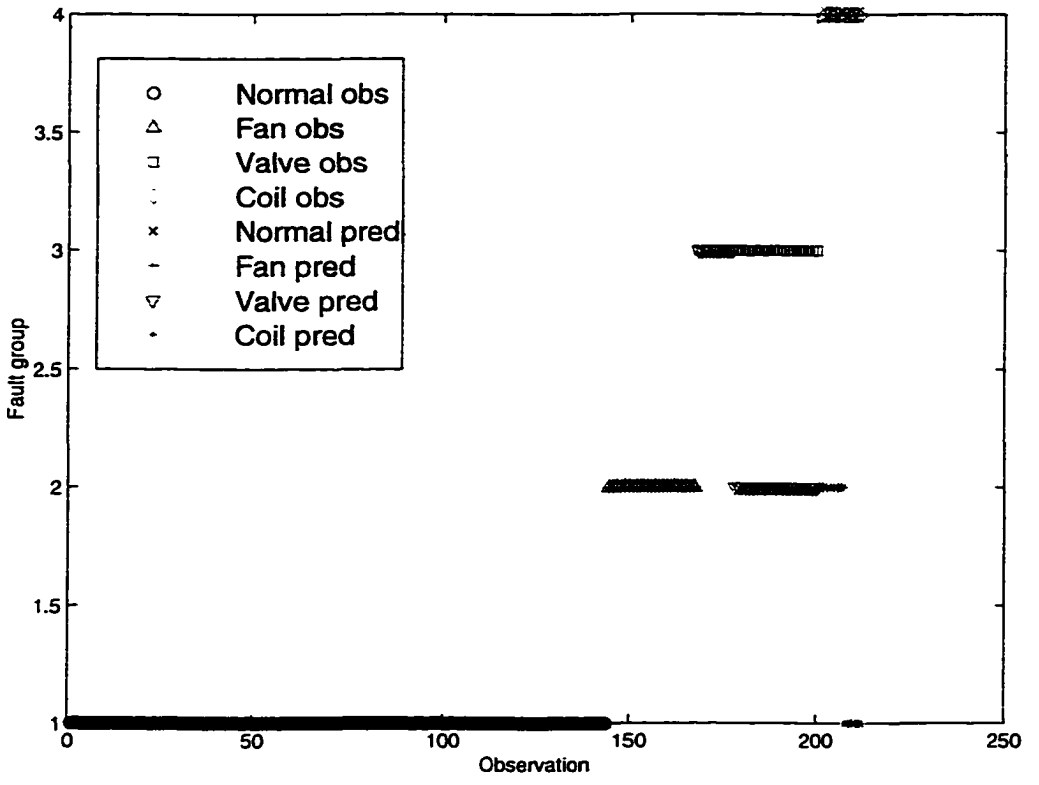


Figure 6.18 Classification on test data with 4 variables using linear discrimination function

Table 6.18 Classification summary using linear classification rule for reduced set of 4 variables on the test set

	Predicted				
Observed group	Normal	Fan	Valve	Coil	Total count
Normal	144	0	0	0	144
Fan	0	23	0	0	23
Valve	0	23	10	0	33
Coil	5	7	0	0	12
Total	149	53	10	0	212
Priors	0.8058	0.0583	0.0777	0.0583	
percent classified					
Normal	100.00	0.00	0.00	0.00	
Fan	0.00	100.00	0.00	0.00	
Valve	0.00	69.70	30.30	0.00	
Coil	41.67	58.33	0.00	0.00	

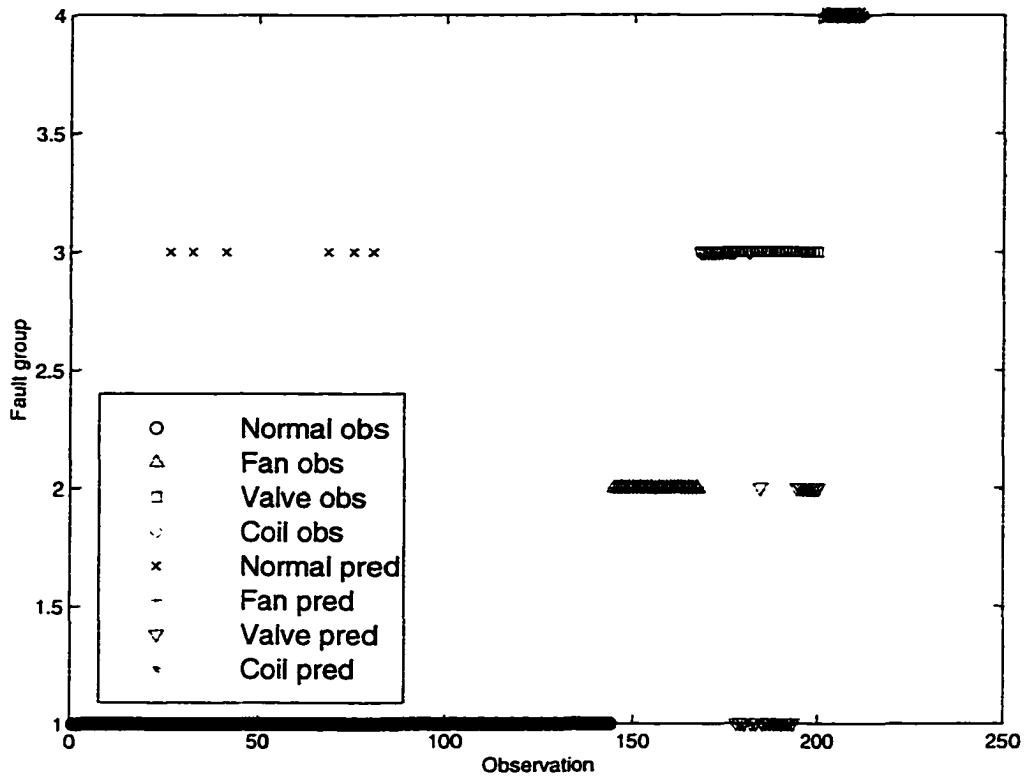


Figure 6.19 Classification on test data with 4 variables using quadratic discrimination function

Table 6.19 Classification summary using quadratic classification rule for reduced set of 4 variables on the test set

	Predicted				
Observed group	Normal	Fan	Valve	Coil	Total count
Normal	138	0	6	0	144
Fan	0	23	0	0	23
Valve	14	8	11	0	33
Coil	0	0	0	12	12
Total	152	31	17	12	212
Priors	0.8058	0.0583	0.0777	0.0583	
percent classified					
Normal	95.83	0.00	4.17	0.00	
Fan	0.00	100.00	0.00	0.00	
Valve	42.42	24.24	33.33	0.00	
Coil	0.00	0.00	0.00	100.00	

on test data. The resulting error rates were 0%, 0%, 70%, and 100% for normal, fan, valve, and coil faults respectively. Using the quadratic classification, the overall rate was 97% on the training set and 78% on the test set.

For the 7 variable combination using the linear classification rule on the training set, overall correct classification rate was 97% with 0%, 33%, 0%, and 17% error rates for normal, fan, valve, and coil respectively. On the test set, the overall correct classification was 94% with 0%, 0%, 39%, and 0% error rates for normal, fan, valve, and coil faults respectively. Using the quadratic rule on the training set, there were no misclassification. However, the overall correct classification rate dropped to 68% with 0%, 100%, 97%, and 100% error rates for the normal, fan, valve, and coil faults.

Ranking the detection of the fault by the high overall classification rate is listed in the Table 6.20. To achieve the minimum variable selection without causing too many false alarms, it is recommended to use the 4 variable combination quadratic discriminant function to classify three faults occurring in AHU.

Logistic regression

Logistic regression was used in this research to determine the effects of the AHU variables, predictors, on the occurrence and non occurrence of the fault conditions for the fan, valve, and heat exchanger coil. The regressor variables consist of 4 of the 9 total measured variables. They are inlet air temperature (Tain), outlet air temperature (Taout), hot water outlet temperature (Twout), and pressure rise (ΔP) across AHU. The regressors are listed in the order of their difficulty in measurability. The 4 variables, Qwater, CFM, RPM, and Pow, were omitted for the analysis to see the effect of the detection of the general faults. Moreover, their measurability surpasses that of the listed variables. Also, additional information increases the complexity of the analysis, i.e. more parameter estimates are required. The response variable Y is binary with 1 indicating occurrence and 0 indicating nonoccurrence of the faults. Assuming Y be a Bernoulli random variable, success probability, π, depends on the value of the AHU regressor variables.

Table 6.20 Ranking of the classification summary by correct classification rate

Variable Combination	Overall Correct Classification	Classification Type	Data	Error Rate Normal	Error Rate Fan	Error Rate Valve	Error Rate Coil
Full set*	100	Quadratic	Train	0	0	0	0
7 variable†	100	Quadratic	Train	0	0	0	0
Full set*	98	Linear	Train	0	17	0	17
7 variable†	97	Linear	Train	0	33	0	17
5 variable‡	96	Quadratic	Train	0	0	37	17
4 variable††	94	Quadratic	Train	2	0	37	17
5 variable‡	91	Linear	Train	0	33	50	50
4 variable††	89	Linear	Train	1	33	50	67
Full set*	97	Linear	Test	0	0	39	0
7 variable†	94	Linear	Test	0	0	39	0
4 variable††	87	Quadratic	Test	4	0	67	0
4 variable††	84	Linear	Test	0	0	70	100
5 variable‡	84	Linear	Test	0	0	70	100
5 variable‡	78	Quadratic	Test	14	0	64	50
7 variable†	68	Quadratic	Test	0	100	97	100
Full set*	68	Quadratic	Test	0	100	100	100

* Tain, Taout, Twin, Twout, Qwater, ΔP, CFM, RPM, POW

† Tain, Taout, Twin, Twout, Qwater, ΔP, RPM

‡ Tain, Taout, Twin, Twout, ΔP

†† Tain, Twin, Twout, ΔP

Table 6.21 lists the first 59 variable combinations that have been sorted out by their smallest deviance. The variable assignments in the table are $G1 = Tain$, $G2 = Taout$, $G3 = Twin$, $G4 = Twout$, $G5 = Qwater$, $G6 = \Delta P$, $G7 = CFM$, $G8 = RPM$, $G9 = Pow$. There were $\binom{9}{4} = 126$ total combinations of 9 variables taken 4 at a time. The 4 variable combination criteria was chosen in accordance with the result from the discriminant stepwise variable selection procedure.

The best 4 variable combination resulted in the variable combination of Twin, ΔP, CFM, and RPM. The associated multiple logistic regression model is given in equation 6.5, and the estimated coefficients for the fitted model equation are listed in Table 6.22. However, since some of these variables were omitted, the next best model that includes the selected variables is given in equation 6.6 and the estimated coefficients for this model equation are listed in Table 6.23.

$$\begin{aligned} logit(\pi_i) &= \log \frac{\pi_i}{1-\pi_i} \\ &= \beta_0 + \beta_1 Twin_i + \beta_2 \Delta P_i + \beta_3 CFM_i + \beta_4 RPM_i, \quad i, \ldots, n \end{aligned} \tag{6.5}$$

$$\begin{aligned} logit(\pi_i) &= \log \frac{\pi_i}{1-\pi_i} \\ &= \beta_0 + \beta_1 Tain_i + \beta_2 Taout_i + \beta_3 Twout_i + \beta_4 \Delta P_i, \quad i, \ldots, n \end{aligned} \tag{6.6}$$

Equivalently, equation 6.6 can be written,

$$\pi_i = \frac{\exp(\beta_0 + \beta_1 Tain_i + \beta_2 Taout_i + \beta_3 Twout_i + \beta_5 \Delta P_i)}{1 + \exp(\beta_0 + \beta_1 Tain_i + \beta_2 Taout_i + \beta_3 Twout_i + \beta_5 \Delta P_i)} \quad i, \ldots, n \tag{6.7}$$

The assessment of significance of the variables in the model is performed under the null hypothesis that the 4 "slope" coefficients for the covariate in the model are equal to zero. The distribution of the statistics is χ^2 with 4 degrees of freedom. The value of the log likelihood is $L = -354.19$. A second model, fit with the constant term only, yields $L = -786.06$. Hence, the statistic, G, which is the difference between deviance of model without the variable and the deviance of model with the variable results in 864. The p-value for the test is $\Pr[\chi^2_4 > 864] < 0.001$; thus the null hypothesis is rejected and we may conclude that at least one and perhaps

Table 6.21 Logistic model selection table

Index	Variable combination	Log-likelihood	Deviance
108	G3 G6 G7 G8	-218.31	436.62
88	G2 G6 G7 G8	-231.14	462.27
89	G2 G6 G7 G9	-234.02	468.05
109	G3 G6 G7 G9	-235.92	471.84
54	G1 G6 G7 G9	-256.55	513.11
22	G1 G3 G4 G5	-269.52	539.05
53	G1 G6 G7 G8	-278.22	556.43
76	G2 G4 G6 G7	-303.18	606.35
16	G1 G2 G6 G7	-305.05	610.09
119	G4 G6 G7 G9	-305.42	610.85
41	G1 G4 G6 G7	-313.39	626.78
126	G6 G7 G8 G9	-319.9	639.8
91	G2 G7 G8 G9	-321.46	642.91
66	G2 G3 G6 G7	-326.9	653.81
118	G4 G6 G7 G8	-327.63	655.26
57	G2 G3 G4 G5	-332.82	665.64
56	G1 G7 G8 G9	-349.71	699.42
90	G2 G6 G8 G9	-350.74	701.48
8	G1 G2 G4 G6	-354.2	708.39
43	G1 G4 G6 G9	-362.44	724.88
42	G1 G4 G6 G8	-364.35	728.69
45	G1 G4 G7 G9	-366.46	732.91
44	G1 G4 G7 G8	-366.8	733.6
92	G3 G4 G5 G6	-366.87	733.75
9	G1 G2 G4 G7	-370.43	740.85
79	G2 G4 G7 G8	-372.11	744.23
80	G2 G4 G7 G9	-376.8	753.59
10	G1 G2 G4 G8	-377.01	754.02
4	G1 G2 G3 G7	-377.3	754.59
46	G1 G4 G8 G9	-379.77	759.54
19	G1 G2 G7 G8	-380.32	760.64
11	G1 G2 G4 G9	-382.41	764.82
20	G1 G2 G7 G9	-382.43	764.87
70	G2 G3 G7 G9	-383.55	767.1
69	G2 G3 G7 G8	-389.22	778.44
121	G4 G7 G8 G9	-391.43	782.86
111	G3 G7 G8 G9	-397.9	795.8
58	G2 G3 G4 G6	-399.4	798.79
78	G2 G4 G6 G9	-401.14	802.29
81	G2 G4 G8 G9	-415.23	830.46

Table 6.21 (Continued)

18	G1 G2 G6 G9	-427.27	854.54
77	G2 G4 G6 G8	-429.1	858.19
68	G2 G3 G6 G9	-429.86	859.72
21	G1 G2 G8 G9	-433.58	867.15
6	G1 G2 G3 G9	-435.82	871.65
71	G2 G3 G8 G9	-438.36	876.72
3	G1 G2 G3 G6	-446.83	893.66
67	G2 G3 G6 G8	-449.42	898.84
17	G1 G2 G6 G8	-452.19	904.38
5	G1 G2 G3 G8	-453.12	906.25
35	G1 G3 G7 G9	-487.1	974.21
55	G1 G6 G8 G9	-506.41	1012.82
120	G4 G6 G8 G9	-514.22	1028.43
31	G1 G3 G6 G7	-516.15	1032.3
32	G1 G3 G6 G8	-543.31	1086.63
34	G1 G3 G7 G8	-545.98	1091.96
33	G1 G3 G6 G9	-562.4	1124.8
110	G3 G6 G8 G9	-575.26	1150.53
36	G1 G3 G8 G9	-576.77	1153.54

Table 6.22 Estimated coefficients for the logistic model equation 6.5

Variables			Log-likelihood	-218.31
	G3 G6 G7 G8		Deviance	436.62
	Parameter	β	Std. error	Wald
		estimate	se(β)	Statistic
Intercept	β_0	-55.5701	4.9796	-11.1596
Twin	β_1	0.2802	0.0318	8.8215
ΔP	β_2	-37.8567	2.6234	-14.4304
CFM	β_3	-0.0180	0.0012	-15.3884
RPM	β_4	0.0992	0.0072	13.8166

Table 6.23 Estimated coefficients for the logistic model equation 6.6

Variables G1 G2 G4 G6			Log-likelihood Deviance	-354.19 708.39
	Parameter	β estimate	Std. error se(β)	Wald Statistic
	β_0	42.4305	2.9932	14.1756
Tain	β_1	-1.7082	0.1635	-10.4481
Taout	β_2	0.9853	0.2181	4.5178
Twout	β_3	-0.6267	0.0552	-11.3474
ΔP	β_4	-3.8884	0.5343	-7.2771

all 4 coefficients are different from zero. Before concluding that any or all of the coefficients are nonzero, the Wald test statistic is applied to test the significance of the β estimates.

A Wald test is obtained by comparing the maximum likelihood estimate of the slope parameter, or β's to an estimate of its standard error. The resulting ratio, under the hypothesis that $H_o : \beta = 0$ follows a standard normal distribution (Hosmer 1989). The Wald statistic is given by Equation 6.8.

$$W_j = \frac{\hat{\beta}_j}{se(\hat{\beta}_j)} \tag{6.8}$$

where $\hat{\beta}_j$ is the parameter estimates and $se(\hat{\beta}_j)$ is the standard error associated with the j^{th} variable. The ratio is used in checking the hypothesis of redundancy of the j^{th} variable. The null hypothesis, H_o is accepted if $|W_j| \leq c$, and rejected otherwise. Here, c is the $(1 - \alpha/2)$ quantile of the standard normal distribution. For $\alpha = .05$, $c = 1.96$.

Under the null hypothesis that an individual coefficient is zero, the univariate Wald test statistic follows the standard normal distribution. The two tailed p-value is Pr[$|Z| >$ WaldStatistic]. The value of these statistics provides an indication of which of the variables in the model may or may not be significant. If the critical value of 2 is chosen, which would lead to an approximate level of significance of 0.05, then we would conclude that none of the variables are insignificant. Although the overall goal is to obtain the best fitting model while minimizing the number of parameters, no further reduction of the parameter is possible with the test of statistical significance.

Table 6.24 Classification summary table based on the logistic regression model in Table 6.23 and Table 6.22

Using Train Data	Model equation 6.6	Model equation 6.5
classified group 1 into group 1	81	81
classified group 1 into group 2	2	2
classified group 2 into group 1	6	2
classified group 2 into group 2	14	18
Using Test Data		
classified group 1 into group 1	144	144
classified group 1 into group 2	0	0
classified group 2 into group 1	17	16
classified group 2 into group 2	51	52

The results of classifying the observations of AHU faults using the fitted models given in Tables 6.22 and 6.23 are presented in Table 6.24. The associated figures for the classification of AHU data are presented in Figures 6.20 through 6.23. Using the model equation 6.6, the overall rate of correct classification is estimated at 92%, with 98% for the normal operation and 70% for the fault operations. The performance of the model on the test data set resulted in 92% for the overall rate of correct classification, no misclassifications for the normal operation, and 75% for the AHU fault modes.

Surprisingly, the model using equation 6.5 and Table 6.22 did not show much improvement over the classification result. Overall performance difference between the selected model, equation 6.6, and the best model, equation 6.5 resulted in 4%. The model correctly classified at 100% for the normal operation and 76% for the fault conditions. The proportion of the faults occurring during normal operation were identified at a satisfactory rate. Hence, the study of reduced number of variables used for classification of the faults resulted in good estimates of the proportion of the faults occurring for the flow loop test conditions. These faults included reduced fan speed, and coil blockage, and sticking hot water coil valve. The predictor variables included 4 variables, namely the inlet and outlet air temperatures, hot water outlet temperature, and pressure increase across the AHU.

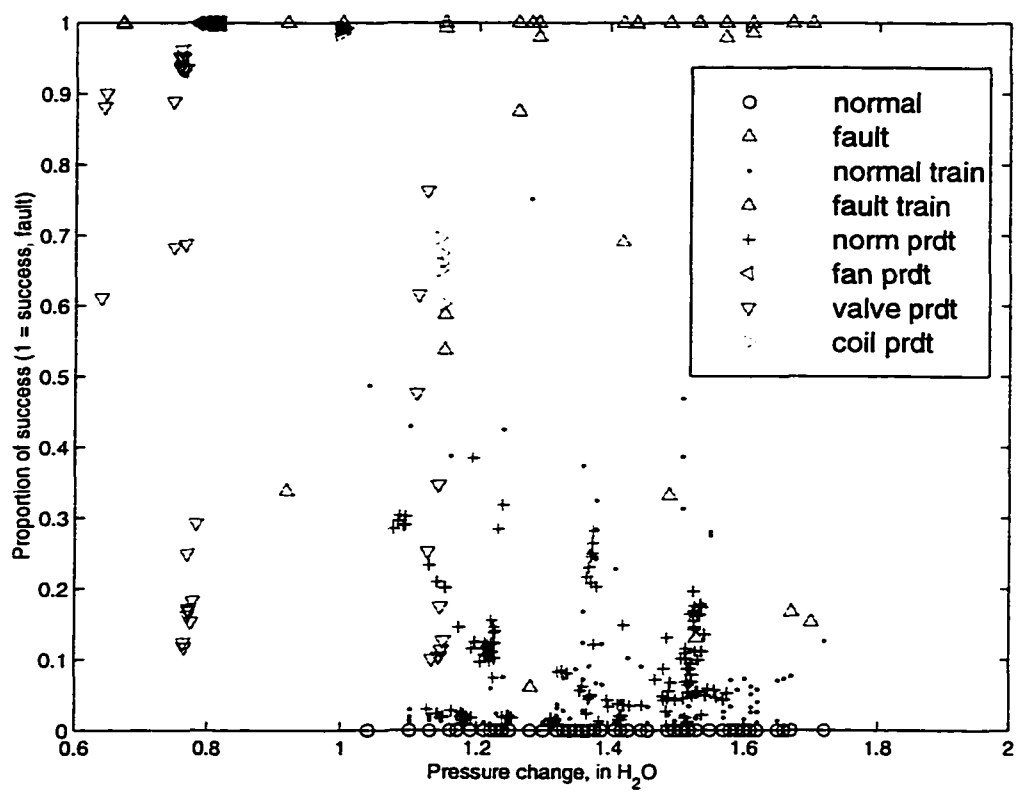


Figure 6.20 Logistic classification by equation 6.6, ΔP

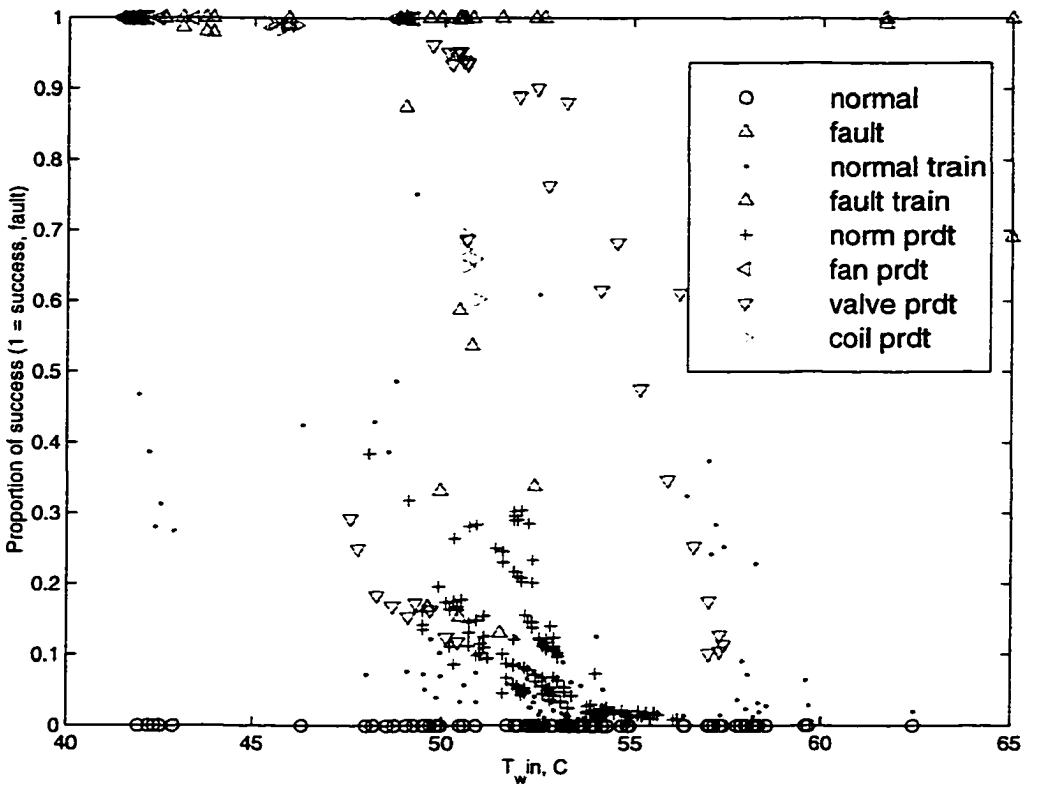


Figure 6.21 Logistic classification by equation, 6.6, Twout

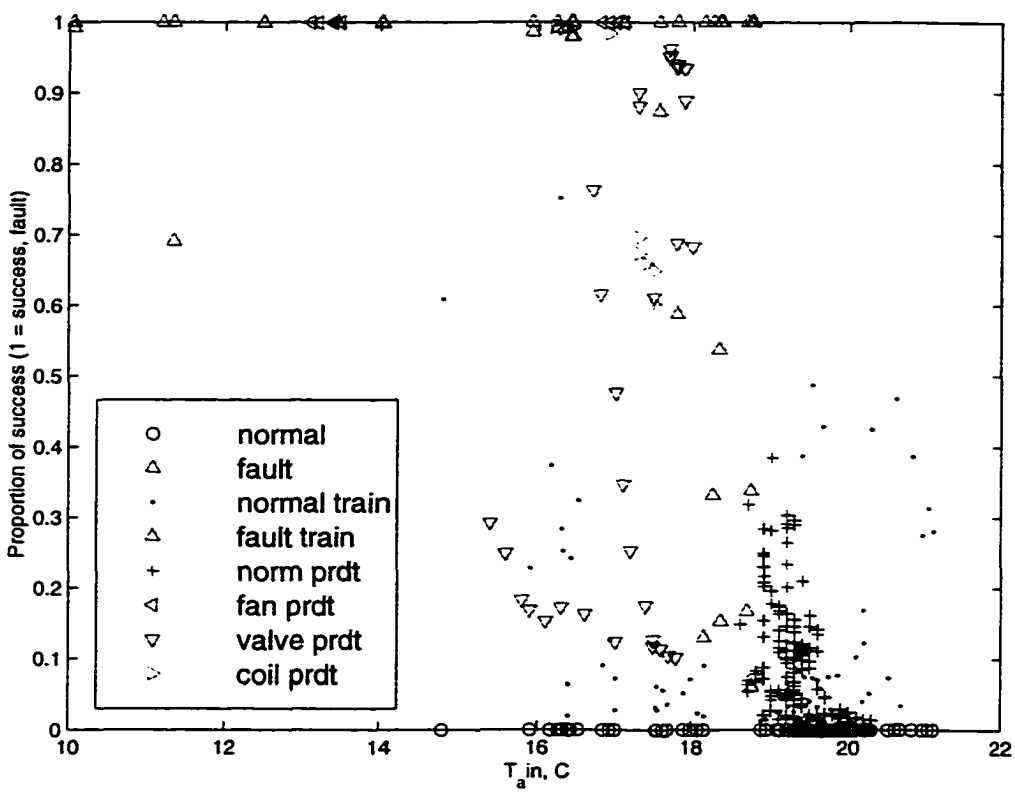


Figure 6.22 Logistic classification by equation 6.6, Taout

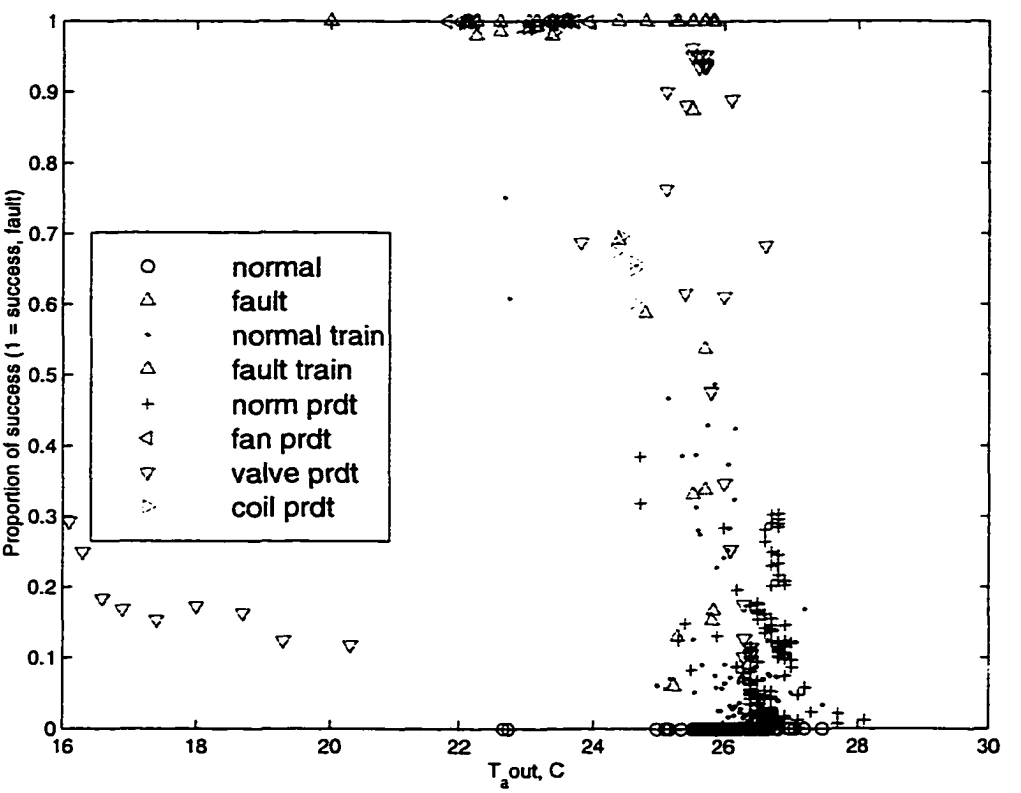


Figure 6.23 Logistic classification by equation 6.6, Twin

7 CONCLUSIONS

Multivariate techniques such as discriminant and classification analysis, principal component analysis, and logistic regression analysis were applied in this research to differentiate and classify three different faults (reduced fan speed, coil blockage, and sticking of hot water valve) occurring in an AHU. The research objective was to reduce the number of the variables needed without jeopardizing the correct classification.

Rational for minimizing the number of variables in the model is that the resultant model is more likely to be numerically stable, and more easily generalized. The more variables included in a model, the greater the estimated standard errors become. Another problem is that of overfitting which produces numerically unstable estimates. Overfitting is typically characterized by unrealistically large estimated coefficients and estimated standard errors.

Because discriminant analysis is exploratory in nature, it is often carried out in one time basis in order to investigate the observed differences when causal relationships are not well understood. On the other hand, the classification process is less exploratory in the sense that it provides a well defined set of rules which can be used to assign new objects. The main concern for this research was to be able to detect those changes in the AHU with a minimal set of variables. The conclusions drawn from this research concur with the initial research hypothesis and a reduced set of variables was selected for classification of the AHU faults.

Logistic regression was utilized to estimate the relationship between one or more predictor variables and the likelihood that an individual belongs to a fault group. The interpretation of the log odds in the logistic regression coefficients was similar to the interpretation given by the process dynamic transfer functions. The procedure also gave the probability associated with each prediction. Discriminant analysis can be used to predict group membership, but it

requires assumptions about the data that are more restrictive than those for logistic regression. It requires that the predictors have a multivariate normal distribution for each category of the grouping variables and that each category have the same variances and covariances for the predictors. This implies that discriminant analysis should not be used with categorical predictors. Unfortunately logistic regression requires many of the same assumptions as linear regression analysis, including independence of observations, and completely specified model-conditions that are often difficult to meet in practice. For hypothesis tests to be accurate, logistic regression requires large samples.

The results of this research on fault detection and classification found that the PCA results gave satisfactory reduction in dimensionality. Discriminant analysis was able to distinguish between each of the AHU faults and normal operation. Stepwise discriminant analysis selected four variables that are most significant to the identification of the three faults. With the reduction in the set of variables, there was a small reduction in the overall correct classification rate from 97% to 87% compared to using the full set of variables.

The following procedure was used in this research for developing fault detection and classification models:

- Approximately 2000 sets of data points were obtained from factorial experiments. Four variables (Tain, Twin, Twout, ΔP) were used to describe the 3 faults (fan, valve, and coil).
- Multivariate techniques (PCA, discrimination and classification, and logistic regression) allowed exploration of the faults involved in classification.
- The method was implemented in a process monitoring scheme.

Practical applications from the results of this study are that the adequacy of fault prediction can be applied as an index for preventative maintenance scheduling and can reduce sensor costs in a monitoring system.

APPENDIX A MULTIVARIATE NORMAL DISTRIBUTION

Many univariate tests and confidence intervals are based on the univariate normal distribution. Similarly, the vast majority of multivariate procedures have as their "underpinning" the multivariate normal distribution. The following are some of the useful features of the multivariate normal distribution.

- Only means, variances, and covariances need be estimated to completely describe the distribution.
- Bivariate plots show linear trends.
- If the variables are uncorrelated, they are independent.
- Linear functions of multivariate normal variables are also normal.
- The convenient form of the density function leads to derivation of many properties and test statistics.
- Even when the data are not multivariate normal, the multivariate normal may serve as a useful approximation, especially in inferences involving sample mean vectors, which are approximately normal by the central limit theorem.

If a random variable y, with mean, μ, and variance, σ^2, then normal distribution has the form of

$$f(y) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp -\frac{(y-\mu)^2}{2\sigma^2} \qquad -\infty < y < \infty \tag{A.1}$$

And if y has the density function of equation A.1, then we say that y is distributed as $N(\mu, \sigma^2)$, or simply y is $N(\mu, \sigma^2)$. If $\vec{y}$ is p variate and has a multivariate normal distribution with mean vector, $\vec{\mu}$, and covariance matrix, Σ, the density is given by

$$g(\vec{y}) = \frac{1}{(\sqrt{2\pi})^p |\Sigma|^{1/2}} \exp(-\frac{(\vec{y}-\vec{\mu})'\Sigma^{-1}(\vec{y}-\vec{\mu})}{2}) \tag{A.2}$$

where p is the number of variables and $|\Sigma|$ is the determinant of Σ. Here A' is transpose of a matrix A. When $\vec{y}$ has the density (A.2), we say that $\vec{y}$ is $N_p(\vec{\mu}, \Sigma)$. The term $(y-\mu)^2/\sigma^2 = (y-\mu)(\sigma^2)^{-1}(y-\mu)$ in the exponent of the univariate normal density measures the squared distance from y to μ in standard deviation units. Similarly, the term $(\vec{y}-\vec{\mu})'\Sigma^{-1}(\vec{y}-\vec{\mu})$ in the exponent of the multivariate normal density is the squared generalized distance from $\vec{y}$ to $\vec{\mu}$, or the Mahalanobis distance, Δ^2.

The effect of generalized population variance $|\Sigma|$ (generalized sample variance is noted as $|\mathbf{S}|$ in this dissertation) on the density is such that a small value of $|\Sigma|$ indicates that the variables are highly intercorrelated and the effective dimensionality is less than p. In general, for any number of variables p, a decrease in intercorrelations among the variables or an increase in the variances will lead to a larger $|\Sigma|$. This is a quick way to see if there is a way to discover reduction of dimensionality to represent the data. In the presence of multicollinearity, one or more eigenvalues of Σ will be near zero and $|\Sigma|$ will be small, since $|\Sigma|$ is the product of the eigenvalues. For any square matrix A with eigenvalues $\lambda_1, \lambda_2, \ldots \lambda_n$, $|A| = \prod_{i=1}^{n} \lambda_i$. Usually graphical usage to display the bivariate distribution is by contour plots.

While real data may not be exactly multivariate normal, the multivariate normal will frequently serve as a good approximation to the true distribution. Other reasons are the availability of tests and graphical procedures for assessing normality and many are used in available software packages. Many of the procedures based on multivariate normality are robust to departures from normality, (Rencher 1995). For the random $p \times 1$ vector $\mathbf{y}$ from a multivariate normal distribution $N_p(\mu, \Sigma)$ Table A.1 lists some of the properties of the multivariate normal random variables.

Maximum Likelihood Estimation (MLE) in the multivariate normal

With the assumption that the multivariate normal holds for a population, parameters are found by the method of *maximum likelihood*. The idea is simple in that the observed vectors $\mathbf{y_1}, \mathbf{y_2}, \ldots, \mathbf{y_n}$ are considered to be known and the values of μ, Σ are iterated so that joint density of the *likelihood functions*, $\mathbf{y}$, are maximized. For the multivariate normal, the

Table A.1 Properties of multivariate normal random variables

Normality of linear combinations of the variables in $\mathbf{y}$
If $\mathbf{a}$ is a vector constants, the linear function $\mathbf{a}'\mathbf{y} = a_1 y_1 + \ldots + a_p y_p$, the mean and variance are given by $E(\mathbf{a}'\mathbf{y}) = \mathbf{a}'\mu$ and $var(\mathbf{a}'\mathbf{y}) = \mathbf{a}'\Sigma\mathbf{a}$ for any random vector $\mathbf{y}$. In addition, $\mathbf{a}'\mathbf{y}$ has a univariate normal distribution if $\mathbf{y}$ is $N_p(\mu, \Sigma)$.
If $\mathbf{A}$ is a constant $q \times p$ matrix of rank q, where $q \leq p$, then $\mathbf{Ay}$ consists of q linear combinations of the variables in $\mathbf{y}$, with distribution $N_q(\mathbf{A}\mu, \mathbf{A}\Sigma\mathbf{A}')$. Similar to univariate sense, $E(\mathbf{Ay}) = \mathbf{A}\mu$ and $cov(\mathbf{Ay}) = \mathbf{A}\Sigma\mathbf{A}'$. In addition, the q variables in $\mathbf{Ay}$ have a multivariate normal distribution.

Standardized variables
If $\mathbf{y}$ is $N_p(\mu, \Sigma)$, a standardized vector $\mathbf{z}$ can be obtained by $\mathbf{z} = (\mathbf{T}')^{-1}(\mathbf{y} - \mu)$ where $\Sigma = \mathbf{T}'\mathbf{T}$ is factored using the Cholesky procedure. This is important in that in the multivariate case, a standardized vector of random variables has all 0 means and unit variances and all correlation = 0

χ^2 distribution
A χ^2 random variable with p degrees of freedom is defined as the sum of squares of p independent standard normal random variables. Thus, if $\mathbf{z}$ is standardized vector, then $\sum_{i=1}^{p} z_i^2 = \mathbf{z}'\mathbf{z}$ has the χ^2-distribution with p degrees of freedom, denoted χ_p^2. Hence, if $\mathbf{y}$ is $N_p(\mu, \Sigma)$, then $(\vec{y} - \vec{\mu})'\Sigma^{-1}(\vec{y} - \vec{\mu})$ is χ_p^2.

maximum likelihood estimates of μ, Σ are as stated in the equation A.5.

$$\hat{\mu} = \bar{\mathbf{y}} \tag{A.3}$$

$$\hat{\Sigma} = 1/n \sum_{i=1}^{n} (\mathbf{y_i} - \bar{\mathbf{y}})(\mathbf{y_i} - \bar{\mathbf{y}})' \tag{A.4}$$

$$= \frac{n-1}{n}\mathbf{S} \tag{A.5}$$

where $\mathbf{S}$ is the sample covariance matrix. Since $\hat{\Sigma}$ has divisor of n instead of $n-1$, it is biased and $\mathbf{S}$ is used inplace of $\hat{\Sigma}$. Proof:

Because the $\mathbf{y}_i$'s, constitute a random sample, they are independent, and the joint density is

the product of the densities of the $\mathbf{y}$. The likelihood function is

$$L(\mathbf{y_1}, \mathbf{y_2}, \cdots, \mathbf{y_n}, \mu, \Sigma) = \prod_{i=1}^{n} f(\mathbf{y_i}, \mu, \Sigma) \quad \text{(A.6)}$$

$$= \prod_{i=1}^{n} \frac{1}{(\sqrt{2\pi})^p |\Sigma|^{1/2}} \exp(-(\mathbf{y_i} - \mu)'\Sigma^{-1}(\mathbf{y_i} - \mu)/2) \quad \text{(A.7)}$$

$$= \frac{1}{(\sqrt{2\pi})^{np} |\Sigma|^{n/2}} \exp(-\sum_{i=1}^{n}(\mathbf{y_i} - \mu)'\Sigma^{-1}(\mathbf{y_i} - \mu)/2) \quad \text{(A.8)}$$

To see that $\hat{\mu} = \hat{\bar{\mathbf{y}}}$ indeed maximize the likelihood function, write the exponent term with addition and subtraction of $\bar{\mathbf{y}}$.

$$-\sum_{i=1}^{n}(\mathbf{y_i} - \bar{\mathbf{y}} + \bar{\mathbf{y}} - \mu)'\Sigma^{-1}(\mathbf{y_i} - \bar{\mathbf{y}} + \bar{\mathbf{y}} - \mu)/2 \quad \text{(A.9)}$$

And after expanding in terms of $(\mathbf{y_i} - \bar{\mathbf{y}})$ and $(\bar{\mathbf{y}} - \mu)$, two of the four resulting terms go away because $\sum_i (y_i - \bar{y}) = 0$. Hence the equation A.8 becomes

$$L = \frac{1}{(\sqrt{2\pi})^{np} |\Sigma|^{n/2}} \exp(-\sum_{i=1}^{n}(\mathbf{y_i} - \bar{\mathbf{y}})'\Sigma^{-1}(\mathbf{y_i} - \bar{\mathbf{y}})/2 - n(\bar{\mathbf{y}} - \mu)'\Sigma^{-1}(\bar{\mathbf{y}} - \mu)/2) \quad \text{(A.10)}$$

Since Σ^{-1} is positive definite, $-n(\bar{\mathbf{y}} - \mu)'\Sigma^{-1}(\bar{\mathbf{y}} - \mu)/2 \leq 0$

and $0 < exp(-n(\bar{\mathbf{y}} - \mu)'\Sigma^{-1}(\bar{\mathbf{y}} - \mu)) \leq 1$, with maximum resulting when the exponent is 0. Therefore, L is maximized when $\hat{\mu} = \hat{\bar{\mathbf{y}}}$. Also the maximum likelihood estimator of the population correlation matrix is the sample correlation matrix, $\hat{\mathbf{P}}_\rho = \mathbf{R}$. Relationships among multinormal variables are linear and the estimators $\mathbf{S}, \mathbf{R}$ is useful because they measure only linear relationships but they are not useful for some nonnormal distributions.

For the distribution of $\bar{\mathbf{y}} = \sum_{i=1}^{n} \mathbf{y_i}/n$, if observations are based on a random sample from a multivariate normal distribution, then $\bar{\mathbf{y}}$ is $N_p(\mu, \Sigma/\mathbf{n})$. If $\bar{\mathbf{y}}$ is based on a random sample from a nonnormal multivariate population with mean vector and covariance matrix, then for large n, $\bar{\mathbf{y}}$ is approximately $N_p(\mu, \Sigma/\mathbf{n})$. This is the result from the multivariate central limit theorem.

Assessing multivariate normality

Many tests and procedures have been suggested for evaluation of the assumption of the multivariate normality. One possible check is to test each variable separately for univariate

normality.

Assessing with univariate normality

Although the multivariate normality implies individual normality, assessment of the individual may not guarantee joint normality due to variables that may be correlated. Thus, if even one of the separate variables is not normal, the vector is not multivariate normal. Hence initial check on the individual variable maybe useful.

A basic graphical check for normality is the $Q-Q$ plot that compares quantiles of a sample against the population quantiles of the univariate normal. If the points are close to a straight line, then there is no indication of departure from normality. On the other hand deviation from the straight line indicates nonnormality. In fact, the type of nonlinear pattern leads to type of departure from normality. Quantiles are similar to the percentiles. For example, a test score of 90th percentile is 0.9 quantile score.

The sample quantiles for the $Q-Q$ plot are obtained as follows.

- Rank the observations $y_1, y_2, \ldots, y_n$ from low to high.
- The point $y(i)$ is the i/n sample quantile.
- For better estimate use $(i-1/2)/n$ sample quantile.

On the same note, if $q_1, q_2, \ldots, q_n$, then q_i is the value below which a proportion $(i-1/2)/n$ of the observations lie. That is $(i-1/2)/n$ is the probability of getting an observation less than or equal to q_i. q_i is found for the standard normal random variable y with distribution $N(0,1)$ by solving

$$\Phi(q_i) = \Pr(y < q_i) = \frac{i-1/2}{n} \tag{A.11}$$

This requires numerical integration or tables of the cumulative standard normal distribution, $\Phi(x)$. The population does not have to have the same mean and variance as the sample, because changes in mean and variance just change the slope and intercept of the plot line in the $Q-Q$ plot. Therefore using the standard normal distribution allows finding the q_i values

from the table of standard normal probabilities. Finally plot $(q_i, y(i))$ and check for linearity. This $Q-Q$ plot provides a good visual check on normality and is considered to be adequate for the study.

APPENDIX B MULTIVARIATE VERSUS UNIVARIATE TESTS

This chapter discusses statistical facts from various sources, Flury (1997), Rencher (1995), Ott (1944), and Johnson (1992), with regard to the statistical inferences for both univariate and multivariate analyses. Most of these facts are well documented in statistical textbooks. However, these facts are presented in this appendix as a handy reference to speed up the computation and the research analysis.

Initial Concepts

Hypothesis testing in a multivariate context is more complex than in a univariate setting. The p-variate normal distribution has p means, p variances, and $\binom{p}{2} = p!/2!(p-2)!$ covariances. Total number of parameters is $p(p+3)/2$. Each parameter corresponds to a hypothesis that could be formulated. The use of p univariate tests inflates the Type I error rate, α, whereas the multivariate test preserves the exact α level. This is because the univariate tests completely ignores the correlations among the variables. In contrast, the multivariate tests make the use of the covariance matrix. The multivariate test is more powerful in many cases. Since the power of a test is the probability of rejecting H_0 when it is false, unlike the univariate tests, the multivariate test produces significant indication with small effects jointly combined. The small effects on the single univariate tests would otherwise fail to reach the significance.

Univariate test of significance

Test of hypothesis for the difference of the means

To investigate the effects of experimental performance differences between existing groups, statistical tests can be performed with sampled data. Two groups of interests are either

independent (observations are not related to one another) or dependent (observations are correlated where the measurements are taken from the same unit). In either case, the data are summarized in the form of sample means that can be compared with F-test, t-test or a z-test.

The most commonly used t-test for two independent samples utilizes a pooled variance in the calculation of the standard error of the difference. The assumption of the equal variances of the two samples allows usage of the pooled variance, (Ott 1994).

Univariate t-Test for one independent samples with unknown variance

- One variable measured on each sampling unit
- Assume random sample $y_1, y_2, \ldots, y_n$ from $N(\mu, \sigma^2)$
- Estimate $\bar{y}, s^2$
- Test the null hypothesis, $H_0 : \mu = \mu_0$ vs alternate hypothesis, $H_a : \mu \neq \mu_0$

$$t_s = \frac{\bar{y} - \mu_0}{s/\sqrt{n}} = \frac{\sqrt{n}(\bar{y} - \mu_0)}{s} \tag{B.1}$$

If H_0 is true, the test statistic, t_s, has a student's t-distribution with $(n-1)$ degrees of freedom. Reject H_0 if $|t_s| \geq t_{\alpha/2,n-1}$ where $t_{\alpha/2,n-1}$ is a critical value from the t-table. In words, the test rejects that the specific μ_0 is a plausible value of μ, if the observed $|t|$ exceeds a specified percentage point of a t-distribution with $n-1$ degrees of freedom. The expression $\frac{\bar{y}-\mu_0}{s/\sqrt{n}}$ is the characteristic form of the t-statistic and it represents a sample standardized distance between $\bar{y}$ and μ_0. In this form, the hypothesized mean is subtracted from $\bar{y}$ and the difference is divided by $s_{\bar{y}} = s/\sqrt{n}$. Since y_i is $N(\mu, \sigma^2), i = 1, 2, \ldots, n$, $\bar{y}$ and s are independent.

Rejecting H_0 when $|t|$ is large is equivalent to rejecting H_0 if its square, t^2, equation B.2 is large.

$$t^2 = \frac{(\bar{y} - \mu_0)^2}{s^2/n} = n(\bar{y} - \mu_0)(s^2)^{-1}(\bar{y} - \mu_0) \tag{B.2}$$

The variable t^2 is the squared distance from the sample mean $\bar{y}$ to the test value μ_0. The units of distance are expressed in terms of $s/\sqrt{n}$ or estimated standard deviations of the sample

mean. $\bar{y}$. The rejection criteria of the H_0 in favor of H_a, at significance level α is as follows.

$$n(\bar{y}-\mu_0)(s^2)^{-1}(\bar{y}-\mu_0) > t^2_{\alpha/2,(n-1)} \tag{B.3}$$

where $t^2_{\alpha/2,(n-1)}$ denotes the upper $100(\alpha/2)^{\text{th}}$ percentile of the t-distribution with $(n-1)$ degrees of freedom.

If H_0 is not rejected, then conclude μ_0 is a plausible value for the normal population mean. The $100(1-\alpha)\%$ confidence interval for those values of μ_0 that would not be rejected by the test $H_0 : \mu = \mu_0$ is as follows.

$$\bar{y} - t_{\alpha/2,(n-1)}\frac{s}{\sqrt{n}} \leq \mu_0 \leq \bar{y} + t_{\alpha/2,(n-1)}\frac{s}{\sqrt{n}} \tag{B.4}$$

The probability that the interval contains μ is $1-\alpha$. That is, among the large number of such intervals, there are $100(1-\alpha)\%$ of them that contain μ. A natural generalization of the squared distance is the statistic T^2 called *Hotelling*'s T^2. This is discussed later in this appendix.

Univariate t-Test for two independent samples with known variance

The test statistic, t_s, formula for the independent samples with the assumption of the equal variance is stated in equation B.5. For the independent sample t-Test, following assumptions hold.

- Observations are randomly sampled from each populations.
- Observations are normally distributed for both populations.
- Each population variances are unknown.
- Sample population is independent of each other.

The null hypothesis, H_0, to be tested is $\mu_1 - \mu_2 = D_0$.

$$t_s = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{S_{\overline{x}_1 - \overline{x}_2}} \tag{B.5}$$

Where the difference between the two sample mean is indicated by $(\overline{X}_1 - \overline{X}_2)$ and the term $(\mu_1 - \mu_2)$ is the hypothesized difference between the population means. The denominator is the

standard error of the difference. The formula for the pooled variance is calculated by equation B.6.

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2} \tag{B.6}$$

S_1^2 and S_2^2 are the variance of the samples 1 and 2 respectively. n_1 and n_2 are the number of observations in the samples groups 1 and 2 respectively. The term weighted average of the sample variance is used frequently for the pooled variance where each variance is weighted by its respective degrees of freedom. The standard error of the difference is given by equation B.7.

$$S_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{S_p^2}{n_1} + \frac{S_p^2}{n_2}} \tag{B.7}$$

This statistic has $n_1 + n_2 - 2$ degrees of freedom. The test result of the t-Test is compared

Table B.1 t-Test rejection region for independent samples, equal variance

H_a	Rejection Region for Type I error α and dof $= n_1 + n_2 - 2$
$\mu_1 - \mu_2 > D_0$	reject H_0 if $t_s > t_\alpha$
$\mu_1 - \mu_2 < D_0$	reject H_0 if $t_s < -t_\alpha$
$\mu_1 - \mu_2 \neq D_0$	reject H_0 if $abs(t_s) > t_{\alpha/2}$

against obtained t_s value with the appropriate t_{α,n_1+n_2-2} critical value. The test results are considered statistically significant if the probabilities are less than the research test level.

Approximate t-Test for independent samples with unknown variance

If the population variances are unknown and unequal, it may not be good to use a pooled variance estimate. Welch (1938) showed that percentage points of a t distribution with modified degrees of freedom can be used to set the rejection region for the null hypothesis, $H_0 : \mu_1 - \mu_2 = D_0$. In estimating the standard error of the differences, two sample variances, s_1^2 and s_2^2 are used. The test statistics, t_s, and the standard error, $s_{\overline{X}_1 - \overline{X}_2}$ are calculated by the equations B.8 and B.9. The null hypothesis, H_0, to be tested is $\mu_1 - \mu_2 = D_0$.

$$t_s = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{s_{\overline{x}_1 - \overline{x}_2}} \tag{B.8}$$

$$s_{\overline{X}_1 - \overline{X}_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \tag{B.9}$$

The degree of freedom are approximated by the equation B.10.

$$\text{dof}^\star = \frac{(n_1 - 1)(n_2 - 1)}{(n_2 - 1)(1 - C)^2 + (n_1 - 1)(1 - C)^2} \tag{B.10}$$

where

$$C^2 = \frac{s_1^2/n_1}{s_1^2/n_1 + s_2^2/n_2} \tag{B.11}$$

Table B.2 t-Test rejection region for independent samples, unequal variance

H_a	Rejection Region for a specified value of α, dof*
$\mu_1 - \mu_2 > D_0$	reject H_0 if $t_s > t_\alpha$
$\mu_1 - \mu_2 < D_0$	reject H_0 if $t_s < -t_\alpha$
$\mu_1 - \mu_2 \neq D_0$	reject H_0 if $abs(t_s) > t_{\alpha/2}$

The test based on the test statistic,t_s, is sometimes known as the *separate-variance t test* because of the separate sample variances, s_1^2, s_2^2 instead of pooled variance. This procedure is seldom used for following reasons. If the sample sizes, n_1 and n_2, are equal, then the pooled variance t-test is robust with respect to violations of the homogeneous variances assumption. If sample sizes are relatively large, say greater than 30, the pooled variances provides a satisfactory approximation.

Paired-Samples t-Test

Paired samples t-test is commonly used for two types of data analysis situations. One of them is used in the before-after design and the another for matching pairs of research subjects with correlated dependent variables. This test is best suited for the situation where the difference shows more information than by a single independent variable approach. The

test of hypothesis is $H_0 : \mu_d = D_0$. the formula for the paired samples test statistics, t, is given by equation B.13.

$$t = \frac{\overline{D} - \mu_d}{s_{\overline{D}}} \tag{B.12}$$

$$s_{\overline{D}} = \frac{s_D}{\sqrt{n}} \tag{B.13}$$

$\overline{D}$ is the mean of the difference in the samples, μ_d is the hypothesized population mean difference, and $s_{\overline{D}}$ is the standard error of the mean difference. s_D is the standard deviation of the difference of the samples and n is the total number of paired observations.

Table B.3 t-Test rejection region for independent samples, unequal variance

H_a	Rejection Region for a specified value of α
$\mu_d > D_0$	reject H_0 if $t > t_\alpha$
$\mu_d < D_0$	reject H_0 if $t < -t_\alpha$
$\mu_d \neq D_0$	reject H_0 if $abs(t) > t_{\alpha/2}$

The paired t-Test has $n - 1$ degrees of freedom. The statistical significance of the obtained t value can be determined by comparing α with the appropriate probability. For example, if the obtained probability is less than that of the significance level, α, a conclusion is made to reject the null hypothesis and conclude that there is statistical significance of the samples being compared. One method to check for the relationships between the paired samples obtained is the pearson correlation. High pearson correlation indicate strong relationship between the pairs of observations.

Analysis of Variance

Research designs often include more than one independent variable whose levels can involve more than two levels. To test the differences among group means, the analysis of variance (ANOVA) statistical tool is utilized. There are One-way Between-Groups ANOVA and Two-way Between-Groups ANOVA.

One-Way Between-Groups ANOVA

In the balanced one-way ANOVA, there are random samples of n observations from each of k normal populations with equal variances. The k samples or the populations from where they come from are sometimes referred to as *groups*. The groups are sometimes referred to as *treatments* in an experiment. The k samples are assumed to be independent and have the common variance to estimate the F-test. Therefore, to test one-way between-groups ANOVA, the F-test is utilized. The test statistic, F, is defined in equation B.14.

$$F = \frac{MST}{MSE} \tag{B.14}$$

MST is the mean squared between group error and MSE is the mean squared within group error. The underlying assumptions for using the F-test is as follows.

- Samples are selected randomly from the k populations.
- Observations are independent from each other.
- Samples are assumed to come from normal distribution.
- Variances of the K populations are unknown but equal to each other.

In testing the hypothesis, H_0: all means are same, if the probability associated with the obtained value of F, p-value, is smaller than the test of significance level, then a conclusion is made for the alternate hypothesis that the test is statistically significant and at least one of the means are not same.

One-way Repeated Measures ANOVA

In One-way repeated measures ANOVA, many measurements are made from the same experimental unit. If observations are made just two times, the paired samples t-test should be utilized. For this method, an additional assumption is made that the population covariances for all pairs of treatment levels are equal. A similar form of analysis is the two-factor without replication.

Two-way Between-Groups ANOVA

This method allows a researcher to analyze two factors and the interaction between the two factors simultaneously. The procedure has three different F-tests. The first test is the test for the main-effect for the first factor, second test is the test for the main-effect of the second factor, and third test is the test of interaction between factors.

Multivariate test of significance

In the multivariate case there are several variables measured on each sampling unit. The following explicitly illustrates the test of hypothesis for $H_0 : \mu = \mu_0$ vs $H_a : \mu \neq \mu_0$.

$$H_0 : [\mu_1 \ \mu_2 \ \dots \ \mu_p]^T = [\mu_{01} \ \mu_{02} \ \dots \ \mu_{0_p}]^T \tag{B.15}$$

$$H_a : [\mu_1 \ \mu_2 \ \dots \ \mu_p]^T \neq [\mu_{01} \ \mu_{02} \ \dots \ \mu_{0_p}]^T \tag{B.16}$$

where each μ_{0i} is a specified target value. The vector equality implies that $H_0 : \mu_i = \mu_{0i}$ for all $i = 1, 2, \dots, p$ and the vector inequality implies at least one $\mu_i \neq \mu_{0i}$.

Hotelling's T^2-test for $H_0 : \mu = \mu_0$ with unknown Σ

This section describes a hypothesis test procedure on a mean vector with unknown population covariance matrix. Assume p variables are measured on each sampling unit.

- Assume random samples, $\mathbf{y_1 y_2 \dots y_n}$ from $N_p(\mu, \Sigma)$
- Each $\mathbf{y_i}$ contains p measurements
- Estimate μ by $\bar{\mathbf{y}}$ and Σ by $\mathbf{S}$

In order to test $H_0 : \mu = \mu_0$ versus $H_a : \mu \neq \mu_0$, an extension of univariate t is applied.

$$t^2 = \frac{n(\bar{y} - \mu_0)^2}{s^2} = n(\bar{y} - \mu_0)(s^2)^{-1}(\bar{y} - \mu_0) \tag{B.17}$$

By replacing $(\bar{y} - \mu_0)$ and s^2 with $(\bar{\mathbf{y}} - \mu_0)$ and $\mathbf{S}$, test statistic, T^2 for the multivariate results. Because the distribution T^2 was obtained by Hotelling in 1931, the multivariate test statistic

is referred as Hotteling's T^2 to this day.

$$T^2 = n(\bar{\mathbf{y}} - \mu_0)'\mathbf{S}^{-1}(\bar{\mathbf{y}} - \mu_0) \tag{B.18}$$

$$Reject\ H_0 :\ \ if\ \ T^2 > T^2_{\alpha,p,n-1} \tag{B.19}$$

The distribution is indexed by two parameters, dimension p, and degrees of freedom $(n-1)$. The saying of "accepting H_0" is used for convenience to describe the decision made for not rejecting the hypothesis. Strictly speaking, we do not accept H_0 in the sense of actually believing it is true. If the sample size were very large and we accepted "H_0", we could be reasonably certain that the true μ is close to the hypothesized value of μ_0. Otherwise "accepting H_0" means only that we have failed to reject it.

The T^2-statistic can be expressed as $T^2 = nD^2$, where D^2, equation B.20 is known as the sample standardized distance. This distance is discussed in the discriminant analysis. Also the test can be viewed from the viewpoint of distance between the observed sample mean vector and the hypothetical mean vector. If the sample mean vector is distinctly distant from the hypothetical mean vector, then there is question of the hypothetical mean vector and reject H_0.

$$D^2 = (\bar{\mathbf{y}} - \mu_0)'\mathbf{S}^{-1}(\bar{\mathbf{y}} - \mu_0) \tag{B.20}$$

The test statistic is a scalar, univariate, quantity. As with the χ^2 distribution of Z^2, the density of T^2 is skewed because there is no upper limit and the lower limit is zero. Following are summary of T^2-test properties, (Rencher 1995).

- The inequality $n - 1 > p$ must be satisfied to avoid singularity in S
- Both one-sample and two-sample degrees of freedom are analogous to univariate t-test, $n - 1$, $n_1 + n_2 - 2$ for single and two samples respectively.
- The alternative hypothesis is two-sided. Because the space is multidimensional, we do not consider one-sided alternative hypothesis, such as $\mu > \mu_0$. However, even though the alternative hypothesis, $\mu \neq \mu_0$ is two-sided, the critical region is one-tailed.
- In the univariate case, $t^2_{n-1} = F_{1,n-1}$. The T^2 is converted to F-statistic by the equation B.21. The degrees of freedom for T^2 in one-sample case is $\nu = n - 1$.

$$\frac{\nu - p + 1}{\nu p} T^2_{p,\nu} = F_{p,\nu-p+1} \tag{B.21}$$

Multivariate One-Way Analysis of Variance Model (MANOVA)

In multivariate case, where several dependent variables are measured instead of just one, there is an assumption that k independent random samples of size n are obtained from p -variate normal populations with equal covariance matrices, see Table B.4.

Table B.4 Observation layout

	Sample 1 from $N_p(\mu_1, \Sigma)$	Sample 2 from $N_p(\mu_2, \Sigma)$	...	Sample k from $N_p(\mu_k, \Sigma)$
	$\mathbf{y_{11}}$	$\mathbf{y_{21}}$	...	$\mathbf{y_{k1}}$
	$\mathbf{y_{12}}$	$\mathbf{y_{22}}$	...	$\mathbf{y_{k2}}$
	$\vdots$	$\vdots$	$\vdots$	$\vdots$
	$\mathbf{y_{1n}}$	$\mathbf{y_{2n}}$	...	$\mathbf{y_{kn}}$
Total	$\mathbf{y_{1.}}$	$\mathbf{y_{2.}}$	...	$\mathbf{y_{k.}}$
Mean	$\bar{\mathbf{y}}_{\mathbf{1.}}$	$\bar{\mathbf{y}}_{\mathbf{2.}}$	...	$\bar{\mathbf{y}}_{\mathbf{k.}}$

The totals and mean are defined as follows: Total of the i^{th} sample: $\mathbf{y_{i.}} = \sum_{j=1}^{n} \mathbf{y_{ij}}$ Overall total: $\mathbf{y_{..}} = \sum_{i=1}^{k} \sum_{j=1}^{n} \mathbf{y_{ij}}$ Mean of the i^{th} sample: $\bar{\mathbf{y}}_{\mathbf{i.}}/n$ Overall mean: $\bar{\mathbf{y}}_{\mathbf{..}} = \mathbf{y_{..}}/kn$

To compare the mean vectors of the k samples for significant differences, the hypothesis testing that all mean vectors are equal,

$$H_0 : \mu_1 = \mu_2 = \ldots = \mu_k, \text{ verses } H_a : \text{at least two } \mu'\text{s are unequal.} \tag{B.22}$$

If two means differ for just one variable, i.e. $\mu_{23} \neq \mu_{43}$, then the null hypothesis, H_0 is rejected. Analogous to the univariate case, the multivariate case has between and within sums of squares but now they are in matrices, **H**, **E**. Equation B.23 is defined as between sums of squares and Equation B.24 is defined as within sums of squares.

$$\mathbf{H} = n \sum_{i=1}^{k} (\bar{\mathbf{y}}_{\mathbf{i.}} - \bar{\mathbf{y}}_{\mathbf{..}})(\bar{\mathbf{y}}_{\mathbf{i.}} - \bar{\mathbf{y}}_{\mathbf{..}})' \tag{B.23}$$

$$\mathbf{E} = \sum_{i=1}^{k}\sum_{j=1}^{n}(\mathbf{y_{ij}} - \mathbf{\bar{y}_{i.}})(\mathbf{y_{ij}} - \mathbf{\bar{y}_{i.}})' \tag{B.24}$$

Tests on covariance matrices

This section discusses tests of hypothesis involving the variance-covariance, Σ, structure. The tests allow check on assumptions relating to other tests covering hypotheses of the covariance matrix with particular structure, hypotheses of two or more equal covariance matrices and hypotheses of zero elements of the covariance matrix with implication of independence of the multivariate normal random variances. The likelihood approach is used for the methods and the resulting test statistics involve the ratio of the determinants of the sample covariance matrix.

Testing Σ pattern

Testing the hypothesis that the variables $y_1, y_2, \ldots, y_p$ in $\mathbf{y}$ are independent and have the same variance can be expressed as $H_o : \Sigma = \sigma^2\mathbf{I}$ versus $H_a : \Sigma \neq \sigma^2\mathbf{I}$ where σ^2 is the unknown common variance.

Tests comparing covariance matrices

An assumption for T^2 or MANOVA tests comparing two or more mean vectors is that the corresponding population covariance matrices are equal. That is $\Sigma_1 = \Sigma_2 = \ldots = \Sigma_k$. Under this assumption, the sample covariance matrices, $S_1, S_2, \ldots, S_k$ reflect a common population Σ and are therefore can be pooled to obtain an estimate of Σ. If $\Sigma_1 = \Sigma_2 = \ldots = \Sigma_k$ is not true, then large differences in $S_1, S_2, \ldots, S_k$ may lead to rejection of $H_o : \mu_1 = \mu_2 = \ldots = \mu_k$. However, Rencher (1995) pointed out that as long as the sample sizes are large and equal, the T^2 or MANOVA tests are fairly robust to heterogeneity of covariance matrices.

For the two sample univariate tests of equal variances, hypothesis $H_o : \sigma_1^2 = \sigma_2^2$ versus $H_a : \sigma_1^2 \neq \sigma_2^2$ is tested with the F statistic, $F = s_1^2/s_2^2$. Here, s_1^2 and s_2^2 are the variances from

the two samples. If H_o is true, F is distributed as F_{ν_1,ν_2} with ν_1, ν_2 degrees of freedom from the two samples.

For k sample case, Bartlett's (1937) test of homogeneity of variance stipulates, null hypothesis $H_o : \sigma_1^2 = \sigma_2^2, \ldots, \sigma_k^2$ versus H_a: non homogeneity is tested with F statistic, $F = a_2 m/(a_1(b-m))$.

$$c = 1 + \frac{1}{3(k-1}\left[\sum_{i=1}^{k}\frac{1}{\nu_i} - \frac{1}{\sum_i}\right], \tag{B.25}$$

$$s^2 = \frac{\sum_{i=1}^{k}\nu_i s_i^2}{\sum_{i=1}^{k}\nu_i}, \tag{B.26}$$

$$m = \left(\sum_{i=1}^{k}\nu_i\right)\ln s^2 - \sum_{i=1}^{k}\nu_i \ln s_i^2, \tag{B.27}$$

$$a_1 = k-1 \qquad a_2 = \frac{k+1}{(c-1)^2} \qquad b = \frac{a_2}{2-c+2/a_2} \tag{B.28}$$

where $s_1^2, s_2^2, \ldots, s_k^2$ are independent sample variances with $\nu_1, \nu_2, \ldots, \nu_k$ degrees of freedom. Reject H_o if $F > F_\alpha$. This test is inappropriate for comparing $s_{11}, s_{22}, \ldots, s_{pp}$ from the diagonals of $\mathbf{S}$ because s_{ii} are correlated.

Multivariate tests of equality of covariance matrices

The Box's M-test for k multivariate populations in the hypothesis of equality of covariance matrices is

$$H_o : \Sigma_1 = \Sigma_2 = \ldots = \Sigma_k. \tag{B.29}$$

The test assume that the samples of size $n_1, n_2, \ldots, n_k$ are from independent multivariate normal distributions. The approximate F test statistics is as follows.

$$c_1 = 1 + \frac{1}{3(k-1)}\left[\sum_{i=1}^{k}\frac{1}{\nu_i} - \frac{1}{\sum_i}\right], \tag{B.30}$$

$$c_2 = \frac{(p-1)(p+2)}{6(k-1)}\left[\sum_{i=1}^{k}\frac{1}{\nu_i^2} - \frac{1}{(\sum_i \nu_i)^2}\right] \tag{B.31}$$

$$a_1 = \frac{1}{2}p(k-1)(p+1) \qquad a_2 = \frac{a_1+2}{|c_2 - c_1^2|} \tag{B.32}$$

$$b_1 = \frac{1 - c_1 - a_1/a_2}{a_1} \qquad b_2 = \frac{1 - c_1 - 2/a_2}{a_2} \tag{B.33}$$

$$\ln M = \frac{1}{2}\sum_{i=1}^{k}\nu_i \ln|\mathbf{S_i}| - \frac{1}{2}\left(\sum_{i=1}^{k}\nu_i\right)\ln|\mathbf{S_{pl}}| \tag{B.34}$$

$$\text{if} \quad c_2 > c_1^2, \quad \text{then} \quad F = -2b_1 \ln M \quad \text{is approximately} \quad F_{a1,a2} \tag{B.35}$$

$$\text{if} \quad c_2 < c_1^2, \quad \text{then} \quad F = -\frac{a_2 b_2 \ln M}{a_1(1 + 2b_2 \ln M)} \quad \text{is approximately} \quad F_{a1,a2} \tag{B.36}$$

Reject H_o if $F > F_\alpha$. It is suggested to perform T^2 or MANOVA tests before performing Box's M-test. Because the test is sensitive to some forms of non-normality, M-test may signal covariance heterogeneity in some cases where it is not damaging to the MANOVA tests. Hence we may not wish to automatically rule out standard MANOVA tests if the M-test leads to rejection of H_o.

APPENDIX C COMPUTER PROGRAMS

This chapter includes programs written for the data acquisition and the data analysis used in this research.

Quick BASIC program

```
DEFDBL A-Z

DECLARE FUNCTION VALCONT (COUNTER%)
DECLARE FUNCTION VALREAD (DVM%)
DECLARE FUNCTION PCONV (NUM%, XVAL)
DECLARE FUNCTION MFCONV (NUM%, XVAL)
DECLARE FUNCTION TCONV (XVL)

' Variable Declarations:

DIM XVAL(149)
DIM AIRIN(8)
DIM AIROUT(8)
DIM PEKI(149)
DIM PRIVR(100)
DIM xpow(200)
DIM Xair(100)

' Opening devices - HP 3488A scanner and the HP 3456A digital voltmeter:

    DEVICE$ = "GPIB0"
    CALL IBFIND(DEVICE$, BOARD%)
    DEVICE$ = "DVM"
    CALL IBFIND(DEVICE$, DVM%)
    DEVICE$ = "SCANNER"
    CALL IBFIND(DEVICE$, SCANNER%)
    DEVICE$ = "SCAN2"
    CALL IBFIND(DEVICE$, SCAN2%)
    DEVICE$ = "COUNTER"
    CALL IBFIND(DEVICE$, COUNTER%)

' Clearing the devices to ensure clean starting status:
```

```
CALL IBCLR(DVM%)
CALL IBCLR(COUNTER%)
CALL IBCLR(SCANNER%)
CALL IBCLR(SCAN2%)

' Setting the parameters on the HP3456A digital voltmeter:
CALL IBWRT(DVM%, "F1 R7 T2 A1 M3 H1")

CLS
path$ = "C:\KEN\DATA\"
PRINT path$
shell$ = "dir " + path$ + "*.*"
PRINT
SHELL shell$

INPUT "Enter Data File name"; OFILE$
PRINT
NFILE$ = OFILE$ + ".txt"
PRINT
OPEN path$ + NFILE$ FOR OUTPUT AS #2

PRINT "Enter time to measure in seconds"
INPUT tmeas
tbegin = TIMER
'PRINT , TBEGIN

PRINT "HOW MANY DATA POINTS"
INPUT icount
ICONT = 1

PRINT "ENTER TYPE OF FAULT [0=NOFAULT 1=COIL 2=HOTWATER 3=RPM 4=DAMPER] "
INPUT IFAULT

DO WHILE INKEY$ <> "S"

      ' MEASURE ROOM TEMPERATURE
      CALL IBWRT(SCANNER%, "C09E")
      CALL IBWRT(SCAN2%, "Close 300")
      CALL IBWRT(SCAN2%, "Close 409")
      Xwba = VALREAD(DVM%)
      TROOM = TCONV(Xwba) - 273.16
      CALL IBCLR(SCANNER%)
      CALL IBCLR(SCAN2%)
      Troomf = 1.8 * TROOM + 32
      PRINT "ROOM TEMPERATURE:", USING "###.#"; TROOM; Troomf

'      Read Air Temperature dry bulb BEFORE AHU, Tdbo (C)
       CALL IBWRT(SCANNER%, "C12E")
       CALL IBWRT(SCAN2%, "Close 300")
       CALL IBWRT(SCAN2%, "Close 409")
       Xdb1 = VALREAD(DVM%)
       Tdb1 = TCONV(Xdb1) - 273.16
```

```
        CALL IBCLR(SCANNER%)
        CALL IBCLR(SCAN2%)
        CALL IBWRT(SCANNER%, "C13E")
        CALL IBWRT(SCAN2%, "Close 300")
        CALL IBWRT(SCAN2%, "Close 409")
        Xdb2 = VALREAD(DVM%)
        Tdb2 = TCONV(Xdb2) - 273.16
        CALL IBCLR(SCANNER%)
        CALL IBCLR(SCAN2%)

        Tdba = (Tdb1 + Tdb2) / 2
        Tdbaf = 1.8 * Tdba + 32

'       Reading wet bulb temperature BEFORE AHU, Twbo (C)
        CALL IBWRT(SCANNER%, "C14E")
        CALL IBWRT(SCAN2%, "Close 300")
        CALL IBWRT(SCAN2%, "Close 409")
        Xwba = VALREAD(DVM%)
        Twba = TCONV(Xwba) - 273.16
        Twbaf = 1.8 * Twba + 32
        CALL IBCLR(SCANNER%)
        CALL IBCLR(SCAN2%)

        PRINT "Return Air Tdb Twb :", USING "###.#"; Tdba; Twba
        PRINT ":                  :", USING "####.#"; Tdbaf; Twbaf
' Measuring AIR STATION Air Temperatures AFTER AHU (C)
'       Read Air Tempearture dry bulb, Tdbaa (C)
        CALL IBWRT(SCANNER%, "C15E")
        CALL IBWRT(SCAN2%, "Close 300")
        CALL IBWRT(SCAN2%, "Close 409")
        Xdb3 = VALREAD(DVM%)
        Tdb3 = TCONV(Xdb3) - 273.16
        CALL IBCLR(SCANNER%)
        CALL IBCLR(SCAN2%)

        CALL IBWRT(SCANNER%, "C16E")
        CALL IBWRT(SCAN2%, "Close 300")
        CALL IBWRT(SCAN2%, "Close 409")
        Xdb4 = VALREAD(DVM%)
        Tdb4 = TCONV(Xdb4) - 273.16
        CALL IBCLR(SCANNER%)
        CALL IBCLR(SCAN2%)
        Tdbaa = (Tdb3 + Tdb4) / 2
        Tdbaaf = 1.8 * Tdbaa + 32
'       Read Air Temperature wet bulb, Twbaa (C)
        CALL IBWRT(SCANNER%, "C17E")
        CALL IBWRT(SCAN2%, "Close 300")
        CALL IBWRT(SCAN2%, "Close 409")
        Xwbi = VALREAD(DVM%)
        Twbaa = TCONV(Xwbi) - 273.16
        Twbaaf = 1.8 * Twbaa + 32
        CALL IBCLR(SCANNER%)
        CALL IBCLR(SCAN2%)
```

```
        PRINT "Air Station Tdb Twb :", USING "###.#"; Tdbaa; Twbaa
        PRINT ":                   :", USING "####.#"; Tdbaaf; Twbaaf
        PRINT

' Measuring Air Flow rate
'       air velocity,  Avel(m/s), FPM(ft/min), DP(in. W.C.)
        CALL IBWRT(SCANNER%, "C56E")
        CALL IBWRT(SCAN2%, "Close 401")
        sumair = 0
        FOR II% = 1 TO 5
         Xair(II%) = VALREAD(DVM%)
         sumair = sumair + Xair(II%)
        NEXT II%
        CALL IBCLR(SCANNER%)
        CALL IBCLR(SCAN2%)
'mean of the collected 5 airflow rate readings
        meanair = sumair / 5
'Calculation of the standard deviation of 5 readings
         skb = 0
        FOR K% = 1 TO 5
         skb = skb + (Xair(K%) - meanair) ^ 2
        NEXT K%
        airstd = SQR(skb / (5 - 1))
'        PRINT meanair; airstd

'       Differential pressure in inches of water column
        dp = 6.3379E-04 + .26892 * meanair
        IF dp < 0 THEN
        dp = .0001
        END IF
        AVEL = 7165 / 2 * .00508 * (dp ^ .5)
        FPM = AVEL / .00508
        CFM = 2 * FPM
        CALL IBCLR(SCANNER%)
        CALL IBCLR(SCAN2%)
        PRINT "Air Flow Rate CFM: ", USING "#####.#"; CFM

' Meauring Fan pressure rise, DELP (in. W.C.)
       CALL IBWRT(SCANNER%, "C45E")
       CALL IBWRT(SCAN2%, "Close 400")
       XPSUM = 0
       FOR IK% = 1 TO 10
        XP = VALREAD(DVM%)
        XPSUM = XPSUM + XP
       NEXT IK%
       XP = XPSUM / 10
       'DELP = (-53.83 + 35.943 * XP) / 200
       DELP = .176484 * XP - .292443
       CALL IBCLR(SCANNER%)
       CALL IBCLR(SCAN2%)
       PRINT "Static Pressure Rise Across Fan inWC: ", USING "###.####"; DELP
       PRINT
```

```
LOOP

IBE = 1

FOR IMEASURE% = 1 TO icount

tbegin = TIMER
ICONT = 1

PRINT TIME$

' MEASURING AIR TEMPERATURE BEFORE REFR COIL

10
        CALL IBWRT(SCANNER%, "C24E")
        CALL IBWRT(SCAN2%, "Close 301")
        CALL IBWRT(SCAN2%, "Close 409")
        X1 = VALREAD(DVM%)
        CALL IBCLR(SCANNER%)
        CALL IBCLR(SCAN2%)

        CALL IBWRT(SCANNER%, "C26E")
        CALL IBWRT(SCAN2%, "Close 301")
        CALL IBWRT(SCAN2%, "Close 409")
        X2 = VALREAD(DVM%)
        CALL IBCLR(SCANNER%)
        CALL IBCLR(SCAN2%)

        CALL IBWRT(SCANNER%, "C28E")
        CALL IBWRT(SCAN2%, "Close 301")
        CALL IBWRT(SCAN2%, "Close 409")
        X3 = VALREAD(DVM%)
        CALL IBCLR(SCANNER%)
        CALL IBCLR(SCAN2%)

        CALL IBWRT(SCANNER%, "C30E")
        CALL IBWRT(SCAN2%, "Close 301")
        CALL IBWRT(SCAN2%, "Close 409")
        X4 = VALREAD(DVM%)
        CALL IBCLR(SCANNER%)
        CALL IBCLR(SCAN2%)

        XAVG = (X1 + X2 + X3 + X4) / 4

        TAIRIN = TCONV(XAVG) - 273.16

        Terr = ABS(TAIRIN - Tdba)

        IF Terr > 20 THEN
PRINT , " ERR IS LARGER THAN 3 TAIRIN Tdba: ", USING "####.#"; TAIRIN; Tdba; Terr
        GOTO 10
        END IF
'        PRINT , "TAIRIN Tdba: ", USING "####.#"; TAIRIN; Tdba; Terr
```

```
'       Read Air Temperature AFTER HOT WATER COIL(C)
        CALL IBWRT(SCAN2%, "Close 304")
        CALL IBWRT(SCAN2%, "Close 409")
        Xdbo = VALREAD(DVM%)
        TAIRO = TCONV(Xdbo) - 273.16
        CALL IBCLR(SCANNER%)
        CALL IBCLR(SCAN2%)

'       Read Air Temperature BEFORE STEAM COIL(C)
'        CALL IBWRT(SCAN2%, "Close 305")
'        CALL IBWRT(SCAN2%, "Close 409")
'        Xdbo = VALREAD(DVM%)
'        TST = TCONV(Xdbo) - 273.16
'        CALL IBCLR(SCANNER%)
'        CALL IBCLR(SCAN2%)

'       Read Air Temperature BEFORE CHILLED WATER COIL(C)
'        CALL IBWRT(SCAN2%, "Close 306")
'        CALL IBWRT(SCAN2%, "Close 409")
'        Xdbo = VALREAD(DVM%)
'        TBCW = TCONV(Xdbo) - 273.16
'        CALL IBCLR(SCANNER%)
'        CALL IBCLR(SCAN2%)

'  Reading temperature after the chilled water coil
'        CALL IBWRT(SCAN2%, "Close 307")
'        CALL IBWRT(SCAN2%, "Close 409")
'        XTacw = VALREAD(DVM%)
'        TACW = TCONV(XTacw) - 273.16
'        CALL IBCLR(SCAN2%)

'        PRINT "CHILLED WATER COIL:", TBCW, TACW

' Measuring Air Flow rate
'       air velocity,  Avel(m/s), FPM(ft/min), DP(in. W.C.)
        CALL IBWRT(SCANNER%, "C56E")
        CALL IBWRT(SCAN2%, "Close 401")
        sumair = 0
        FOR II% = 1 TO 10
         Xair(II%) = VALREAD(DVM%)
         sumair = sumair + Xair(II%)
        NEXT II%
        CALL IBCLR(SCANNER%)
        CALL IBCLR(SCAN2%)
'mean of the collected 10 airflow rate readings
        meanair = sumair / 10
'Calculation of the standard deviation of 10 readings
         skb = 0
        FOR K% = 1 TO 10
         skb = skb + (Xair(K%) - meanair) ^ 2
        NEXT K%
```

```
        airstd = SQR(skb / (10 - 1))
'        PRINT meanair; airstd

'       Differential pressure in inches of water column
        dp = 6.3379E-04 + .26892 * meanair
        IF dp < 0 THEN
        dp = .0001
        END IF
        AVEL = 7165 / 2 * .00508 * (dp ^ .5)
        FPM = AVEL / .00508
        CFM = 2 * FPM
        CALL IBCLR(SCANNER%)
        CALL IBCLR(SCAN2%)

' Measuring Hot Water temperatures and Flow rate
'       Hot water inlet, Thwi (C)
       CALL IBWRT(SCANNER%, "C18E")
       CALL IBWRT(SCAN2%, "Close 300")
       CALL IBWRT(SCAN2%, "Close 409")
       X19 = VALREAD(DVM%)
       Thwi = TCONV(X19) - 273.16
       CALL IBCLR(SCANNER%)
       CALL IBCLR(SCAN2%)

'       Hot Water outlet, Thwo (C)
       CALL IBWRT(SCANNER%, "C19E")
       CALL IBWRT(SCAN2%, "Close 300")
       CALL IBWRT(SCAN2%, "Close 409")
       X20 = VALREAD(DVM%)
       Thwo = TCONV(X20) - 273.16
       CALL IBCLR(SCANNER%)
       CALL IBCLR(SCAN2%)

'       Hot Water Flow rate, Qhot (l/s)
        CALL IBWRT(SCAN2%, "Close 408")
        XHOT = VALREAD(DVM%)
        QHOT = .3475949041# * XHOT + .0039688188#
        CALL IBCLR(SCAN2%)

'       Fan Power Consumption
        sumpow = 0
        CALL IBWRT(SCAN2%, "Close 402")
        FOR II% = 1 TO 50
         xpow(II%) = VALREAD(DVM%)
         sumpow = sumpow + xpow(II%)
        NEXT II%
        CALL IBCLR(SCANNER%)
        CALL IBCLR(SCAN2%)
'mean of the collected 50 power readings
        meanp = sumpow / 50
        fpow = 40559 * meanp - 79.868
'Calculation of the standard deviation of 50 readings
 skb = 0
```

```
 FOR K% = 1 TO 50
  skb = skb + (xpow(K%) - meanp) ^ 2
 NEXT K%
 powstd = SQR(skb / (50 - 1))
'PRINT meanp; powstd

' Measuring Fan Speed, RPM (rpm)
  FRPM = VALCONT(COUNTER%)
' PRINT "RPM at main: ", USING "#####.##"; FRPM

' Meauring Fan pressure rise, DELP (in. W.C.)
       CALL IBWRT(SCANNER%, "C45E")
       CALL IBWRT(SCAN2%, "Close 400")
       XPSUM = 0
       FOR IK% = 1 TO 20
        XP = VALREAD(DVM%)
        XPSUM = XPSUM + XP
       NEXT IK%

       XP = XPSUM / 20

       'DELP = (-53.83 + 35.943 * XP) / 200
       DELP = .176484 * XP - .292443
       CALL IBCLR(SCANNER%)
       CALL IBCLR(SCAN2%)

       ' MEASURE ROOM TEMPERATURE
       CALL IBWRT(SCANNER%, "C09E")
       CALL IBWRT(SCAN2%, "Close 300")
       CALL IBWRT(SCAN2%, "Close 409")
       Xwba = VALREAD(DVM%)
       TROOM = TCONV(Xwba) - 273.16
       CALL IBCLR(SCANNER%)
       CALL IBCLR(SCAN2%)

       PRINT " DATA POINT# "; IBE; TIME$; "   TROOM"; USING "###.#"; TROOM

'       Read Air Temperature dry bulb BEFORE AHU, Tdbo (C)
        CALL IBWRT(SCANNER%, "C12E")
        CALL IBWRT(SCAN2%, "Close 300")
        CALL IBWRT(SCAN2%, "Close 409")
        Xdb1 = VALREAD(DVM%)
        Tdb1 = TCONV(Xdb1) - 273.16
        CALL IBCLR(SCANNER%)
        CALL IBCLR(SCAN2%)
        CALL IBWRT(SCANNER%, "C13E")
        CALL IBWRT(SCAN2%, "Close 300")
        CALL IBWRT(SCAN2%, "Close 409")
        Xdb2 = VALREAD(DVM%)
        Tdb2 = TCONV(Xdb2) - 273.16
        CALL IBCLR(SCANNER%)
        CALL IBCLR(SCAN2%)
```

```
        Tdba = (Tdb1 + Tdb2) / 2

'       Reading wet bulb temperature BEFORE AHU, Twbo (C)
        CALL IBWRT(SCANNER%, "C14E")
        CALL IBWRT(SCAN2%, "Close 300")
        CALL IBWRT(SCAN2%, "Close 409")
        Xwba = VALREAD(DVM%)
        Twba = TCONV(Xwba) - 273.16
        CALL IBCLR(SCANNER%)
        CALL IBCLR(SCAN2%)

' PRINT "Air station Tdb Twb:", USING "###.##"; Tdbaa; Twbaa
 PRINT #2, USING "#####"; IFAULT;
 PRINT #2, USING "#######"; TIMER;
 PRINT #2, USING "####.#"; TROOM; Tdba; Twba; TAIRIN; TAIRO; Thwi; Thwo;
 PRINT #2, USING "##.####"; QHOT; DELP;
 PRINT #2, USING "#####.#"; CFM; FRPM; fpow
 PRINT , "Fault:", USING "#####"; IFAULT;
 PRINT , "Timer:", USING "#######"; TIMER
 PRINT , "  Tdba  Twba TAIRIN TAIRO Thwi  Thwo "
 PRINT , USING "####.#"; Tdba; Twba; TAIRIN; TAIRO; Thwi; Thwo
 PRINT , " English Units"
 Tdbaf = 1.8 * Tdba + 32
 Twbaf = 1.8 * Twba + 32
 TAIRINf = 1.8 * TAIRIN + 32
 TAIROf = 1.8 * TAIRO + 32
 Thwif = 1.8 * Thwi + 32
 Thwof = 1.8 * Thwo + 32
 PRINT , USING "####.#"; Tdbaf; Twbaf; TAIRINf; TAIROf; Thwif; Thwof

 PRINT , "QH2O DELP "; USING "##.####"; QHOT; DELP
 PRINT , "CFM RPM POW "; USING "#####.#"; CFM; FRPM; fpow

 DO WHILE ABS(TIMER - tbegin) < tmeas
 ICONT = ICONT + 1
 LOOP

 IBE = IBE + 1
 NEXT IMEASURE%
 CLOSE
 END

DEFINT I-N
SUB INIT (VDEAD, UDEAD, HDEAD, SDEAD)
 VDEAD = 0
 UDEAD = 0
 HDEAD = 0
 SDEAD = 0
END SUB

DEFDBL I-N
FUNCTION MFCONV (NUM%, VOLT)
```

```
IF VOLT < 0 THEN LET VOLT = 0
SELECT CASE NUM%
  CASE 1
    'Converts voltage to total air volumetric flow rate (CFM).
     MFCONV = 7165 * (VOLT * .055) ^ .5

  CASE 2
    'Converts voltage to refrigerant mass flow rate(kg/min).
     MFCONV = (VOLT / .25063 * 1.625 - 6.5) * .45359

  CASE 3
    'Converts voltage to cooling water mass flow rate (kg/min).
     MFCONV = (15.1511 * (VOLT * 1000 / 7.5443) ^ .50193) * 3.77883

END SELECT
END FUNCTION

' The following function converts the given voltage from a specified
' pressure transducer to pressure.
FUNCTION PCONV (NUM%, VOLT)
  SELECT CASE NUM%
    CASE 1
      'Differential pressure transducer (P7D+50 PSIA)
       PCONV = (-.045625 + 5.0188 * VOLT) * 6.8948

    CASE 2
      'Differential pressure transducer (DP30-0015-111)
       PCONV = (-.033114 + 1.5052 * VOLT) * 6.8948

    CASE 3
      'Differential pressure transducer (DP30-0010-111)
       PCONV = (.060604 + .99678 * VOLT) * 6.8948

    CASE 4
      'Pressure transducer (PX 304-300)
      ' PCONV = (1.3548 + 2982.7 * VOLT) * 6.8948
      ' New pressure transducer for condenser side 11/26/96
      ' Setra S/N 424032
      PCONV = -15.8381 + 690.4486 * VOLT

    CASE 5
      'Pressure transducer (SN 6008 PLC)
      'PCONV = (.81424 + 1992.5 * VOLT) * 6.8948
      ' New evaporator side pressure transducer 11/26/96
      ' Setra S/N 293209
      PCONV = -27.297 + 344.421 * VOLT

     CASE 6
      PCONV = 33.577 + 20.581 * 1000 * VOLT

     CASE 7
      PCONV = 8.9717 + 13.765 * 1000 * VOLT
  END SELECT
```

```
END FUNCTION

FUNCTION TCONV (XVL)

' Voltage signal to temperature conversion
' Constants for thermocouple temperature conversion are defined:

    A0 = .10086091#
    A1 = 25727.94369#
    A2 = -767345.8295#
    A3 = 78025595.81#
    A4 = -9247486589#
    A5 = 697688000000#
    A6 = -26619200000000#
    A7 = 394078000000000#
' Voltage is converted to temperature in degrees C for an initial guess:
    TEMP = (A4 + XVL * (A5 + XVL * (A6 + XVL * A7)))
    TEMP = (A2 + XVL * (A3 + XVL * TEMP))
    TEMP = (A0 + XVL * (A1 + XVL * TEMP))

    TCONV = TEMP + 273.15

END FUNCTION

' Valcont function takes reads counter value and
' returns numeric value of RPM
FUNCTION VALCONT (COUNTER%)
  VONE$ = SPACE$(20)
  XCSUM = 0
  'PRINT "COUNTER VALUE "; COUNTER%
  FOR J = 1 TO 10
   TWAIT = .2
   TIME.START = TIMER
   CALL IBTRG(COUNTER%)
   IF (TIMER - TIME.START) < TWAIT THEN
    CALL IBRD(COUNTER%, VONE$)
   ' PRINT "AT READING COUNTER"
   ' INPUT K
   END IF
   'PRINT "VONE$:"; VONE$
   'INPUT K
   X$ = MID$(VONE$, 9, 11)
   XC = VAL(X$)
   IF J = 1 THEN
    XC = 0
   END IF

   XCSUM = XCSUM + XC
   CALL IBCLR(COUNTER%)
 ' PRINT J;
 ' PRINT X$;
 ' PRINT XC
 ' PRINT
```

```
 NEXT J
  XC = XCSUM / 9
 ' PRINT "FINAL MEASURE "; XC
 ' PRINT "RPM MEASURED: "; XC * 20
  VALCONT = XC * 20
END FUNCTION

FUNCTION VALREAD (DVM%)
    TWAIT = .2
'  time.start = TIMER: DO: LOOP WHILE (TIMER - time.start) < TWAIT
    CALL IBTRG(DVM%)
    VONE$ = SPACE$(15)
'    time.start = TIMER: DO: LOOP WHILE (TIMER - time.start) < TWAIT
    CALL IBRD(DVM%, VONE$)
    VALREAD = VAL(VONE$)
'    PRINT VAL(VONE$)
END FUNCTION
```

MATLAB program

Following are the MATLAB programs for discrimination and classification analysis.

```
% This program is for classification and discrimination for AHU
% Written by Kyung Jang
clear all; close all;
kmenu = menu('Testing for','ahgroup','test group');
if kmenu == 1
   load f:\phdwork\data\ahgroup.dat ; rawdat = ahgroup ;
elseif kmenu == 2
   load f:\phdwork\data\ahgroup_test.dat ; rawdat = ahgroup_test ;
end
x = rawdat(:, 1) ; k = 4 ;
tkmenu = menu('fullset?','yes','no');
if tkmenu == 1
   Y = rawdat(:, 6:14) ;
elseif tkmenu == 2
   %    Qwater       Twout       Twin        CFM          pow
   Y = [rawdat(:,10) rawdat(:,9) rawdat(:,8) rawdat(:,12) rawdat(:,14)];
end
load train_disc;
ylam = input('Enter power to use: ');
Y = 1/ylam*(Y.^ylam - 1);
[r c] = size(Y);
xones = ones(r,1);
Xmean = xones*xmean;
Y = (Y - Xmean)./(xones*xstd);
load discrimx;
Z1 = Y * Gam1 ;          % canonical discriminant function scores
Ntot = r;
clasfun = zeros(Ntot, k) ;  % matrix for k classification functions
for j = 1:k ;               % loop over groups
   clasfun(:, j) = (Z1 - ones(Ntot, 1)*(nu(j, :))/2 ) ...
      * nu(j, :)'  + prior(j)  ;  % evaluate classification functions
```

```
   post = exp(clasfun) ./(sum(exp(clasfun)')' * ones(1, k)) ;
   % evaluate posterior probabilities
end ;          % end of loop
[sortord, sortind] = sort(post') ;
xhat = sortind(k, :)'          % predicted group membership
%plot the predicted group membership by classification function
plot(Z1(1:51,1),Z1(1:51,2),'o') % plot normal groups
hold on
plot(Z1(52:171,1),Z1(52:171,2),'x') % plot fault group 2
hold on
plot(Z1(172:311,1),Z1(172:311,2),'s') % plot fault group 3
hold on
plot(Z1(312:431,1),Z1(312:431,2),'^') % plot fault group 4

% plot of prediction comparison
figure
plot((1:r),x,'.',(1:r),xhat,'x') % plot normal groups
legend('test data','predicted')
xlabel('Observation number')
ylabel('group membership')
print -deps classfy.eps
%hold on
%plot(x(52:171),xhat(52:171),'x') % plot group 2
%hold on
%plot(x(172:311),xhat(171:311),'s') % plot group 2
%hold on
%plot(x(312:431),xhat(312:431),'s') % plot group 2
clear all; close all;
load f:\phdwork\data\ahgroup.dat ; rawdat = ahgroup ;
x = rawdat(:, 1) ; k = 4 ;
kmenu = menu('Use fullset?','yes','no')
if kmenu == 1
   Y = rawdat(:, 6:14) ;
elseif kmenu == 2
   %    Qwater       Twout       Twin        CFM          pow
  Y = [rawdat(:,10) rawdat(:,9) rawdat(:,8) rawdat(:,12) rawdat(:,14)];
end

ylam = input('Enter power to use: ');
y = 1/ylam*(Y.^ylam - 1);

Y = standardize(y);
prior = zeros(k,1) / 3 ;
[Within, Between, Gam1, lambda1, post, xhat, prior, nu] =
  candisc(Y, k, x, prior) ;
save discrimx nu Gam1 prior;
% This program utilizes the linear discriminant function
% for training between 2 groups  7/18/98

% Input data are defined as follows
% X1 = group1 vector
% X2 = group2 vector
```

```
% Output data are defined as follows
% b = linear coefficient vector of discriminant function
% D = Multivariate standard distance

close all
clear all
dir('f:\research\*.m')
kmenu = menu('Select type of data','Flury','Research','Other file')
disp('Loading a Training data')
if kmenu == 1
   flmenu = menu('Select the data','flee beetle Tab5.3.2',...
'Electrode Tab5.3.5','Midge Tab1.3','Turtle Tab1.4','Microtus Tab5.4.1')
   if flmenu == 1
      load f:\Academic_by_author\flury\tab53_2.dat;
      Xx = tab53_2;
      disp('There are total of 5 variables including group class')
      T1 = 'HOleracea';
      T2 = 'HCarduorum';
   elseif flmenu == 2
      load f:\Academic_by_author\flury\tab53_5.dat;
      Xx = tab53_5;
      disp('There are total of 6 variables including group class')
      T1 = 'Machine_1';
      T2 = 'Machine_2';
   elseif flmenu == 3
      load f:\Academic_by_author\flury\midge.dat;
      Xx = midge;
      disp('Midge data have 3 variables including group class')
      T1 = 'Af';
      T2 = 'Apf';
   elseif flmenu == 4
      load f:\Academic_by_author\flury\tab1_4.dat;
      Xx = tab1_4;
      disp('Turtle data have 4 variables including group class')
      T1 = 'Male';
      T2 = 'Female';
   elseif flmenu == 5
      load f:\Academic_by_author\flury\tab54_1.dat;
      Xx = tab54_1;
      disp('Microtus data have 9 variables including group class')
      T1 = 'MMultiplex';
      T2 = 'MSubterraneus';
   end
   ktrain = 4;
elseif kmenu == 2
   callahu_fault;
   ktrain = menu('Train on what ?','Norm vs RPM','Norm vs Valve','Norm vs Coil')
   if ktrain == 1
      Xx = [N_type;R_type];
      T1 = 'Normal';
      T2 = 'RPMFaults';
   elseif ktrain == 2
      Xx = [N_type;V_type];
```

```
      T1 = 'Normal';
      T2 = 'ValveFaults';
   elseif ktrain == 3
      Xx = [N_type;C_type];
      T1 = 'Normal';
      T2 = 'CoilFaults';
   end
elseif kmenu == 3
   num_var = input('How many variables including group id: ');
%   Teststring = 'f:\Train_data\ncombo.txt'
%   Teststring = 'f:\Multivar\steel.dat'
   Teststring = input('Enter the file name with suffix: ','s');
   A = read_file(Teststring,num_var);
% fid = fopen('e:\Train_rep1\tst1.txt')
%fid = fopen(str(Teststring))
%A = fscanf(fid,'%f%f%f%f%f%f%f%f%f%f%f%f%f%f%f%f',[14,inf]);
%A = A';
   Xx = A; %[A(:,6:14)]
   T1 = 'Group1';T2 = 'Group2';
   ktrain = 4;
end
[rowt,colt] = size(Xx)
fprintf('\n Total data has (row: %i) (col: %i)',rowt,colt)
true = 0;
while true == 0
   if kmenu == 1 | kmenu == 3
   icol = input('\n\nHow many variables to be used include group id: ');
      for i = 1:icol,
         ival(i) = input('Enter the variable column number used: ');
      end
   else
      ksmenu = menu('Select Variable combination',...
'Full 6 7 8 9 10 11 12 13 14','6 7 8 9 10 11 12 13', ...
         '6 7 8 9 10 11 12','6 7 8 9 10 11','6 7 8 9 10'...
,'6 7 8 9','6 7 8','6 7','6','7','7 8', ...
         '7 8 9','7 8 9 10','1 7 9 10 11 13','Other')
      if ksmenu == 1
         icol = 10;
         ival = [1     6     7     8     9     10    11    12    13    14];
      elseif ksmenu == 2
         icol = 9;
         ival = [1     6     7     8     9     10    11    12    13];
      elseif ksmenu == 3
         icol = 8;
         ival = [1     6     7     8     9     10    11    12];
      elseif ksmenu == 4
         icol = 7;
         ival = [1     6     7     8     9     10    11];
      elseif ksmenu == 5
         icol = 6;
         ival = [1     6     7     8     9     10];
      elseif ksmenu == 6
         icol = 5;
```

```
      ival = [1     6     7     8     9];
   elseif ksmenu == 7
      icol = 4;
      ival = [1     6     7     8];
   elseif ksmenu == 8
      icol = 3;
      ival = [1     6     7];
   elseif ksmenu == 9
      icol = 2;
      ival = [1     6];
   elseif ksmenu == 10
      icol = 2;
      ival = [1     7];
   elseif ksmenu == 11
      icol = 3;
      ival = [1     7     8];
   elseif ksmenu == 12
      icol = 4;
      ival = [1     7     8     9];
   elseif ksmenu == 13
      icol = 5;
      ival = [1     7     8     9    10];
   elseif ksmenu == 14
      icol = 6;
      ival = [1  7  9  10  11  13];
   elseif ksmenu == 15
      icol = input('Enter total variable used for analysis including group id: ')
      for i = 1:icol,
         ival(i) = input('Enter the variable column number used: ');
      end
   end
end
i_f = 1;
while i_f <= icol
   X(:,i_f) = Xx(:,ival(i_f));
   i_f = i_f + 1;
end

[row col] = size(X);
fprintf('\n training samples (rows: %i)   (columns: %i)',row,col)
x = Xx(:,1);
Y = Xx(:, 2:icol);

X1 = Y(find(x==1),:);
[r1 c1]=size(X1);
fprintf('\n Group 1 (rows: %i) (column: %i)',r1,c1)

X2 = Y(find(x==2),:);
[r2 c2]=size(X2);
fprintf('\n Group 2 (rows: %i) (column: %i)\n\n',r2,c2)

if r1 == 0
   fprintf(' No group membership found for 1st group\n');
```

```
      igroup = input('Enter the 1st group membership id: ')
      X1 = Y(find(x==igroup),:);
      [r1 c1]=size(X1);
      fprintf('\n Group 1 (rows: %i) (column: %i)\n',r1,c1)
   end
   if r2 == 0
      fprintf(' No group membership found for the 2nd group\n');
      igroup = input('Enter the 2st group membership id: ')
      X2 = Y(find(x==igroup),:);
      [r2 c2]=size(X2);
      fprintf('\n Group 2 (rows: %i) (column: %i)\n\n',r2,c2)
   end
% estimate groupwise sample covariance matrices of the data vectors X1 and X2
disp('       Groupwise sample covariance matrices');
disp('=============================================================')
X1cov = cov(X1)
X2cov = cov(X2)

% estimate correlation matrix
disp('       Groupwise sample correlation matrices');
disp('=============================================================')
X1cor = corrcoef(X1cov)
X2cor = corrcoef(X2cov)

% estimate sample mean of the data vectors X1 and X2
X1mean = mean(X1);
X2mean = mean(X2);

X1std = std(X1);
X2std = std(X2);

disp('       Pooled covariance matrix')
disp('=============================================================')
S = 1/(r1+r2-2)*( (r1-1)*X1cov + (r2-1)*X2cov )
Sinv = inv(S);

% Univariate pooled variance
Suni = sqrt(1/(r1+r2-2)*( (r1-1)*X1std.^2 + (r2-1)*X2std.^2 ));

%vector of difference of group mean
d = X1mean' - X2mean';

%coefficient of the linear discriminant function
b = Sinv*d;
b1 = inv(X1cov)*d;
b2 = inv(X2cov)*d;

%multivariate standard distance between 2 groups with respect to
% 1) pooled covariance matrices D
% 2) group1 covariance matrice  D1
% 3) group2 covariance matrice  D2
D = sqrt(d'*Sinv*d);
D1 = sqrt(d'*inv(X1cov)*d);
```

```
D2 = sqrt(d'*inv(X2cov)*d);

%Univariate standard distance
Duni = abs(d)./Suni';

disp('================================================================')
disp('Univariate Summary Statistics');
disp('================================================================')
disp('   X1mean,    X2mean,     X1std,     X2std,     Std Distance');
[X1mean', X2mean', X1std', X2std', Duni]

%optimal linear combination of the variable, linear discriminant function
V1 = X1*b;
V2 = X2*b;
V11 = X1*b1;
V12 = X2*b1;
V21 = X1*b2;
V22 = X2*b2;

%lets attempt classification using a heuristic approach
% V group mean and standard deviation
V1mean = mean(V1);
V2mean = mean(V2);
V1std = std(V1);
V2std = std(V2);
disp('Half distance of the group means')
m = (V1mean + V2mean)/2;
disp('Normal Theory Classification rule')
% Classify into Group 1 if b'Y' - 1/2b'(y1mean - y2mean) + log(pi1/pi2) > 0
% Classify into Group 2 if b'Y' - 1/2b'(y1mean - y2mean) + log(pi1/pi2) < 0
N1 = r1;
N2 = r2;

disp('If the training samples were drawn from the mixture and not
conditional distribution ') disp('then it is reasonable to assume
prior probabilities by the relative frequencies') disp('Now assume a
prior probabilities and that the samples are from normal density')

%pi_1 = 0.90115;
%pi_2 = 0.09885;
pi_1 = N1/(N1 + N2);
pi_2 = N2/(N1 + N2);
if ktrain == 1
   save lin_disc_func_nor b -ascii -double
   save gmean_nor m -ascii -double
elseif ktrain == 2
   save lin_disc_func_val b -ascii -double
   save gmean_val m -ascii -double
elseif ktrain == 3
   save lin_disc_func_col b -ascii -double
   save gmean_col m -ascii -double
elseif ktrain == 4
   save flury_b b -ascii -double
```

```
   save flury_m m -ascii -double
end
fprintf('\n============================================================\n')
disp(' Multivariate summary statistics')
disp('============================================================')
fprintf('  1) Multivariate standard distance using pooled covariance D: %7.2f\n',D)
fprintf('  2)                              group1 covariance matrice D1: %7.2f\n',D1)
fprintf('  3)                              group2 covariance matrice D2: %7.2f\n',D2)
fprintf('\n\n Group 1 Sample size is: %5i\n',N1)
fprintf(' Group 2 Sample size is: %5i\n',N2)
fprintf('\n Group 1 Prior probabilities pi_1 is: %10.5f\n',pi_1)
fprintf(' Group 2 prior probabilities pi_2 is: %10.5f\n',pi_2)
fprintf('\n  Coefficient of the linear discriminant function b:\n')
fprintf('%10.3f',b)
fprintf('\n  Coefficient of the lin discrim function Group 1 covariance b1:\n')
fprintf('%10.3f',b1)
fprintf('\n  Coefficient of the lin discrim function Group 2 covariance b2:\n')
fprintf('%10.3f',b2)
fprintf('\n\n  V,Disc function, Means and standard deviations\n')
fprintf(' Group%3i %10.3f %10.3f\n',1,V1mean,V1std)
fprintf(' Group%3i %10.3f %10.3f\n',2,V2mean,V2std)
fprintf('\n\n')

% A simple classification rule could be, since V1mean > V2mean
% Assign to Group 1 if V > m
%           Group 2 if V < m

%if we assume b(3) = b(4) = 0
% then cutoff line for classification is straight line with
% b(1)*x11 + b(2)*x21 + b(3)*x31 + b(4)*x41 = m
%xcut1 = m/b(1) - b(2)/b(1)*xflow - b(3)/b(1)*x31 - b(4)/b(1)*x41;

[xcut,xout,post] = call_classification(V1,V2,X1,X2,X,b,T1,T2);
fprintf('\n Half distance of the group means is: %14.4f\n',m)
fprintf(' Normal Theory Classification cutoff point is: %14.4f\n',-xcut)

/*This figure is an attempt to show that a linear discriminant
function may be uniquely defined, up to multiplication by a constant,
even in cases whre the covariance matrices are distinctly different In
practical applications, one might compute two vectors of discriminant
function coefficients b1=inv(S1)*(x1mean - x2mean) and
b2=inv(S2)*(x1mean-x2mean) to assess the effect of differences between
the covariance matrices on the linear discriminant function.  Then we
can compute the values of V1 and V2 for all observations and study the
joint disribution of V1 and V2 in a scatter plot.  Ideally if the
differences between S1 and S2 do not affect the linear discriminant
function at all, the correlation between V1 and V2 would be 1. */

fprintf('\n\n'); disp('From the Figure 1 we can see that although the
two linear combinations appear to be quite different,'); disp('they
are highly correlated and yield about the same group separtion.');
disp('The Plot suggests that a single linear combination would be
adequate'); figure plot(V11,V21,'o',V12,V22,'d'); xlabel('V1 =
```

```
discriminant function using S1'); ylabel('V2 = discriminant function
using S2'); title('Plot of joint distribution of two linear
discriminant functions'); legend(T1,T2)

for i = 1:c1 xtemp1(:,i) = X1(:,i) - X1mean(i); xtemp2(:,i) = X2(:,i)
   - X2mean(i); end f_1 =
   (2*pi)^(-c1/2)*det(S)^(-1/2)*exp(xtemp1*inv(S)*xtemp1'); f_2 =
   (2*pi)^(-c2/2)*det(S)^(-1/2)*exp(xtemp2*inv(S)*xtemp2'); p1f =
   pi_1*f_1; p2f = pi_2*f_2; fprintf('\n xout = Actual GroupID
   Grp_Score,z GID_hat Post_Prob_1 Post_prob_2\n'); figure
   plot(xout(:,2),post(:,1),'o',xout(:,2),post(:,2),'x') xlabel('z
   (Value of discriminant function)') ylabel('Posterior probability')
   title('Normal theory classification based on the distribution of
   the discriminant function') legend(T1,T2) figure
   plot(V1,p1f(:,1),'o') figure plot(V2,p2f(:,1),'o') fprintf('\n\n')
   disp('Figure 2 shows distribution of the linear discriminant
   function V') fprintf('\n\n') disp('Figures 3,4 show Histogram of
   the Discriminant function') [joe1, joe2] =
   frequency(V1,V2,T1,T2,-xcut);

% Estimate error rate
[error_rate] = err_rate(x,Y,T1,T2);

% Statistical Inference for means
% For Variables X1 and X2

kk = menu('More Training ?','Yes','No');
if kk==1
   true = 0;
   figure
elseif kk==2
   true = 1;
end
end
```

PCA analysis

The following function program performs principal component analysis.

```
function Ystdz = pca_standardize(Y)
% This function standardizes observation in the Y matrix
% Assume Y = [ var_11 var_12 ... var_1k
%              var_21 var_22 ... var_2k
%              ...    ...    ... ...
%              var_n1 ...    ... var_nk]
%
%clear all
%close all
```

```
%format compact
%filest = 'f:\MULTIVAR\sons.dat';num_var = 4;
%A = read_file(filest,num_var);
%Y = [A(:,1) A(:,2)]; %for sons.dat

[n k]=size(Y);
kones = ones(1,n); %row vector of 1xn =[ 1 1 1 1 ..]
Ymean = mean(Y);   %row vector of 1xk
Ymv   = Ymean'*kones ;% mean matrix of kxk
Yd    = Y - Ymv';   %  distance from mean

Ycov  = cov(Y);
Ystd  = sqrt(diag(Ycov)); % Standard deviation column vector of kx1

for ii = 1:k,
   for ij = 1:n,
      Ystdz(ij,ii) = Yd(ij,ii)/Ystd(ii);
   end
end
% Ystdz
% Ystd  = inv(sqrt(Ycov))*(Y'-Ymean')
% PCA analysis
%
% Written By: Jang, Kyung-Jin
% October 1998
%
% subfunctions: pca_standardize.m, collect_ahu_data.m, read_file.m
%
% Data configuration
% Sons.dat format is in columns and has 2 groups with 2 variables.
% Football.dat has row vectors with group id in the 1st column
% 1st PC: z1 = a'y = 0.207 y1 + 0.873 y2 + 0.261 y3 + 0.326 y4 + 0.066
% y5 + 0.128 y6 Research data has column format
%
%
% References: Flury, Rencher

clear all; close all; format compact; format short e

% read file
dir('f:\Train_data\*.txt')
kmenu = menu('Select type of data','Flury','Rencher','Research','Other
file')
disp('Loading a Training data')
if kmenu == 1
   flurymenu = menu('What data ?','Flee
```

```
    beetle','Microtus','Head','Water Strider','turtle shell');
    if flurymenu == 1
       filest = 'f:\data\flury\tab53_2.dat';num_var = 5;
    elseif flurymenu == 2
       filest = 'f:\data\flury\tab54_1.dat';num_var = 9;
    elseif flurymenu == 3
       filest = 'f:\data\flury\tab1_2.dat';num_var = 6;
    elseif flurymenu == 4
       filest = 'f:\data\flury\tab85_3.dat';num_var = 6;
    elseif flurymenu == 5
       filest = 'f:\data\flury\tab1_4.dat';num_var = 4;
    end
elseif kmenu == 2
    renchmenu = menu('What data ?','Sons','football')
    if renchmenu == 1
       filest = 'f:\data\rencher\sons.dat';num_var = 4;
    elseif renchmenu == 2
       filest = 'f:\data\rencher\football.dat';num_var = 7;
    end
elseif kmenu == 3
    research = menu('What data ?','pca.dat','tst1','normal','exploring
    normal fault using rep1 data')
    if research == 1
       filest = 'f:\phdwork\data\pca.dat'; num_var=9;
    elseif research == 2
       filest = 'f:\Train_rep1\tst1.txt'; num_var=14;
    elseif research == 3
       filest = 'f:\Train_data\ncombo.txt'; num_var=14;
    elseif research == 4
       collect_ahu_data
    end
elseif kmenu == 4
    filest = input('Enter the path and filename: ','s')
    num_var = input('Enter total number of variables (include group id)
    ') gp_id = input('Enter group number ID to be selected(ie 1 or 2
    etc) if none enter 0 ') end

if kmenu == 3
    if research == 4
       A = N_type; num_var = 14;
    else
       A = read_file(filest,num_var);
    end
else
    A = read_file(filest,num_var);
end
```

```
% Assign variable y to the observation of interest
[row col] = size(A);

if kmenu == 1
   if flurymenu == 1
      x = A(:,1);
      yp = A(:,2:col);
      x1 = yp(find(x==1),:);
      y = x1; %for group 1 of flury data
   elseif flurymenu == 2
      x = A(:,1);
      yp = A(:,2:4);
      x1 = yp(find(x==1),:);
      y = x1; % for group 1 of microtus data
   elseif flurymenu == 3 | flurymenu == 4
      y = A;
   elseif flurymenu == 5
      x = A(:,1);
      yp = A(:,2:col);
      yp = 10*log(yp);
      x1 = yp(find(x==2),:);
      y = x1 % for group 2(female) of the turtle shell
   end
end
if kmenu == 2
   if renchmenu == 1
      y = [A(:,1) A(:,2)]; %for sons.dat
   end
   if renchmenu == 2
      y = [A(31:60,2:7);A(61:90,2:7)]; %for football.dat
   end
end
if kmenu == 3
   A = log(A);
   if research == 1
      y = A;
   else
      y = [A(:,6:14)]; %for research data
      plot(A(:,2)/60,A(:,14),'o');legend('Power');
      figure
      plot(A(:,2)/60,A(:,6),'o');legend('T air in');
      figure
      rvar = menu('Use ratio combination ?','yes','no');
      if rvar == 1
         x1 = (A(:,7)-A(:,6))./(A(:,8)-A(:,6));  % Temperature ratio
```

```
            x2 = (A(:,9));                    % Temperature of water out
            x3 = (A(:,10));                          % Water flow rate
            x4 = (A(:,12));                          % Air flow rate
            x5 = (A(:,13));                          % RPM
            x6 = (A(:,11));                          % Del P
            x7 = (A(:,14));                          % Power
            y = [x1 x2 x3 x4 x5 x6 x7];
         elseif rvar == 2
         end
      end
   end
   if kmenu == 4
      if gp_id == 0
         y = A;
         x = ones(row,1);
      else
         x = A(:,1);
         yp = A(:,2:col);
         kcheck = menu('Does the data contain multiple group
         id?','yes','no');
         if kcheck == 1
            x1 = yp(find(x==gp_id),:);
            y = x1; %for group id (gp_id)
         elseif kcheck == 2
            y = [A(:,2:col)];
         end
      end
   end
   [r c]=size(y);
   p = c;
   num_va = c;

   plot(y(:,1),y(:,2),'o')
   xlabel('y_1')
   ylabel('y_2')

   fprintf('\n==== Analysis of PCA with out standardizing the variable
   ====\n')

   ybar = mean(y)
   ystd = std(y)
   S_y = cov(y)                              % Covariance Matrix
   R_y = corrcoef(y)
   %inv(diag(diag(S_y)))*S_y*inv(diag(diag(S_y)))  % Correlation Matrix
   ymean = ones(r,1)*ybar;
   gv = det(S_y);tv = trace(S_y);
```

```
fprintf('\n Generalized sample variance det(S_y) = (%f)',gv)
fprintf('\n Total sample variance trace(S_y) = (%f)\n',tv)
fprintf('\n det(R) = (%f)\n',det(R_y))
fprintf('\n Generalized sample variance has a geometric
interpretation.')
fprintf('\n det|S_y| = 0 indicate a redundancy in the form of a linear
relationship among the variables\n')
fprintf('\n Relative large measure of either overall variance, reflect
a broad scatter about the mean')
fprintf('\n for case of |S_y| an extensive scatter maby be masked by
small eigenvalues that reduce |S|')
fprintf('\n A very small value of |S| |R| indicate small scatter or
multicollinearity')
fprintf('\n Multicollinearity maybe due to high pairwise correlations
or to a high multiple correlation')
fprintf('\n between one variable and several of the other variables\n')
fprintf('\n When the variables are highly collinear, then S or R
becomes nearly singular and inverse is unstable')
fprintf('\n Large changes in inv(S) result from minor change in S, In
this case |R| is close to zero')
fprintf('\n this |R| is a measure of the amount of intercorrelation
among the variables\n\n') pause
[eigen_vec,Dlambda]=eig(S_y); %Estimate the eigenvalues and
eigenvectors of Covariance matrix

for i=1:num_va, eigen_val(i) = Dlambda(i,i); end eigen_val eigen_vec
   eigen_non_vec = eigen_vec; eigen_non_val = eigen_val;
[B, lambda, stdeB, stdelam] = lpca(y)
pause
fprintf('\n\n Because the eigenvalues are variances of the principal
components\n');
fprintf(' we can speak of "the Proportion of Variance explained" by
the first k component\n');
lsum_bot = trace(S_y);
Prov = eigen_val./lsum_bot; %Proportion of variance explained
var_no = 1:num_va;
pca_summary = [eigen_val' Prov' var_no'];
% pca_sort = sort(pca_summary);
pca_sortrows = sortrows(pca_summary);

psum = 0;
for i=num_va:-1:1,
   psum = psum + pca_sortrows(i,2);
   cum_pro(i) = psum;
end
```

```
% fprintf('\nNOTE! The 1st largest eigenvalue is at the BOTTOM !\n')
 fprintf('NOTE! Last two PC account for %f of the total
 variance\n',cum_pro(num_va-1))

pca_summary = [pca_sortrows cum_pro']; fprintf('\n PCA Summary\n\n')
fprintf('Variable Eigenvalue Proportion Cumulative\n')
fprintf('Identity of Variance Proportion\n') for i=num_va:-1:1,
fprintf('%f  %f  %f      %f\n',pca_summary(i,3),pca_summary(i,1:2),
pca_summary(i,4))
end
fprintf('\nNOTE! First two PC account for %f percent of the total
variance\n',cum_pro(num_va-1)*100)
fprintf('\nSaving the output to pca1.out\n');
fid = fopen('pca1.out','wt');
fprintf(fid,'\n==== Analysis of PCA with out standardizing the
variable ====\n');
fprintf(fid,' Mean Vector\n');
fprintf(fid,'%f ',ybar);
fprintf(fid,'\n Standard Deviation\n');
fprintf(fid,'%f   ',ystd);
fprintf(fid,'\n Covariance Matrix\n');
for ii = 1:c,
   ssy = num2str(S_y(ii,:));
   fprintf(fid,'%s\n',ssy);
end

fprintf(fid,'\n');
fprintf(fid,' Correlation Matrix\n');
for ici = 1:c,
   sscr = num2str(R_y(ici,:));
   fprintf(fid,'%s\n',sscr);
end

fprintf(fid,'\n');
fprintf(fid,' Eigen Values\n');
fprintf(fid,'%f ',eigen_val);
fprintf(fid,'\n Eigen Vectors\n');
for ii = 1:num_va,
   ssy = num2str(eigen_vec(ii,:));
   fprintf(fid,'%s\n',ssy);
end
fprintf(fid,'\n\n PCA Summary\n\n');
fprintf(fid,'Variable Eigenvalue  Proportion  Cumulative\n');
fprintf(fid,'Identity             of Variance Proportion\n');
for i=num_va:-1:1,
   fprintf(fid,'%f  %f  %f      %f\n',pca_summary(i,3),
```

```
pca_summary(i,1:2),pca_summary(i,4));
end
fprintf(fid,'\nNOTE! First two PC account for %f percent of
the total variance\n',cum_pro(num_va-1)*100);
fclose(fid)
fprintf('\nScree Plot of PCA for non-standardized variables')
xs = c:-1:1;
ys = pca_summary(:,1);
figure
plot(xs,ys,'-o');legend('Eigen Value')
xlabel('Principal Component')
ylabel('Eigen Value')
%
fprintf('\n==== Analysis of PCA with standardizing the variable ====
\n')

Y = pca_standardize(y);
figure
plot(Y(:,1),Y(:,2),'o')
xlabel('y_1')
ylabel('y_2')

ybar = mean(Y)
S_y = cov(Y)
[eigen_vec,Dlambda]=eig(S_y);

for i=1:num_va,
   eigen_val(i) = Dlambda(i,i);
end
eigen_val
eigen_vec

% Because the eigenvalues are variances of the principal components
% we can speak of "the Proportion of Variance explained" by the
first k component
lsum_bot = trace(S_y);
Prov = eigen_val./lsum_bot; % Proportion of variance explained
var_no = 1:num_va;
pca_summary = [eigen_val' Prov' var_no'];
% pca_sort = sort(pca_summary);
pca_sortrows = sortrows(pca_summary);

psum = 0;
for i=num_va:-1:1,
   psum = psum + pca_sortrows(i,2);
   cum_pro(i) = psum;
```

```
end

% fprintf('\nNOTE! The 1st largest eigenvalue is at the BOTTOM !\n')
% fprintf('NOTE! Last two PC account for %f of the total variance\n'
,cum_pro(num_va-1))
pca_summary = [pca_sortrows cum_pro'];
fprintf('\n PCA Summary\n\n')
fprintf('Variable Eigenvalue  Proportion  Cumulative\n')
fprintf('Identity             of Variance Proportion\n')
for i=num_va:-1:1,
   fprintf('%f  %f  %f      %f\n',pca_summary(i,3),
pca_summary(i,1:2),pca_summary(i,4))
end
fprintf('\nNOTE! First two PC account for %f percent of the total
 variance\n',cum_pro(num_va-1)*100)

fprintf('\nScree Plot')
xs = c:-1:1;
ys = pca_summary(:,1);
figure
plot(xs,ys,'-o');legend('Eigen Value')
xlabel('Principal Component')
ylabel('Eigen Value')

fprintf('\nSaving the output to pca2.out\n');
fid = fopen('pca2.out','wt');
fprintf(fid,'\n==== Analysis of PCA with standardizing the variable
 ====\n');
fprintf(fid,' Mean Vector\n');
fprintf(fid,'%f ',ybar);
fprintf(fid,'\n Covariance Matrix\n');
for ii = 1:c,
   ssy = num2str(S_y(ii,:));
   fprintf(fid,'%s\n',ssy);
end

fprintf(fid,'\n');
fprintf(fid,' Eigen Values\n');
fprintf(fid,'%f ',eigen_val);
fprintf(fid,'\n Eigen Vectors\n');
for ii = 1:num_va,
   ssy = num2str(eigen_vec(ii,:));
   fprintf(fid,'%s\n',ssy);
end
fprintf(fid,'\n\n PCA Summary\n\n');
fprintf(fid,'Variable Eigenvalue  Proportion  Cumulative\n');
```

```
fprintf(fid,'Identity              of Variance Proportion\n');
for i=num_va:-1:1,
   fprintf(fid,'%f  %f  %f      %f\n',pca_summary(i,3),
   pca_summary(i,1:2),pca_summary(i,4));
end
fprintf(fid,'\nNOTE! First two PC account for %f percent
 of the total variance\n',cum_pro(num_va-1)*100);
fclose(fid);

% Principal component scores U(i) = B'(X(i) - Xmean), i=1,...,N
U_non = (y-ymean)*eigen_non_vec; [u_r u_c]=size(U_non);
% PCA scores without standardizing
U_std = Y*eigen_vec;  % PCA scores with standardizing
disp('To view PCA scores for nonstandardized score type U_non')
disp('To view PCA scores for    standardized score type U_std')
[U_y, Y_y,proj_matx] = pcscore(eigen_non_vec(:,1),y);
%[U_y, Y_y,proj_matx] = pcscore(eigen_non_vec(:,9),y);
figure;plot(U_non(:,(u_c)), U_non(:,u_c-1),'x');xlabel('U_1');
ylabel('U_2');legend('PC')
% print -depsc2 -adobecset f:\Figures\scatter_pca_U1_U2.eps
figure;plot(U_non(:,(u_c)), U_non(:,u_c-2),'x');xlabel('U_1');
ylabel('U_3');legend('PC')
%  axis([-.8 .8 -.4 .4]); print -depsc2 -adobecset
% f:\Figures\scatter_pca_U1_U3.eps
figure;plot(U_non(:,(u_c-1)), U_non(:,u_c-3),'x');xlabel('U_2');
ylabel('U_3');legend('PC')
%  axis([-.4 .4 -.2 .2]); print -depsc2 -adobecset
% f:\Figures\scatter_pca_U2_U3.eps

% Average eigenvalue
eigen_mean = sum(eigen_non_val)/num_va
disp('in deciding how many component count the component that
 has higher value then eigen_mean')

% carry out the significant tests
u_test_statistic(1) = 0;
kk(1) = 1;
test_menu = menu('I need you to look up Chi^2 values and enter
 them here','Do it ?','no')
for k_i = 2:num_va,
   i_k = num_va-k_i+1;
   lam_mean_sum = 0;
   for k_s = i_k:num_va,
      lam_mean_sum = lam_mean_sum + lambda(k_s);
   end
   lam_mean = lam_mean_sum/k_i;
```

```
   log_lam_mean = log(lam_mean);

   log_lam_sum = 0;
   for k_s = i_k:num_va,
      log_lam_sum = log_lam_sum + log(lambda(k_s));
   end
   u_test_statistic(k_i) = (r - (2*num_va + 11)/6)*(k_i*log_lam_mean
     - log_lam_sum);
   df_u(k_i) = 1/2*(k_i-1)*(k_i+2);
   if test_menu == 1
      fprintf('alpha(0.05) df:(%f) \n',df_u);
      X_2(k_i) = input('Enter X^2 value for above alpha and df : ');
   elseif test_menu == 2
      X_2(k_i) = 0;
   end
   kk(k_i) = k_i;
end

k_i = num_va;
for ik = 1:num_va,
   k(ik) = kk(k_i);
   u_sta(ik) = u_test_statistic(k_i); % statistic to be tested with
 %nu = 1/2(k-1)(k+2) deg freedom chi square.
   df(ik) = df_u(k_i);
   X2(ik) = X_2(k_i);
   k_i = k_i - 1;
end
disp('Eigenvalue    k          u            df
 x^2(0.05)(df)\n');
[lambda k' u_sta' df' X2']
```

Quadratic discrimination function program

Following MATLAB program performs quadratic discriminant analysis.

```
load  f:\data\flury\tab53_5.dat ; rawdat = tab53_5 ;
groups = rawdat(:, 1) ; Y = rawdat(:, [2 6]) ;

Y1 = Y(find(groups == 1), :) ;
N1 = size(Y1, 1) ; ybar1 = mean(Y1)' ; S1 = cov(Y1) ;
Y2 = Y(find(groups == 2), :) ;
N2 = size(Y2, 1) ; ybar2 = mean(Y2)' ; S2 = cov(Y2) ;
pi1 = 0.5 ; pi2 = 0.5 ;
S1inv = inv(S1) ; S2inv = inv(S2) ;
```

```
A1 = -S1inv / 2 ; b1 = S1inv * ybar1 ;
c1 = log(pi1) - log(det(S1))/2 - ybar1' * S1inv * ybar1 / 2 ;
A2 = -S2inv / 2 ; b2 = S2inv * ybar2 ;
c2 = log(pi2) - log(det(S2))/2 - ybar2' * S2inv * ybar2 / 2 ;

q1 = sum(((Y * A1) .* Y)')'  +  Y * b1  +  c1 ;
q2 = sum(((Y * A2) .* Y)')'  +  Y * b2  +  c2 ;

post1 = exp(q1) ./ (exp(q1) + exp(q2)) ;
post2 = 1 - post1 ;
Q = q1 - q2 ;

disp([post1(41:60) post2(41:60)])  ;
load f:\data\flury\tab53_5.dat ; rawdat = tab53_5 ;
x = rawdat(:, 1) ;
Y = rawdat(:, [2 6]) ;
%Choose equal prior probabilities by
prior1 = 0.5 ;
%and then call the qda function:
[qdf, xhat, post1] = qda(x, Y, prior1) ;
% display predicted group membership and posterior
% probabilities %for selected observations (numbers 41 to 60):

[(41:60)'  xhat(41:60)  100*post1(41:60)]
```

Logistic function program

Logistic function subprogram estimates the maximum likelihood estimate, β, using likelihood equations.

```
%  function LOGISTIC
%  Reference: Rencher 1995
%  Some modification is made to show loglikelihood and deviance
%
% input:        X   design matrix, with N rows, assumed to have full
%                   column rank. Note: If you want the model to contain
%                   an intercept term, the first column of X should be
%                   a vector of 1's.
%               y   N-vector, numbers of successes
%               M   N-vector, numbers of trials
%
% output:    beta   estimated logistic regression coefficients
%         Sigma   estimated covariance matrix of the parameter estimates
%         deviance   deviance of the fitted model
```

```
%            prob   N-vector of estimated success probabilities


function [beta, Sigma, deviance, prob] = logistic(X, y, M) ;

[N, k] = size(X) ;  % dimension of design matrix
eps = 10^(-12) ;    % set convergence criterion
diff = 1 ;          % initial test value for convergence
beta = zeros(k, 1); % initialize parameter vector
it = 0 ;            % initialize counter for iterations
while diff > eps ;  % start iterations of Newton-Raphson
prob = exp(X*beta) ./ (1 + exp(X*beta)) ;  % update probabilities
loglik = sum( y .* log(prob) + (M-y) .* log(1-prob) ) ;
                                  % evaluate log-likelihood function
%disp([it beta' loglik]) ;  % display current numerical values
it = it + 1;               % increase iteration counter
betaold = beta;            % store old value of parameter vector
e = M .* prob ;            % estimated expected frequencies
score = X' * (y-e) ;       % score function
Wvect = M .* prob .* (1-prob) ; % diagonal of matrix W
info = X' * ( (Wvect*ones(1,k)) .* X) ; % information function
beta = beta + inv(info) * score ;       % update beta
diff = max(abs(beta-betaold)) ;     % max. difference between old and
                                    % new parameter values
end ;                               % end of iterations
Sigma = inv(info) ;                 % estimated covariance matrix
lmax = sum( log( (y./M).^y)  + log( ((M-y)./M).^(M-y)) ) ; % value of
                          % log-likelihood function for saturated model
deviance = 2 * (lmax - loglik) ;    % deviance
fprintf(' log likelihood: %10.3f\n\n',loglik);
```

Discriminant plots and scatter plot programs

Disc plot function program performs matrix scatter plots for given data matrix. This function also plots discriminant frequency plots used in the two group identification.

```
% function disc_plot
%
% input: kd = number of discriminant functions
%        k  = number of group
%        YY = observation matrix including group membership
%        B  = p by kd eigenvector of discriminant function
%
% output: scatter plots of discriminant function values
```

```
%
% function disc_plot(kd, k, YY, B);

%function disc_plot(kd, k, YY,B);

load f:\phdwork\data\ahgroup.dat; rawdat = ahgroup;
YY = [rawdat(:,1) standardize(rawdat(:,6:14))];
% Total sample standardized canonical coefficients from sas prog
B = [ -0.884146189      -0.718535970      -0.241119080
   1.165119626       0.639065334       0.530151373
   1.913324873      -2.094050549      -1.516111255
   -2.337049569       1.597850082       0.838617280
   2.815969240      -3.031161754      -1.775960924
   0.768593328      -0.755990962       2.957552928
   1.741375174       0.663783856       1.446261733
   1.003168524       2.376422029      -4.238542409
   -1.274973873      -0.922267407       1.530014664];
k = 4;
kd = 3;

[N_y p] = size(YY) ;
x = YY(:, 1); Y = YY(:, 2:p);
% separate the group membership and data matrix

for jk = 1:kd,
   for j = 1:k,
      Yj = Y(find(x==j),:) ; % find the j group
      N(j) = size(Yj,1)    ; % sample size of j group
      mu(j,:) = mean(Yj)   ; % mean vector of j group
   end

   Z(:,jk) = Y*B(:,jk)     ; % discriminant function values
end

ipx = 1;
for ip = 1:kd,
   for ip2 = 1:kd,
      if ip~=ip2
         subplot(kd,kd,ipx),plot(Z(:,ip2),Z(:,ip),'.');grid
      end
      ipx = ipx + 1;
   end
end

for ip = 1:kd,
   xb2 = num2str(ip)
```

```
   st1 = strcat('Z_',xb2);
   gtext(st1);
end

print -depsc2 -adobecset f:\mcodes\disc_scatter.eps
disp('');
disp('File is saved in f:\mcodes\disc_scatter.eps');

% figure
s_mat =[ '^' 's' '+' 'x' ];
%for j = 1:kd-1,
%   x_begin = 1
%   for jj = 1:k,
%      s_plot = input('select plot symbol (o, ^, s, +, x, *)','s');
%      s_plot = s_mat(jj)
%      plot(Z(x_begin:N(jj),j),Z(x_begin:N(jj),j+1),s_plot)
%      hold on;       x_begin = x_begin + N(jj)
%   end
%   xb1 = num2str(j); xb2 = num2str(j+1)
%   st1 = strcat('Z_',xb1); st2 = strcat('Z_',xb2);
%   xlabel(st1)
%   ylabel(st2)
%   figure
%end

figure
plot(Z(1:N(1),1),Z(1:N(1),2),'.')
xlabel('Z_1');ylabel('Z_2');
hold on;
plot(Z(N(1)+1:N(1)+N(2),1),Z(N(1)+1:N(1)+N(2),2),'^')
hold on
plot(Z(1+N(1)+N(2):sum(N(1:3)),1),Z(1+N(1)+N(2):sum(N(1:3)),2),'s')
hold on
plot(Z(1+N(1)+N(2)+N(3):sum(N(1:4)),1),Z(1+N(1)+N(2)+N(3):sum(N(1:4)),
 2),'+')
legend('Normal','Fan','Valve','Coil')

print -depsc2 -adobecset f:\figures\disc_scatter12.eps
disp('');
disp('File is saved in f:\figures\disc_scatter.eps');

figure
plot(Z(1:N(1),1),Z(1:N(1),3),'.')
xlabel('Z_1');ylabel('Z_3');
hold on;
plot(Z(N(1)+1:N(1)+N(2),1),Z(N(1)+1:N(1)+N(2),3),'^')
```

```
hold on
plot(Z(1+N(1)+N(2):sum(N(1:3)),1),Z(1+N(1)+N(2):sum(N(1:3)),3),'s')
hold on
plot(Z(1+N(1)+N(2)+N(3):sum(N(1:4)),1),Z(1+N(1)+N(2)+N(3):sum(N(1:4)),
 3),'+')
legend('Normal','Fan','Valve','Coil')

print -depsc2 -adobecset f:\figures\disc_scatter13.eps
disp('');
disp('File is saved in f:\figures\disc_scatter.eps');

figure
plot(Z(1:N(1),2),Z(1:N(1),3),'.')
xlabel('Z_2');ylabel('Z_3');
hold on;
plot(Z(N(1)+1:N(1)+N(2),2),Z(N(1)+1:N(1)+N(2),3),'^')
hold on
plot(Z(1+N(1)+N(2):sum(N(1:3)),2),Z(1+N(1)+N(2):sum(N(1:3)),3),'s')
hold on
plot(Z(1+N(1)+N(2)+N(3):sum(N(1:4)),2),Z(1+N(1)+N(2)+N(3):sum(N(1:4)),
 3),'+')
legend('Normal','Fan','Valve','Coil')

print -depsc2 -adobecset f:\figures\disc_scatter23.eps
disp('');
disp('File is saved in f:\figures\disc_scatter.eps');
```

Frequency estimation and plot function

This program plots and performs class interval estimation for given matrix input.

```
function [xfreq1,xfreq2] = frequency(V1,V2,T1,T2,xcut)
% Plots and output with row vectors of 2 variables
% inputs are two row vector V1 and V2
% outputs are two matrices each including class and frequencies
% example:
%       X1 = [1 2 3 4 5 6]';
%       X2 = [9 8 7 6 5]';
%       [xr,xs] = frequency(X1,X2)

V1min = min(V1);
V1max = max(V1);
V2min = min(V2);
V2max = max(V2);
```

```
[r1 c1] = size(V1);
[r2 c2] = size(V2);

xmin = min(V1min,V2min);
xmax = max(V1max,V2max);

xrange = xmax - xmin;

subinterval = 20;

xinterval = xrange/subinterval; % this is the class interval width

xbegin = xmin - xinterval/2;
xfinal = xmax + xinterval/2;

clss(1) = xbegin;
for i = 2:subinterval+2,
   clss(i) = clss(i-1) + xinterval;
end

class1 = zeros(subinterval+1,1);
class2 = zeros(subinterval+1,1);

for i = 1:r1,
   for j = 1:subinterval+1,
      if V1(i) > clss(j) & V1(i) < clss(j+1)
         class1(j) = class1(j) + 1;
      end
   end
end
for i = 1:r2,
   for j = 1:subinterval+1,
      if V2(i) > clss(j) & V2(i) < clss(j+1)
         class2(j) = class2(j) + 1;
      end
   end
end

ymax = max(class1);
ymin = min(class1);
xx = [xcut xcut];
yy = [ymin ymax];

discrim_score = (xmin:xinterval:xmax);
figure
```

```
plot(discrim_score',class1,'-o',discrim_score',class2,'-x')
hold on
plot(xx,yy)
title('Frequency plot of the linear discriminant function')
xlabel('Discriminant Score')
ylabel('Frequency')
legend(T1,T2,'Classification Cut off')

figure
bar(discrim_score',class1,'y')
hold on
bar(discrim_score',class2,'g')
title('Show Histogram of the Discriminant Scores')
xlabel('Discriminant Score')
ylabel('Frequency')
%legend('Group 1','Group 2')
legend(T1,T2)

tot_sum = sum(class1) + sum(class2);
rel_freq1 = class1/tot_sum;
rel_freq2 = class2/tot_sum;

figure
bar(discrim_score',rel_freq1,'y')
hold on
bar(discrim_score',rel_freq2,'g')
xlabel('Discriminant Score')
ylabel('Relative Frequency')
%legend('Group 1','Group 2')
legend(T1,T2)

xfreq1 = [discrim_score' rel_freq1];
xfreq2 = [discrim_score' rel_freq2];
```

MANOVA function program

This program performs MANOVA for matrix data.

```
%function [r,c,Yitot,Y1t,Ybar1] = H1(Y1);
clear all; close all; format compact; format short e;
% multivariate One-Way Analysis of Variance Model (MANOVA),
% Data = Y1
% {\bf y}_{1n} = Y1 = [var1 var2 var3 var4 varp]
% Y1total = sum(Y1)1
```

```
kmenu = menu('select the data','Fish or Rootstock(4var,6group)',
 'Fish(4var3grp) ,guinea pig(6var3grp)','Research')
if kmenu == 1
      load f:\data\rencher\root.dat; YY=root; k = 6;
elseif kmenu == 2
   k4menu = menu('select','fish 4 variable 3 group',
 'guinea pig 6 variable 3 group')
   if k4menu ==1
      load f:\data\rencher\fish.dat; YY=fish; k=3;
   elseif k4menu ==2
      load f:\data\rencher\guinea.dat; YY = guinea; k=3;
   end
elseif kmenu == 3
   load f:\phdwork\data\ahgroup.dat; Yraw=ahgroup; k = 4;
%   YY = [Yraw(:,1) standardize(Yraw(:,6:14))];
   YY = [Yraw(:,1) Yraw(:,6:14)];
end
[N p] = size(YY) ;
grpID = YY(:, 1); Y = YY(:, 2:p);ybar = mean(Y)' ;p=p-1;

if kmenu ==1
   Y1 = Y(find(grpID==1), :) ; N1 = size(Y1, 1), ybar1 = mean(Y1)' ;
 Psihat1 = (N1-1) * cov(Y1) / N1 ;
   Y2 = Y(find(grpID==2), :) ; N2 = size(Y2, 1), ybar2 = mean(Y2)' ;
 Psihat2 = (N2-1) * cov(Y2) / N2 ;
   Y3 = Y(find(grpID==3), :) ; N3 = size(Y3, 1), ybar3 = mean(Y3)' ;
 Psihat3 = (N3-1) * cov(Y3) / N3 ;
   Y4 = Y(find(grpID==4), :) ; N4 = size(Y4, 1), ybar4 = mean(Y4)' ;
 Psihat4 = (N4-1) * cov(Y4) / N4 ;
   Y5 = Y(find(grpID==5), :) ; N5 = size(Y5, 1), ybar5 = mean(Y5)' ;
 Psihat5 = (N5-1) * cov(Y5) / N5 ;
   Y6 = Y(find(grpID==6), :) ; N6 = size(Y6, 1), ybar6 = mean(Y6)' ;
 Psihat6 = (N6-1) * cov(Y6) / N6 ;
   k = 6;
   H1 = N1*(ybar1 - ybar)*(ybar1 - ybar)';
   H2 = N2*(ybar2 - ybar)*(ybar2 - ybar)';
   H3 = N3*(ybar3 - ybar)*(ybar3 - ybar)';
   H4 = N4*(ybar4 - ybar)*(ybar4 - ybar)';
   H5 = N5*(ybar5 - ybar)*(ybar5 - ybar)';
   H6 = N6*(ybar6 - ybar)*(ybar6 - ybar)';
   H = H1 + H2 + H3 + H4 + H5 + H6
% "hypothesis matrix H has "Between" SS on diagonal for each p var
% and Sum of products on off diagonal for each pair of variables
   E1 = (Y1 - ones(N1,1)*ybar1');
   E2 = (Y2 - ones(N2,1)*ybar2');
   E3 = (Y3 - ones(N3,1)*ybar3');
```

```
   E4 = (Y4 - ones(N4,1)*ybar4');
   E5 = (Y5 - ones(N5,1)*ybar5');
   E6 = (Y6 - ones(N6,1)*ybar6');
   Ex = [E1;E2;E3;E4;E5;E6];
   E = Ex'*Ex
%pxp error matrix E has "Within" sums of square on
% diagonal and sum of products off diagonal
   T = E + H
elseif kmenu ==2
   Y1 = Y(find(grpID==1), :) ; N1 = size(Y1, 1), ybar1 = mean(Y1)' ;
 Psihat1 = (N1-1) * cov(Y1) / N1 ;
   Y2 = Y(find(grpID==2), :) ; N2 = size(Y2, 1), ybar2 = mean(Y2)' ;
 Psihat2 = (N2-1) * cov(Y2) / N2 ;
   Y3 = Y(find(grpID==3), :) ; N3 = size(Y3, 1), ybar3 = mean(Y3)' ;
 Psihat3 = (N3-1) * cov(Y3) / N3 ;
   H1 = N1*(ybar1 - ybar)*(ybar1 - ybar)';
   H2 = N2*(ybar2 - ybar)*(ybar2 - ybar)';
   H3 = N3*(ybar3 - ybar)*(ybar3 - ybar)';
   H = H1 + H2 + H3
% "hypothesis matrix H has "Between" SS on diagonal for each p var
% and Sum of products on off diagonal for each pair of variables
   E1 = (Y1 - ones(N1,1)*ybar1')
   E2 = (Y2 - ones(N2,1)*ybar2')
   E3 = (Y3 - ones(N3,1)*ybar3')
   Ex = [E1;E2;E3];
   E = Ex'*Ex
% pxp error matrix E has "Within" sums of square on diagonal and sum
% of products off diagonal
   T = E + H
elseif kmenu == 3
   Y1 = Y(find(grpID==1), :) ; N1 = size(Y1, 1), ybar1 = mean(Y1)' ;
 Psihat1 = (N1-1) * cov(Y1) / N1 ;
   Y2 = Y(find(grpID==2), :) ; N2 = size(Y2, 1), ybar2 = mean(Y2)' ;
 Psihat2 = (N2-1) * cov(Y2) / N2 ;
   Y3 = Y(find(grpID==3), :) ; N3 = size(Y3, 1), ybar3 = mean(Y3)' ;
 Psihat3 = (N3-1) * cov(Y3) / N3 ;
   Y4 = Y(find(grpID==4), :) ; N4 = size(Y4, 1), ybar4 = mean(Y4)' ;
 Psihat4 = (N4-1) * cov(Y4) / N4 ;
   k = 4;
   H1 = N1*(ybar1 - ybar)*(ybar1 - ybar)';
   H2 = N2*(ybar2 - ybar)*(ybar2 - ybar)';
   H3 = N3*(ybar3 - ybar)*(ybar3 - ybar)';
   H4 = N4*(ybar4 - ybar)*(ybar4 - ybar)';
   H = H1 + H2 + H3 + H4
% "hypothesis matrix H has "Between" SS on diagonal for each p var
% and Sum of products on off diagonal for each pair of variables
```

```
   E1 = (Y1 - ones(N1,1)*ybar1');
   E2 = (Y2 - ones(N2,1)*ybar2');
   E3 = (Y3 - ones(N3,1)*ybar3');
   E4 = (Y4 - ones(N4,1)*ybar4');
   Ex = [E1;E2;E3;E4];
   E = Ex'*Ex
% pxp error matrix E has "Within" sums of square on diagonal and sum
% of products off diagonal
   T = E + H
end

%Next compute the "Total", "Within", and "Between" covariance matrices:
Total = (N-1) * cov(Y) / N
if kmenu == 1
   E_Within = (N1*Psihat1 + N2*Psihat2 + N3*Psihat3 + N4*Psihat4 +
 N5*Psihat5 + N6*Psihat6) / N
elseif kmenu == 2
   E_Within = (N1*Psihat1 + N2*Psihat2 + N3*Psihat3) / N
elseif kmenu == 3
   E_Within = (N1*Psihat1 + N2*Psihat2 + N3*Psihat3 + N4*Psihat4) / N
end
H_Between = Total - E_Within

% Wilks' LAMBDA
disp(' need to look up Wilks Lambda distribution with the following 4
 Table A.9 Rencer');
[n p] = size(Y1);
LAMBDA = det(E_Within)/det(Total)
vH = k - 1;
vE = N-k;
fprintf('\n Alpha (0.05): p_number of variables (%f)\n',p);
fprintf(' df_deg freedom for hypothesis: (%f)\n',vH);
fprintf(' df_deg freedom for      error: (%f)\n',vE);
W_lambda = input(' Enter the value after look up: ')
if LAMBDA < W_lambda
   fprintf(' Reject Ho: mu_1 = mu_2 = ... = mu_p\n\n\n')
end

fprintf(' An indication of the pattern of the mean vectors is given
 by the eigenvalues of E^(-1)H\n');
fprintf(' if there is one large eigenvalue and the others are small,
 the mean vectors lie close to\n');
fprintf(' a line in space.  If there are two large eigenvalues,
 the mean vectors lie mostly in two\n');
fprintf(' dimensions, and so on ...\n');
fprintf(' Because Roy test uses only the largest eigenvalue of
```

```
 inv(E)H it is more powerful than\n');
fprintf(' others if the mean vectors are collinear\n');
fprintf(' If the mean vectors are spread out in several dimension,
 Others test has more power\n');

eigen_val = sort(eig(inv(E)*H));
[re rc] = size(eigen_val);
e_sum = 0;
for ir = re:-1:1,
   e_sum = e_sum + eigen_val(ir);
   e_prop(ir) = e_sum/sum(eigen_val);
end
disp('Eig_val  Proportion');
[eigen_val e_prop'] % put in a column vector
fprintf('\n\nNote: largest two eigenvalues account for a proportion:
 (%f) percent\n',e_prop(re-1));
fprintf('  Also if the largest two account for more than 90 percent
 then p mean vectors\n');
fprintf('in two dimensions\n');
```

Outlier function program

This program performs normality assessment and detects possible outliers.

```
% Assessing Multivariate Normality and detection of Outliers
% written by: Kyung-Jin Jang  Oct 1998
%
% function [D2] = standardized_d(y);
clear all
close all
format compact
% if ix = 1 then group identity in first column in data file
% if ix = 2 then non group identy in first column in data file
% if ix = 3 then research file format
kmenu = menu('Data available','hematol','calcium','bone','probe word',
 'AHU normal','AHU RPM','AHU Valve','AHU Coil','Other ahu data')
if kmenu == 1
   load f:\Data\Rencher\HEMATOL.DAT;yc=hematol;ix=2;
elseif kmenu == 2
   load f:\Data\Rencher\CALCIUM.DAT;yc=calcium;ix=1;
elseif kmenu == 3
   load f:\Data\Rencher\BONE.DAT;yc=bone;ix=1;
elseif kmenu == 4
   load f:\Data\Rencher\PROBE.DAT;yc=probe;ix=1;
```

```
elseif kmenu == 5
   nmenu = menu('Normal data available','normal_combo','normal_13',
 'normal_rep1_raw','normal_rep1')
   if nmenu == 1
      load f:\Data\Train_data\normal_combo_train.txt;
 yc=normal_combo_train;ix=3;
   elseif nmenu == 2
      load f:\Data\Train_data\normal_13.txt;yc=normal_13;ix=3;
   elseif nmenu == 3
      load f:\Data\Train_data\normal_rep1_raw.txt;yc=normal_rep1_raw;
 ix=3;
   elseif nmenu == 4
      load f:\Data\Train_data\normal_rep1.txt;yc=normal_rep1;ix=3;
   end
elseif kmenu == 6
   nmenu = menu('RPM data available','rpm_combo_train','rpm_combo')
   if nmenu == 1
      load f:\Data\Train_data\rpm_combo_train.txt;yc=rpm_combo_train;
 ix=3;
   elseif nmenu == 2
      load f:\Data\Train_data\rpm_combo.txt;yc=rpm_combo;ix=3;
   end
elseif kmenu == 7
   nmenu = menu('Valve data available','valve_combo_train',
 'valve_combo')
   if nmenu == 1
      load f:\Data\Train_data\valve_combo_train.txt;
 yc=valve_combo_train;ix=3;
   elseif nmenu == 2
      load f:\Data\Train_data\valve_combo.txt;yc=valve_combo;ix=3;
   end
elseif kmenu == 8
   nmenu = menu('Coil data available','coil_combo_train','coil_combo')
   if nmenu == 1
      load f:\Data\Train_data\coil_combo_train.txt;
 yc=coil_combo_train;ix=3;
   elseif nmenu == 2
      load f:\Data\Train_data\coil_combo.txt;yc=coil_combo;ix=3;
   end
elseif kmenu == 9
   filest = input('Enter the path and filename','s')
   num_var = input('Enter total number of variables
 (include group id) ')
   yc = read_file(filest,num_var); ix=3;
end
[n k]=size(yc);fprintf('\nTotal Observation(%i) Variable(%i)\n',n,k)
```

```
if ix == 1
 y=yc(:,2:k);
elseif ix == 2
   y=yc;
elseif ix == 3
   y=yc(:,6:k);
   itrans = menu('Transform with','log','sqrt','lambda')
   if itrans == 1
      y = log(y);
   elseif itrans == 2
      y = sqrt(y);
   elseif itrans == 3
      ylam = input('Enter power to use: ');
      y = 1/ylam*(y.^ylam - 1);
   end
%   y = pca_standardize(y);

   format short e
end
% This function calculates standardized distance from each y_i
% to y_mean by relation
% D^2_i = (y_i - ybar)'S^(-1)(y_i - ybar)
%
% y_i = (ith observation) = f(y_i1, y_i2, y_i3, ..., y_ik)
% S   = covariance matrix of the sample observations
% This may show possible outliers by how big the calculated D is

[n k]=size(y);
ipx = 1;
for ip = 1:k
   for p = 1:k
      if ip~= p
         subplot(k,k,ipx),plot(y(:,p),y(:,ip),'.')
      end
      ipx = ipx + 1;
   end
end
print -depsc2 -tiff f:\Figures\scatter.eps
ybar = mean(y);
S  = cov(y);
Sinv = inv(S);
for j = 1:n,
 for i = 1:k,
  yd(i) = y(j,i) - ybar(i);
 end
 D2(j) = yd*Sinv*yd';
```

```
end
% find u
for i = 1:n,
 u(i) = n*D2(i)/(n-1)^2;
% Rencher eq(4.28) has beta distribution related to F dist
end

xindex = 1:n;          % Rank the u(1) < u(2) < ... < u(n)
uav = [u' D2' xindex'];
uav2 = sortrows(uav);
p = k;
alpha = (p - 2)/(2*p);
beta = (n-p-2)/2/(n-p-1);
for i = 1:n
 v(i) = (i-alpha)/(n-alpha-beta+1);
end

figure
plot(v,uav2(:,1),'o');xlabel('v_i');ylabel('u_i')
print -depsc2 -tiff f:\Figures\uvQQplot.eps

fprintf('Nonlinear pattern would indicate a departure fronormality\n');
nfirst = 3;fprintf('\nFirst %i terms\n',nfirst);
for nfi = 1:nfirst,
   fprintf('Ob No(%5i) D2(%4.4f) u(%.5f) v(%.5f)\n',uav2(nfi,3),
            uav2(nfi,2), uav2(nfi,1), v(nfi))
end
nlast = 5;fprintf('\nLast %i terms\n',nlast);
for nfi = n-nlast:n,
   fprintf('Ob No(%5i) D2(%4.4f) u(%.5f) v(%.5f)\n',uav2(nfi,3),
            uav2(nfi,2), uav2(nfi,1), v(nfi))
end

fprintf('\n Saving the output of the outlier index to outlier.out\n');
fid = fopen('outlier.out','wt');
fprintf(fid,'\n==== Assessing Multivariate Normality ====\n');
fprintf(fid,'\nObservation Distance^2 u              v    \n');
for ii = 1:n,
   fprintf(fid,'%i %f %f %f\n',uav2(ii,3),uav2(ii,2),uav2(ii,1),v(ii));
end

S_MLE = (n-1)/n*S;
SMinv = inv(S_MLE);

for j = 1:n,
 for i = 1:k,
```

```
  yd(j,i) = y(j,i) - ybar(i);
 end
end
yd2 = yd;

g = yd*SMinv*yd';
g3 = g.^3;
g2 = g.^2;

b1p_sum = sum(g3);
b1p = sum(b1p_sum)/n^2;

b2p_sum = diag(g2);
b2p = sum(b2p_sum)/n;

fprintf('\n\n b_1p = (%f) b_2p = (%f)\n',b1p,b2p);
fprintf(fid,'\n\n b_1p = (%f) b_2p = (%f)\n',b1p,b2p);

% For other values of p and n > 50
if n > 50 | p > 4
   b1p_50 = (p+1)*(n+1)*(n+3)/6/((n+1)*(p+1)-6)*b1p;
   chi_df = 1/6*p*(p+1)*(p+2);
   %[chi] = chidist(0.05,chi_df);
   fprintf('\n There are n>50 points or p>4 hence use b_1p =
            (%f)\n',b1p_50);
   fprintf('\n reject normality assumption if b_1p >= chi^2(0.05)\n');
   b2p_50_U = (b2p - p*(p+2))/sqrt( 8*p*(p+2)/n );
   if n <=400
      b2p_50_L = (b2p - p*(p+2)*(n+p+1)/n)/sqrt( 8*p*(p+2)/(n-1) );
   else
      b2p_50_L = (b2p - p*(p+2))/sqrt( 8*p*(p+2)/n );
   end
   fprintf('\n For upper 2.5 percent points b_2p =(%f)\n',b2p_50_U);
   fprintf(' For lower 2.5 percent points b_2p =(%f)\n',b2p_50_L);
   fprintf(' which is approximately N(0,1)\n');
   fprintf(fid,'\n There are n>50 points or p>4 hence use b_1p =
           (%f)\n',b1p_50);
   fprintf(fid,'\n reject normality assumption if b_1p >=
           chi^2(0.05)\n');
   fprintf(fid,'\n For upper 2.5 percent points b_2p =
           (%f)\n',b2p_50_U);
   fprintf(fid,' For lower 2.5 percent points b_2p =(%f)\n',b2p_50_L);
   fprintf(fid,' which is approximately N(0,1)\n');
end
fclose(fid)
```

REFERENCES

Baruch, M. 1984. Methods of Reference Basis for Identification of Linear Dynamic Systems. *AIAA Journal.* Vol.22, No.4, p.561-564.

Berman, A., and E.I. Nagy. 1983. Improvement of a Large Analytical Model Using Test Data. *AIAA Journal.* Vol.21, No.8, p.1168-1173.

Box, E.P. George, William G. Hunter, and Stuart J. Hunter. 1978. *Statistics for Experimenters.* NY: Wiley and Sons Inc.

Box, E.P. George, D. R. Cox. 1964. An Analysis of Transformations. *Journal of the Royal Statistical Society.* 26, No. 2, p.211-252.

Cattell, R. 1966. The meaning and strategic use of factor analysis. *Handbook of multivariate experimental psychology.* Chicago: Rand McNally, p.174-243.

Collins, J.D., G.C. Hart, T.K. Hasselman, and B. Kennedy. 1974. Statistical Identification of Structures. *AIAA Journal.* Vol.12, No.2, p.185-190.

Cochran, G. William, and Gertrude M. Cox. 1992. *Experimental Designs.* NY, NY: Wiley and Sons Inc.

Desborough, L., and T. Harris. 1992. Performance Assessement Measures for Univariate Feedback Control. *Can. J. Chem. Eng.* 70: 1186-1197.

Desborough, L., and T. Harris. 1993. Performance Assessement Measures for Univariate Feedforward/Feedback Control. *Can. J. Chem. Eng.* 71: 605-616.

Dretzke, J. Beverly, and A. Kenneth Heilman. 1998. *Statistics with Microsoft®Excel* Englewood Cliffs, NJ: Prentice Hall.

Dexter, A.L., and M. Benouarets. 1996. A Generic Approach to Identifying Faults in HVAC Plants. *ASHRAE technical data bulletin.* Vol. 12, No. 2, p.33-39.

Dorfman, H. Jeffrey. 1997. *Bayesian Economics through Numerical Methods a Guide to Econometrics and Decision-Making with Prior Information* NY, NY: Springer.

Everitt, B.S., and G. Dunn. 1991. *Applied Multivariate Data Analysis.* London: Edward Arnold.

Fasolo, P.S., D.E. Seborg. 1995. Monitoring and Fault Detection for an HVAC Control System. *HVAC&R Research.* Vol. 1, No. 3, p.177-193.

Fisher, R.A. 1936. The use of multiple measurements on taxonomic problems. *Annals of Eugenics.* 7: 179-188.

Fisher, R.A. 1938. The Statistical Utilization of Multiple Measurements. *Annals of Eugenics.* 8: 376-386.

Flury, B. 1997. *A First Course in Multivariate Statistics.* New York, NY: Springer-Verlag.

Freund, E.John. 1992. *Mathematical Statistics.* 5^{th} Edition, Englewood Cliffs, NJ: Prentice-Hall, Inc.

Gangadharan, S.N., E. Nikolaidis, and R.T. Haftka. 1991. Probabilistic System Identification of Two Flexible Joint Models. *AIAA Journal.* Vol.29, No.8, p.1319-1326.

Gourieroux, Christian. 1997. *Time Series and Dynamic Models.* Cambridge: Cambridge University Press.

Grimm, G. Laurence, and Paul R. Yarnold. 1997. *Reading and Understanding Multivariate Statistics.* Washington DC: American Psychological Assoc.

Hand, D.J. 1981. *Discrimination and Classification.* London: John Wiley and Sons.

Haves, P., T.I. Salsbury, and J.A. Wright. 1996. Condition Monitoring in HVAC Subsystems Using First Principal Models. *ASHRAE technical data bulletin.* Vol. 12, No. 2, p.2-10.

Hosmer, D.W. 1989. *Applied Logistic Regression.* New York, NY: John Wiley and Sons.

Hotelling, H. 1931. The generalization of Student's ratio. *Annals of Mathematical Statistics.* 2, p.360-378.

Hotelling, H. 1933. Analysis of a complex of statistical variables into principal components. *Journalof Educational Psychology.* 24, p.417-441.

Incropera, P. Frank and David P. DeWitt. 1981. *Fundamentals of Heat and Mass Transfer.* 2^{nd} Edition New York, NY: John Wiley and Sons.

Isermann, R. 1984. Process Fault Detection Based on Modelling and Estimation Methods: A Survey. *Automatica.* Vol. 20, p.387-404.

Johnson, R. 1992. *Applied Multivariate Statistical Analysis.* 3^{rd} Edition, Englewood Cliffs, NJ: Prentice Hall.

Kaiser, H.F., G.C. Hart, T.K. Hasselman, and B. Kennedy. 1960. The application of electronic computers to factor analysis. *Educational and Psychological Measurement.* No.20, p.141-151.

Kalman, R.E., 1960. A New Approach to Linear Filtering and prediction Problems. *Trans. ASME Journal Basic Engrg.*, Series D, 82, 35-45.

Karlov, V.I., D.W. Miller, W.E. Vander Velde, and E.F. Crawley. 1994. Identification of Model Parameters and Associated Uncertainties for Robust Control Design. *Journal of Guidance, Control, and Dynamic.* Vol. 17, No. 3, p.495-503.

Kelly, R.J. 1992. MLS System Error Model Identification and Synthesis. *IEEE Transactions on Aerospace and Electronic Systems.* Vol. 28, No. 1, p.164-172.

Kolmogorov, A. 1939. Sur l'interpolation et l'extrapolation des suites stationnaires, *C.R. Acad. Sci.* Paris, 208, 2043-2045

Kolmogorov, A. 1941. Interpolation und extrapolation von stationaren Zufalligen Folgen, *Bull. Acad. Sci.* (Nauk), USSR, Ser. Math., 5, 3-14

Kuehl, O. Robert. 1994. *Statistical Principles of Research Design and Analysis.* Menlo Park, California: Wadsworth Publishing Company.

Lee, W.Y., C. Park, and G.E. Kelly. 1997. Fault Diagnosis and Temperature Sensor Recovery for an Air-Handling Unit. *ASHRAE Transactions.* 103(1).

Lee, W.Y., C. Park, and G.E. Kelly. 1996. Fault Detection in an Air-Handling Unit Using Residual and Recursive Parameter Identification Methods. *ASHRAE technical data bulletin.* Vol. 12, No. 2, p.11-22.

Lee, W.Y., C. Park, and G.E. Kelly. 1996. Fault Diagnosis of an Air-Handling Unit Using Artificial Neural Networks. *ASHRAE technical data bulletin.* Vol. 12, No. 2, p.23-32.

Li, X., H. Vaezi-Nejad, and J.C. Visier. 1996. Development of a fault Diagnosis Method for Heating Systems Using Neural Networks. *ASHRAE technical data bulletin.* Vol. 12, No. 2, p.48-55.

Mendenhall, W., Dennis D. Wackerly, and Richard L. Scheaffer. 1990. *Mathematical Statistics with Applications.* Boston: PWS-Kent Publishing Company.

Mili, L., T. Van Cutsem, and M. Ribbens-Pavella. 1984. Hypothesis Testing Identification. *IEEE Transactions on Power Apparatus and Systems.* Vol. Pas-103, No. 11, p.3239-3251.

Moran, J. Michael. 1989. *Availability Analysis.* New York, NY: ASME Press.

Nise, S. Norman. 1995. *Control Systems Engineering.* 2^{nd} Edition, Menlo Park, California: Addsion Wesley Publishing Company.

Ott, R. Lyman. 1994. *An Introduction to Statistical Methods and Data Analysis.* 4^{th} Edition, Belmont, California: Wadsworth Publishing Company.

Patton, Ron, P.Frank, and Robert Clark. 1989. *Fault Diagnosis in Dynamic Systems Theory and Application.* New York, NY: Prentice Hall.

Peitsman, H.C., and V.E. Bakker. 1996. Application of Black-Box Models to HVAC Systems for Fault Detection. *ASHRAE technical data bulletin.* Vol. 12, No. 2, p.69-81.

Peitsman, H.C., and L.L. Soethout. 1997. ARX Models and Real-Time Model-Based Diagnosis. *ASHRAE Transactions.* 103(1)

Pole, A., M. West,and J. Harrison. 1994. *Applied Bayesian Forecasting and Time Series Analysis.* New York, NY: Chapmann & Hall.

Rencher, C. Alvin 1995. *Methods of Multivariate Analysis.* New York, NY: Wiley and Sons.

Searle, S. R. 1997. *Linear Models.* New York, NY: John Wiley and Sons.

Seborg, D.E., T.F. Edgar, and D.A. Mellichamp. 1989. *Process Dynamics and Control.* New York, NY: John Wiley and Sons.

Shapiro, S.S., A.J. Gross. 1981. *Statistical Modeling Technique.* New York, NY: Marcel Dekker.

Shapiro, N. Howard., Michael.J. Moran. 1988. *Fundamentals of Engineering Thermodynamics.* New York, NY: John Wiley and Sons.

Snedecor, W. George. 1956. *Statistical Methods.* Ames: Iowa State College Press.

Stevens, J.P. 1986. *Applied multivariate statistics for the social science.* Hillsdale, NJ: Erlbaum.

Stylianou, M., and D. Nikanpour. 1996. Performance Monitoring, Fault Detection, and Diagnosis of Reciprocating Chillers. *ASHRAE technical data bulletin.* Vol. 12, No. 2, p.56-68.

Stylianou, M. 1997. Application of Classification Functions to Chiller Fault Detection and Diagnosis. *ASHRAE Transactions.* 103(1).

Sveshnikov, A. A. 1968. *Problems in Probability Theory, Mathematical Statistics and Theory of Random Functions.* New York, NY: Dover Publications, Inc.

Tugnait, J.K. 1992. Stochastic System Identification with Noisy Input Using Cumulant Statistics. *IEEE Transaction on Automatic Control.* Vol.37, No.4, p476-484.

Tsutsui, H., and K. Kamimura. 1996. Chiller Condition Monitoring Using Topological Case-Based Modeling. *ASHRAE technical data bulletin.* Vol. 12, No. 2, p.82-89.

Vardeman, B. Stephen. 1994. *Statistics for Engineering Problem Solving.* Boston: PWS Publishing Company.

Wei, W.S. 1990. *Time Series Analysis: Univariate and Multivariate Methods.* New York, NY: Addison-Wesley.

Weiner, N. 1949. *The Extrapolation, Interpolation to the theory of Stationary Random Functions.* New York, NY: Wiley and Sons.

Welch, B.L. 1938. The Significance of the Differences Between Two Means When the Population Variances are Unequal. *Biometrika.* 29: 350-362.

Whittle, P. 1983. *Prediction and Regulation by Linear Least-Square Methods*, 2^{nd} Edition, Minneapolis, University of Minnesota.

Yaglom, A.M. 1962. *An Introduction to the Theory of Stationary Random Functions* NJ: Prentice Hall, Englewood Cliffs.

Yoshida, H., T. Iwami, H. Yuzawa, and M. Suzuki. 1996. Typical Faults of Air-Conditioning Systems and Fault Detection by ARX Model and Extended Kalman Filter. *ASHRAE technical data bulletin.* Vol. 12, No. 2, p.40-47.

Yun, C.B., and M. Shinozuka. 1980. Identification of Non-linear Structural Dynamic Systems. *Journal of Structural Mechanics.* Vol.8, No.2, p.187-203.

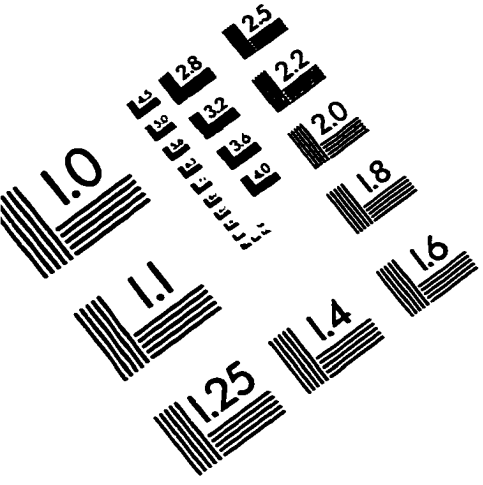
ACKNOWLEDGMENTS

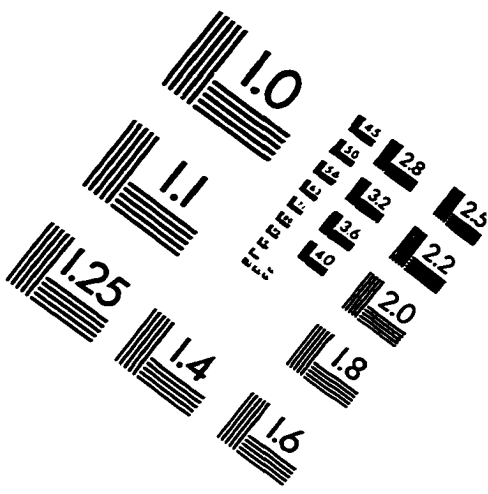
I would like to thank my major professor, Dr. Ron Nelson who has given me unconditional support and ideas. Through the balance of criticism and encouragement he has provided, I have gained much insights and understanding. I thank Professors R. Brown, H.T. David, G. Maxwell, and L. Zachary for serving on my graduate committee. I also want to express my appreciation to Jim Deutromont, Gaylord Scandrett, and Hap Steed for the technical support they have provided for the experimental work. I have thoroughly enjoyed my stay at Iowa State University for the program of Mechanical Engineering as well as Statistical program. I have never felt so compelled to learn and yet there is so much more to be learned.

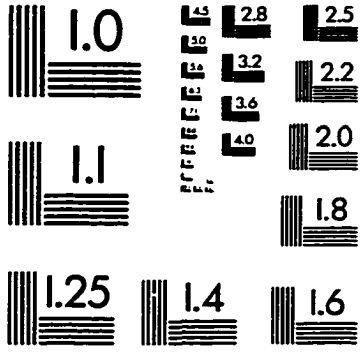
I especially wish to thank my wife Yeon-Sook, for her continual support. I also want to thank my daughters Eunice and Sylvia for being patient with me. I owe them much love and time which I could not provide during my academic journey. I thank my father Shi-hung and mother Eugenia for the support and care throughout my life's journey.

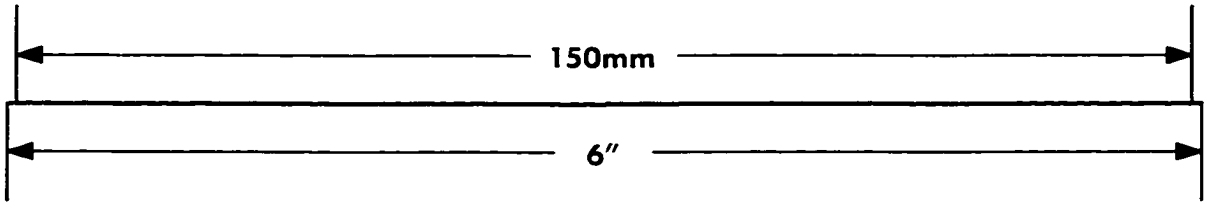
Above all, I thank the Lord God for all that He has created and challenges He makes on our daily lives.

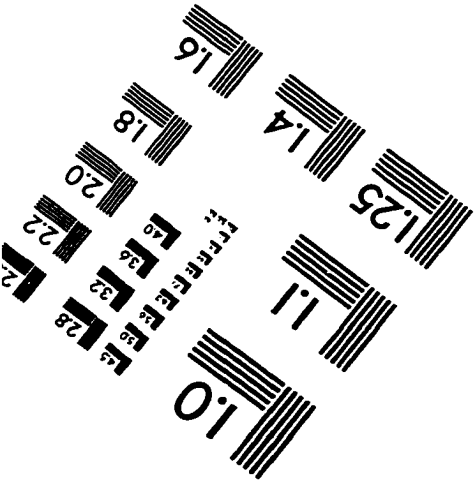
IMAGE EVALUATION
TEST TARGET (QA-3)











APPLIED IMAGE, Inc
1653 East Main Street
Rochester, NY 14609 USA
Phone: 716/482-0300
Fax: 716/288-5989