

國立交通大學

統計學研究所

碩士論文

利用多變量變異數分析與半母數線性模型

辨識行進中車輛

Using MANOVA and
Semi-parametric Linear Mixed Effects Model
for Traveling Vehicles Identification

研究生：楊菡慈

指導教授：周幼珍 博士

中華民國九十六年六月

利用多變量變異數分析與半母數線性模型
辨識行進中車輛

Using MANOVA and
Semi-parametric Linear Mixed Effects Model
for Traveling Vehicles Identification

研究生：楊菡慈
指導教授：周幼珍

Student : Chai-Tzu Yang
Advisor : Yow-Jen Jou

國立交通大學

統計學研究所

碩士論文

A Thesis
Submitted to Institute of Statistics
College of Science
National Chiao Tung University
in Partial Fulfillment of the Requirements
for the Degree of
Master
in
Statistics
June 2007

Hsinchu, Taiwan, Republic of China

中華民國九十六年六月

利用多變量變異數分析與半母數線性模型 辨識行進中車輛

研究生：楊菡慈

指導教授：周幼珍 博士

國立交通大學統計學研究所



道路上的即時資料是交通管理系統中重要的一環，為了測得在眾多道路上車輛所在的車道與種類，我們利用架設在路旁的無線電射頻系統晶片來收集車輛的資訊。當車子進入偵測區時，晶片所記錄的資料是雷達回波的強度。而原始的資料經過適當的截取及平移後，可視為一筆函數資料。本文中提出兩種模型方法來辨識車種，其中一種是利用多變量分析方法來說明車種與車道之間主變因與其交互作用的影響，另一種則是用半母數線性模型來突顯函數資料中各群的特徵。若分的群數不多，不論用哪種模型分析，都可以得到不錯的結果；但當群數過多時，則會降低此兩種模型分析的能力。

Using MANOVA and Semi-parametric Linear Mixed Effects Model for Traveling Vehicles Identification

Student: Chai-Tzu Yang

Advisor: Yow-Jen Jou

Institute of Statistics

National Chiao Tung University



In order to make the detecting of the lanes and the types of the vehicles traveling on various roadways affordable, radio-frequency (RF) system-on-chip is designed and will be mounted on the roadside to collect vehicle information. The data originally collected by the chip is the intensity of the back wave of the vehicle entering the range of detection. The raw data is registered by landmark and then treated as functional data. In order to classify the types of the vehicles, two models are proposed to model the data. One is multivariate analysis of variance model to account for the main effect and the interaction effect between type and lane, the other is the semi-parametric linear model to emphasize the functional characteristic of the data. Both models work well when the number of groups is small but deteriorate when the number of groups increases.

誌 謝

首先我要誠摯地感謝我的指導教授周幼珍老師，由於老師耐心的指導，使我對研究的方向與內容越來越有更深入的了解，論文才能順利完成；也謝謝老師在生活與待人處事上給予我許多寶貴的建議。接著，感謝運管所的卓訓榮老師、學長與同學，因為有你們才能順利地收集到實測的車輛資料，且在後續的資料處理上給予我最大的支援與幫助。另外，也要感謝口試委員洪志真、徐南蓉及胡殿中老師熱心地指正我的論文且提供我許多珍貴的意見，感謝各位老師。

這篇論文能順利的完成，除了老師的協助外，我還要感謝班上的許多同學，感謝雪芳、小米、建威在程式上給我很大的幫助。尤其是永在，謝謝你幫我解決了許許多多、各式各樣的問題，我很高興能與你同在周老師的指導下做研究。另外也結交了兩個好姊妹，素梅與雅莉，不論快樂或傷心的時刻都有你們陪伴在我身邊，能夠交到這樣的好朋友是我最幸福的事情。當然還要感謝所上的老師，教導我許多有用的學識；更感謝所上的同學們，大家一起在研究室煮飯、吃飯、聊天，三不五時的班遊、聚餐，讓平淡的生活中添加了許多樂趣，謝謝你們讓我在交大這兩年短短的研究生生活過得多采多姿。

我還要感謝一路陪我走來的阿蒲，謝謝你總是陪在我身邊，在我感到挫折時給我加油打氣，與我分享我的開心與喜悅，沒有你的體諒與包容，這兩年的生活一定無法過的如此的豐富。最後更要感謝我的家人，給我最大的鼓勵、關心與支持，讓我能專心地完成研究所的學業。

在此，僅以此篇論文獻給我親愛的家人與好友們，謝謝你們。

楊菝慈 謹誌于
國立交通大學統計研究所
中華民國九十六年六月

Contents

| | |
|--|----|
| 1. Introduction | 2 |
| 2. Background and Literature Review | 4 |
| 3. Model Specification and Methodology | 9 |
| 3.1 Multivariate Analysis of Variance (MANOVA) | 9 |
| 3.2 Semi-parametric Linear Mixed Effects Model(SLM). | 13 |
| 3.2.1 Linear Mixed Effects Model. | 13 |
| 3.2.2 Semi-Parametric Model. | 15 |
| 3.2.3 Semi-parametric Linear Mixed Effects Model. | 16 |
| 4. Empirical Illustration | 22 |
| 4.1 Data Collection. | 22 |
| 4.2 Model Description and Data Analysis. | 27 |
| 4.2.1 Multivariate Analysis of Variance (MANOVA) | 28 |
| 4.2.2 Semi-parametric Linear Mixed Effects Model(SLM). | 35 |
| 5. Conclusion and Discussion | 41 |
| References | 43 |



List of Tables

| | | |
|---|--|----|
| 1 | The table of MANOVA by using Wilks' test statistic..... | 32 |
| 2 | The criteria for model selection..... | 38 |
| 3 | The correct classification rate of MANOVA model and the three SLM models for 8 groups over one run..... | 39 |



List of Figures

| | | |
|----|--|----|
| 1 | The miniature of the vehicle radar microwave detector..... | 22 |
| 2 | The sketch of the relative positions of the device and the range covered by the transmitter..... | 23 |
| 3 | A typical curve of the intensity of back wave for small vehicle. (a) Lane 1; (b) Lane 2; (c) Lane 3; (d) Lane 4..... | 24 |
| 4 | A typical curve of the intensity of back wave for large vehicle. (a) Lane 1; (b) Lane 2; (c) Lane 3; (d) Lane 4..... | 25 |
| 5 | The typical curves for small and large vehicles plotted on the same scale. (a) small vehicles; (b) large vehicles..... | 26 |
| 6 | The 8 learning samples of the back wave data of each type of vehicles on Lanes 1, 2..... | 29 |
| 7 | The 8 learning samples of the back wave data of each type of vehicles on Lanes 3, 4..... | 30 |
| 8 | The 8 learning samples of the back wave data of small and large vehicles on 4 lanes..... | 31 |
| 9 | The 8 black curves as the learning sample of small vehicles and the red lines represent the fitted value of each group (a) Lane 1; (b) Lane 2; (c) Lane 3; (d) Lane 4..... | 33 |
| 10 | The 8 black curves as the learning sample of large vehicles and the red lines represent the fitted value of each group (a) Lane 1; (b) Lane 2; (c) Lane 3; (d) Lane 4..... | 34 |
| 11 | One random sample selected from the learning sample of the large vehicle on Lane 1 and the predicted curves of this sample from these three models..... | 38 |

12 The box-plot of vehicle correct classification rate over 100 runs.....40



Using MANOVA and Semi-parametric Linear Mixed Effects Model for Traveling Vehicles Identification

Student: Chai-Tzu Yang

Advisor: Yow-Jen Jou

Institute of Statistics

National Chiao Tung University



ABSTRACT

In order to make the detecting of the lanes and the types of the vehicles traveling on various roadways affordable, radio-frequency (RF) system-on-chip is designed and will be mounted on the roadside to collect vehicle information. The data originally collected by the chip is the intensity of the back wave of the vehicle entering the range of detection. The raw data is registered by landmark and then treated as functional data. In order to classify the types of the vehicles, two models are proposed to model the data. One is multivariate analysis of variance model to account for the main effect and the interaction effect between type and lane, the other is the semi-parametric linear model to emphasize the functional characteristic of the data. Both models work well when the number of groups is small but deteriorate when the number of groups increases.

Key words: Radar recognition; Radar Cross Section; Fast Fourier Transform; Multivariate analysis of variance (MANOVA); Semi-parametric linear mixed effects model (SLM); Classification.

1. Introduction

The vehicle detector plays an important role in collecting traffic information which in turn is essential to Advanced Traffic Management System (ATMS). Since there are numerous roadways in our country and the detectors are usually imported from abroad with high cost. As a result, it would have some sort of difficulties on setting up detectors intensively, and the follow-up maintenance will be high-priced. A radio frequency system-on-chip (RF SoC) using the theory of frequency modulated continuous waves (FMCW) is a promising solution. Traditional radars collect and transform the information to form images such as inverse aperture radar (ISAR) or a sequence of one-dimensional range profile from the raw data, and then classify targets based on these images. Herman (2003) bypassed the image formation and attempted target recognition directly from the received data, which is labeled as Automatic Target Recognition (ATR).



The information we obtained from the vehicle detector is a continuous curve, which is the signal of the radar back wave. We transform the continuous curve into a finite-dimensional vector of time index. Our approach to recognize our targets (the vehicles on the road) adapts the same idea of Herman's work, i.e., attempt automatic classification directly from the received data. We put our data into a form of a data matrix which rows correspond to sample individuals and columns to the time index of a vector, as a repeated measures study. As long as it displays a form of multivariate data set, we can utilize multivariate analysis to analyze it. Furthermore, since the data is a continuous curve measured at different time points, we also regard the data as functional data and use nonparametric approach to build mixed effects model for all vehicles to classify the vehicles with different characteristics.

Therefore the subject matters of this thesis are organized as follows. In Section 2, there has three parts; the first one is a brief introduction of radar, explaining why it can be used extensively and how to deal the information collected from radar. The second is the background of the multivariate analysis and how to use multivariate analysis to conduct a repeated measures study. The last part is a summarized account of the nonparametric statistical procedures. Section 3 shows two different statistical analyses: the traditional parametric method, the model of multivariate analysis of variance (MANOVA) and the representative table; and the other viewpoint, the nonparametric aspect, the construction of the semi-parametric linear mixed effects model. An empirical illustration uses two types of statistical analyses would be described in Section 4. Comparison the results of using different methods on the same data set and a few concluding remarks are made in Section 5.



2. Background and Literature Review

The word “radar” was originally an acronym, RADAR, for “RADio Detection And Ranging.” Today, the technology is so common that the word has become a standard English noun. Early radar development was driven by military necessity, and the military is still the dominant user and developer of radar technology. In the recent years, radar has an increasing range of applications, such as the police traffic radar used for enforcing speed limits and the “color weather radar” familiar to every viewer of local television news. One of the most common radar applications is used for collision avoidance and buoys detection by ships, and is now beginning to serve the same role for the automobile and trucking industries.

RCS, Radar Cross Section, the amount of power reradiated by the target back toward the radar transmitter, is an equivalent area that can be used to relate the incident power density at the target to the reflected power density that results at the receiver. Many researches, such as Bennett and Toomey (1981), Herman (2003), Ehrman and Lanterman (2003) treated RCS as one of the criterions of target recognition. They used “frequency space trajectories”, which was a database of the past data, compared to the multi-frequency RCS to differentiate the variety and the size of cars.

Roe, H. and Hobson, G. S. (1992) used the reflected signal of FMCW (frequency-modulated continuous wave) and Doppler to draw the reflected signal graph, and observed the characteristics of the targets. Songhua, H., Hui, Z. and Bo, H. (1995) utilized polarization identification of High Resolution Range Profiles to identify the features of objective. Chun, J. C., Kim, T. S., Kim, J. M., Lim, Z.S. and

Park, W.S. (2001) used the “Beat Signals” in the FMCW Radar Level Meter as a basis to get the distance between the antenna and the target. Weber, N., Moedl, S. and Hackner, M. (2002) employed a novel signal processing approach in microwave Doppler speed sensing. Park, S. J., Kim, T. Y., Kang, S. M. and Koo, K. H. (2003) brought up a new signal processing technique for vehicle detection radar. Ehrman, L. M. and Lanterman, A. D. (2003) added automatic target recognition (ATR) components to passive radar systems to correctly identify aircraft in the target class with exceptional accuracy at the anticipated noise levels. Greneker, E. F. and Rausch, E. O. (2004) applied the X-band radar to detect the speed of trucks on the highway, measured their speed and issued a warning when they exceeded the speed limit to help mitigate truck overturn incidents on US interstate highways.

The signal of the back wave radar could be recorded in various ways and it is very useful in the traffic and engineering. In this thesis, we transform the signal of the radar back wave into the numerical data, and utilize it by the statistical approach to discriminate cars of different types and traveling on the different lanes.

Statistical data arise whenever any responses are either measured or observed on a set of individuals. Each particular response is referred to generally as a “variable” or “variate” ; if just a single observation or measurement is made on each individual then the data are said to be “univariate”, whereas if more than one observation or measurement is made on each individual then the data are said to be multivariate.

The popularity of multivariate analysis has spread rapidly over the past thirty years, and the use of multivariate procedures in diverse fields of application has generated a growth in the development of theoretical results in multivariate analysis.

Recent studies suggest that there is no clear advantage to either the univariate or multivariate approach. Indeed, for some sets of data the univariate approach is more powerful, whereas in other instances, data with sphericity violated, the multivariate approach may be more efficient.

The data that we collected from the radar microwave detector can be treated as the repeated measurements of the same group of individuals. The simplest approach of repeated measurement is to reduce the vector of multiple measurements from each experimental unit to a single measurement. Thus, a multivariate response is reduced to a univariate response. Wishart (1938) appears to have been the first researcher to document the use of this approach. Pocock (1983), Matthew *et al.* (1990), Dawson and Lagakos (1991, 1993), and Frison and Pocock (1992) refer to these types of method as the “summary-statistic approach”. Although this permits the use of simple analysis methods, the resulting loss of information may not be desirable. Another alternative to a univariate analysis of data of a repeated measures study is to perform a multivariate analysis. It uses the multivariate nature of a subject’s observations. Thus, rather than reduce the vector of repeated measurements from each subject to a single summary measurement, observations at different measurements are treated as a vector of random variables; all of the data are used. A step-by-step procedure for conducting a MANOVA on repeated measures data is presented by O’Brien and Kaiser (1985). Therefore, we will analyze our back wave radar data by the multivariate approach.

The typical introductory course is parametric statistical procedures. A characteristic of these procedures is the fact that the appropriateness of their use for purposes of inferences depends on certain assumptions, for example, assume that samples have been drawn from normally distributed populations with equal variances.

Since populations do not always meet the assumptions underlying parametric test, therefore we need inferential procedures whose validity does not depend on rigid assumptions. Nonparametric statistical procedures fill this need in many instances, since they are valid under very general assumptions. As a result, the inference conclusion reached with nonparametric methods need not be tempered by qualifying statements as strong as, “If the distribution is normal, then ...”. The qualifications are always much less restrictive for nonparametric model than for classical (parametric) one.

Since our radar data can also be treated as a real function for a travelling car at time t , which satisfies the reasons for considering data analysis from a functional perspective given by Ramsay and Dalzell (1991). Familiar examples of functional data include growth curves, weather measured over time, and satellite data. Ramsay and Dalzell used the term “functional data analysis” to describe the statistical methods for analyzing such data. Smoothing spline has long been used for fitting curves to data. The monograph by Wahba (1990) can be consulted for more details concerning the mathematical basis for splines.

To take care of various fixed and random elements, mixed effects models provide flexible approach to fit models. For example, for longitudinal data, Wypij *et al.* (1993) and Wang and Taylor (1995) used regression splines to model the fixed effects. Anderson and Jones (1995) used smoothing spline structure to model the random effects. Shi *et al.* (1996) used regression splines to model both the fixed effects and the random effects. For functional data, Rice and Silverman (1991) used smoothing splines to model the fixed effects; the random effects were also modeled nonparametrically by expanding the covariance function in terms of eigenfunctions.

Under certain conditions, the integrated random walk model is equivalent to a smoothing polynomial spline model. Wecker and Ansley (1983) and Kohn and Ansley (1985, 1987) devise the spline model as an integrated random walk observed with error. They used a state-space model which allowed them to calculate likelihoods using the Kalman filter. Hence, stochastic process can also be applied into the random effects in the nonparametric model. Wang (1998) applied the stochastic process as the nonparametric part in his mixed effects model and provided program in R to fit the model (<http://www.pstat.ucsb.edu/faculty/yuedong/ASSIST/manual/node36.html>).



3. Model Specification and Methodology

Two diverse procedures that are applicable to build models for the transformed radar back wave data are described in this section. One is the traditional parametric statistical approach, multivariate analysis of variance, which can provide the argument for the significance of differences between groups. The other uses nonparametric approach to build mixed effects model to classify data with different characteristics.

3.1 Multivariate Analysis of Variance (MANOVA)

A standard method of displaying a set of multivariate data is in the form of a data matrix in which rows correspond to sample individuals and columns to variables, so that the entry in the i th row and j th column gives the values of the j th variate as measured or observed on the i th individual. And, if the data are a random sample from some populations and the objective is to describe these populations, then the first step is to choose an appropriate probability model for them. There are many different multivariate distributions that can be used to build a probability model, however, by the reason of tractability on the mathematical side and the central limit theorem on the statistical side, the multivariate normal distribution is always the underlying model.

Multivariate linear model

The univariate linear model may be written

$$\mathbf{y}_{n \times 1} = \mathbf{X}_{n \times (q+1)} \boldsymbol{\beta}_{(q+1) \times 1} + \boldsymbol{\varepsilon}_{n \times 1}$$

where \mathbf{y} is a vector of dependent variables on n individuals, and the aim is the prediction of a random variables \mathbf{y} as a linear function of (x_1, x_2, \dots, x_q) . The matrix \mathbf{X} contains the values of these variables, and $x_0 = 1$, giving the constant term in the

linear model. $\boldsymbol{\beta}$ is a vector of unknown parameters, to be estimated, and $\boldsymbol{\varepsilon}$ is a vector of random error terms.

The multivariate linear model is the extension of the univariate linear model, it has the form

$$\mathbf{Y}_{n \times p} = \mathbf{X}_{n \times (q+1)} \boldsymbol{\beta}_{(q+1) \times p} + \boldsymbol{\varepsilon}_{n \times p}$$

and

$$\begin{aligned} E(\boldsymbol{\varepsilon}) &= \mathbf{0}_{n \times p} \\ \text{cov}(\boldsymbol{\varepsilon}) &= \boldsymbol{\Sigma}_{p \times p} \otimes \mathbf{I}_{n \times n} \end{aligned}$$

where \mathbf{Y} has p dependent variates. The random error terms of the p -variate observations are independent, and have an unknown covariance matrix $\boldsymbol{\Sigma}$. Different responses in the same individual item may be correlated.

The matrix of residual sums of squares and products is

$$(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})$$

and the estimator of $\boldsymbol{\beta}$ is $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$. Gauss-Markov theorem on least squares shows that the estimates $\hat{\boldsymbol{\beta}}$ and the predictions $\hat{\mathbf{Y}}$ are unbiased and have minimum variance in the class of unbiased linear estimates. Therefore the predicted values are given by

$$\hat{\mathbf{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$$

and the residuals are

$$\hat{\boldsymbol{\varepsilon}} = \mathbf{Y} - \hat{\mathbf{Y}}$$

The covariance matrix of the residuals is

$$\hat{\boldsymbol{\Sigma}} = \frac{\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}}{n - q - 1}$$

and the elements of $\hat{\boldsymbol{\Sigma}}$ are the unbiased estimates of $\boldsymbol{\Sigma}$.

The further assumption that the errors have a multivariate normal distribution,

$$\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \boldsymbol{\Sigma})$$

implies that $\hat{\boldsymbol{\beta}}$ and $\hat{\mathbf{Y}}$ also have multivariate normal distribution, and significance test and confidence intervals based on this assumption. Least squares estimation is also maximum likelihood estimate under multivariate normality and likelihood ratio test are available.

MANOVA

The acronym “MANOVA” stands for “multivariate analysis of variance”. As the name implies, it is just the extension of ANOVA to the multivariate case of vector observations. With multivariate methods, we have seen those univariate variances are replaced by multivariate covariance matrices. In MANOVA, the sums of squares of ANOVA are replaced by the corresponding SSP (sums of squares and cross products) matrices. Thus, one can construct an ANOVA-type table in which the former SS column is replaced by one containing SSP matrices corresponding to the various effects. The effects referred to here are only fixed since any random effects have been subsumed into the covariance structure.

Now consider a two-way layout. Let us suppose we have nrc independent observations generated by the model.

$$\mathbf{Y}_{ijk} = \boldsymbol{\mu} + \boldsymbol{\alpha}_i + \boldsymbol{\beta}_j + (\boldsymbol{\alpha}\boldsymbol{\beta})_{ij} + \boldsymbol{\varepsilon}_{ijk}, \quad i = 1, \dots, r, \quad j = 1, \dots, c, \quad k = 1, \dots, n$$

where $\boldsymbol{\mu}$ is the overall effect, $\boldsymbol{\alpha}_i$ is the i th row effect, $\boldsymbol{\beta}_j$ is the j th column effect, $(\boldsymbol{\alpha}\boldsymbol{\beta})_{ij}$ is the interaction effect between the i th row and the j th column, and $\boldsymbol{\varepsilon}_{ijk}$ is the error term which is assumed to be independent $N_p \sim (\mathbf{0}, \boldsymbol{\Sigma})$ for all i, j, k . The table of MANOVA is list as follow:

| Source of variation | Sum of squares and cross product (SSP) | Degree of freedom (d.f.) |
|---------------------|---|--------------------------|
| Rows treatment | $\mathbf{R} = c n \sum_{i=1}^r (\bar{\mathbf{Y}}_{i..} - \bar{\mathbf{Y}}_{...}) (\bar{\mathbf{Y}}_{i..} - \bar{\mathbf{Y}}_{...})'$ | r-1 |
| Columns treatment | $\mathbf{C} = r n \sum_{j=1}^c (\bar{\mathbf{Y}}_{.j.} - \bar{\mathbf{Y}}_{...}) (\bar{\mathbf{Y}}_{.j.} - \bar{\mathbf{Y}}_{...})'$ | c-1 |
| Interaction | $\mathbf{I} = n \sum_{i=1}^r \sum_{j=1}^c (\bar{\mathbf{Y}}_{ij.} - \bar{\mathbf{Y}}_{i..} - \bar{\mathbf{Y}}_{.j.} + \bar{\mathbf{Y}}_{...}) (\bar{\mathbf{Y}}_{ij.} - \bar{\mathbf{Y}}_{i..} - \bar{\mathbf{Y}}_{.j.} + \bar{\mathbf{Y}}_{...})'$ | (r-1)(c-1) |
| Residual | $\mathbf{E} = \sum_{i=1}^r \sum_{j=1}^c \sum_{k=1}^n (\mathbf{Y}_{ijk} - \bar{\mathbf{Y}}_{ij.}) (\mathbf{Y}_{ijk} - \bar{\mathbf{Y}}_{ij.})'$ | rc(n-1) |
| Total | $\mathbf{T} = \mathbf{R} + \mathbf{C} + \mathbf{I} + \mathbf{E} = \sum_{i=1}^r \sum_{j=1}^c \sum_{k=1}^n (\mathbf{Y}_{ijk} - \bar{\mathbf{Y}}_{...}) (\mathbf{Y}_{ijk} - \bar{\mathbf{Y}}_{...})'$ | rcn-1 |

where $\bar{\mathbf{Y}}_{i..} = \frac{1}{cn} \sum_{j=1}^c \sum_{k=1}^n \mathbf{Y}_{ijk}$, $\bar{\mathbf{Y}}_{.j.} = \frac{1}{rn} \sum_{i=1}^r \sum_{k=1}^n \mathbf{Y}_{ijk}$, $\bar{\mathbf{Y}}_{ij.} = \frac{1}{n} \sum_{k=1}^n \mathbf{Y}_{ijk}$,

and $\bar{\mathbf{Y}}_{...} = \frac{1}{rcn} \sum_{i=1}^r \sum_{j=1}^c \sum_{k=1}^n \mathbf{Y}_{ijk}$.

In univariate ANOVA, the F-test is based on ratios of the sums of squares, which represent the differences between groups. However in MANOVA, Wilk's lambda is defined as $\Lambda = \frac{|T_B|}{|T_B + T_W|} = \frac{|T_B|}{|T|}$ and it is equivalent to the likelihood ratio test of the hypothesis of equal group means.

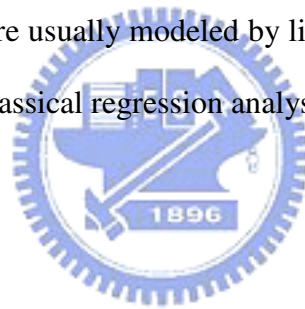
After building the MANOVA model for our data, the model will be used to identify the group which a new observation comes from. The objective of discriminant analysis is that when given the existence of several distinct groups of individuals, and a sample of observations from each group, find the functions of these

observations that can distinguish the groups and enable future unidentified individuals to be classified to their correct group. Multivariate analysis of variance provides the basis for inferential argument about the significance of differences between groups and the importance of particular variables. Therefore, we use the result of multivariate analysis of variance as the criterion to discriminate the radar back wave data.

3.2 Semi-parametric Linear Mixed Effects Model (SLM)

3.2.1 Linear Mixed Effects Model

Our radar response data could also be regarded as a repeated measurement data. Repeated measurement data are usually modeled by linear mixed effects model which can be generalized from the classical regression analysis through a two-stage analysis.



Two-Stage Analysis

Let the random variable Y_{ij} denote the response for the i th individual measured at time t_{ij} , $i=1, \dots, N$, $j=1, \dots, n_i$, N is the number of individuals, n_i is the number of observations for i th individual, and Y_i be the n_i -dimensional vector of all repeated measurements for the i th subject, that is $Y_i = (Y_{i1}, Y_{i2}, \dots, Y_{in_i})'$.

The first stage of the two-stage approach assumes that Y_i satisfies the linear regression model

$$Y_i = Z_i \alpha_i + \varepsilon_i \quad (3.2.1)$$

where Z_i is a $(n_i \times q)$ matrix of known covariates, represents how the response

evolves over time for the i th subject. α_i is a q -dimensional vector of unknown regression coefficients, and ε_i is a vector of residual components ε_{ij} , $j=1, \dots, n_i$. We assume that all ε_i are independent and normally distributed with mean zero and covariance matrix $\sigma^2 I_{n_i}$, where I_{n_i} is the n_i -dimensional identity matrix.

The next step, a multivariate regression model of the form

$$\alpha_i = K_i \alpha + b_i \quad (3.2.2)$$

is used to explain the observed variability between the subjects and their regression coefficients α_i . K_i is a $(q \times p)$ matrix of known covariates, and α is a p -dimensional vector of unknown regression parameters. And b_i s are assumed to be independent random terms, following a q -dimensional normal distribution with mean zero and covariance matrix D .



The General Linear Mixed Effects Model

Combining the models from two-stage analysis, we replace α_i in (3.2.1) by the equation (3.2.2), it becomes

$$Y_i = X_i \alpha + Z_i b_i + \varepsilon_i \quad (3.2.3)$$

where Y_i is the n_i -dimensional response vector for the subject i , $X_i = Z_i K_i$ and Z_i are $(n_i \times p)$ and $(n_i \times q)$ dimensional matrices of known covariance. α and b_i are p -dimensional and q -dimensional vectors containing the fixed and random effects, and ε_i is an n_i -dimensional vector of residual components.

Therefore the equation (3.2.3) is called a linear mixed effect model with fixed

effect α and random effect b_i . And we assume that

$$\begin{cases} b_i \sim N(0, D) \\ \varepsilon_i \sim N(0, \Sigma_i) \end{cases}$$

where D is a $(q \times q)$ covariance matrix with (i, j) element $d_{ij} = d_{ji}$ and Σ_i is a $(n_i \times n_i)$ covariance matrix which will not depend upon i .

As a result, the model (3.2.3) has the ability to model the mean structure (fixed effects) and the covariance structure (random effects and random residuals) simultaneously.

3.2.2 Semi-Parametric Model

To illustrate semi-parametric regression, let us take a look at a specific example. Let Y be *log wages* and consider the explanatory variables *schooling*, and labor market *experience*. If we assume *log wages* are linearly related to these explanatory variables, then the linear regression model is as follow:

$$E(Y|school, exp) = \beta_0 + \beta_1 \cdot school + \beta_2 \cdot exp \quad (3.2.4)$$

Equation (3.2.4) is a parametric regression model that we have already known.

Suppose we want to estimate

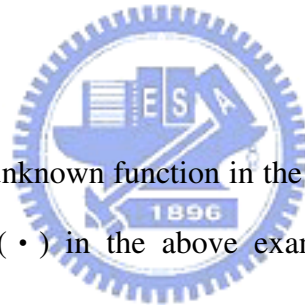
$$E(Y|school, exp) = m(school, exp) \quad (3.2.5)$$

and assume that $m(\cdot)$ is a smooth function. Then equation (3.2.5) is a nonparametric regression model. And now, suppose the regression function of *log wages* on *schooling* and *experience* has the following shape:

$$E(Y|school, exp) = \alpha + g_1(school) + g_2(exp) \quad (3.2.6)$$

Here $g_1(\cdot)$ and $g_2(\cdot)$ are two unknown smooth functions of main effects and α is an unknown parameter. Therefore equation (3.2.6) combines the additive structure of the parametric regression model with the flexibility of the nonparametric approach, which is also called the generalized additive model (GAM). Having both parametric and nonparametric components means the models are semi-parametric. Hence, equation (3.2.6) is a semi-parametric regression model.

In brief, semi-parametric model is the model that combines the parametric estimator and nonparametric function. But some scholars consider that as long as the model has nonparametric parts then it should be called nonparametric models.



In order to estimate the unknown function in the semi-parametric model, such as the functions $g_1(\cdot)$ and $g_2(\cdot)$ in the above example, nonparametric regression estimators have to be utilized. That means when estimating semi-parametric models we usually have to use nonparametric skills. Therefore in this thesis, we would illustrate how to use nonparametric techniques to estimate the unknown functions in the semi-parametric model.

3.2.3 Semi-parametric Linear Mixed Effects Model

Most previous papers, e.g. Rice (1991), Wang (1998), ... etc., considered on special models for special data. In this thesis, we suggest a general family of semi-parametric mixed effects models for most data sets. The fixed effects are modeled by general smoothing spline models which are estimated by maximizing the

penalized likelihood. The random effects are modeled parametrically by assuming that the covariance function depends on a parsimonious set of parameters. We estimate these parameters and the smoothing parameter simultaneously by the generalized maximum likelihood (GML) method.

A general form of semi-parametric mixed effects model:

$$Y = f + Zb + \varepsilon \quad (3.2.7)$$

where $Y = (Y_1, Y_2, \dots, Y_N)'$, $f = (f(t_1), f(t_2), \dots, f(t_N))'$, $\varepsilon = (\varepsilon_1, \varepsilon_2, \dots, \varepsilon_N)'$, $t_i \in \mathcal{T}_1$, \mathcal{T}_1 is any arbitrary domain, and $f \in \mathcal{H} = \mathcal{H}_0 \oplus \mathcal{H}_1$. $b \sim N(0, \sigma^2 D)$, $\varepsilon \sim N(0, \sigma^2 \Lambda)$, where D and Λ are some symmetric positive definite matrices, and they are mutually independent.

$f \in \mathcal{H}$, where \mathcal{H} is a reproducing kernel Hilbert space (RKHS) of real-valued functions on \mathcal{T}_1 . \mathcal{H} can be decomposed into $\mathcal{H}_0 \oplus \mathcal{H}_1$, where \mathcal{H}_0 is a finite dimensional space containing terms which will not be penalized. The penalized likelihood estimate of f in \mathcal{H} is the solution of

$$\min_{f \in \mathcal{H}} \left(\frac{1}{N} \|Y - f\|^2 + \lambda \|P_1 f\|^2 \right)$$

where P_1 is the orthogonal projection of f onto \mathcal{H}_1 in \mathcal{H} . λ is a smoothing parameter. See Wahba (1990) for details on RKHS and general smoothing spline models.

Model (3.2.7) is very general. $f \in \mathcal{H}$ is a function on any arbitrary domain \mathcal{T}_1 . For example, \mathcal{H} are polynomial splines on $\mathcal{T}_1 = [0,1]$ for growth curves. \mathcal{H} may also be a subset of the tensor product of some RKHSs. For example, let $\mathcal{T}_1 = [0,1] \otimes \mathcal{T}_2 \otimes \dots \otimes \mathcal{T}_d$, where $\mathcal{T}_2, \dots, \mathcal{T}_d$ are covariates such as treatment group, sex, age

etc. : we can build a smoothing spline ANOVA model (Wahba, 1990) for longitudinal data.

Smoothing Spline Analysis of Variance Decompositions

Smoothing spline structure can be used to model mixed effects by using smoothing spline ANOVA decompositions. We would describe how to use smoothing spline ANOVA decompositions to build main effects and interactions that can be interpreted as in classical ANOVA.

Our sample consists of curves of the intensity of the back wave of cars from different groups. The aims of our analysis are to model two main fixed effects, time and group, and one random effect of the individual cars in each group and then use the model to classify a new observed curve into a specific group.

There are three factors involved in our model: two categorical covariates *group* and *car* and a continuous covariate *time*. From the design of the experiment, the *time* and *group* factors are fixed, and the *car* factor is nested within the group factor which will be treated as random effect. For *group* k , denote \mathcal{B}_k as the population from which the cars in *group* k were drawn with sampling distribution $P_{w|k}$.

Assume the model

$$y_{kwj} = f(k, w, t_j) + \epsilon_{kwj} ; k = 1, \dots, g ; w \in \mathcal{B}_k ; t_j \in [0,1] ,$$

where y_{kwj} is the back wave intensity at *time* t_j of *car* w in the population \mathcal{B}_k , $f(k, w, t_j)$ is the “true” mean intensity at *time* t_j of *car* w in the population \mathcal{B}_k , and $\epsilon_{wk} \stackrel{\text{iid}}{\sim} N(0, \sigma^2)$ and is independent of $f(k, w, t_j)$. $f(k, w, t_j)$ is a function defined on $\{\{1\} \otimes \mathcal{B}_1, \dots, \{g\} \otimes \mathcal{B}_g\} \otimes [0,1]$. Note that $f(k, w, t_j)$ is a random

variables since w is a random sample from \mathcal{B}_k . What we observe are realizations of this “true” mean function plus random errors. We use label i to denote the car we actually observe.

For the time effects, the reproducing kernel Hilbert space

$$W_2 = \{f : f^{(v)} \text{ absolutely continuous, } v = 0, 1, f^{(2)} \in \mathcal{L}_2[0,1] \}$$

is decomposed into $W_2 = \{1\} \oplus \{t - 0.5\} \oplus H_1$ where $\{1\}$ and $\{t - 0.5\}$ represent linear spaces spanned by the constant 1 and linear basis $t - 0.5$. H_1 , the orthogonal complement of span $\{1, t - 0.5\}$ in W_2 , is a RKHS with RK

$$R_1(s, t) = k_2(s)k_2(t) - k_4(s - t) \quad (3.2.8)$$

where $k_v(x) = B_v(x)/v!$ and $B_v(x)$ is the v th Bernoulli polynomial.

For the group effect, the RKHS G is decomposed into $G = \{1\} \oplus H_2$, where $\{1\} = \{z \in G: z(1) = z(2) = \dots = z(g)\}$ and $H_2 = \{z \in G: \sum_{i=1}^g z(i) = 0\}$ with RK

$$R_2(i, j) = I_{[i=j]} - 11'/g, \quad (3.2.9)$$

where I is the identity matrix and 1 is the vector with 1 in every entry. Clearly, $\{1\}$ and H_2 are orthogonal.

Define the following averaging operators:

$$\left\{ \begin{array}{l} A_1 f = \sum_{k=1}^g A_2 f(k, w, t)/g \\ A_2 f = \int_{\mathcal{B}_k} f(k, w, t) dP_{w|k} \\ A_3 f = \int_0^1 f(k, w, t) dt \\ A_4 f = \left\{ \int_0^1 \frac{\partial}{\partial t} f(k, w, t) dt \right\} (t - 0.5) \end{array} \right.$$

Therefore, an AVOVA-like decomposition can be defined as

$$\begin{aligned}
f &= [(A_1 + (A_2 - A_1) + (I - A_2)][A_3 + A_4 + (I - A_3 - A_4)]f \\
&= A_1A_3f + A_1A_4f + A_1(I - A_3 - A_4)f \\
&\quad + (A_2 - A_1)A_3f + (A_2 - A_1)A_4f + (A_2 - A_1)(I - A_3 - A_4)f \\
&\quad + (I - A_2)A_3f + (I - A_2)A_4f + (I - A_2)(I - A_3 - A_4)f \\
&= \mu + \beta(t - 0.5) + s_1(t) + \xi_k + \delta_k(t - 0.5) + s_2(k, t) \\
&\quad + \alpha_{w(k)} + \gamma_{w(k)}(t - 0.5) + s_3(k, w, t) \tag{3.2.10}
\end{aligned}$$

where μ is the grand mean, $\beta(t - 0.5)$ is the linear main effect of *time*, $s_1(t)$ is the smooth main effect of *time*, ξ_k is the main effect of *group*, $\delta_k(t - 0.5)$ is the linear interaction between *group* and *time*, $s_2(k, t)$ is the smooth interaction between *group* and *time*, $\alpha_{w(k)}$ is main effect of *car*, $\gamma_{w(k)}(t - 0.5)$ is the linear interaction between *car* and *time*, $s_3(k, w, t)$ is the smooth interaction between *car* and *time*.

The first three terms in equation (3.2.10) are fixed time effects. They are orthogonal components in the RKHS

$$W_2 = \{f : f^{(v)} \text{ absolutely continuous, } v = 0, 1, f^{(2)} \in \mathcal{L}_2[0, 1] \}.$$

The fourth to sixth terms in (3.2.10) are the fixed group effects. The identifiability will be fulfilled because of the orthogonality of the subspaces. The last three terms in (3.2.10) are random effects. The last two terms are interactions between the individuals and the time.

We will assume that $\begin{pmatrix} \alpha_{w(k)} \\ \gamma_{w(k)} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \sigma^2 D_k\right)$ and $s_3(k, w, t)$, independent of $\alpha_{w(k)}$ and $\gamma_{w(k)}$, is a realization of some stochastic process on $G \times [0, 1]$ with mean 0 and covariance function $\sigma_1^2 R_2(i, j) R_1(s, t)$, where $R_1(s, t)$ and $R_2(i, j)$ are defined in (3.2.8) and (3.2.9), respectively. The comment by Wang and Wahba (1998)

for Brumback and Rice (1998) can be consulted for more details concerning the mathematical basis for model structure.



4. Empirical Illustration

As an example of our methods in the previous section, here we present analyses of a data set recorded by a radar microwave detector.

4.1 Data Collection

The practical data was recorded by the vehicle radar microwave detector which is a side-looking device. Figure 1 shows the miniature of the radar microwave detector. The data set was collected near the section 1 of Singlong Road, Jhubei City, Hsinchu Country (新竹縣竹北市興隆路一段). There are four lanes on the road and the instrument is set up by the side of the Singlong Road. The sketch of the relative positions of the device and the range covered by the transmitter is shown in Figure 2.

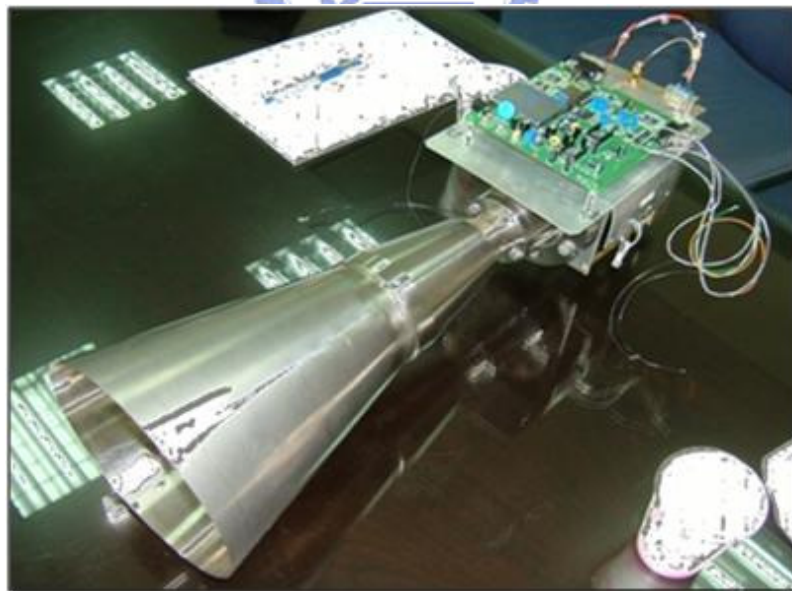


Figure 1: The miniature of the vehicle radar microwave detector.

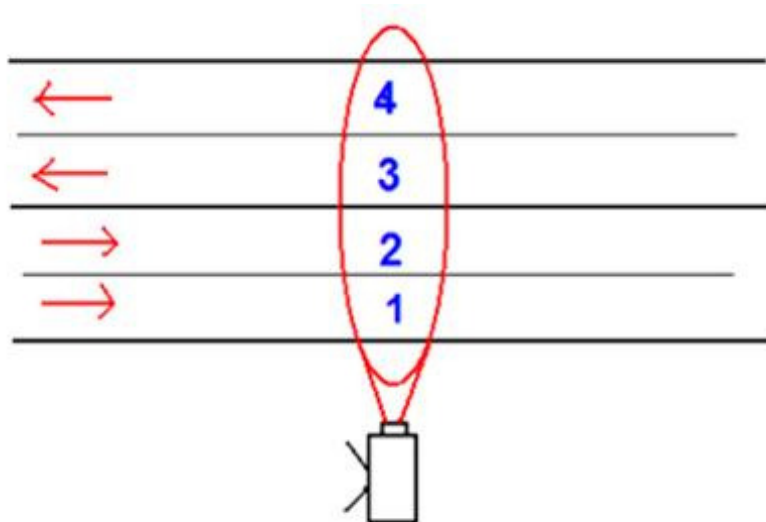


Figure 2: The sketch of the relative positions of the device and the range covered by the transmitter.

The data was collected from 10:00 AM to 5:00 PM on October 11, 2006. The original data consists of radar back waves of small and large vehicles; sedans belong to small vehicles, whereas cranes, buses, trucks and goods wagons, etc. were classified as the large vehicles. There dataset comprises of 248 observed curve data of different vehicles. At this stage of our experiment, the speed of small cars is controlled at 40 km/hour on the four lanes of Singlong Road.

Nevertheless, the data was recorded at the unit time (2×10^{-5} sec) and the recorder made a note every 512 data points. We take the maximum (the most powerful intensity of the radar back wave) of each 512 data as the data to be analyzed and the frequency is one observation for every $512 \times (2 \times 10^{-5} \text{ sec}) = 1.024 \times 10^{-2}$ second. After adjusting to the new time unit, typical curves of intensity of back waves for small and large vehicles on lane 1, 2, 3, 4 are shown in Figure 3 and Figure 4, respectively. And they are plot on the same scale in Figure 5.

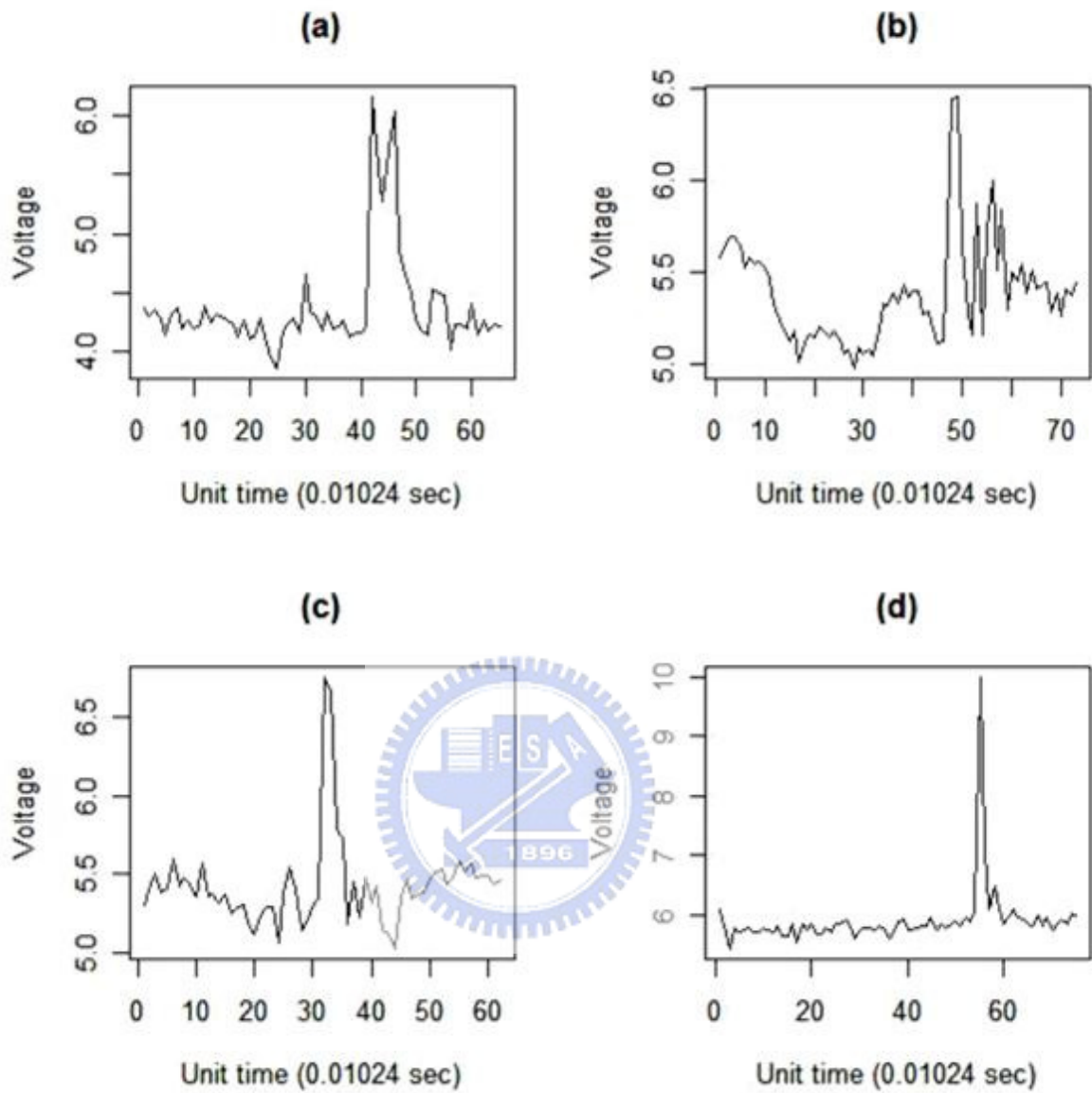


Figure 3: A typical curve of the intensity of back wave for small vehicle. (a) Lane 1; (b) Lane 2; (c) Lane 3; (d) Lane 4.

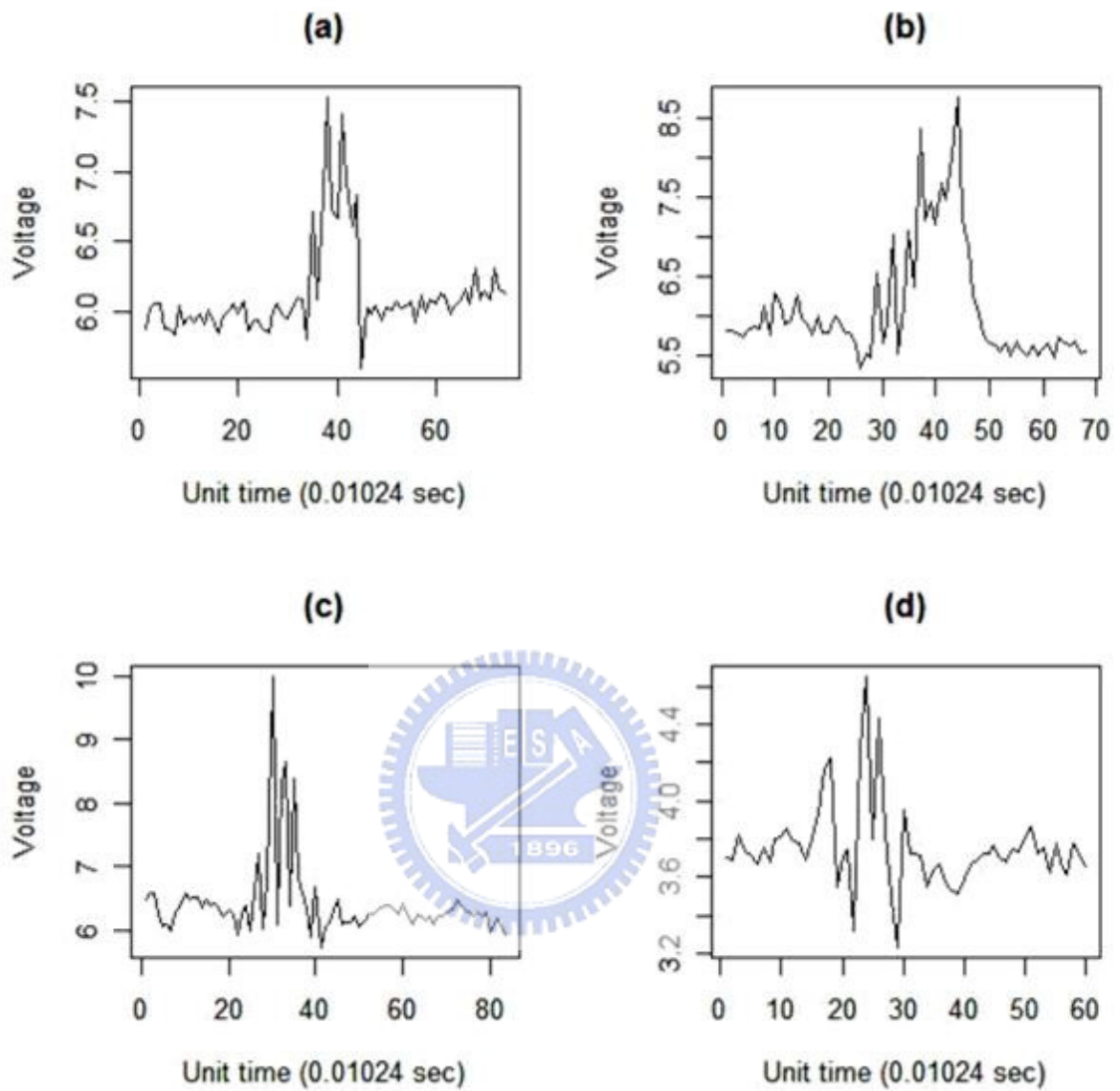


Figure 4: A typical curve of the intensity of back wave for large vehicle. (a) Lane 1; (b) Lane 2; (c) Lane 3; (d) Lane 4.

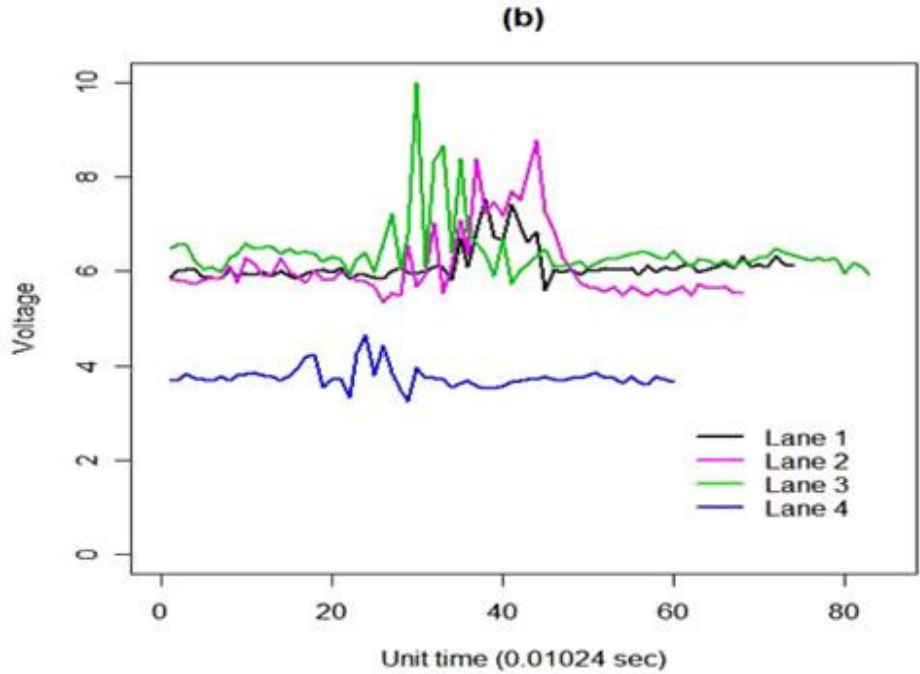
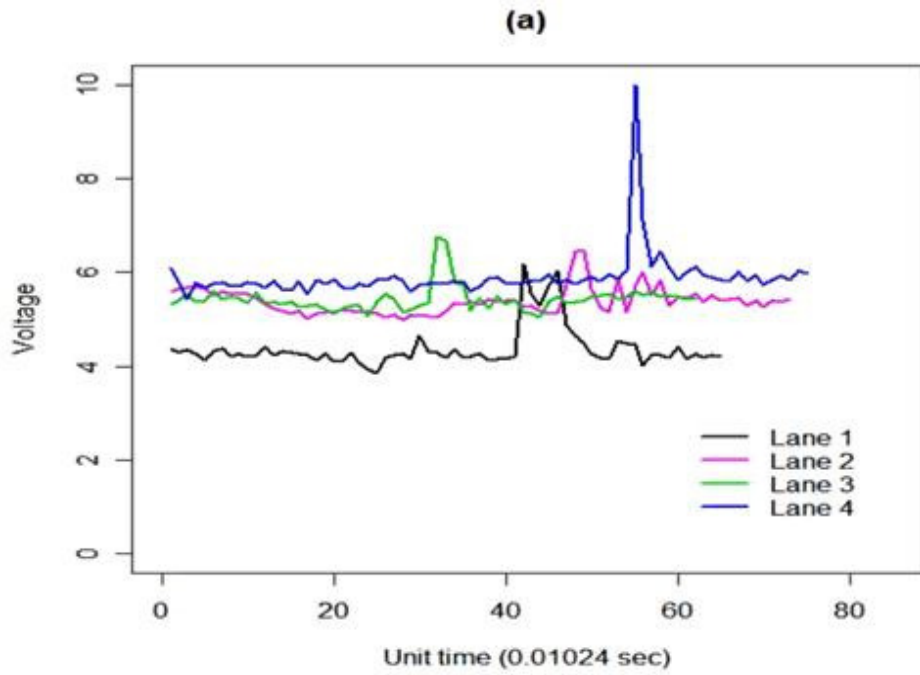


Figure 5: The typical curves for small and large vehicles plotted on the same scale.
 (a) small vehicles; (b) large vehicles.

It can be seen from Figure 5 that the magnitudes and the peaks are different for different types of vehicles. The relatively low magnitudes at the beginning are caused by the time lag of manual operation. We are interested in classify the types of vehicles according to the intensity. Because of the lengths of the data are different, we have to regulate the curves of the back wave in order to analyze them under the same criterions, so called registration¹ of functional data (Ramsay and Li, 1998). Here, the peak of the curve is regarded as the marker. We make the maximum intensity as the center of the data by taking 15 points of the new unit time before the peak and 14 points of the new unit time after the peak. The resulting 30 numbers observed at discrete time are the data that we use to analyze.

4.2 Model Description and Data Analysis

Our main goal is to classify the vehicles according to their types and the lanes they are running. There are two classes of vehicles: large and small vehicles, and four lanes of the Singlong Road, therefore there are 8 groups in total. In order to examine the results of our discrimination analyses, we separate the dataset into two parts: one comprises the learning sample, and the other is the testing sample. The former is set up for building our models and the later is taken for authenticating the power of discrimination. We use our models to get the predicted groups of the testing samples, and then the correct classification rate is computed as the proportion of correctly

¹ If the observed responses from experimentation were viewed as continuous curves rather than scalars or vectors, this kind of data are called functional data sets. An essential preliminary to a functional data analysis is often the registration or alignment of a salient curve features. Marker registration is the process of aligning curves by identifying the timing of certain salient features in the curves. Using this strategy, curves are aligned by transforming time so that marker events occur at the same values of the transformed time.

predicted groups. Here we select 16 curves from each group randomly, 8 curves of them would be the learning sample set, and the other 8 curves are the testing sample set. Figure 6 and Figure 7 display the 8 learning samples of the back wave data for each type of vehicles on lanes 1, 2 and lanes 3, 4, respectively. And the 8 learning samples of the back wave data of small and large vehicles on 4 lanes is showed in Figure 8.

This thesis will apply two different statistical approaches to analyzing the same learning sample. Due to the data is a 30-dimensional vector of time index, we can utilize the traditional parametric statistical approach, multivariate analysis of variance, to analyze the learning sample. Furthermore, the same sample would also be dealt with the semi-parametric model which combines the parametric and nonparametric function.



4.2.1 Multivariate Analysis of Variance (MANOVA)

By transforming the back wave data into a 30-dimensional vector of time index data set, we could assume the multivariate linear model as follow:

$$y_{ijkt} = \mu_t + \alpha_{it} + \beta_{jt} + (\alpha\beta)_{ijt} + \varepsilon_{ijkt}$$

where α represents the lane effect, $i = 1, \dots, 4$; β indicates the type of vehicle effect, $j = 1, 2$ (“1” represents the large vehicle, and “2” stands for the small vehicle); $(\alpha\beta)$ is the interaction effect; $k = 1, \dots, n_{ij}$ is the numbers of observation in i th lane and j th type of vehicle; $t = 1, \dots, 30$ represents the 30 unit time index; and we assume that $\varepsilon_{ijkt} \stackrel{\text{iid}}{\sim} N(0, \sigma^2)$.

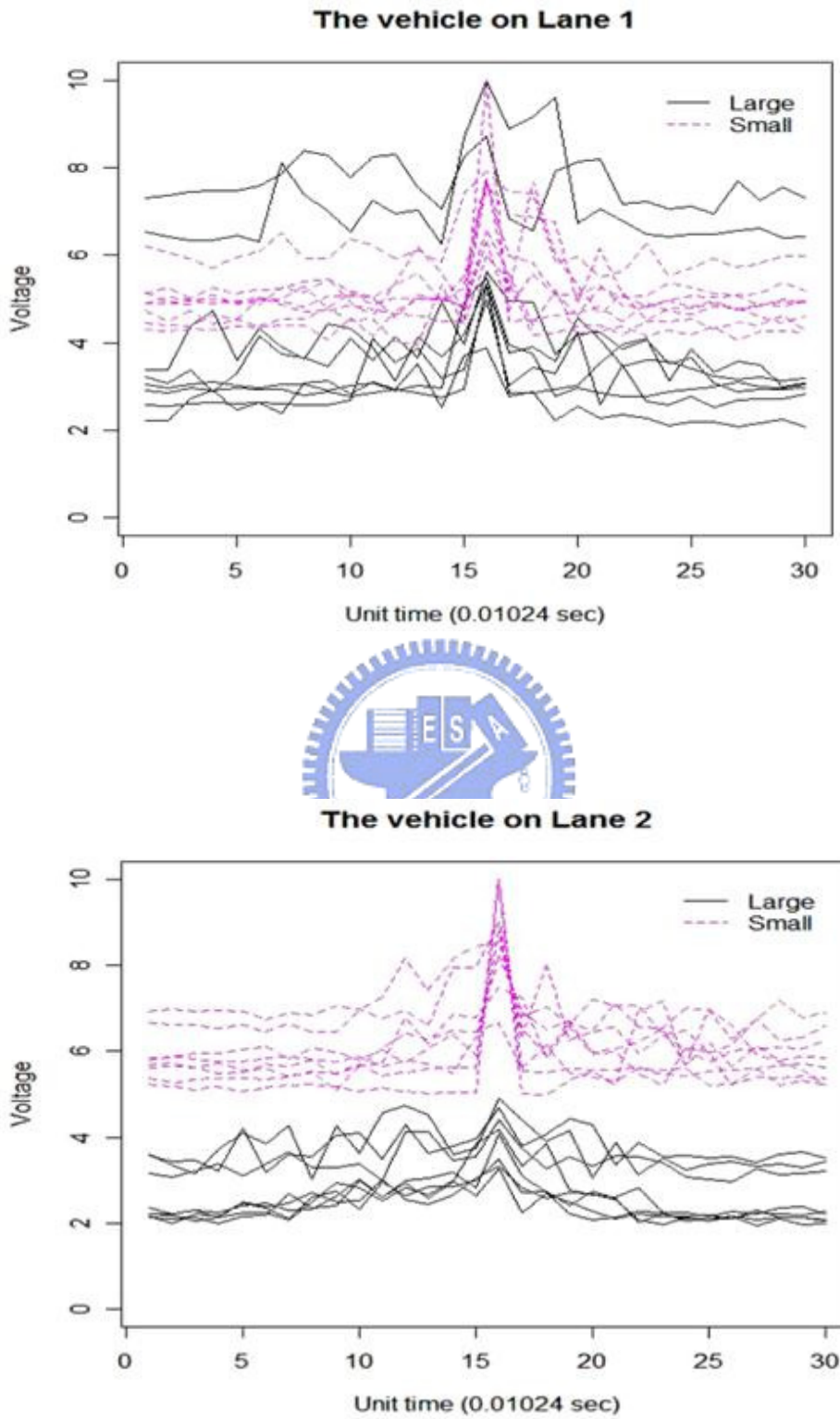


Figure 6: The 8 learning samples of the back wave data of each type of vehicles on Lanes 1, 2.

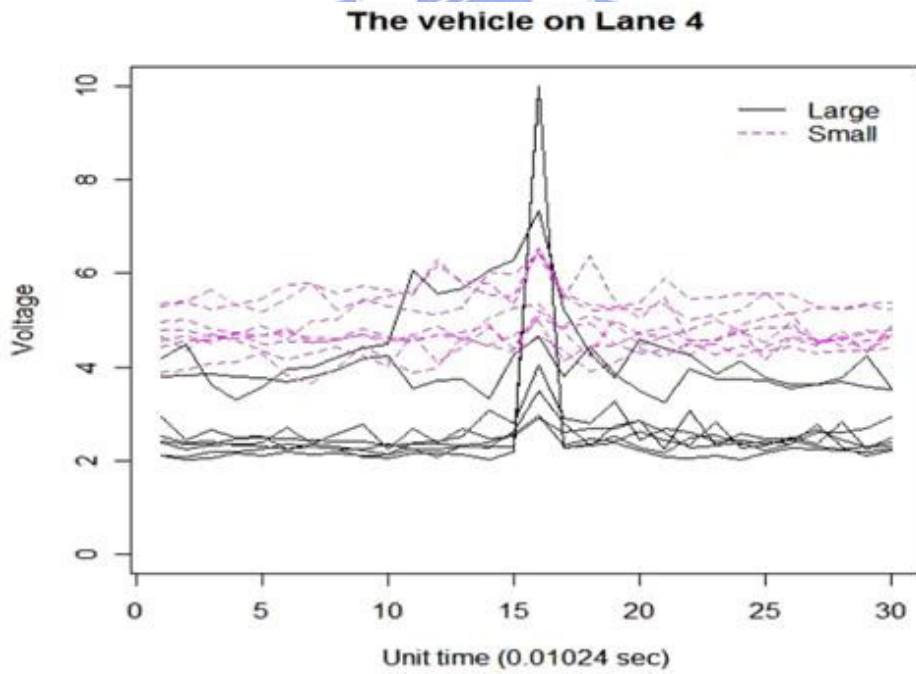
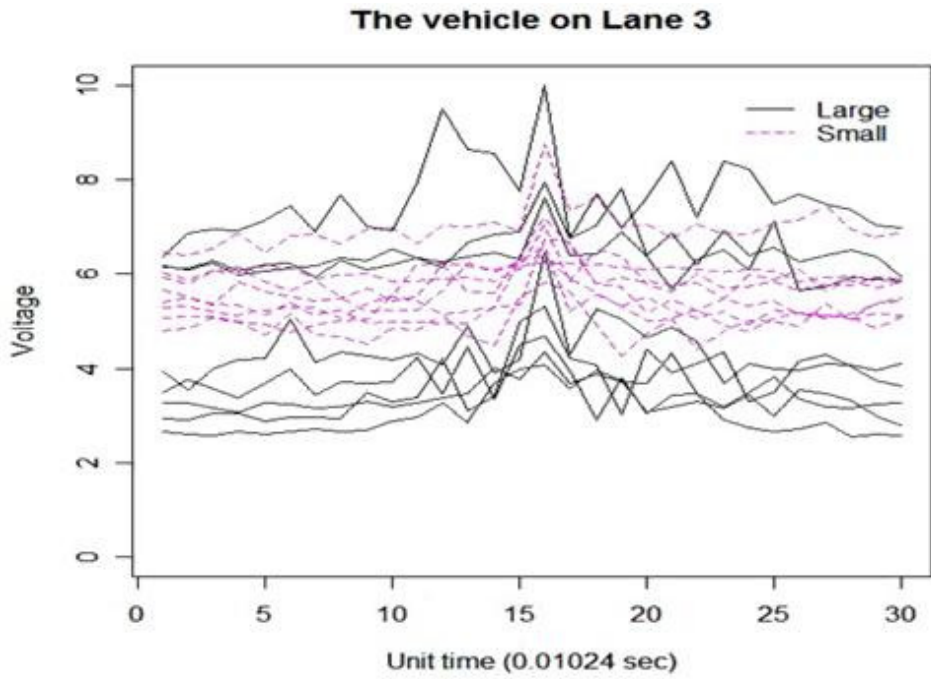


Figure 7: The 8 learning samples of the back wave data of each type of vehicles on Lanes 3, 4.

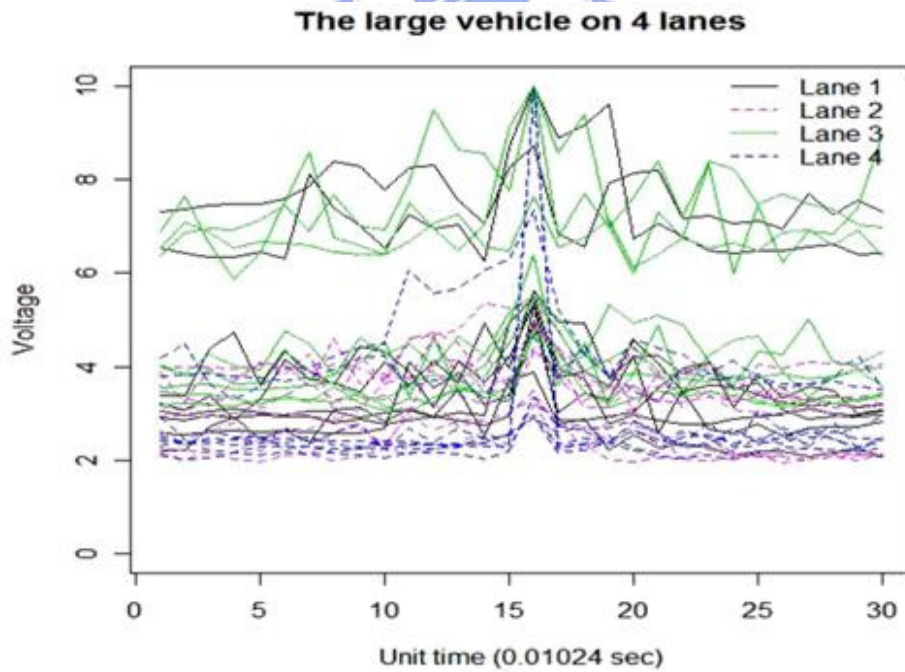
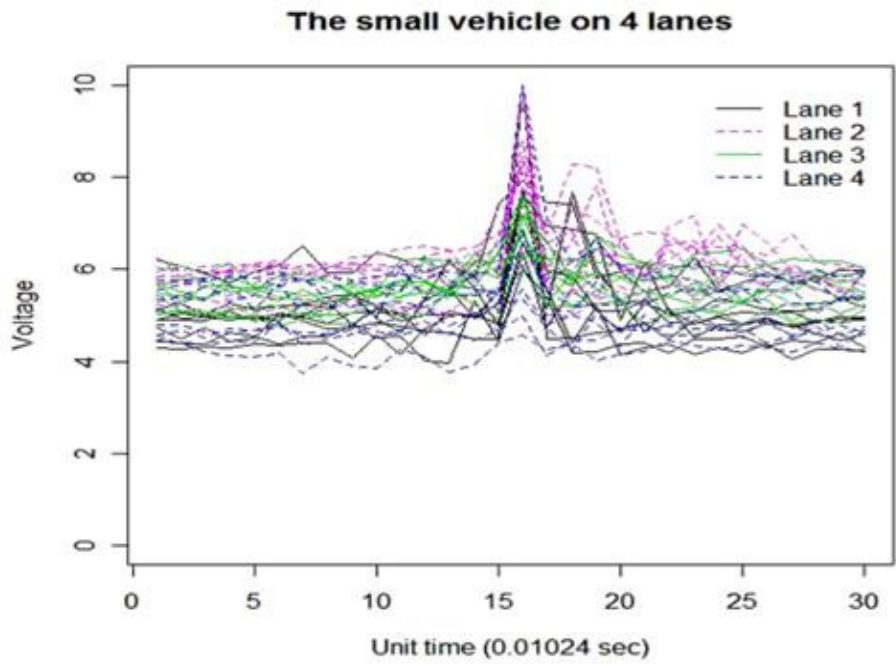


Figure 8: The 8 learning samples of the back wave data of small and large vehicles on 4 lanes.

In multivariate analysis of variance, Wilks' lambda statistic can be regarded as the likelihood ratio test of the hypothesis of equal group means. Therefore, the following Table 1 shows the multivariate table that uses Wilks' lambda as the test statistic of equal group means. Furthermore, the randomly selected 8 curves as the learning sample of small and large vehicles on each lane and the fitted value calculated by multivariate analysis of each group are displayed in Figure 9 and Figure 10.

Table 1: The table of MANOVA by using Wilks' test statistic

| Source of variation | df | Wilks | approx F | num df | den df | Pr(>F) |
|---------------------|----|--------|----------|--------|--------|---------------|
| lane | 3 | 0.0503 | 1.5579 | 90 | 81.69 | 0.02135 * |
| car | 1 | 0.1251 | 6.2926 | 30 | 27.00 | 3.307e-06 *** |
| lane × car | 3 | 0.0488 | 1.5822 | 90 | 81.69 | 0.01802 * |
| residuals | 56 | | | | | |

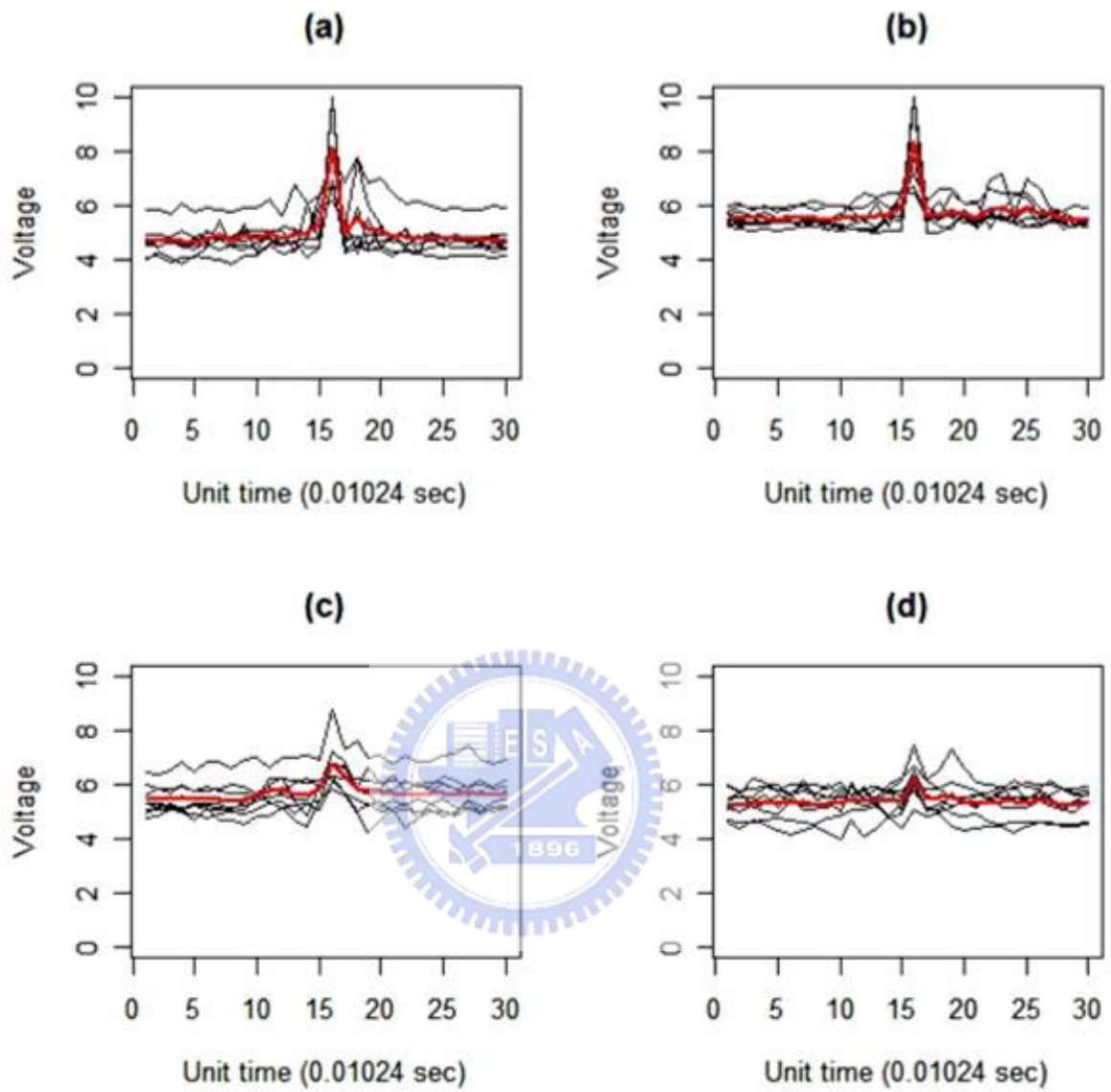


Figure 9: The 8 black curves as the learning sample of small vehicles and the red lines represent the fitted value of each group (a) Lane 1; (b) Lane 2; (c) Lane 3; (d) Lane 4.

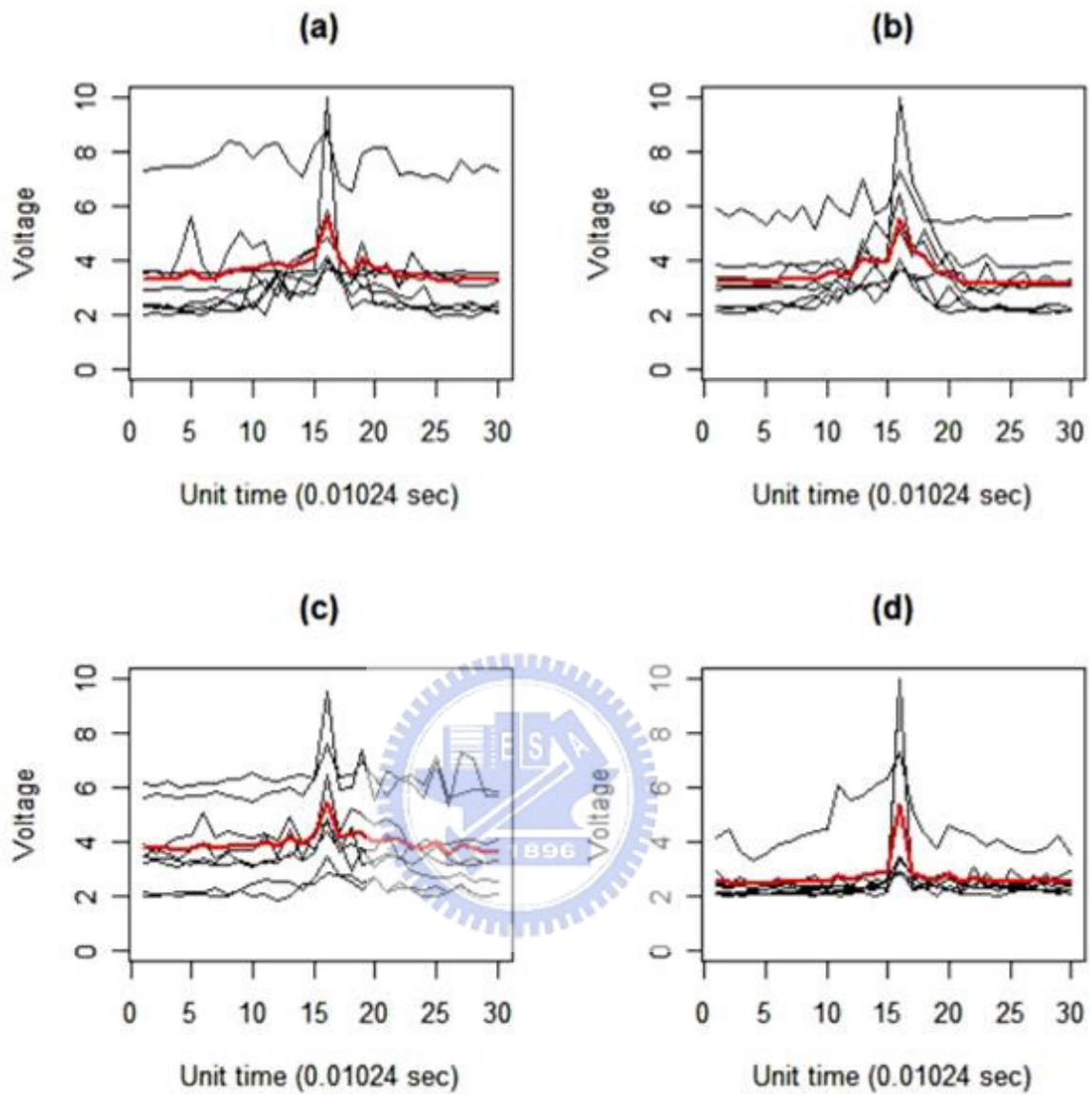


Figure 10: The 8 black curves as the learning sample of large vehicles and the red lines represent the fitted value of each group (a) Lane 1; (b) Lane 2; (c) Lane 3; (d) Lane 4.

4.2.2 Semi-parametric Linear Mixed Effects Model (SLM)

Since we randomly select 8 curves from each 8 group as the learning sample set, that means there have 64 car curves would be assigned to 8 groups. There are two categorical covariates *group* and *car* and a continuous covariate *time*. We code the *group* factor as 1 to 8 and the observed *car* factor as 1 to 64, and transform the *time* variable into $[0,1]$. There are two fixed factors, *group* and *time*, and the *car* factor is nested within the *group* factor therefore we treat *car* as a random factor. For *group* k , denote \mathcal{B}_k as the population from which the cars in *group* k were drawn. Assume the following model:

$$y_{kwj} = f(k, w, t_j) + \epsilon_{kwj} ; k = 1, \dots, 8 ; w \in \mathcal{B}_k ; t_j \in [0,1] ; j=1, \dots, 30 ,$$

where y_{kwj} is the back wave intensity at *time* t_j of *car* w in the population \mathcal{B}_k , $f(k, w, t_j)$ is the “true” mean intensity at *time* t_j of *car* w in the population \mathcal{B}_k , and ϵ_{kwj} ’s are random errors. $f(k, w, t_j)$ is a function defined on $\{\{1\} \otimes \mathcal{B}_1, \{2\} \otimes \mathcal{B}_2, \{3\} \otimes \mathcal{B}_3, \{4\} \otimes \mathcal{B}_4, \{5\} \otimes \mathcal{B}_5, \{6\} \otimes \mathcal{B}_6, \{7\} \otimes \mathcal{B}_7, \{8\} \otimes \mathcal{B}_8\} \otimes [0,1]$.

Note that $f(k, w, t_j)$ is a random variables since w is a random sample from \mathcal{B}_k .What we observe are realizations of this “true” mean function plus random errors.

Then by the SS ANOVA decomposition

$$\begin{aligned} f = & \mu + \beta(t - 0.5) + s_1(t) + \xi_k + \delta_k(t - 0.5) + s_2(k, t) \\ & + \alpha_{k(w)} + \gamma_{k(w)}(t - 0.5) + s_3(k, w, t) \end{aligned} \quad (4.2.1)$$

where μ is a constant, $\beta(t - 0.5)$ is the linear main effect of *time*, $s_1(t)$ is the smooth main effect of *time*, ξ_k is the main effect of *group*, $\delta_k(t - 0.5)$ is the linear interaction between *time* and *group*, $s_2(k, t)$ is the smooth interaction between *time* and *group*, $\alpha_{k(w)}$ is the main effect of *car*, $\gamma_{k(w)}(t - 0.5)$ is the linear interaction between *time* and *car*, and $s_3(k, w, t)$ is the smooth interaction between *time* and *car*.

We calculate the main effect of *time* as $\beta(t - 0.5) + s_1(t)$, the interaction between *time* and *group* as $\delta_k(t - 0.5) + s_2(k, t)$, and the interaction between *time* and *car* as $\gamma_{k(w)}(t - 0.5) + s_3(k, w, t)$. The first six terms in the equation (4.2.1) are fixed effects, and the last three terms in equation (4.2.1) are random effects since they depend on the random variable w .

In the equation (4.2.1), the first three terms, depending on *time* only, represent the mean curve for all cars. The middle three terms measure the departure of a particular group from the population mean curve. The last three terms measure the departure of a particular car from the mean curve of a population from which the car was chosen.

Therefore, by the equation (4.2.1), we can fit three models.

Model 1: assuming there is no interaction between cars and time.

$$\begin{aligned} y_{kwj} &= f_1(k, w, t_j) + \epsilon_{kwj} \\ &= \mu + \beta(t - 0.5) + s_1(t) + \xi_k + \delta_k(t - 0.5) + s_2(k, t) + \alpha_{k(w)} + \epsilon_{kwj} \end{aligned}$$

It has a different population mean curve for each group plus a random intercept for each car. We assume that $\alpha_i \stackrel{\text{iid}}{\sim} N(0, \sigma_\alpha^2)$, $\epsilon_{kij} \stackrel{\text{iid}}{\sim} N(0, \sigma^2)$, and they are mutually independent.

Model 2: assuming there is no smooth interaction between cars and time.

$$\begin{aligned} y_{kwj} &= f_2(k, w, t_j) + \epsilon_{kwj} \\ &= \mu + \beta(t - 0.5) + s_1(t) + \xi_k + \delta_k(t - 0.5) + s_2(k, t) \\ &\quad + \alpha_{k(w)} + \gamma_{k(w)}(t - 0.5) + \epsilon_{kwj} \end{aligned}$$

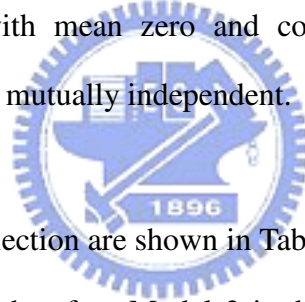
It has a different population mean curve for each group plus a random intercept and a

random slope for each car. We assume that $(\alpha_i, \gamma_i) \stackrel{\text{iid}}{\sim} N((0, 0), \text{diag}(\sigma_\alpha^2, \sigma_\gamma^2))$, $\epsilon_{kij} \stackrel{\text{iid}}{\sim} N(0, \sigma^2)$, and they are mutually independent.

Model 3:

$$\begin{aligned} y_{kwj} &= f_3(k, w, t_j) + \epsilon_{kwj} \\ &= \mu + \beta(t - 0.5) + s_1(t) + \xi_k + \delta_k(t - 0.5) + s_2(k, t) \\ &\quad + \alpha_{k(w)} + \gamma_{k(w)}(t - 0.5) + s_3(k, w, t) + \epsilon_{kwj} \end{aligned}$$

It has a different population mean curve for each group plus a random intercept, a random slope and a smooth random effect for each car. We assume that $(\alpha_i, \gamma_i) \stackrel{\text{iid}}{\sim} N((0, 0), \text{diag}(\sigma_\alpha^2, \sigma_\gamma^2))$, $s_3(k, w, t)$'s are stochastic process which are independent between cars with mean zero and covariance function $\sigma_2^2 R_1(s, t)$ ², $\epsilon_{kij} \stackrel{\text{iid}}{\sim} N(0, \sigma^2)$, and they are mutually independent.



The criteria for model selection are shown in Table 2. As we can see that the AIC of Model 3 is the minimum, therefore Model 3 is the most favorable. Furthermore, Figure 11 shows one random sample selected from the learning sample of the large vehicle on Lane 1 and the predicted curves of this sample from these three models. And from Figure 11, we find that the predicted curve from Model 2 is closer to the observed curve than Model 1, but the predicted curve from Model 3 is even closer than Model 2. This plot also shows that Model 3 is the most suitable model among these three models.

² $\sigma_2^2 = \sigma_1^2 R_2(i, j)$, $R_1(s, t) = [k_2(s)k_2(t) - k_4(s - t)]$, where $k_v(x) = B_v(x)/v!$ and $B_v(\cdot)$ is the v th Bernoulli polynomial.

Table 2: The criteria for model selection

| Model | AIC | BIC | logLikelihood | Test | L.Ratio | p-value |
|-------|----------|----------|---------------|--------|----------|---------|
| 1 | 2877.512 | 2921.984 | -1430.756 | | | |
| 2 | 2795.187 | 2850.777 | -1387.593 | 1 vs 2 | 86.32490 | <.0001 |
| 3 | 2710.626 | 2771.775 | -1344.313 | 2 vs 3 | 86.56081 | <.0001 |

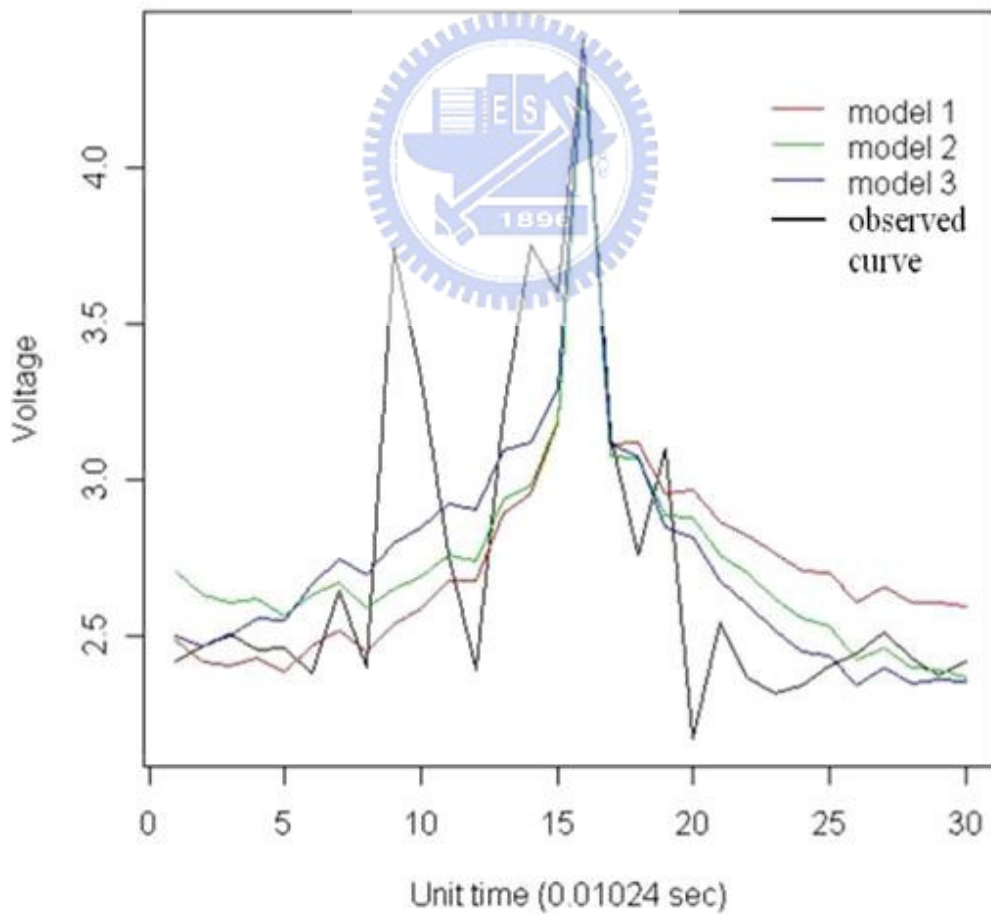


Figure 11: One random sample selected from the learning sample of the large vehicle on Lane 1 and the predicted curves of this sample from these three models.

After using the learning sample set to build models, we use the models built from the learning samples to acquire the mean curve of each group. And calculate the distance between the mean curve and the curve of the testing sample. Assign the testing sample to the group that has the smallest distance measure between the group mean curve and the testing sample. Then compute the correct classification rate by evaluating the proportion of correctly predicted groups. Table 3 displays the correct classification rate using MANOVA model and the three models of SLM for 8 groups over one run. Although we have already known that Model 3 of SLM is more favorable than Model 1 and Model 2 in the sense of goodness-of-fit, but the correct classification rate of Model 3 is not much higher than Model 1 and 2. Nevertheless, it takes much more time to fit Model 3, therefore under the limited time and the restriction of resources; we only use Model 1 and 2 to fit our real data. After repeated sampling the learning and testing samples from our data over 100 runs, we would obtain 100 correct classification rates from each model. Figure 12 displays the box-plot of the vehicle correct classification rate over 100 runs for three models (MANOVA model and the Model 1, Model 2 of SLM).

Table 3: The correct classification rate of MANOVA model and the three SLM models for 8 groups over one run

| | MANOVA | SLM | | |
|-----------------------------|-----------|-----------|-----------|-----------|
| | model | Model 1 | Model 2 | Model 3 |
| correct classification rate | 0.3392857 | 0.3928571 | 0.3571429 | 0.4285714 |

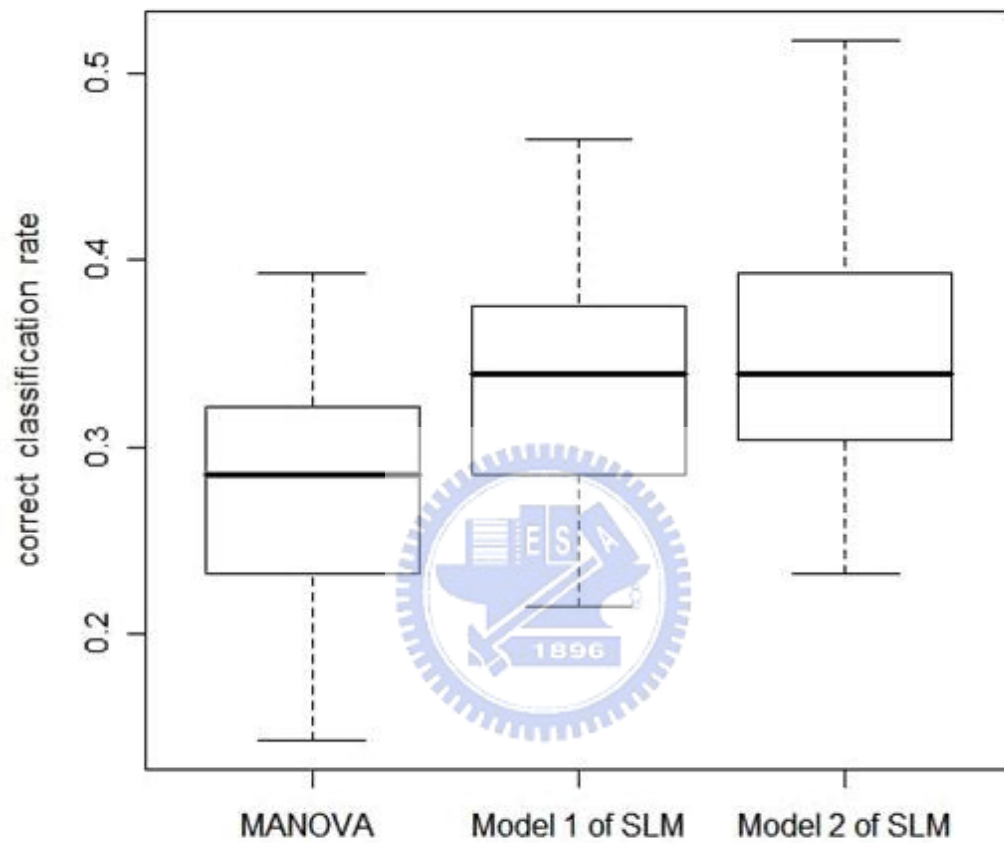
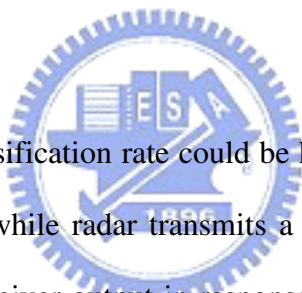


Figure 12: The box-plot of vehicle correct classification rate over 100 runs.

5. Conclusion and Discussion

This study proposed two methodologies to deal with radar back wave data and our goal is to classify the sizes of vehicles. The parametric multivariate analysis presents a traditional statistical approach and only takes very short time to do the classification. But as we can see from the fitted value of large vehicles on Lane 1 in Figure 8 and the predicted curve of one large vehicle on Lane 1 in Figure 9, the fitted curve of MANOVA is not as good as the curve predicted from SLM model. However, from Table 3, the correct classification rate of MANOVA doesn't seem too awful. Additionally, the SLM model costs too much time to predict the curve, it only shows a higher correct classification rate. Therefore these two models have respective advantages.



However the correct classification rate could be higher if we take account of the following two factors. First, while radar transmits a controlled, well-defined signal, the signal measured at the receiver output in response is the superposition of several major components. The major components are the target, clutter, noise and jamming. Noise and jamming are interference signals; they degrade the ability to measure targets. In some cases, clutters may be interference, too. However, our data were collected from the real-time back wave of the radar microwave detector, the data certainly include the target, clutter, noise and jamming. But we didn't eliminate those interference signals before analyzing data; therefore the result obtained from the analyses may have inaccuracy.

The other factor is that the number of observations is relatively few to classify them into eight groups. We have tried to classify the observations into only two

groups, ignoring the lane effect, and the correct classification rate is doubled. This phenomenon occurs for both models, i.e., both of these two models would work well when the number of groups are small but deteriorate rapidly when the number of groups increases.

The methodologies presented in this thesis offer some views of recognizing vehicles. Since the form, size and color of the vehicles on the road are all distinct, it is worth noting that the model of dissimilar vehicles may be different. Therefore, if the number of observations is enough, the model can be expanded for color effect or other possible effects.



Reference

1. Anderson, S. J. and Jones, R. H. (1995) Smoothing splines for longitudinal data. *Statistics in Medicine*, Vol. **14**, 1235-1248.
2. Bennett, C. L. and Toomey, J. P. (1981) Target classification with multiple frequency illumination. *IEEE Transactions on Antennas and Propagation*, Vol. **29**, No. 2, 352-358.
3. Brumback, B. A. and Rice, J. A. (1998) Smoothing spline models for the analysis of nested and crossed samples of curves. *Journal of the American Statistical Association*, Vol. **93**, No. 443. 961-994.
4. Chun, J. C., Kim, T. S., Kim, J. M., Lim, Z. S. and Park, W. S. (2001) Spectrum correlation of beat signals in the FM-CW radar level meter and application for precise distance. *IEEE*, Vol. **3**, 2251-2254.
5. Crowder, M. J. and Hand, D. J. (1990) Analysis of repeated measures. London/ Chapman & Hall.
6. Daniel, Wayne W. (1990) Applied nonparametric statistics. Boston/ PWS-KENT Pub. Co.
7. Davis, Charles S. (2002) Statistical methods for the analysis of repeated measurements. New York/ Springer.
8. Dawson, J. D. and Lagakos, S. W. (1991) Analyzing laboratory marker changes in AIDS clinical trials. *Journal of Acquired Immune Deficiency Syndromes*, Vol. **4**, 667-676.
9. Dawson, J. D. and Lagakos, S. W. (1993) Size and power of two-sample tests of repeated measures data. *Biometrics*, Vol. **49**, 1022-1032.
10. Ehrman, L. M. and Lanterman, A. D. (April 2003) Automated target recognition using passive radar and coordinated flight models. *Automatic Target Recognition XIII, Proceedings of SPIE, the International Society for Optical Engineering*. Vol. **5094**, (Orlando, FL) 196-207.

11. Ehrman, L. M. and Lanterman, A. D. (October 2003) Target identification using modeled radar cross section and a coordinated flight model. *Proceedings from the Third Multi-National Conference on Passive and Covert Radar*, (Seattle, WA)
12. Frison, L. and Pocock, S. J. (1992) Repeated measures in clinical trials: Analysis using mean summary statistics and its implications for design. *Statistics in Medicine*, Vol. **11**, 1685-1704.
13. Girden, E. R. (1992) ANOVA: repeated measures. Newbury Park, Calif. / Sage Publications.
14. Gibbons, J. D. (1993) Nonparametric statistics: an introduction. Newbury Park, Calif. / Sage Publications.
15. Greneker, E. F. and Rausch, E. O. (2004) Using radar to help mitigate truck overturn incidents on US interstate highways. *Proceedings of SPIE*, Vol. **5410**, 122-132.
16. Härdle, W., Müller, M., Sperlich, S. and Werwatz, A. (2004) Nonparametric and semiparametric models. Berlin/ Springer.
17. Herman, S. M. (2003) Joint passive radar tracking and target classification using radar cross section. *Proceedings of SPIE*, Vol. **5204**, 402-417.
18. Kohn, R. and Ansley, C. F. (1985) A structured state space approach to computing the likelihood and its derivatives. *Journal of Statistical Computing and Simulation*, Vol. **21**, 135-169.
19. Kohn, R. and Ansley, C. F. (1987) A new algorithm for spline smoothing based on smoothing a stochastic process. *SIAM Journal on Scientific and Statistical Computing*, Vol. **8**, 33-48.
20. Krzanowski, W. J. and Marriott, F. H. C. (1994) Multivariate analysis. Part 1. Distributions, ordination and inference. London/ Edward Arnold.
21. Krzanowski, W. J. and Marriott, F. H. C. (1994) Multivariate analysis. part 2. Classification, covariance structures and repeated measurements. London/

Edward Arnold.

22. Matthews, J. N. S., Altman, D. G., Campbell, M. J. and Royston, P. (1990) Analysis of serial measurements in medical research. *British Medical Journal*, Vol. **300**, 230-235.
23. Mardia, K. V., Kent, J. T. and Bibby, J. M. (1979) *Multivariate analysis*. Academic Press. A subsidiary of Harcourt Brace Jovanovich, Publishers.
24. O'Brien, R. G. and Kaiser, M. K. (1985) MANOVA method for analyzing repeated measures designs: An extensive primer. *Psychological Bulletin*, Vol. **97**, No. 2, 316-333.
25. Park, S. J., Kim, T. Y., Kang, S. M. and Koo, K. H. (2003) A novel signal processing technique for vehicle detection radar. *IEEE*, Vol. **1**, 607- 610.
26. Pocock, S. J. (1983) *Clinical Trials: A practical approach*. New York / John Wiley and sons.
27. Ramsay, J. O. and Dalzell, C.J. (1991) Some tools for functional data analysis. *Journal of the Royal Statistical Society. Series B (Methodological)*, Vol. **53**, No. 3, 539-572.
28. Ramsay, J. O. and Li, X. (1998) Curve registration. *Journal of the Royal Statistical Society. Serier B (Statistical Methodology)*, Vol. **60**, No. 2, 351-363.
29. Rice, J. A. and Silverman, B. W. (1991) Estimating the mean and covariance structure nonparametrically when the data are curves. *Journal of the Royal Statistical Society. Series B (Methodological)*, Vol. **53**, No. 1, 233-243.
30. Richards, Mark A. (2005) *Fundamentals of Radar Signal Processing*. McGraw-Hill electronic engineering.
31. Roe, H. and Hobson, G. S. (1992) Improved discrimination of microwave vehicle profiles. *IEEE*, Vol. **2**, 717-720.
32. Ruppert, D., Wand, M. P. and Carroll, R. J. (2003) *Semiparametric regression*. Cambridge; New York/ Cambridge University Press.

33. Shi, M., Weiss, R. E. and Taylor, J. M. G. (1996) An analysis of paediatric CD4 counts for acquired immune deficiency syndrome using flexible random curves. *Applied Statistics*, Vol. **45**, 151-163.
34. Songha, H., Hui, Z. and Bo, H. (1995) Polarization identification of high resolution range profiles. *IEEE*, Vol. **1**, 53-55.
35. Verbeke, G. and Molenberghs, G. (2000) Linear mixed models for longitudinal data. New York/ Springer.
36. Wahba, G. (1990) Spline models for observational data. CBMS-NSF *Regional Conference Series In Applied Mathematics*, Vol. **59**, Philadelphia/ SIAM.
37. Wang, Y. (1998) Mixed effects smoothing spline analysis of variance. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, Vol. **60**, No.1, 159-174.
38. Wang, Y. and Taylor, J. M. G. (1995) Inference for smooth curves in longitudinal data with application to an AIDS clinical trial. *Statistics in Medicine*, Vol. **14**, 1205-1218.
39. Weber, N., Moedl, S. and Hackner, M. (2002) A novel signal processing approach for microwave Doppler speed sensing. *IEEE*, Vol. **3**, 2233-2235.
40. Wecker, W. E. and Ansley, C. F. (1983) The signal extraction approach to nonlinear regression and spline smoothing. *Journal of the American Statistical Association*, Vol. **78**, 81-89.
41. Wishart, J. (1938) Growth rate determination in nutrition studies with the bacon pig, and their analysis. *Biometrika*, Vol. **30**, 16-28.
42. Wypij, D., Pugh, M. and Ware, J. H. (1993) Modeling pulmonary function growth with regression splines. *Statistica Sinica*, Vol. **3**, 329-350.