FACULTEIT ECONOMISCHE EN
TOEGEPASTE ECONOMISCHE
WETENSCHAPPEN

KATHOLIEKE
UNIVERSITEIT
LEUVEN

ROBUST DISCRIMINANT ANALYSIS

Proefschrift Voorgedragen tot
het Behalen van de Graad van
Doctor in de Toegepaste
Economische Wetenschappen

door

Kristel JOOSSENS

NUMMER 215 2006

Committee

Prof. Dr. Christophe Croux (Advisor)	*Katholieke Universiteit Leuven*
Prof. Dr. Geert Dhaene	*Katholieke Universiteit Leuven*
Prof. Dr. Peter Filzmoser	*Technische Universität Wien*
Prof. Dr. Gentiane Haesbroeck	*Université de Liège*
Prof. Dr. Ana M. Pires	*Universidade Técnica Lisboa*
Prof. Dr. Martina Vandebroek	*Katholieke Universiteit Leuven*

Daar de proefschriften in de reeks van de Faculteit Economische en Toegepaste Economische Wetenschappen het persoonlijk werk zijn van hun auteurs, zijn alleen deze laatsten daarvoor verantwoordelijk.

Acknowledgements

"De laatste loodjes wegen het zwaarst" is surely an appropriate Dutch expression, regarding the work of the last few months. My years of research were great as well as very tough. This is the moment to thank all people that made these years as they were.

First and for the most, I owe a lot to my advisor, Christophe Croux. He gave me the opportunity to start a Ph.D. here in Leuven and taught me research at the best. It has been a great pleasure and experience to work under his guidance and get his support, encouragement and a share of his wisdom. I also would like to thank him for allowing me to make plenty of international contacts, they provided me with lots of knowledge and inspiration. This gain and much more I will retain for the rest of my life. Many thanks also go to the other members of my doctoral committee: Ana Pires, Peter Filzmoser, Gentiane Haesbroeck, Martina Vandebroek and Geert Dhaene for kindly accepting to be a member of my jury, and for their many valuable comments and suggestions on the whole improving the quality and readability of my dissertation.

For Chapters 1 and 4, I gratitude Gentiane Haesbroeck for the help with the calculations and the programming. Besides all this she provided me with lots of interesting ideas, suggestions and comments. Chapter 3 grew out of my master thesis that I obtained almost 4 years ago in Brussels, under the guidance of Christophe. For Chapter 5 special thanks go to Peter Filzmoser whom I have worked with in Austria. The pleasant working environment, his nice family and snow made it unforgettable. For Chapter 6 I thank Marnik Dekimpe for his interesting comments. For the whole dissertation, I thank the K. U. Leuven for their trust and financial support of the OT-project "Robuuste discriminant analyse" (0T/02/10) through *bijzonder onderzoeksfonds*.

I would also like to thank all colleagues from ORSTAT, OM and accounting for the nice working environment. The "fifth floor" is the top, especially for the long late evenings we needed to introduce the new members in the group. Much respect to my colleague Aurélie Lemmens. It was nice working with her all these years. Of course, I cannot forget to thank my present officemate Sarah Gelper for putting up with me for almost two years by now. She became a good friend with whom I can share almost everything.

Finally, I have to say 'thank you' to all my friends and family. They have given me strength during these years of the doctoral project. The long evenings of going out and playing games were really good at providing distraction from the many days of hard work. Sincere thanks to my parents and my sisters for all they have ever done for me. My parents gave me the opportunity to study and supported me in many ways. Therefore this dissertation is also part of their work. I might not always shown it, but I really appreciate their interest, support and involvement in everything I do. They are the best!

At last but certainly not least I want to express many loving thanks to Koen for his support as best friend and future husband. The last months were not easy, neither for me nor for you, but you were always at my side. Thanks a lot for everything!

To end, "Thank you" everyone for every support you gave and ever will give!

Kristel Joossens Leuven, januari 2006.

Summary

Discriminant analysis, and associated classification rules, are often used in practice. Take for example the marketing division of a bank trying to sell investment funds to new costumers. Because they want to avoid overloading their new clients, they would like to give the advertisements only to those who might be interested in investments. Discriminant analysis can help here, but one has to take into account that the database of the manager can be huge and can contain a lot of atypical observations, also known as outliers.

In discriminant analysis one tries to construct a rule that allows to categorise multivariate observations into different groups or populations. This rule is constructed on the basis of a training sample, being a collection of observations for which the source population is known. As an example of a training sample, a collection of bank clients can be considered. One part of the clients having investment funds and the other part not. For all these clients certain relevant characteristics have been measured, such as saving money, family income, number of children, etc. Using this information, a discrimination rule can be constructed, allowing to assign new clients, whose characteristics are known, into one of the groups. Only the clients who are assigned to the group interested in investment funds, will receive publicity on these funds. Many other applications of discriminant analysis can be found in economics, biology, medicine, chemometrics, etc.

However, the classical discriminant rules can be strongly influenced by the presence of outliers in the training sample, through which the results can become unreliable. This creates a need for robust alternatives that behaves more stable in the presence of outliers in the data. Existing literature provides results for robust discriminant analysis, although these results were mainly restricted to the linear discriminant analysis and in the case of only two groups. In this dissertation non-linear discriminant analysis is investigated, using quadratic and logistic rules. Moreover, an extension to discriminant analysis for multiple groups is provided. In economic applications, it might for instance be useful to distinguish among groups of investors, depending on the characteristics of the persons of the different groups.

In this dissertation new discriminant procedures are developed, that behave robust in presence of outliers and give the smallest possible probability of mis-

classification. Statistical properties are derived for the various methods and are represented in the different chapters. The chapters of this thesis are published or submitted papers in the area of robust statistics. The first five chapters contemplate about robust discriminant analysis while the last chapter concerns robust time series analysis. An overview of the different chapters is provided below.

Chapter 1 offers an introduction to the theory of robust statistics, illustrated by various examples of business economics. Basic concepts, e.g. influence functions measuring the influence of observations on statistics, which are used throughout the whole thesis, are explained in an easy way.

Chapter 2 provides an empirical comparison of linear and quadratic discriminant analysis for two groups, of which the discriminant rules are respectively of linear and quadratic form. Both can be easily robustified. Classical and robust versions are compared in absence and presence of outliers. Using the probability of misclassification as criterion, it is shown that robust methods behave much more stable in the presence of outliers than classical methods. In absence of outliers there is no clear difference between the classical and robust methods.

In Chapter 3, quadratic discriminant analysis for two groups is studied. Influence functions and probabilities of misclassification are derived theoretically, allowing to study the influence of observations in the training sample on the probability of misclassification. A similar theory has already been derived for linear discriminant analysis. It turned out to be much more difficult and complex for the quadratic version. This chapter contains a lot of calculations where partial influence functions are used. This variant of the influence functions for multiple populations where population sizes remain unchanged and one population is contaminated. These partial derivatives can be used to define diagnostics in an easy way, allowing to create figures to identify influential observations.

While previous chapters investigate linear and quadratic discriminant analysis with, respectively, linear and quadratic rules; in Chapter 4 discriminant analysis with rules estimated by logistic regressions is considered. The method is called logistic discriminant analysis and is used in many applications. The advantage of logistic discriminant analysis is that, only a condition on conditional distributions is needed to get an optimal classification rule (so with minimal probability on misclassification) and not the assumption of multivariate normality of the whole distribution; as with linear discriminant. In this chapter, the theoretical analysis for robust logistic discriminant analysis is studied at the normal discrimination model, making the formal derivations feasible. Using an optimal rule makes the first order influence function of the probability of misclassification vanish. Hence it is appropriate to switch over to second order influence functions.

Chapter 5 extends robust discriminant analysis to the multiple group case (i.e. more than two). Starting from the well-known Fisher discriminant rule, the probability of misclassification for several groups is calculated. Working at a special setting with normally distributed populations with means on one line and equal covariance matrices, we can develop the theorie and the Fisher rule is optimal. It

is of course possible to apply this rule when these assumptions are not fulfilled. As in Chapter 4, it is appropriate to consider the second order influence functions. We demonstrate the possibility to compute classification efficiencies by means of these second order influence functions. In particular, we compute the classification efficiency of the robust estimator with respect to the classical one, when no outliers are present. Expected values of the probabilities of misclassification are computed, as well theoretical as by means of Monte carlo simulations. Note that the distribution of the probability of misclassification is not normal, but a mixture of chi-squared distributions.

Chapter 6 involves multivariate time series analysis and is based on robust multivariate regression. In this chapter, classical time series analysis is shown to be very influenceable by atypical observations, pointing out the need for robustness. In this chapter an estimation methode is proposed on the basis of a trimmed least squares estimator. This estimator can be computed fast and can be used as starting values for other procedures. It is also illustrated how the order of the vector autoregressive model can be determined and how the confidence bounds around the robustly estimated impulse response functions can be constructed.

Of course, research is a never-ending process and this doctoral thesis can be seen as a starting point for further research. For the time being many mathematical questions remain unanswered. For example, the most obtained results used the assumption of normality of the underlying groups, but what happens if the distributions deviate from the normal one? Does this result in a loss of robustness? For linear discriminant analysis was assumed that the covariance matrices of the different populations are equal and the means are collinear, but how doe the result change if the underlying distributions deviate from these assumptions?

The research for this dissertation gave us new ideas for development of new methods for robust discriminant analysis in other situations. Not much research has been done yet on discriminant analysis for multiple groups. As for linear discriminant analysis, one can think of quadratic discriminant analysis for multiple groups by using more than one discrimination rules. As an extension of (robust) logistic discriminant analysis one might think of (robust) multinomial and ordinal logit (and probit) for unordered and ordered categories, respectively.

In this dissertation classification efficiencies for robust methods have been computed for the first time, where the techniques we developed, by means of the second order influence functions, provides many other applicabilities. In particular we think of computation of classification efficiencies for non-parametric discriminant rules, e.g. the nearest neighbour method.

Samenvatting

Discriminantanalyse, en de bijhorende classificatieregels, wordt vaak gebruikt in de praktijk. Denk bijvoorbeeld aan de marketing afdeling van een bank die tracht beleggingsfondsen te verkopen aan nieuwe klanten. Vermits men uiteraard niet onnodig nieuwe klanten wenst lastig te vallen, wil men ervoor zorgen dat de reclame alleen aan mogelijk geïnteresseerden gegeven wordt. Discriminantanalyse kan hier helpen, maar er moet rekening mee gehouden worden dat de gegevensbank waarover de bank beschikt erg groot kan zijn, en dat deze vele atypische observaties kan bevatten, ook wel uitschieters genaamd.

In discriminantanalyse tracht men een regel op te stellen die toelaat om multivariate observaties aan verschillende groepen toe te wijzen. Deze regel wordt geconstrueerd op basis van een oefensteekproef, wat een verzameling observaties is waarvan men reeds weet tot welke groep ze behoren. Als voorbeeld kan als oefensteekproef een verzameling cliënten beschouwd worden. Een deel hiervan zijn mensen die reeds beleggingsfondsen hebben en een ander deel niet. Voor al deze cliënten worden enkele relevante karakteristieken gemeten, zoals het spaargeld, het gezinsinkomen, informatie over lopende leningen, het aantal kinderen, enz. Gebruikmakende van deze informatie kan men dan een discriminant regel opstellen, die toegepast kan worden op nieuwe cliënten waarvan men enkel de karakteristieken kent, doch die niet in de oefensteekproef zaten. Deze cliënten kunnen dan toegekend worden aan één van de groepen. Enkel de nieuwe cliënten die toegekend worden aan de groep van mensen geïnteresseerd in beleggingsfondsen, zullen de reclame ontvangen. Vele andere toepassingen van discriminantanalyse kunnen uiteraard gevonden worden in economie, biologie, geneeskunde, enz.

De klassieke discriminant regels kunnen echter erg sterk beïnvloed worden door aanwezigheid van enkele uitschieters in de oefensteekproef, waardoor de resultaten onbetrouwbaar kunnen worden. Daarom is er nood aan robuuste alternatieven die zich stabieler gedragen in aanwezigheid van uitschieters in de data. In de literatuur werden reeds resultaten voor robuuste discriminantanalyse gegeven, doch dit was meestal beperkt tot lineaire discriminantanalyse en in het geval van slechts twee groepen. In dit proefschrift worden ook robuuste niet-lineaire discriminant regels bestudeerd, zoals kwadratische en logistische regels. Tevens wordt in dit proefschrift een uitbreiding naar discriminantanalyse voor meerdere

groepen voorzien. Het kan bijvoorbeeld zeer interessant zijn om groepen van beleggers te onderscheiden, afhankelijk van de karakteristieken van de personen in die verschillende groepen.

In dit proefschrift worden nieuwe discriminant procedures ontwikkeld, die zich robuust gedragen in aanwezigheid van uitschieters en een zo klein mogelijke kans op foutieve classificatie geven. Statistische eigenschappen worden afgeleid voor de verscheidene methodes en voorgesteld in de verschillende hoofdstukken. De hoofdstukken van dit proefschrift zijn artikels die reeds gepubliceerd werden of ingestuurd werden voor publicatie. Ze situeren zich in het domein van robuuste statistiek. De eerste vijf hoofdstukken handelen over robuuste discriminantanalyse, terwijl het laatste hoofdstuk handelt over robuuste tijdreeksenanalyse. Een kort overzicht van de verschillende hoofdstukken wordt hieronder gegeven.

Hoofdstuk 1 voorziet een inleiding over de theorie van robuuste statistiek, geïllustreerd aan de hand van verscheidene voorbeelden uit de bedrijfseconomie. Basisbegrippen, zoals invloedsfuncties, die doorheen mijn hele proefschrift gebruikt worden, worden hier op een eenvoudige wijze uitgelegd.

Hoofdstuk 2 voorziet een empirische vergelijkende studie van lineaire en kwadratische discriminantanalyse voor twee groepen, waarbij de discriminant regels respectievelijk van lineaire en kwadratische vorm zijn. Beiden kunnen op eenvoudige wijze gerobustifieerd worden. Klassieke en robuuste versies worden in af- en aanwezigheid van uitschieters met elkaar vergeleken. Gebruikmakend van kansen tot foutieve classificatie als criterium, wordt aangetoond dat robuuste methodes zich veel stabieler gedragen in aanwezigheid van uitschieters dan klassieke methodes. In afwezigheid van uitschieters is er echter geen beduidend verschil tussen de klassieke en robuuste methodes.

In Hoofdstuk 3 wordt kwadratische discriminantanalyse voor twee groepen grondiger bestudeerd. Invloedsfuncties en kansen tot foutieve classificatie worden theoretisch afgeleid, wat toelaat om te bestuderen hoe observaties van de oefensteekproef de analyses beïnvloeden. Gelijkaardige theorie werd reeds afgeleid voor lineaire discriminantanalyse. Voor de kwadratische versie bleek dit veel moeilijker en complexer te zijn. Dit hoofdstuk omvat dan ook veel berekeningen waarbij partiële invloedsfuncties gebruikt worden. Dit is een variant van de gewone invloedsfuncties voor meerdere populaties, waar de groottes van populaties onveranderd blijven en één populatie gecontamineerd wordt. Deze partiële invloedsfuncties kunnen gebruikt worden om op eenvoudige wijze diagnostieken te definiëren die toelaten om figuren te maken waarop invloedrijke observaties geïdentificeerd kunnen worden.

Voorgaande hoofdstukken beschouwen lineaire en kwadratische discriminantanalyse, waarbij de regels van lineaire en kwadratische vorm zijn. In Hoofdstuk 4 wordt discriminantanalyse beschouwd, waarbij de regels geschat zijn door logistieke regressie. Deze methode wordt logistieke discriminantanalyse genoemd en wordt in vele toepassingen gebruikt omdat ze met verschillende types van vari-

abelen overweg kan. Het voordeel van logistieke discriminant analyse is, dat enkel een voorwaarde op de conditionele verdeling nodig is om tot een optimale classificatie regel (dus met minimale kans op foutieve classificatie) te komen en niet de aanname van multivariate normaliteit van de volledige verdeling; zoals bij met lineaire discriminant analyse. In dit hoofdstuk wordt de theoretische analyse voor robuuste logistieke discriminantanalyse beschouwd voor het normale discriminant model, om formele afleidingen doenbaar te maken. Door gebruik te maken van een optimale regel is de eerste orde invloedsfunctie van de kans tot foutieve classificatie nul, waardoor het gepast is over te gaan naar tweede orde invloedsfuncties.

Hoofdstuk 5 breidt robuuste discriminantanalyse uit naar het geval met meerdere groepen (d.w.z. meer dan twee). Startende van de alom gekende Fishers discriminant regel wordt de kans tot foutieve classificatie berekend voor het geval met meerdere groepen. Werkende met een speciaal geval van normaal verdeelde populaties met gemiddelden die op een lijn liggen en met gelijke covariantie matrices, kunnen we de theorie ontwikkelen en is de Fisher regel optimaal. Het is uiteraard ook mogelijk om deze regel toe te passen wanneer aan deze voorwaarden niet voldaan is. Net zoals in Hoofdstuk 4, is het gepast om over te gaan tot tweede orde invloedsfuncties. We tonen aan dat het mogelijk is om classificatie efficiënties te berekenen met behulp van deze tweede orde invloedsfuncties. In het bijzonder berekenen we de classificatie efficiëntie van de robuuste methode ten opzichte van een klassieke. Er wordt aangetoond dat het verlies aan classificatie efficiëntie beperkt is, wat wil zeggen dat de robuuste methode slechts zeer beperkt verliest in vergelijking met de klassieke methodes, als er geen uitschieters aanwezig zijn. Verwachte waarden van de kans tot foutieve classificatie worden berekend, zowel theoretisch als met behulp van Monte carlo simulaties. Merk op dat de verdeling van de kans tot foutieve classificatie niet normaal is, maar een mengsel van chi-kwadraat verdelingen.

Hoofdstuk 6 handelt over meervoudige tijdreeksanalyse en is gebaseerd op robuuste multivariate regressie. Hierin wordt aangetoond dat de klassieke tijdreeksanalyse zeer beïnvloedbaar is door atypische observaties, wat duidt op nood aan robuustheid. In dit hoofdstuk wordt een schattingsmethode voorgesteld op basis van een getrimde kleinste kwadraten schatter. Deze schatter is snel uit te rekenen, en kan ook als startwaarde dienen voor andere procedures. Tevens wordt in dit hoofdstuk geïllustreerd hoe de orde van een autoregressief model robuust kan bepaald worden en hoe betrouwbaarheidsbanden rond de robuust geschatte impuls-response functies geconstrueerd kunnen worden.

Uiteraard is onderzoek een nooit eindigend proces en kan dit proefschrift gezien worden als een startpunt voor verder onderzoek. Vooreerst blijven er nog vele wiskundige vragen open waarop een antwoord dient gevonden te worden. Zo gebruiken de meeste bekomen resultaten de aanname van normaliteit van de onderliggende groepen, maar wat gebeurt er als hiervan wordt afgeweken? Resulteert dit in verlies aan robuustheid? Voor de theoretische resultaten bekomen

voor lineaire discriminant analyse werd aangenomen dat de covariantie matrices van de verschillende populaties gelijk zijn en dat hun gemiddeldes collineair zijn, maar hoe veranderen de bekomen resultaten indien de onderliggende verdelingen van deze aannames afwijken?

Het onderzoek verricht voor dit proefschrift gaf ons ook ideëen voor het ontwikkelen van nieuwe methodes voor robuuste discriminantanalyse in andere situaties. Zo is er voor discriminantanalyse met meer dan twee groepen nog niet zo veel onderzoek verricht. Zoals voor lineaire discriminant analyse, kan gedacht worden aan kwadratische discriminant analyse voor het geval met meerdere groepen, door gebruik te maken van meerdere discriminatie regels. Als een uitbreiding van (robuuste) logistieke discriminantanalyse kan gedacht worden aan (robuuste) multinomiale en ordinale logit (en probit) voor ongeordende en geordende categorieën, respectievelijk.

In dit proefschrift worden voor de eerste keer classificatie efficiënties voor robuuste methodes berekend, waarbij de door ons ontwikkelde techniek, gebruik makend van tweede orde invloedsfuncties, vele andere toepassingmogelijkheden biedt. In het bijzonder denken we aan het berekenen van classificatie efficiënties voor niet-parametrische discriminant regels, zoals bijvoorbeeld de “nearest neighbour” methode.

Table of contents

Chapter 1

Inleiding tot robuuste statistiek:
Elementen van theorie en bedrijfseconomische toepassingen

Co-Auteur: C. Croux

Samenvatting Vele vaak gebruikte statistische methoden geven onbetrouwbare resultaten in de aanwezigheid van uitschieters. Robuuste statistische methoden blijven goed werken wanneer er atypische observaties aanwezig zijn of wanneer er niet perfect aan andere modelvoorwaarden voldaan is. Ofschoon de theorie van de robuuste statistiek zich reeds sinds enkele decennia ontwikkeld heeft, is het pas recentelijk dat robuuste schatters ook snel uitgerekend kunnen worden en in algemene statistische software pakketten opgenomen zijn. Toegepaste economen kunnen nu dan ook zonder problemen gebruik maken van robuuste schatters wanneer ze vrezen dat er uitschieters aanwezig zijn in hun gegevensbestanden. In dit hoofdstuk geven we een korte inleiding tot de theorie van de robuuste statistiek, aangevuld met verschillende voorbeelden uit de bedrijfseconomie.

1.1 Inleiding

De wetenschap van de statistiek tracht bruikbare informatie te distilleren uit empirisch beschikbare gegevens. Om dit te realiseren, wendt men zich gedurende al meer dan twee eeuwen tot statistische modellen. Deze methode kende zijn apotheose in de eerste helft van de 20ste eeuw, vooral onder impuls van R.A. Fisher die een groot aantal statistische procedures introduceerde. Zijn werk vormt de

basis van de inferentiële statistiek die men dagdagelijks gebruikt en die gebaseerd is op een parametrische specificatie van het statistische model.

Deze klassieke benadering van de statistiek veronderstelt dat de statistische modellen goed gespecificeerd zijn. Sinds geruime tijd beseft men echter dat de reële wereld zich niet gedraagt zoals in de meeste vooropgestelde modellen. De performantie en validiteit van de toepassingen van parametrische procedures vereisen echter dat er strikt aan de hypotheses van het model voldaan is. Daarom werd de niet-parametrische statistiek geïntroduceerd en sommige van deze methodes zijn heel populair geworden in de toegepaste statistiek. Niettegenstaande het feit dat sommige problemen zeer bevredigend opgelost kunnen worden met een niet-parametrische methode, heeft de parametrische aanpak nog steeds een dominante rol omdat zij vaak meer precies is en de geschatte parameters vaak een (economische) interpretatie hebben. Bovendien zijn parametrische procedures in een veel groter gamma van situaties toepasbaar.

De robuuste statistiek combineert de kracht van beide benaderingen. Zij doet niet alleen dienst in parametrische modellen, maar zij gebruikt ook procedures die minder essentieel steunen op de hypotheses waaraan het gekozen model moet voldoen. Bovendien laten robuuste methodes toe om afwijkende observaties te identificeren. De robuuste statistiek gaat ervan uit dat de meest voorkomende hypotheses in de statistiek (zoals normaliteit, lineariteit, ...) enkel bij benadering juist zijn. Haar doel is dus het creëren van procedures die weerstand bieden aan zulke modelafwijkingen.

Laten we een voorbeeld geven van een dataset waarin uitschieters (outliers) voorkomen. Voor de periode 1950-1973 werd jaarlijks de duurtijd in minuten van internationale telefoonoproepen in België gemeten (zie Rousseeuw en Leroy 1987). Deze gegevens worden voorgesteld in Figuur 1.1. Deze tijdreeks bevat een aantal zeer atypische observaties van 1964 tot 1969, wat te wijten is aan het feite een ander registratiesysteem gebruikt werd in die periode. In die periode werd immers het aantal telefoongesprekken gemeten en niet de totale duurtijd. In dit voorbeeld komen er dus "grove fouten" voor, die uitschieters genereren. Uitschieters zijn observaties die zich anders gedragen dan de grote meerderheid van de andere gegevens en waarvan het zeer onwaarschijnlijk is dat ze door hetzelfde proces gegenereerd zijn als de grote meerderheid van de andere observaties. In een robuuste procedure gaat men deze uitschieters dan ook een kleiner gewicht geven, of soms zelfs helemaal weglaten, zodat zij weinig of zelfs geen invloed hebben op de analyse.

In dit voorbeeld kan men de uitschieters gemakkelijk grafisch detecteren, ze bevinden zich immers ver van de meerderheid van de gegevens in de grafiek van Figuur 1.1. Het detecteren van uitschieters is niet altijd zo eenvoudig. Indien we werken met multivariate gegevens, bijvoorbeeld observaties in 5 dimensies, dan wordt het onmogelijk de data grafisch voor te stellen en kunnen uitschieters niet meer visueel gedetecteerd worden. Daarom is nuttig om robuuste procedures te gebruiken en om detectieprocedures voor uitschieters te ontwikkelen. Merk op dat

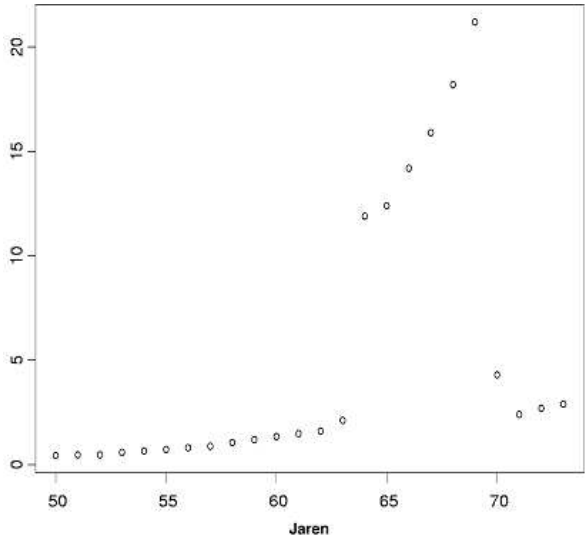

Figure 1.1: *Duurtijd telefoonoproepen van 1950 tot 1973 in België.*

de uitschieters soms juist de interessantste observaties zijn, omdat ze met speciale gebeurtenissen overeenkomen.

Het probleem van robuustheid is reeds lang gekend en statistici zijn zich sterk bewust van de gevaren van uitschieters. Op de middelbare school leren scholieren reeds dat de mediaan meer bestand is tegen uitschieters dan het gemiddelde. Toch heeft het relatief lang geduurd vooraleer men een meer formele benadering van het probleem vond. Pionierswerk van Huber (1964) en Hampel (1971) introduceerde maten om de robuustheid van een schatter te meten. Sindsdien zijn er tal van theoretische ontwikkelingen gebeurd en werden vele nieuwe technieken geïntroduceerd die resistent zijn tegen uitschieters.

Dit hoofdstuk geeft een eerste kennismaking met robuuste statistiek. In Sectie 1.2 komen de basisbegrippen invloedsfunctie en breekpunt van een schatter aan bod. Dit zijn maten van robuustheid die toelaten om de robuustheid van een schatter te evalueren. In deze sectie beperken we ons tot ééndimensionale gegevens, om zodoende de uiteenzetting eenvoudig te kunnen houden. In Sectie 1.3 wordt het lineaire regressiemodel behandeld. We zullen aantonen dat de klassieke kleinste kwadraten schatter niet robuust is, en een alternatieve schattingsmethode bespreken. In Sectie 1.4 illustreren we de voordelen van een robuuste aanpak met enkele bedrijfseconomische toepassingen. In een laatste sectie maken we de nodige verdere verwijzingen naar de wetenschappelijke literatuur in dit onderzoeksdomein.

1.2 Basisbegrippen van robuuste statistiek

Beschouwen we een steekproef van univariate gegevens, die we noteren als $X = \{x_1, \ldots, x_n\}$. Onderstel dat de populatie waaruit deze gegevens getrokken worden normaal verdeeld is met gemiddelde μ en een variantie σ^2. We willen nu de parameter μ, die hier de centrale waarde van de verdeling aangeeft, schatten. De klassieke schatter voor μ is het rekenkundig gemiddelde, gedefinieerd als

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i.$$

Deze schatter wordt echter sterk beïnvloed door één of meerdere uitschieters in onze steekproef. Beschouw volgende reeks van bruto maandinkomens van 12 werknemers van een zeker bedrijf (in euro).

$$X = \{1513, 1834, 2112, 2160, 2288, 2375, 2424, 2647, 3156, 3908, 4233, 9961\}.$$

Het rekenkundig gemiddelde van deze steekproef van inkomens bedraagt $\bar{x} = 3217.58$ euro per maand, wat duidelijk afwijkt van het centrum van de data, zoals gemeten door de mediaan. Dit wordt geïllustreerd in Figuur 1.2. Het gemiddelde werd hier sterk aangetast door de atypische extreme waarde, 9961, en is hier dus geen goede schatter voor de parameter μ van de vooropgestelde normale verdeling. Een meer robuuste schatter is nodig .

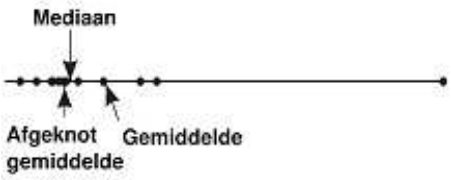

Figure 1.2: *Inkomensreeks van 12 werknemers van een firma.*

Indien we in een gegeven steekproef uitschieters verwachten, kunnen we verschillende strategieën toepassen. Het rekenkundig gemiddelde is immers niet de enig mogelijke schatter voor μ, vele alternatieven bestaan. Een van de meest gebruikte is zonder twijfel de mediaan, die gelijk is aan de "middelste" observatie in de steekproef. Een formele definitie wordt gegeven door

$$\underset{i}{\text{med}}\ x_i = \frac{1}{2}\left(x_{(\lfloor n/2 \rfloor)} + x_{(\lfloor n/2 \rfloor + 1)}\right),$$

waar $x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$ de geordende observaties zijn en $\lfloor z \rfloor$ correspondeert met het grootste geheel getal kleiner of gelijk aan z. Een andere mogelijke schatter

is het afgeknot gemiddelde: voor een reële waarde α tussen 0 en 0.5, wordt het afgeknot gemiddelde met drempel α gedefinieerd als het rekenkundig gemiddelde berekend op basis van een "afgeknotte" steekproef. De afgeknotte steekproef is de steekproef waaruit we de $\lfloor \alpha n \rfloor$ kleinste en de $\lfloor \alpha n \rfloor$ grootste observaties laten wegvallen. We noteren dit afgeknot gemiddelde met $\bar{x}_\alpha$ en een wiskundige definitie is

$$\bar{x}_\alpha = \frac{1}{n - 2\lfloor \alpha n \rfloor} \sum_{i=\lfloor \alpha n \rfloor + 1}^{n - \lfloor \alpha n \rfloor} x_{(i)}.$$

De waarde α dient door de statisticus zelf gekozen te worden. Als verwacht wordt dat er in de dataset veel atypische observaties kunnen voorkomen, is het aangewezen om α groot te nemen. Hoe groter de waarde van α, hoe meer efficiëntie men echter verliest. Een goede keuze voor α wordt gegeven door $\alpha = 0.25$ (cfr. Croux en Haesbroeck 2002). Een afgeknot gemiddelde met drempelwaarde 25% resulteert in een schatter met een grote robuustheid, en tegelijk een precisie die bijna zo groot is als die van het steekproefgemiddelde (in afwezigheid van uitschieters).

Merk op dat een afgeknot gemiddelde met drempel $\alpha \approx 0.5$ overeen komt met de mediaan. In het voorbeeld van Figuur 1.2 is de mediaan van de inkomens 2399.5 terwijl het afgeknot gemiddelde met drempel 25% gegeven wordt door $\bar{x}_{0.25} = 2508.33$.

Omdat er vele alternatieven zijn voor het rekenkundig gemiddelde, is het belangrijk dat we hun performanties kunnen vergelijken aan de hand van verschillende criteria. Vaak wordt als criterium de efficiëntie van de schatter genomen. Hoe efficiënter een schatter, hoe preciezer hij de onbekende μ zal schatten. Men kan aantonen dat het rekenkundig gemiddelde, onder de assumptie van normaliteit, de meest efficiënte schatter is. Het is echter zo dat het gemiddelde deze eigenschap snel verliest en helemaal niet meer zo precies is wanneer er afwijkingen van het model zijn. Daarom is het ook nodig om andere maten van performantie van een schatter te bekijken. In de volgende twee paragrafen worden twee manieren voorgesteld om de robuustheid van een schatter te meten.

1.2.1 De empirische invloedsfunctie

De empirische invloedsfunctie (EIF) laat toe het effect van een afwijkende observatie op de schatter te visualiseren. Gegeven is een steekproef $x_1, \ldots, x_n$ en een schatter T. Voor elke mogelijke waarde van x berekenen we dan

$$\mathrm{EIF}(x; T) = n\{T(x_1, \ldots, x_n, x) - T(x_1, x_2, \ldots, x_n)\}.$$

(In bovenstaande formule is de vermenigvuldiging met de steekproefgrootte n enkel een herschaling). De empirische invloedsfunctie laat toe het effect op de

schatter T te meten, wanneer een observatie x aan de steekproef wordt toegevoegd. Wanneer de schatter robuust is, zou dit effect beperkt moeten blijven. We willen immers niet dat individuele observaties, die mogelijke uitschieters kunnen zijn, teveel invloed op onze schatter uitoefenen.

Voor het voorbeeld met de inkomens, waar we als steekproef de eerste 11 observaties zonder de uitschieter nemen, hebben we empirische invloedsfuncties uitgerekend voor het rekenkundig gemiddelde $\bar{x}$, de mediaan, en het 25% afgeknot gemiddelde $\bar{x}_{0.25}$ (Figuur 1.2). We stellen onmiddellijk vast dat de EIF van het rekenkundig gemiddelde onbegrensd is. Grote waarden hebben een onbegrensde invloed op $\bar{x}$, wat de niet-robuustheid van het gemiddelde aantoont. Noem

$$\gamma(T) = \sup_x |\text{EIF}(x; T)| \tag{1.1}$$

de maximale waarde die de EIF kan aannemen. Dit getal wordt ook de *gross-error sensitiviteit* genoemd, en is een maat voor de robuustheid van een schatter. Hoe kleiner deze waarde is, hoe beter. Uit Figuur 1.3 blijkt dat de empirische invloedsfunctie van de mediaan en het afgeknot gemiddelde begrensd zijn. De waarden ervan zijn respectievelijk $\gamma(\text{med}) = 478.5$ en $\gamma(\bar{x}_{0.25}) = 1291.19$, terwijl $\gamma(\bar{x}) = \infty$.

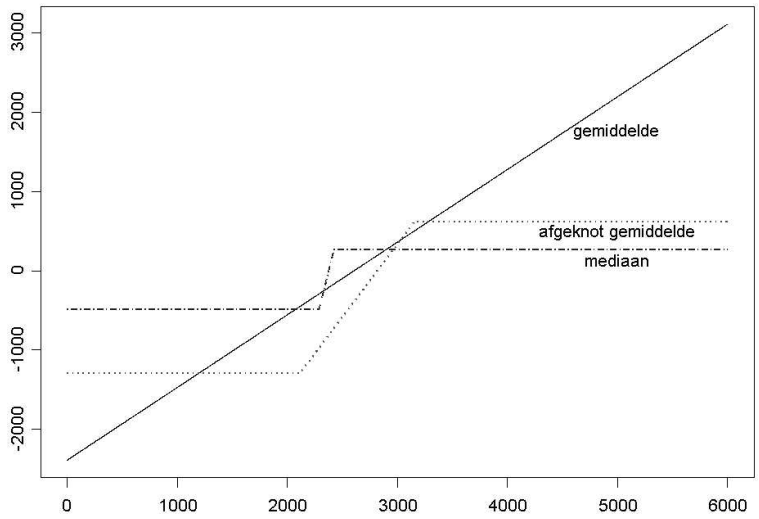

Figure 1.3: *Empirische invloedsfuncties voor het rekenkundig gemiddelde (volle lijn), de mediaan (gestreepte lijn) en het afgeknot gemiddelde met drempel* 25% *(stippellijn).*

Indien men enkel de gross-error sensitiviteit als een maat voor robuustheid neemt, is de mediaan te verkiezen boven het afgeknot gemiddelde met drempel

25%. Het rekenkundig gemiddelde heeft geen begrensde invloedsfunctie, wat zijn niet-robuustheid aantoont.

De empirische invloedsfunctie en bijhorende gross-error sensitiviteit zijn gemakkelijk te berekenen, maar een nadeel van formule (1.1) is dat deze waarde nog afhangt van de gegevens $x_1, \ldots, x_n$. Met behulp van statistische functionalen en verdelingen, kan men theoretisch meer werkbare definities van invloedsfuncties en gross-error sensitiviteit introduceren (Hampel et al 1986). Het achterliggend idee is echter hetzelfde als hierboven beschreven. De empirische invloedsfunctie meet de gevoeligheid van een schatter voor individuele observaties. Het blijkt nu dat in datasets vaak meerdere uitschieters tegelijk voorkomen; men spreekt dan van clusters van atypische observaties. Een meer geschikte maat van robuustheid in dit kader is dan het breekpunt, dat in de volgende paragraaf gedefinieerd wordt.

1.2.2 Breekpunt

Het breekpunt van een schatter T is de kleinste fractie observaties die we moeten wijzigen opdat de schatter willekeurig grote waarden kan aannemen. Om het breekpunt van een schatter T te vinden voor een steekproef $X = \{x_1, \ldots, x_n\}$, gaat men als volgt te werk. Vertrekkende van de initiële steekproef X creëren we een gecontamineerde steekproef X' door m observaties van X te veranderen in willekeurig (grote) waarden. Dit creëert dan een vertekening of bias die gegeven wordt door $|T(X) - T(X')|$.

Bedoeling is nu om de m observaties dusdanig te veranderen zodat deze *bias* zo groot mogelijk wordt. Deze maximale bias van de schatter T die men kan verkrijgen door m observaties te wijzigen is dan

$$\text{maxbias}(m, T, X) = \sup_{X'} - + |T(X) - T(X')|. \tag{1.2}$$

Als deze maxbias oneindig groot is, zegt men dat de schatter "breekt", hij neemt een volstrekt onbetrouwbare waarde aan. Het breekpunt $\varepsilon^*(T)$ van de schatter T is nu de kleinste fractie m/n van observaties die men moet veranderen alvorens de bias oneindig groot wordt, en een wiskundige definitie is

$$\varepsilon^*(T) = \frac{1}{n} \min\{m : \text{maxbias}(m, T, X) = \infty\}.$$

Het is nu niet moeilijk om de breekpunten van de beschouwde schatters te berekenen. Kijken we naar Figuur 1.1 en beelden we ons in dat we één enkele observatie naar oneindig verplaatsen. Dan zal het rekenkundig gemiddelde ook mee oneindig groot worden, en we krijgen dus $\varepsilon^*(\bar{x}) = 1/12$. Het verplaatsen van deze observatie naar oneindig gaat echter de mediaan en het afgeknot gemiddelde niet laten breken. Voor de mediaan moeten we maar liefst 6 observaties naar oneindig laten gaan, terwijl het voor een 25% afgeknot gemiddelde slechts 4 observaties gecontamineerd moeten worden om deze schatter te laten breken en dus een oneindig grote waarde te laten aannemen.

In ons voorbeeld hadden we een kleine steekproef. Het is echter niet moeilijk om in te zien dat voor zeer grote steekproeven geldt $\varepsilon^*(\bar{x}) \approx 0$, $\varepsilon^*(\text{med}) \approx 0.5$ en $\varepsilon^*(\bar{x}_{0.25}) \approx 0.25$. Indien we het breekpunt van een schatter als maat voor robuustheid nemen, is de mediaan weer te verkiezen boven het 25% afgeknot gemiddelde. Het gewone gemiddelde heeft een breekpunt gelijk aan nul, wat nogmaals de niet-robuustheid van deze schatter aantoont.

Merken we tot slot nog op dat definitie (1.2) afhangt van de gekozen steekproef. Om een theoretisch meer werkbare definitie van de maximale bias te krijgen, zal het weer nodig zijn om met verdelingstheorie en statistische functionalen te werken. We gaan hier niet verder op in. In deze sectie werden verschillende maten voor robuustheid besproken. Bij de keuze van een geschikte schatter zal men echter niet enkel zijn robuustheid beschouwen maar ook zijn precisie en berekenbaarheid.

1.3 Robuuste lineaire regressie

Het afgeknot gemiddelde en dus ook de mediaan, die we in de vorige sectie bespraken, zijn goed gekende robuuste schatters om de centrale positie van een (symmetrische) univariate verdeling te schatten. Het is echter minder duidelijk hoe men een robuuste schatter kan bekomen voor meer complexe modellen, zoals het lineaire regressiemodel. Voor een steekproefgrootte n meten we hier waarden y_i van een te verklaren variabele, en waarden $x_{i1}, \ldots, x_{ip}$ van de verklarende variabelen voor elke observatie $i = 1, \ldots, n$. Er wordt verondersteld dat de relatie tussen de te verklaren variabelen en de verklarende variabelen lineair is, dus

$$y_i = \alpha + x_{i1}\beta_1 + \ldots + x_{ip}\beta_p + e_i \qquad \text{voor } i = 1, \ldots, n. \tag{1.3}$$

De storingtermen $e_1, \ldots, e_n$ worden verondersteld om onafhankelijk en identiek verdeeld te zijn. Vaak wordt daarboven de hypothese van normaliteit voor deze storingstermen opgelegd. De onbekende parameters in het regressiemodel zijn de constante term α, en de richtingscoëfficiënten $\beta_1, \ldots, \beta_p$. We noteren nu de vector van ongekende parameters als

$$\theta = (\alpha, \beta_1, \beta_2, \ldots, \beta_p)$$

en de bedoeling is om deze ongekende parametervector te schatten met behulp van de beschikbare data. Het residu van de i-de observatie wordt gegeven door

$$r_i(\theta) = y_i - (\alpha + x_{i1}\beta_1 + \ldots + x_{ip}\beta_p).$$

Bedoeling is nu om een schatter $\hat{\theta}$ zo te kiezen dat deze residuen zo klein mogelijk zijn. Met zo "klein" mogelijk, wordt bedoeld dat voor een gekozen doelfunctie f, de waarde van de doelfunctie uitgerekend in de residuen minimaal wordt, m.a.w.

$$\hat{\theta} = \underset{\theta}{\text{argmin}} f(r_1(\theta), \ldots, r_n(\theta)). \tag{1.4}$$

Als doelfunctie wordt meestal de som van de gekwadrateerde residuen genomen, wat resulteert in

$$\hat{\theta}_{\text{LS}} = \underset{\theta}{\operatorname{argmin}} \sum_{i=1}^{n} r_i^2(\theta). \tag{1.5}$$

Dit geeft dan de bekende methode van de kleinste kwadraten, of *Least Squares* (LS). Het is echter belangrijk te weten dat dit niet de enige mogelijke schatter voor het lineaire regressiemodel is. Andere doelfuncties f zullen resulteren in andere schatters.

De reden van de populariteit van de kleinste kwadraten schatter is historisch te verklaren: toen men rond 1800 lineaire modellen begon te beschouwen, was de kleinste kwadraten schatter de enige die men vrij eenvoudig kon uitrekenen. Gauss schreef: "Van alle principes is de kleinste kwadraten het eenvoudigste: voor de anderen moeten we complexe berekeningen maken." Daarna introduceerde Gauss de normale verdeling als zijnde de verdeling waarvoor de kleinste kwadraten schatter optimaal is, in de zin van maximale efficiëntie. Sindsdien is de combinatie van de hypothese van normaliteit en het gebruik van de kleinste kwadraten schatter standaard. Door de beschikbaarheid van computers is het nu echter mogelijk geworden om (1.4) ook te berekenen voor andere doelfuncties f. Bovendien hebben statistici zich gerealiseerd dat gegevens vaak niet aan de klassieke normaliteitshypothese voldoen, en dat optimaliteit dus niet gegarandeerd is. In het bijzonder is het geweten dat de kleinste kwadraten methode zeer kwetsbaar is voor uitschieters. In de volgende paragraaf besteden we aandacht aan de verschillende soorten uitschieters die in een regressie analyse kunnen optreden.

1.3.1 Verticale uitschieters en hefboompunten

In de context van regressie komen twee soorten uitschieters voor, namelijk verticale uitschieters en hefboompunten. Als illustratie, beschouw een eenvoudig regressiemodel $y_i = \alpha + \beta x_i + e_i$ met slechts één verklarende variabele. Een fictieve dataset, die in Figuur 1.4(a) wordt voorgesteld, werd gegenereerd volgens dit model. De gegevens kunnen hier in het vlak worden voorgesteld en we zien dat er geen uitschieters aanwezig zijn. Na schatting met de kleinste kwadraten methode bekomen we de regressie rechte $y = \hat{\alpha} + \hat{\beta}x$, en in Figuur 1.4(a) zien we dat deze een goede fit geeft voor de puntenwolk.

Indien we nu één van de observaties in verticale richting verschuiven, dan krijgen we een *verticale uitschieter*, zoals in Figuur 1.4(b). We merken onmiddellijk op dat de geschatte regressie rechte nu een veel minder goede fit geeft. Het toont reeds aan dat kleinste kwadraten schatter door slechts één enkele uitschieter sterk kan veranderen.

Wanneer de uitschieter zodanig is dat de waarde van x_i atypisch is in de ruimte van de verklarende variabelen, dan spreekt men van een *hefboompunt*.

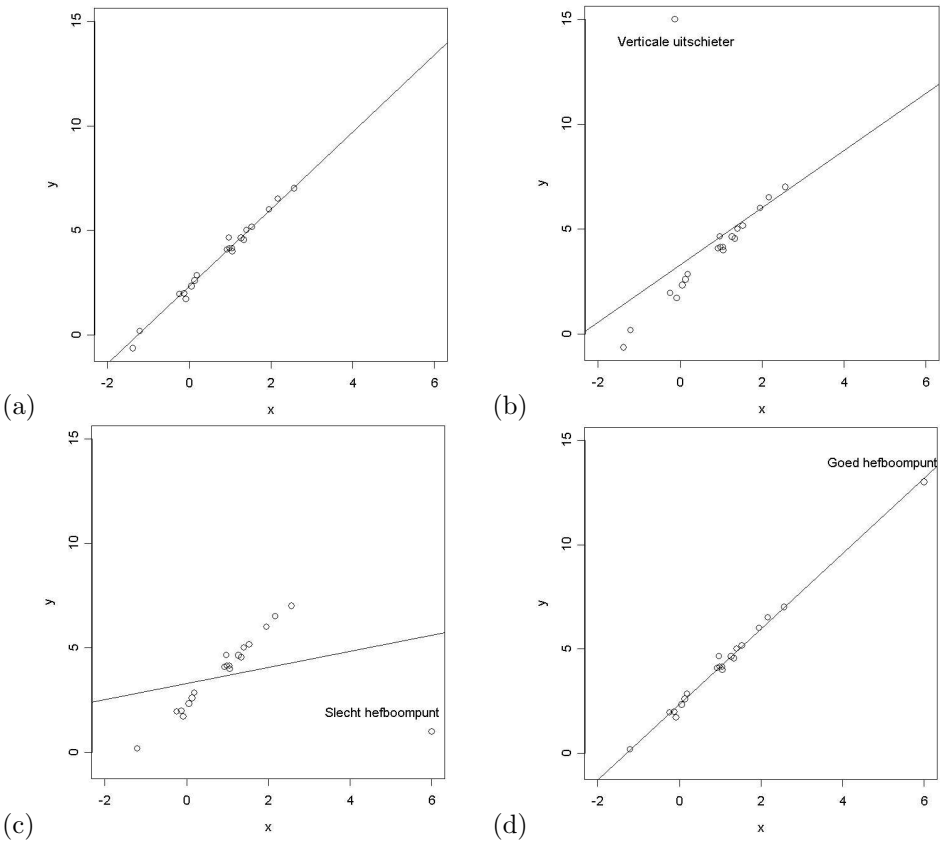

Figure 1.4: *Effect van uitschieters op de kleinste kwadraten schatter: (a) geen uitschieters (b) verticale uitschieter (c) slecht hefboompunt (d) goed hefboompunt.*

In Figuur 1.4(c) en Figuur 1.4(d) zien we zulk een hefboompunt: de corresponderende x-waarde is inderdaad ver weg van de grote meerderheid van andere punten op de x-as. In Figuur 1.4(c) zien we dat het hefboompunt erin slaagt om de geschatte regressie rechte naar zich toe te trekken: de rechte kantelt zoals een hefboom. In Figuur 1.4(d) heeft het hefboompunt schijnbaar zo goed als geen effect op de kleinste kwadraten schatting. Dit komt omdat de uitschieter nog steeds de lineaire relatie volgt die de andere punten ook volgen, en de regressie rechte dus niet doet kantelen. We noemen dit een *goed hefboompunt*, terwijl we in Figuur 1.4(c) spreken van een *slecht hefboompunt*. Slechte hefboompunten zijn het meest gevaarlijk, en oefenen meer invloed uit dan verticale uitschieters.

Keren we nu terug naar ons eerste voorbeeld van de telefoondata. In Figuur 1.5 zien we de gegevens met de geschatte kleinste kwadraten regressie rechte. Er is duidelijk aanwezigheid van verticale uitschieters, en de LS schatter wordt hierdoor sterk vertekend. We kunnen nog moeilijk zeggen dat we een goede fit bekomen voor de data.

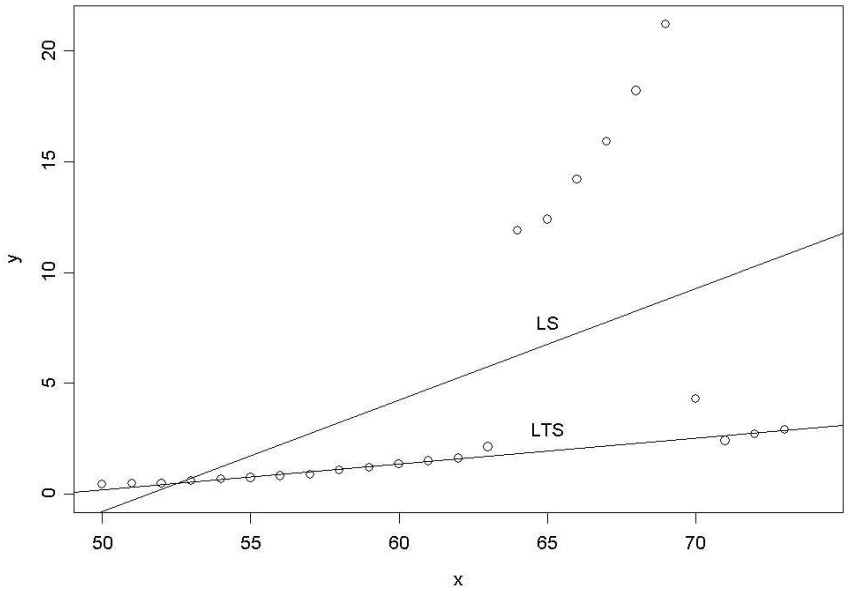

Figure 1.5: *Regressie rechte voor de telefoondata geschat met de kleinste kwadraten methode (LS) en met een robuuste schatter (LTS).*

Merk ook op dat meerdere residuen voor de goede observaties groter zijn dan de residuen voor de verticale uitschieters. Dit betekent dat gebruik maken van de grootte van de residuen -berekend ten opzichte van de regressie rechte- om uitschieters te detecteren geen goede techniek is. Residuële analyse kan erg misleidend zijn: uitschieters kunnen kleine residuen hebben (dit noemt men *mask-*

ing) en goede observaties grote residuen (dit noemt men *swamping*). Wanneer de residuen echter berekend worden ten opzichte van een robuust geschatte regressie rechte, kan men de grootte van de residuen wel gebruiken voor detectie van uitschieters. In Figuur 1.5, waar we ook een robuust geschatte regressie rechte getekend hebben, zien we onmiddellijk dat de uitschieters veel grotere residuen hebben dan de goede observaties.

Uit het bovenstaande kunnen we besluiten dat één uitschieter voldoende is om de kleinste kwadraten schatter zeer sterk te beïnvloeden. Men kan aantonen dat LS een onbegrensde invloedsfunctie en een breekpunt van nul heeft, net zoals het rekenkundig gemiddelde.

Hefboompunten en verticale uitschieters kunnen ook voorkomen in economische data. KBC Bank en Verzekeringen (2001) bestudeerde het "Economisch profiel van de Europese Unie", en we vinden in hun rapport verschillende dispersiediagrammen (of *scatterplots*) terug. Een selectie hiervan presenteren we in Figuur 1.6.

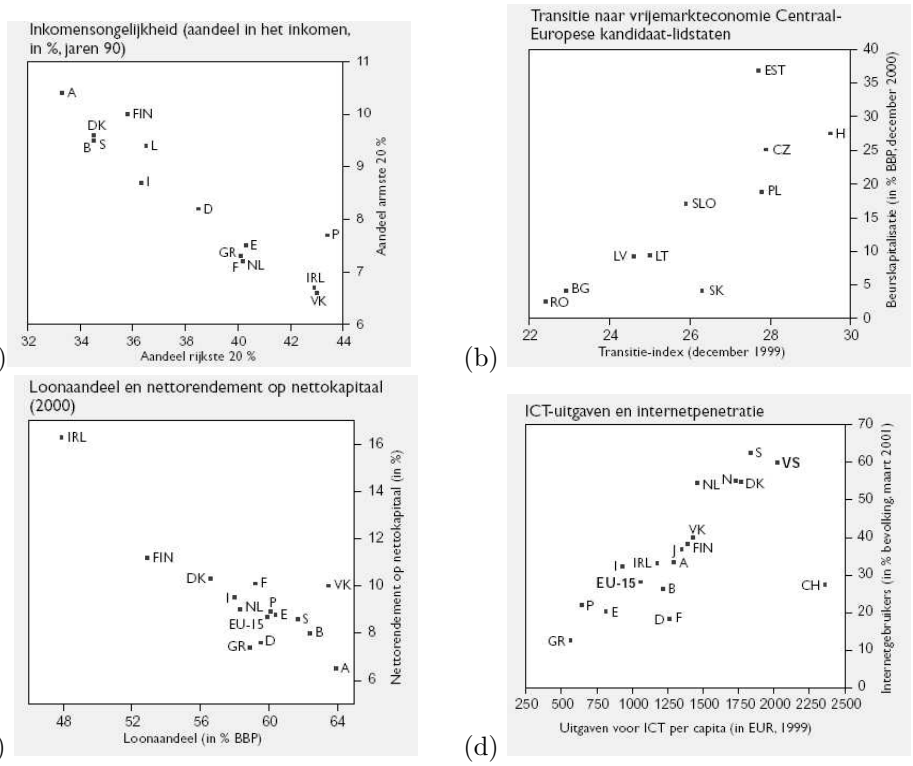

Figure 1.6: *Voorbeelden van dispersiediagrammen met verschillende types uitschieters (bron: KBC-studiedienst).*

In Figuur 1.6(a) zien we een duidelijk dalende relatie, en geen opvallende uitschieters. Figuur 1.6(b) toont een stijgende relatie, maar Estland (EST) is hier een verticale uitschieter. Dit diagram stelt Oost-Europese landen voor, en de beurskapitalisatie in Estland is veel groter dan men op basis van zijn transitie index mag verwachten. In Figuur 1.6(c) zien we een voorbeeld van een goed hefboompunt: Ierland (IRL) volgt de lineaire relatie die de andere landen ook volgen, maar heeft een extreem lage waarde voor de verklarende variabele loonaandeel. Tot slot zien we in Figuur 1.6(d) dat Zwitserland (CH) hier een slecht hefboompunt is. Het heeft de grootste waarde voor de x-variabele, maar volgt de lineaire relatie tussen "uitgaven voor ICT per capita" en "internetgebruikers" niet. In meerdere empirische studies blijkt het dat landen als Zwitserland en Luxemburg vaak als uitschieter gedetecteerd worden. Ze gedragen zich anders dan de meerderheid van andere Europese landen, en de statistische analyse mag hierdoor niet teveel beïnvloed worden.

1.3.2 Robuuste schatters

Door een geschikte doelfunctie f te kiezen in (1.4) is het mogelijk om robuuste schatters te bekomen voor het lineaire regressiemodel. Een reden waarom de kleinste kwadraten schatter erg beïnvloed wordt door uitschieters is dat het kwadraat van de residuen in de doelfunctie optreedt, waardoor hun effect nog vergroot wordt. In plaats van het kwadraat kan men ook de absolute waarden van de residuen opnemen in de doelfunctie. Zo bekomt men de kleinste absolute waarde of *Least Absolute Value* schatter

$$\hat{\theta}_{\text{LAV}} = \underset{\theta}{\operatorname{argmin}} \sum_{i=1}^{n} |r_i(\theta)|. \tag{1.6}$$

In tegenstelling tot de kleinste kwadraten methode biedt $\hat{\theta}_{\text{LAV}}$ bescherming tegen de aanwezigheid van verticale uitschieters, maar blijft gevoelig voor slechte hefboompunten. Men kan aantonen dat het breekpunt van deze schatter ook 0% is. Om een werkelijk robuuste methode met een hoog breekpunt te bekomen kan men in (1.5) de som door een mediaan vervangen. De bekomen schatter is dan de *Least Median of Squares* regressie schatter van Rousseeuw (1984)

$$\hat{\theta}_{\text{LMS}} = \underset{\theta}{\operatorname{argmin}} \underset{i}{\operatorname{med}}\, r_i^2(\theta). \tag{1.7}$$

Een ander voorstel bestaat erin om in de doelfunctie van de kleinste kwadraten schatter niet de som over alle residuen in het kwadraat te nemen, maar enkel de som over de kleinste h. Men kiest dan $h = \lfloor n(1-\alpha) \rfloor$, met α een drempel waarde tussen 0 en 0.5. De doelfunctie is dan een afgeknotte som van residuen in het kwadraat, en men bekomt de *Least Trimmed Squares* (LTS) schatter

$$\hat{\theta}_{\text{LTS}} = \underset{\theta}{\operatorname{argmin}} \sum_{i=1}^{h} r_{(i)}^2(\theta), \tag{1.8}$$

met $r^2_{(1)} \leq \ldots \leq r^2_{(n)}$. Als keuze voor α kan men bijvoorbeeld $\alpha = 0.25$ nemen, wat betekent dat men het grootste kwart van de gekwadrateerde residuen niet laat meespelen in de doelfunctie. Een andere mogelijke keuze is $\alpha = 0.50$, wat leidt tot het hoogst mogelijke breekpunt van 50%.

Ofschoon de definitie van de LMS en LTS schatter vrij eenvoudig is, zijn beide moeilijk uit te rekenen. Sinds enkele jaren zijn echter snelle algoritmes beschikbaar die deze schatters kunnen uitrekenen en die geïmplementeerd werden in statistische software. Er is een voorkeur voor de LTS schatter omdat deze statistisch efficiënter is en sneller te berekenen. De LMS heeft echter een kleinere maximale bias en is in die zin robuuster dan de LTS.

Om de robuustheid van deze schatters te illustreren keren we terug naar de telefoondata (Figuur 1.5), waar ook de rechte bekomen door de LTS schatter (met $\alpha = 0.5$) weergegeven is. We zien dat de robuuste methode de lineaire relatie, die de grote meerderheid van de observaties volgt, terugvindt. Grote residuen ten opzichte van deze regressie rechte geven ons dan de uitschieters. Eens deze uitschieters gedetecteerd zijn, kan men trachten op te sporen waarom deze observaties zich vreemd gedragen. Ook op de artificiële data van Figuur 1.4 kunnen we een robuuste schatter toepassen. We geven hier 3 configuraties van de gegevens (die zonder uitschieters, die met een verticale uitschieter en die met een slecht hefboompunt), samen met drie door LTS (met $\alpha = 0.5$) geschatte regressie rechten op één enkele tekening in Figuur 1.7. De drie regressie rechten voor deze 3 configuraties zijn praktisch niet verschillend, waardoor ze op de tekening niet te onderscheiden zijn. De uitschieters hebben dus nauwelijks effect op de geschatte LTS regressie rechte.

De robuuste regressie schatters LMS en LTS hebben ook nadelen. Door te werken met medianen en afgeknotte sommen in de doelfunctie van (1.5) in plaats van met de volledige som der gekwadrateerde residuen, zullen deze schatters aan efficiëntie inboeten. Ze zijn met andere woorden minder precies dan de LS schatter wanneer er geen uitschieters zijn en de hypothese van normaliteit geldt. Daarom werden alternatieve robuuste schatters voorgesteld die efficiënter zijn dan LMS of LTS. Vaak kan men zulke schatters interpreteren als herwogen kleinste kwadraten schatters, waar de gewichten afhangen van de grootte van de residuen ten opzichte van een initiële LTS fit. Definities en implementaties van deze schatters vindt men bijvoorbeeld in Marazzi (1993).

1.4 Voorbeelden

Lineaire regressie is een van de meest gebruikte statistische technieken. Het wordt op courante wijze gebruikt in de toegepaste economie. Toch wordt er weinig aandacht besteed aan het probleem van uitschieters wanneer men regressie toepast. Hieronder bespreken we kort twee voorbeelden van auteurs die in hun werk ro-

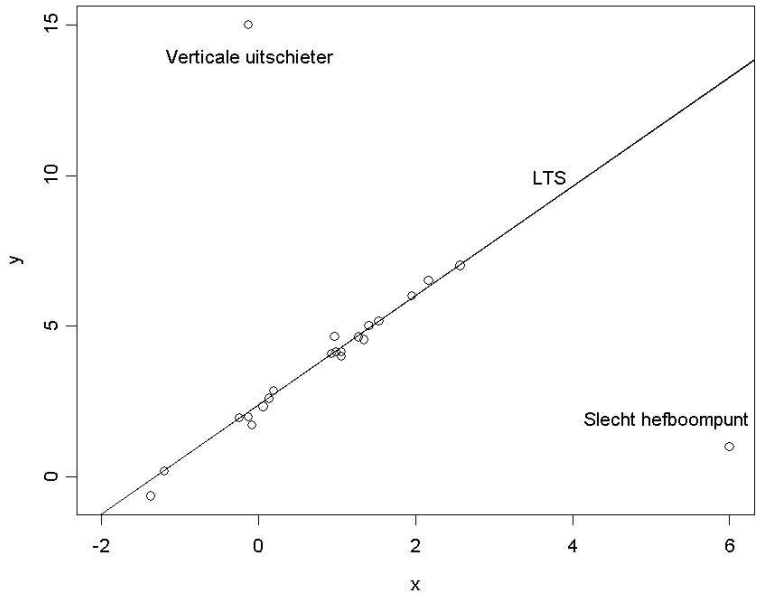

Figure 1.7: *Effect van uitschieters op de LTS schatter.*

buuste methoden gebruikten. Het eerste voorbeeld is afkomstig van De Leval (2001) die actuariële pensioenplannen bestudeert. De berekeningen van deze pensioenplannen steunen op sterftetafels, die op regelmatige basis aangepast worden. Het doel van de auteur was om waarden van toekomstige sterftetafels te voorspellen, door de sterftetrend te modelleren. We beperken ons hier tot de kans dat een 30-jarige man zal sterven in het daaropvolgende jaar, voorgesteld als q_{30}. Deze kans verandert natuurlijk doorheen de tijd, en we definiëren q_{30}^t als de sterftekans in jaar t, waar het aantal jaren gemeten wordt vanaf het referentiejaar 1885. Volgend log-lineair model werd dan geponeerd

$$q_{30}^t = ab^t$$

voor $t = 1, \ldots, T$. Hier staat a voor het initiële overlijdensniveau in het referentiejaar en b staat voor het jaarlijks percentage verandering van de kans op overlijden. Op logaritmische schaal, en na toevoeging van een storingsterm bekomen we dan een eenvoudig lineaire model

$$y_t = \alpha + \beta t + \varepsilon_t$$

met $\alpha = \log a$, $\beta = \log b$ en $y_t = \log q_{30}^t$.

Sterftetafels worden niet jaarlijks aangepast, maar worden constant gehouden gedurende bepaalde periodes. Voor twaalf verschillende periodes werden uit sterftetafels de kansen op overlijdens q_{30}^t bekomen, waar t overeenkomt met het

midden van zo een periode. De laatste beschouwde periode was 1995-1997. We hebben dus geen jaarlijkse observaties, maar slechts 12 observaties die (hopelijk) representatief zijn voor de periodes. Figuur 1.8 toont de puntenwolk (t, y_t) met twee geschatte regressie rechten gebaseerd op LS respectievelijk LTS-regressie. We zien dat de LTS een betere fit geeft voor de meer recente waarnemingen, dus deze met een grote waarden voor t. Voor de andere observaties, met kleine waarde van t, geeft LTS een minder goede fit, gezien LTS vooral de meerderheid van de data goed wil fitten. De robuuste methode zal de observaties, die het lineaire model minder goed volgen, een kleiner gewicht geven en in dit voorbeeld zijn dat de minst recente observaties. Merk op dat de eerste observatie een hefboompunt is. De reden waarom de oudste waarnemingen uitschieters zijn, is waarschijnlijk omdat er in de periode voor de tweede wereldoorlog een andere relatie tussen de kansen op overlijden en de tijd bestond. De robuuste analyse brengt aan het licht dat er twee structuren in de gegevens zijn, waar de klassieke analyse dit veel minder duidelijk zichtbaar maakt.

Stelt men zich even voor dat de meerderheid van de observaties zouden komen van de periode voor 1945. Dan zou de LTS schatter een goede fit geven voor de oudste observaties, en een minder goede fit voor de recentere. Maar, deze meer recente observaties zouden dan wel als een groep van uitschieters gedetecteerd worden. Indien men dan een voorspelling zou willen maken van een toekomstige sterftekans, is het duidelijk dat het model herschat moet worden op basis van enkel de meest recente observaties.

Een ander voorbeeld van toepassing van robuuste regressie vinden we in het artikel van Knez en Ready (1997). Er blijkt enige controverse te zijn over het effect van de grootte van een bedrijf op het verwachtte rendement van zijn aandeel op de beurs. Gegevens kwamen van niet-financiële bedrijven, genoteerd op de New York Stock Exchange (NYSE), de American Stock Exchange (AMEX), en het Nasdaq-register van het Center for Research in Security Prices (CRSP) gedurende de periode van juli 1963 tot december 1990. Fama en French (1992) identificeren meerdere risico factoren die rendementsverschillen kunnen uitleggen, maar we beperken ons hier tot de factor "grootte", gemeten als het logaritme van de totale marktwaarde van de aandelen van het bedrijf. Terwijl Fama en French hun onderzoek baseren op kleinste kwadraten schattingen van het lineaire model, gebruiken Knez en Ready (1997) de robuuste LTS schatter.

Voor de maand maart 1989, die representatief is voor andere maanden, stelt Figuur 1.9, genomen uit het artikel van Kenz en Ready (1997), de puntenwolk van de rendementen van de firma's in functie van hun groottes voor. De (lichtjes stijgende) volle rechte correspondeert met de LTS methode, terwijl de (lichtjes dalende) stippellijn correspondeert met de regressie rechte gebaseerd op de LS methode. Gezien de hellingscoëfficiënten zeer klein zijn, is het verschil tussen de twee rechten klein, maar het blijkt significant verschillend te zijn en voor de meeste maanden voor te komen. De interpretatie is natuurlijk erg verschillend: hebben we een dalend of een stijgend rendement in functie van de grootte?

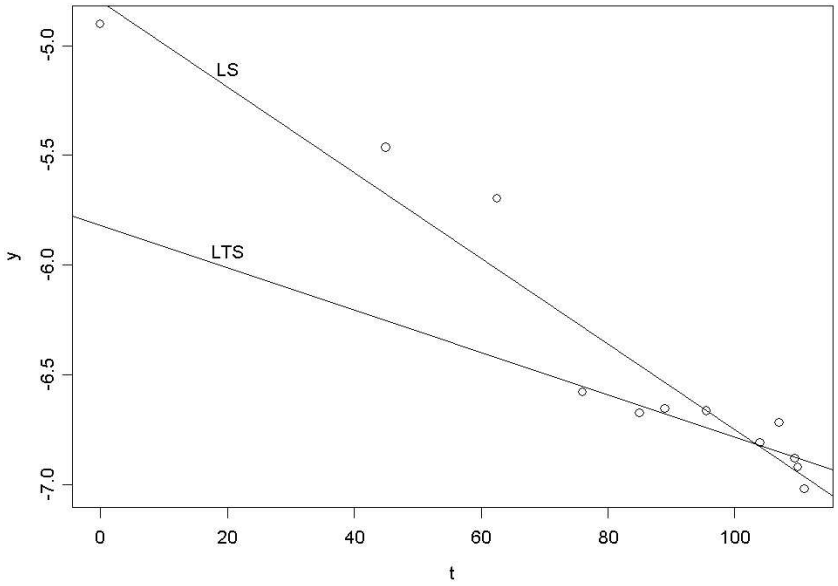

Figure 1.8: *Hierbij is* $y = \log q_{30}^t$ *, met* q_{30}^t *de kans op overlijden voor de 30-jarige Belgische mannen in jaar* t*, t.o.v. de tijd* t *sinds 1885, samen met een klassieke (LS) en robuuste (LTS) regressiefit.*

Er werd hier een 5% afgeknotte som van gekwadrateerde residuen als doelfunctie genomen voor de LTS-schatter. Er worden dus 5% van de observaties afgeknot, en deze worden met een '+' symbool in het diagram van Figuur 1.9 voorgesteld. Men merkt op dat deze observaties vooral voorkomen bij de eerder kleine bedrijven die een groot rendement halen. Merk op dat de gegevens hier niet foutief ingevoerd zijn, maar ongewone waarden aannemen. Er zijn hier dus een aantal, niet erg extreme, verticale uitschieters. De schatting bekomen met LS wordt naar boven vertekend door deze kleine bedrijven die grote positieve rendementen hebben, maar die toch maar minder dan 1% van de gegevens vertegenwoordigen. We verwijzen naar het artikel van Kenz en Ready (1997) voor meer detail.

1.5 Conclusies

In dit hoofdstuk hebben we getracht enkele basisbegrippen uit de theorie van de robuuste statistiek, zoals breekpunt en invloedsfunctie, op eenvoudige wijze uit te leggen. Verder werd een robuuste regressie schatter besproken die dan op twee economische voorbeelden werd geïllustreerd. We verwijzen naar Zaman et al (2001) voor nog andere econometrische toepassingen van robuuste methoden.

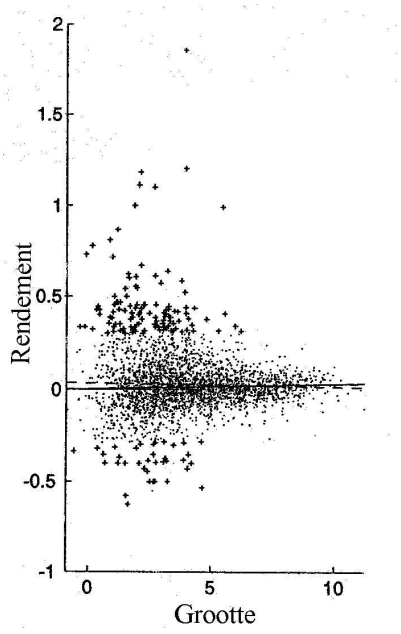

Figure 1.9: *Rendement versus "grootte" van beursgenoteerde bedrijven, met een kleinste kwadraten (stippellijn) en een robuust LTS fit met* $\alpha = 5\%$ *(volle lijn). De* 5% *"afgeknotte" observaties zijn aangeduid met een '+'.*

Laat het duidelijk zijn dat we hier slechts kort hebben kunnen kennismaken met de theorie en praktijk van de robuuste statistiek. Basiswerken in het domein zijn Huber (1981) en Hampel et al (1986), die nog steeds verplichte literatuur zijn voor iedereen die in dit vakgebied werkt. Deze twee basiswerken zijn echter weinig op toepassingen gericht, en soms niet genoeg wiskundig rigoreus. Meer wiskundige werken, waar vooral aandacht is voor het limietgedrag van robuuste schatters en toetsen, vindt men in Rieder (1994) en Jureckova en Sen (1996). Een eerste boek in robuuste statistiek dat zich tot een breed publiek richtte, en zeker heeft bijgedragen tot een verdere doorbraak en verspreiding van het onderzoeksgebied, is Rousseeuw en Leroy (1987). Aan de hand van vele voorbeelden wordt hier op eenvoudige wijze de robuuste regressie problematiek behandeld. Een ander toegankelijk werk is Staudte en Seather (1990), dat naast het regressiemodel ook nog veel aandacht besteed aan één- en twee-steekproef-problemen. Recentere werken zijn Wilcox (1997) en McKean en Hettmansperger (1998). Het eerste boek is erg praktijkgericht, vaak met een eigenzinnige keuze van de aangewende methoden, en het tweede focust op het gebruik van rang-methoden. Vermelden we nog het handboek van Madalla en Rao (1997), waarin meerdere robuuste statistische inferentie procedures behandeld worden.

Verder bestaat er ook een literatuur die procedures beschrijft om uitschieters en afwijkingen van een geponeerd regressiemodel te detecteren (bvb. Riani en Atkinson 2000, Chatterjee et al 2000, en Cook en Weisberg 1999). We spreken hier van het domein van *Regression Diagnostics*. Merk op dat sommige van de voorgestelde procedures in deze literatuur enkel kunnen gebruikt worden indien er slecht één enkele uitschieter aanwezig is. Ze laten niet toe om het model te valideren indien er meerdere uitschieters aanwezig zijn. Het is hier niet de bedoeling om schatters te berekenen of toetsen uit te voeren. In het bekende boek van Draper en Smith (1998) over toegepaste regressieanalyse komt deze aanpak, samen met robuuste regressie, aan bod.

Een ander domein, gerelateerd tot robuustheid, is exploratieve gegevensanalyse. Hier worden grafieken en beschrijvende statistiek gebruikt om inzicht te krijgen in de structuur van de gegevens. Merk op dat we met beschrijvende statistiek geenszins het gebruik van eenvoudige schatters bedoelen, maar veeleer het berekenen van beschrijvende maten zonder expliciete referentie naar een statistisch model. Robuuste methodes, die de structuur van de meerderheid van de data zoekt en toelaat uitschieters te detecteren, is hier een natuurlijk hulpmiddel (zie Hoaglin et al 1982).

Zoals reeds vermeld, is het mogelijk om robuuste regressie uit te voeren met bekende statistische softwarepakketten. Baanbrekend hierin is het pakket Splus, wat reeds vele jaren over een uitgebreide bibliotheek van robuuste procedures beschikt (Marazzi 1993). Meer recent werden robuuste schatters in Stata en SAS opgenomen, we verwijzen hiervoor naar de webdocumenten van Chen et al (2003) en SAS (2002).

Terwijl robuuste regressie goed bestudeerd is, zijn er nog vele andere statis-

tische technieken waar de robuustheid nog verder voor ontwikkeld moet worden. Vooral in het domein van de multivariate statistiek, niet-lineaire veralgemeende regressie en tijdreeksmodellen is er nog veel werk te verrichten. Door de auteurs van dit hoofdstuk werden bijdragen geleverd in onder andere principaalcomponentenanalyse (Croux en Haesbroeck 2000), factor modellen (Croux et al 2003), logistische regressie (Croux en Haesbroeck 2003), en discriminantanalyse (Croux en Dehon 2001, Croux en Joossens 2005). Verder onderzoek is nog steeds lopend binnen onze onderzoeksgroep.

Chapter 2

Empirical comparison of the classification performance of robust linear and quadratic discriminant analysis

Co-Author: C. Croux

Summary The aim of this chapter is to look at the behaviour of the total probability of misclassification of robust linear and quadratic discriminant analysis. The effect of outliers on the discriminant rules is studied by comparing their total probabilities of misclassification in presence of outliers.

2.1 Introduction

In discriminant analysis one observes two groups of multivariate observations forming together the training sample. Using these data a discriminant rule is determined, that is used afterwards to classify new observations into one of the two groups. In this chapter we restrain us to the case of two multivariate normal distributed populations. We observe p-variate observations $x_{11}, \ldots, x_{1n_1}$ coming from a first population $\wp_1 \sim H_1 = N_p(\mu_1, \Sigma_1)$ and $x_{21}, \ldots, x_{2n_2}$ coming from a second population $\wp_2 \sim H_2 = N_p(\mu_2, \Sigma_2)$.
Supposing that the observations are generated from two multivariate normal distributions, it is known that the optimal discriminant rule (i.e. the one minimizing the misclassification probability) is a quadratic function given by

$$Q(x) = -\frac{1}{2}x^t(\Sigma_1^{-1} - \Sigma_2^{-1})x + (\mu_1^t\Sigma_1^{-1} - \mu_2^t\Sigma_2^{-1})x - k(\mu_1, \mu_2, \Sigma_1, \Sigma_2), \quad (2.1)$$

where

$$k(\mu_1, \mu_2, \Sigma_1, \Sigma_2) = \frac{1}{2}\log\left(\frac{|\Sigma_1|}{|\Sigma_2|}\right) + \frac{1}{2}(\mu_1^t\Sigma_1^{-1}\mu_1 - \mu_2^t\Sigma_2^{-1}\mu_2).$$

We assign a new p-variate observation x to $\wp_1$ if

$$Q(x) > \log\left(\frac{c_2\pi_2}{c_1\pi_1}\right) = \tau, \tag{2.2}$$

where c_1 and c_2 are the costs of misclassifying a unit of, respectively, $\wp_1$ and $\wp_2$ and π_1 and π_2 are the prior probabilities that x will belong to, respectively, $\wp_1$ and $\wp_2$. In practice these parameters are unknown and therefore we set $\tau = 0$ throughout this chapter.

If we assume that the covariance matrices are equal ($\Sigma := \Sigma_1 = \Sigma_2$), we get the familiar Fisher's linear discriminant rule

$$L(x) = (\mu_1 - \mu_2)^t\Sigma^{-1}x - \frac{1}{2}(\mu_1 - \mu_2)^t\Sigma^{-1}(\mu_1 + \mu_2). \tag{2.3}$$

Since the primary goal of discriminant analysis is to classify data, we are particulary interested in the total probability of misclassification of a particular discriminant rule.

Robust linear discriminant analysis has been considered in several papers (e.g. Hawkins and McLachlan 1997; He and Fung, 2000; Croux and Dehon, 2001). The first ones to consider robust quadratic discriminant analysis seem to be Randles et al. (1978), who used M-estimators for the means μ_1 and μ_2, and covariance matrices Σ_1 and Σ_2 in (2.1) and a rank based rule to estimate the cut-off value τ in (2.2). One of the Editors of this book also pointed out a forthcoming paper of Hubert and Van Driessen (2003), using the MCD-estimators. Note that this approach is extendable to the multiple group case. In this chapter we compare the performance of robust linear and quadratic discriminant analysis using both S- and MCD-estimators. Based on simulation experiments, our main founding is that it seems to be profitable to use the quadratic over the linear discriminant rule, also in presence of outliers. Also, and not surprisingly, the robust rules outperform the classical ones when deviating from the model, while performing almost as good at the model distribution.

2.2 Robustification of classical discriminant analysis

Outliers and atypical observations might have an influence on the results of classical discriminant analysis, since the discriminant rules are based on estimates of the population parameters. The outliers and atypical observations might shift the estimated means and they might blow up the dispersion matrices. To prevent this

we make use of robust estimators of the population parameters. For our study we will look at two robust estimators, the MCD-estimator and the S-estimator.

The MCD-estimators were introduced by Rousseeuw (1985). The estimator is given by the subset of size h for which the determinant of its covariance matrix is minimal. The MCD-estimator of location is then given by the mean of these h observations and the MCD-estimator of covariance is given by their covariance matrix.

The S-estimators of location and multivariate dispersion were jointly introduced by Davies (1987) and Rousseeuw and Leroy (1987). To define these estimators, let X be a sample of p-variate observations and let n_X the number of observations in the sample. The S-estimators of location and dispersion of this sample are defined as

$$(\hat{\mu}, \hat{\Sigma}) = \underset{\mu,\ \Sigma}{\operatorname{argmin}}\{det\Sigma\} \quad \text{such that} \quad \frac{1}{n_X}\sum_{x \in X} \rho\left(\sqrt{(x-\mu)^t \Sigma^{-1}(x-\mu)}\right) = b,$$

where $\mu \in \mathbb{R}^p$ and $\Sigma \in \mathbb{R}^{p \times p}$ is a symmetric positive definite matrix. The function ρ needs to satisfy the following condition

(R) $\rho : \mathbb{R} \to \mathbb{R}$ is a symmetric, continuous, non decreasing function on $[0, \infty)$. Moreover, $\rho(0) = 0$ and ρ has a continuous derivative in all but finite number of points.

The constant b is set equal to $E_{F_0}[\rho(\|Z\|)]$ for $Z \sim F_0$, the central model distribution, being $N(0, I_p)$ in our case. The most commonly used choice for ρ is the Biweight function which is defined as

$$\rho(u) = \begin{cases} \frac{u^2}{2} - \frac{u^4}{2c^2} + \frac{u^6}{6c^4} & \text{if } |u| < c \\ \frac{c^2}{6} & \text{if } |u| \geq c \end{cases},$$

where c is a tuning constant to achieve the desired value of the breakdown point.

The breakdown point of an estimator is the fraction of the data that can be completely contaminated without destroying the estimator. A breakdown point is a measure of robustness and resides between 0% and 50%. In this chapter, we will use the MCD- and S-estimators with breakdown point 25%. The choice of a 25% breakdown point gives a good compromise between efficiency and robustness of the estimators (see e.g. Croux and Haesbroeck 1999).

Another way to robustify the classical method is by detecting the influential observations and deleting them from the samples. Measures to diagnose influential observations in the context of discriminant analysis have been proposed by Fung (1995a, 1996a). By deleting these influential observations from the sample and using the classical discriminant analysis based on the remaining observations, the classical method becomes robust. Note that for detecting these observations, it is also advised to use robust estimates to avoid the masking effect. Indeed, it is

well known that diagnostics based on non-robust estimates do not always detect all outliers.

As a measure of performance, we will look at the total probability of misclassification. Formally, the total probability of misclassification (TPM) is given by

$$\text{TPM} = \pi_1 P(Q(x) < \tau \mid x \sim \wp_1) + \pi_2 P(Q(x) > \tau \mid x \sim \wp_2).$$

The total probability of misclassification according to a rule can be estimated by classifying observations of which the source population is known and look at the fraction of misclassified observations. Under the normality assumption the probabilities of misclassification for the optimal linear and quadratic discriminant rules can be computed theoretically as function of the population parameters of location and dispersion. In the case of equal covariance matrices, the total probability of misclassification in the linear case is given by

$$\text{TPM}_{\text{linear}} = \Phi\left(\frac{-\Delta}{2}\right), \tag{2.4}$$

where Φ is the cumulative standard normal distribution function and Δ is the Mahalanobis distance between the populations, namely $\sqrt{(\mu_1-\mu_2)^t\Sigma^{-1}(\mu_1-\mu_2)}$. Since the discriminant rule in the linear case is of the form $L(x) = a^t x + b$, we get in the case of populations with different covariance matrices

$$\begin{aligned}
\text{TPM}_{\text{linear}} &= \pi_1 P(L(x) < 0 \mid x \in \wp_1) + \pi_2 P(L(x) > 0 \mid x \in \wp_2) \\
&= \pi_1 P(a^t x + b < 0 \mid x \sim N_p(\mu_1, \Sigma_1)) \\
&\quad + \pi_2 P(a^t x + b > 0 \mid x \sim N_p(\mu_2, \Sigma_2)) \\
&= \pi_1 P\left(z < \frac{-b - a^t\mu_1}{\sqrt{a^t\Sigma_1 a}} \mid z \sim N(0, I)\right) + \\
&\quad \pi_2 P\left(z > \frac{-b - a^t\mu_2}{\sqrt{a^t\Sigma_2 a}} \mid z \sim N(0, I)\right) \\
&= \pi_1 \Phi\left(\frac{-b - a^t\mu_1}{\sqrt{a^t\Sigma_1 a}}\right) + \pi_2\left(1 - \Phi\left(\frac{-b - a^t\mu_2}{\sqrt{a^t\Sigma_2 a}}\right)\right),
\end{aligned} \tag{2.5}$$

where Φ is standard normal cumulative distribution function and a and b are as in (2.3), where Σ can be estimated by the pooled covariance matrix. A theoretical expression of $\text{TPM}_{\text{quadratic}}$ can also be obtained, but it needs to be evaluated numerically. For more details and expressions, we refer to Croux and Joossens (2005).

2.3 Simulation experiment

We generate 1000 random normal distributed observations from two populations $\wp_1 \sim N_p(\mu_1, \Sigma_1)$ and $\wp_2 \sim N_p(\mu_2, \Sigma_2)$ of dimension $p = 3$, which are constructing the training sample. First, we consider the uncontaminated samples and afterwards we contaminate them by adding outliers. We will compute the classic and the robust discriminant rules for both the linear and the quadratic case.

As already mentioned for linear discriminant analysis the discriminant rule (2.3) is of the following form

$$L(x) = a^t x + b,$$

where a is a p-dimensional vector and b a scalar. In the case of quadratic discriminant analysis the discriminant rule (2.1) is of the following form

$$Q(x) = x^t A x + d^t x + e,$$

where A is a p-dimensional matrix , d is a p-dimensional vector and e is a scalar. These parameters need to be estimated, and this will be done classically and robust, using the MCD-estimators and S-estimators.

For the MCD-estimators we use the fastmcd algorithm by Rousseeuw and Van Driessen (1999) and for the S-estimators we used the algorithm developed by Ruppert (1992). For estimating the sample covariance matrix $\Sigma = \Sigma_1 = \Sigma_2$ in the linear case a pooled covariance matrix estimated is used.

Programs for computing robust linear and quadratic discriminant analysis can be retrieved from the website *http://www.econ.kuleuven.ac.be/kristel.joossens*. Note that our primary interest is not in the parameter estimates of the linear and quadratic rule, but only in the probability of misclassification of the rules.

After constructing the discriminant rules we generate 5000 random normally distributed observations for each population, without contamination. These validation samples have the same distribution as the training samples (if we do not take the outliers into account). These observations are classified by all the discriminant rules. In linear discriminant analysis we assign an observation x to the first population if and only if $L(x) > 0$ and in quadratic discriminant analysis if and only if $Q(x)?0$. Since the source populations of these observations are known, we are able to decide whether an observation is then misclassified. The fraction of misclassified observations is then an estimate of the total probabilities of misclassification using the specific discriminant rule. By taking a number of 5000 observations in the validation sample, we aim at attaining an accurate estimate of the population misclassification rate. Indeed, for a given discriminant rule, the standard error of this estimate is always less then 0.71%.

Since the classification rules depend strongly on the generated data of the training sample, we generate 500 different training data sets, from which we apply both linear and quadratic discriminant rules based on classic and robust (MCD-

and S-) estimators. Working with 500 different training sets allows us to take the estimation variability of the discriminant rules into account. Indeed, we can compute the mean and the standard deviation of the total probability of misclassification over the 500 runs for linear and quadratic, classic and robust discriminant analysis.

2.4 Simulation results

We denote the total probability of misclassification in percents and put the associated standard deviations between parenthesis (also in percents). The theoretical values of the TPM, in absence of contamination, are also mentioned. Let us consider 3 cases and take rather extreme cases to illustrate the effect of outliers. For simplicity of notation, a stands for a vector $(a, a, a)^t$ and I for the 3-dimensional identity matrix. Note that similar simulation experiments, but only for the linear case, were constructed in He and Fung (2000) and Croux and Dehon (2001).

2.4.1 Unequal means and equal covariance matrices

The two populations have the same covariance matrix, but different means. Because of the equality of the covariance matrices, the linear method should be preferred because it is much easier then the quadratic and it should lead (asymptotically) to the same rules. Note that in the case of equal covariance matrices, the theoretical values of the total probability of misclassification for the quadratic rule are the same as for the linear rule, computed as in (2.4). The two populations consist of 1000 observations, where $\wp_1 \sim N_3(-1; I)$ and $\wp_2 \sim N_3(1; I)$. Let us now contaminate 10% of observations of each population, deviating in location from the original distributions. The outliers of the first populations follow a $N_3(9; I)$ distribution and those of the second population a $N_3(-9; I)$ distribution. As a second type of contamination, 10% of the data are switched from one group to the other.

From Table 2.1 it seems clear that the results coming from the linear discriminant analysis are close to those of the quadratic discriminant analysis. When the populations are contaminated by extreme outliers, the robust estimators are obviously much better than the classic estimators. Note that in this case the S-estimator behaves more robust than the MCD-estimator, yielding lower values for the simulated TPM. But, taking the standard errors into account, this difference is not significant. For the second type of contamination all discriminant rules performs equally well. Note that this simulation experiment is limiting to equal size groups.

Table 2.1: *Average TPM with standard deviation between parenthesis over* 500 *runs, in case of equal covariance matrices. Linear and quadratic discriminant rules base on the classical, the MCD- and the S-estimators are considered.*

	Theoretical	*Classic*	*MCD*	*S*
Linear	4.16	4.09 (0.02)	4.13 (0.05)	4.09 (0.02)
Quadratic	4.16	4.10 (0.02)	4.16 (0.06)	4.10 (0.02)
	in presence of 10% outliers in location of type 1			
Linear		49.95 (3.60)	4.12 (0.04)	4.09 (0.02)
Quadratic		49.88 (2.85)	4.14 (0.05)	4.10 (0.02)
	in presence of 10% outliers in location of type 2			
Linear		4.21 (0.04)	4.20 (0.03)	4.20 (0.02)
Quadratic		4.24 (0.04)	4.23 (0.04)	4.21 (0.04)

2.4.2 Equal means and unequal covariance matrices

Let us consider now populations with the same mean, but with different covariance matrices. For the computation of the theoretical values of the total probability of misclassification in the linear case we use formula (2.5) and in the quadratic case we use the formula proposed by Croux and Joossens (2005). For the linear discriminant rule, the pooled sample covariance matrix is used.

Let $\wp_1$ and $\wp_2$ be two populations consisting each of 1000 observations following a $N_3(0; 100I)$ and a $N_3(0; I)$ distribution, for respectively the first and the second population. We contaminate again 10% of the observations, but now to get outliers in dispersion. We generate them as they would come from the wrong population. This means that 100 observations have $100I$ as covariance matrix instead of I, for the first population and vice versa for the second population. This leads to the following results.

Table 2.2: *As in Table 2.1, but in case of equal means and unequal covariance matrices.*

	Theoretical	*Classic*	*MCD*	*S*
Linear	50.00	45.73 (1.89)	43.48 (2.66)	45.56 (1.89)
Quadratic	0.82	0.83 (0.01)	0.82 (0.18)	0.83 (0.01)
	in presence of 10% outliers			
Linear		46.64 (2.21)	44.95 (2.16)	45.98 (1.70)
Quadratic		7.24 (0.46)	1.75 (0.13)	1.11 (0.03)

In case of equal means and unequal covariance matrices we notice a significant difference between the linear and the quadratic discriminant analysis. The quadratic method is much better, which is logical because the linear method

assumes that the covariance matrices are equal. In case of contamination the robust method is again better than the classic and the S-estimator is better than the MCD-estimator in this case.

2.4.3 Unequal means and unequal covariance matrices

Consider now two populations with different means and different covariance matrices. Since this is the more representative case, we consider three different sampling schemes. Three different sampling schemes will be considered, with different degrees of overlap between the two populations. This results in increasing values for the total probability of misclassification. From a practical point of view, it means that Scheme 1 corresponds to "easy"' classification problems, while the last scheme is a more difficult one.

Scheme 1: Let $\wp_1 \sim N_3(-1; I)$ and $\wp_2 \sim N_3(1; 0.25I)$ be two populations each consisting of 1000 observations. First we contaminate the samples, by creating 10% outliers in location such that the classical mean estimators should become close: 100 observations of $\wp_1$ follow a $N_3(9; I)$ distribution and 100 observations of $\wp_2$ follow $N_3(-9; 0.25I)$ distribution. Secondly, we change these outliers, such that they deviate in location and dispersion. This is done by changing their corresponding covariance matrices I and $0.25I$ into $0.25I$ and I, for respectively, the outliers in the first and the second population.

Scheme 2: Let $\wp_1 \sim N_3(0; 2.25I)$ and $\wp_2 \sim N_3(1; 0.25I)$. We contaminate 100 observations in each of those two populations such that they deviate in location and dispersion. The outliers of the first population follow a $N_3(3; 9I)$ distribution and those of the second population follow a $N_3(-1; I)$ distribution.

Scheme 3: Let us consider two populations, $\wp_1 \sim N_3(0; 4I)$ and $\wp_2 \sim N_3(1; 16I)$, consisting each of 1000 observations. The first type of outliers are generated by replacing 100 observations of $\wp_1$ by observations coming from a $N_3(4; I)$ distribution and 100 from $\wp_2$ by observations coming from a $N_3(-16; I)$ distribution. A second type of outliers is generated by replacing 100 observations from each population as if they would come from the wrong population. In other words, the first population consists of 900 observations coming from a $N_3(0; 4I)$ distribution and 100 coming from a $N_3(1; 16I)$ distribution and vice versa for the second population.

In all contaminated cases (see Table 2.3), the quadratic rule outperforms the linear one. Classification based on robust estimates is much better than based on classical estimates in presence of outliers. (Note that for Scheme 3, with less extreme outliers as in the previous contamination schemes, the classical quadratic rule is not loosing much.) In the first scheme, for the linear case in the second scheme and for the second type of outliers in the third scheme, the robust discriminant analysis based on the S-estimator is slightly better, however not significantly, than the one based on the MCD-estimator in presence of outliers (linear and quadratic). But for the first type of outliers in the third scheme and

Table 2.3: *Average TPM with standard deviation between parenthesis over* 500 *runs, in the most general case of unequal means and unequal covariance matrices. Linear and quadratic discriminant rules base on the classical, the MCD- and the S-estimators are considered. Three different sampling schemes are considered.*

SCHEME 1	*Theoretical*	*Classic*	*MCD*	*S*
Linear	1.92	2.05 (0.08)	2.09 (0.13)	2.05 (0.08)
Quadratic	0.74	0.75 (0.08)	0.86 (0.08)	0.75 (0.09)
	in presence of 10% outliers in location			
Linear		49.89 (3.61)	2.13 (0.11)	2.12 (0.08)
Quadratic		26.44 (0.06)	0.80 (0.04)	0.87 (0.04)
	in presence of 10% outliers in location and dispersion			
Linear		49.72 (3.60)	2.05 (0.11)	2.04 (0.08)
Quadratic		26.47 (0.12)	0.78 (0.05)	0.83 (0.04)
SCHEME 2				
Linear	18.51	16.20 (0.06)	16.29 (0.14)	16.20 (0.06)
Quadratic	6.82	6.87 (0.03)	8.80 (0.08)	6.87 (0.03)
	in presence of 10% outliers in location and dispersion			
Linear		16.65 (0.45)	16.29 (0.10)	16.23 (0.01)
Quadratic		13.04 (0.26)	8.00 (0.24)	8.27 (0.15)
SCHEME 3				
Linear	38.82	37.29 (0.31)	37.40 (0.48)	37.29 (0.32)
Quadratic	20.67	20.30 (0.05)	23.40 (0.34)	20.31 (0.05)
	in presence of 10% outliers of type 1			
Linear		56.05 (0.82)	37.48 (0.44)	39.82 (0.63)
Quadratic		23.74 (0.19)	21.85 (0.23)	23.03 (0.17)
	in presence of 10% outliers of type 2			
Linear		37.86 (0.33)	38.20 (0.33)	38.12 (0.35)
Quadratic		20.73 (0.11)	20.42 (0.10)	20.36 (0.08)

for the quadratic case in the second scheme the robust classification based on the discriminant rules using the MCD-estimator is slightly better then the one based on the S-estimator in presence of outliers.

If no outliers are present, the S-estimator yields lower misclassification rates than the MCD-estimator. From these simulations in the case of unequal means and covariance matrices, we can conclude that quadratic discriminant analysis is always preferred to linear discriminant analysis and that the robust method is better then the classical method. We notice that sometimes the robust discriminant analysis based on the MCD-estimators is better then the robust discriminant analysis based on the S-estimators and sometimes it is vice versa. Therefore we cannot say which of the robust estimators is preferred.

2.5 Conclusions

It is shown that quadratic discriminant analysis is needed when the covariance matrices of the populations are different. Robust estimators are needed if there are outliers in the populations. If no outliers are present in the training sample there is practical no loss of the robust procedure in classification. Therefore, the only major loss when using robust discriminant rules is computational cost. Note however that fast algorithms have been used, and that computer software is available to compute the robust estimators considered in this chapter. Another motivation for using MCD- and S-estimators their high breakdown point.

While this chapter focuses on a simulation study, a more formal and theoretical treatment is provided in Croux and Joossens (2005), where the influence of outliers on the quadratic discriminant rule and the estimated total probability of misclassification is studied.

Chapter 3

Influence of observations on the misclassification probability in quadratic discriminant analysis

Co-Author: C. Croux

Summary In this chapter it is studied how observations in the training sample affect the misclassification probability of a quadratic discriminant rule. An approach based on partial influence functions is followed. It allows to quantify the effect of observations in the training sample on the performance of the associated classification rule. Focus is on the effect of outliers on the misclassification rate, merely than on the estimates of the parameters of the quadratic discriminant rule. The expression for the partial influence function is then used to construct a diagnostic tool for detecting influential observations. Applications on real data sets are provided.

3.1 Introduction

In discriminant analysis one observes two groups of multivariate observations, forming together the *training sample*. For the data in this training sample, it is known to which group they belong. On the basis of the training sample a discriminant function Q will be constructed. Such a rule is used afterwards to classify new observations, for which the group membership is unknown, into one of the two groups. Data are generated by two different distributions, having densities $f_1(x)$ and $f_2(x)$. The higher the value of Q the more likely the new observation

has been generated by the first distribution. Taking the log-ratio of the densities yields

$$Q(x) = \log \frac{f_1(x)}{f_2(x)}.$$

For f_1 a normal density with mean μ_1 and covariance matrix Σ_2, and for f_2 another normal density with parameters μ_2 and Σ_2, one gets

$$Q(x) = \frac{1}{2}\left\{(x-\mu_2)^t\Sigma_2^{-1}(x-\mu_2) - (x-\mu_1)^t\Sigma_1^{-1}(x-\mu_1)\right\} + \frac{1}{2}\log\frac{|\Sigma_2|}{|\Sigma_1|}. \quad (3.1)$$

Here, $|\Sigma|$ stands for the determinant of a square matrix Σ. The above equation can be written as a quadratic form

$$Q(x) = x^tAx + b^tx + c, \quad (3.2)$$

where

$$A = \frac{1}{2}(\Sigma_2^{-1} - \Sigma_1^{-1}), \quad (3.3)$$

$$b = \Sigma_1^{-1}\mu_1 - \Sigma_2^{-1}\mu_2, \quad (3.4)$$

$$c = \frac{1}{2}\log\frac{|\Sigma_2|}{|\Sigma_1|} + \frac{1}{2}(\mu_2^t\Sigma_2^{-1}\mu_2 - \mu_1^t\Sigma_1^{-1}\mu_1). \quad (3.5)$$

The function $Q(x)$ is called the quadratic discriminant function. Although it has been derived from normal densities it can also be applied as such without making distributional assumptions.

Future observations will now be classified according to the following discriminant rule: if $Q(x) > \tau$, where τ is a selected cut-off value, then assign x to the first group. On the other hand if $Q(x) < \tau$, then assign x to the second group. Now let π_1 be the prior probability that an observation to classify will be generated by the first distribution, and set $\pi_2 = 1 - \pi_1$. For normal source distributions it is known that the optimal discriminant rule, in the sense of minimising the expected probability of misclassification, is given by the above quadratic rule with $\tau = \log(\pi_2/\pi_1)$, e.g. Johnson and Wichern (2002, Chapter 11). In practice, the prior probabilities π_1 and π_2 are often unknown and one uses $\tau = 0$.

The discriminant function (3.1) still depends on the unknown population quantities μ_1, μ_2, Σ_1 and Σ_2, and needs to be estimated from the training sample. So let $x_1, \ldots, x_{n_1}$ be a sample of p-variate observations coming from the first distribution H_1^0 and $x_{n_1+1}, \ldots, x_n$ a second sample drawn from H_2^0. These samples together constitute the training sample. An observation in the training sample will influence the sample estimates of location and covariance, and hence the discriminant rule. In Quadratic Discriminant Analysis (QDA) the primary interest is not in knowing or interpreting the parameter values in (3.2). The aim is to

use QDA for classification purposes. Focus in this chapter is on how observations belonging to the training sample affect the total probability of misclassification, and this effect will be quantified by the influence function. Influence functions in the multi-sample setting were already considered by several authors, e.g. Fung (1992,1996b). In this chapter, the formalism of partial influence functions (Pires and Branco, 2002) as an extension of the traditional influence function concept to the multi-sample setting will be followed.

In the case of equal covariance matrices $\Sigma_1 = \Sigma_2 = \Sigma$ the linear discriminant rule of Fisher results as a special case of (3.1):

$$L(x) = (\mu_1 - \mu_2)^t \Sigma^{-1} (x - \frac{\mu_1 + \mu_2}{2}). \tag{3.6}$$

Influence analysis for Linear Discriminant Analysis (LDA) has been studied by Campbell (1978), Critchley and Vitiello (1991) and Fung (1992, 1995a). The quadratic case seems to be much harder. Some numerical experiments have been conducted to assess the influence of outliers in the training sample on QDA (e.g. Lachenbruch, 1979), while Fung (1996a) proposes several influence measures based on the leave-one-out approach. A more formal approach to influence analysis for quadratic discriminant analysis seems not to exist yet in the literature.

In Section 3.2 of the chapter, a population expression for the total probability of misclassification is presented. The latter is then used as a starting point to compute the partial influence functions for the classification errors in Section 3.3. The expressions obtained for the partial influence function are not only valid when the classical sample averages $\hat{\mu}_1$, $\hat{\mu}_2$ and sample covariance matrices $\hat{\Sigma}_1$ and $\hat{\Sigma}_2$ are used to estimate the unknown population parameters in the discriminant function Q, but also when robust estimators are used. Computations are tedious here and most details have been moved to the Appendix. Besides being of theoretical interest, measuring the influence of an observation in the training sample on the future classification error can be used as a diagnostic tool to detect influential observations. Section 3.4 presents such a diagnostic tool for diagnosing influential points in a classical discriminant analysis, based on the usual sample averages and covariances. However, to make this diagnostic measure robust, i.e. not suspect to masking effects, robust estimates of the population parameters need to be plugged in the theoretical expressions of the influence functions. Several examples in Section 3.4 illustrate the use of this diagnostic tool. Finally, some conclusions are made in Section 3.5.

3.2 Total probability of misclassification

In this section a population version of the Total Probability of Misclassification (TPM) is presented. Denote $H^0 = (H_1^0, H_2^0)$, where H_1^0 and H_2^0 are the distributions having generated the training samples. The population version of the

quadratic discriminant rule is then, by analogy with (3.2),

$$Q(x; H^0) = x^t A(H^0)x + b(H^0)^t x + c(H^0), \tag{3.7}$$

where the population values of the coefficient of the discriminant rule are

$$\begin{aligned} A(H^0) &= \frac{1}{2}\left\{C_2(H^0)^{-1} - C_1(H^0)^{-1}\right\} && (3.8)\\ b(H^0) &= C_1(H^0)^{-1}T_1(H^0) - C_2(H^0)^{-1}T_2(H^0) && (3.9)\\ c(H^0) &= \frac{1}{2}\log\left(\frac{|C_2(H^0)|}{|C_1(H^0)|}\right) && (3.10)\\ &\quad + \frac{1}{2}\left\{T_2(H^0)^t C_2(H^0)^{-1}T_2(H^0) - T_1(H^0)^t C_1(H^0)^{-1}T_1(H^0)\right\}. \end{aligned}$$

In the above formula $T_1(H^0)$ and $T_2(H^0)$ are the values of a location functional T at the distributions H_1^0 and H_2^0. When performing classical discriminant analysis one gets the population averages, i.e. $T_1(H^0) = E_{H_1^0}(X)$ and $T_2(H^0) = E_{H_2^0}(X)$. Similarly, $C_1(H^0)$ and $C_2(H^0)$ are the values of a scatter matrix functional C at the distributions H_1^0 and H_2^0. For classical discriminant analysis, C yields the population covariance matrix, i.e. $C_1(H^0) = \mathrm{Cov}_{H_1^0}(X)$ and $C_2(H^0) = \mathrm{Cov}_{H_2^0}(X)$. In this chapter, focus is on *classical* quadratic discriminant analysis, where one uses the conventional population averages and population covariances, resulting in $Q = Q_{Cl}$. However, it is also possible to use *robust* measures of location for T and robust measures of scatter for C, yielding a different discriminant rule denoted by Q_R. For information on robust estimators of location and scatter we refer to Hampel et al. (1986) and Maronna and Yohai (1998).

The distribution generating the future data is supposed to be a normal mixture $H = \pi_1 H_1 + \pi_2 H_2$, with $H_1 = N_p(\mu_1, \Sigma_1)$ and $H_2 = N_p(\mu_2, \Sigma_2)$. The probability of classifying observations from the first group in the second is given by

$$\Pi_{2|1}(H^0, H) = P(Q(X; H^0) < 0 \mid X \sim H_1), \tag{3.11}$$

and the probability of misclassification for observations following H_2 is

$$\Pi_{1|2}(H^0, H) = P(Q(X; H^0) > 0 \mid X \sim H_2).$$

The total probability of misclassification, or the error rate for classifying observations from H using a discriminant rule Q estimated from H^0, is then defined as

$$\mathrm{TPM}(H^0, H) = \pi_1 \Pi_{2|1}(H^0, H) + \pi_2 \Pi_{1|2}(H^0, H). \tag{3.12}$$

If we want to emphasise that we work with the classical discriminant rule Q_{Cl}, we will use the notation TPM_{Cl}. It is important to distinguish between H^0 and H. In the above definitions, no parametric assumptions are made on the

distribution generating the training data. The quadratic discriminant rule can be applied to any data set, although it might be expected that the rule performs poor if the data are far from normally distributed. For example, they might contain a few outliers. However, to compute a misclassification rate for future data, a parametric assumption is needed to obtain computable expressions. The normality assumption on H is taken here and the results obtained in this chapter all make use of this assumption. The next proposition gives an expression for the TPM.

Proposition 3.1. *With the notations above, for* $H = \pi_1 N_p(\mu_1, \Sigma_1) + \pi_2 N_p(\mu_2, \Sigma_2)$, *and for the quadratic discriminant rule* $Q(X; H^0)$ *defined in* (3.7),

$$\Pi_{2|1}(H^0, H) = P\left(\sum_{j=1}^{p} \lambda_j (W_j - d_{2|1}^t v_j)^2 < k\right), \tag{3.13}$$

where $W_1, \ldots, W_p$ *are i.i.d. univariate standard normal. Furthermore,* $d_{2|1}$ *is a p-variate vector given by*

$$d_{2|1} = d_{2|1}(H^0, H) = \Sigma_1^{-1/2}\left(-\frac{1}{2}A(H^0)^{-1}b(H^0) - \mu_1\right), \tag{3.14}$$

$$k = k(H^0) = \frac{1}{4}b(H^0)^t A(H^0)^{-1}b(H^0) - c(H^0), \tag{3.15}$$

and $\lambda_j = \lambda_j(H^0, H)$ *and* $v_j = v_j(H^0, H)$ *are the eigenvalues and eigenvectors of the matrix*

$$\bar{A}_{2|1}(H^0, H) = \Sigma_1^{1/2} A(H^0) \Sigma_1^{1/2}. \tag{3.16}$$

The expression for $\Pi_{1|2}(H^0, H)$ *is given by*

$$\Pi_{1|2}(H^0, H) = P\left(\sum_{j=1}^{p} \lambda_j (W_j - d_{1|2}^t v_j)^2 > k\right), \tag{3.17}$$

with λ_j *and* v_j *now the eigenvalues and eigenvectors of* $\bar{A}_{1|2}(H^0, H)$. *Here,* $d_{1|2}(H^0, H)$ *and* $\bar{A}_{1|2}(H^0, H)$ *are given by replacing the index 1 by 2 in the definitions of* $d_{2|1}(H^0, H)$ *and* $\bar{A}_{2|1}(H^0, H)$. *The total probability of misclassification is then* $\mathrm{TPM}(H^0, H) = \pi_1 \Pi_{2|1}(H^0, H) + \pi_2 \Pi_{1|2}(H^0, H)$.

When performing a discriminant analysis, one expects that the data to be classified come from the same distribution as the training data, although the proportions of data coming from the first or second group may be different. In this case, where $H^0 = (H_1^0, H_2^0) = (H_1, H_2)$, we say that we the training data follow the model distribution (and in particular contain no outliers). So at the model, the training data follow a normal distribution as well and $T_1(H^0) = \mu_1$, $T_2(H^0) = \mu_2$,

$C_1(H^0) = \Sigma_1$ and $C_2(H^0) = \Sigma_2$. (When we work with Q_R instead of Q_{Cl}, we require consistency of the robust location and covariance measures at the normal distribution.) Hence at the model, the total probability of misclassification is a function of the population parameters of location and covariance. Numerical computation of this TPM requires evaluation of the cumulative distribution function of a linear combination of p chi-squared distributions with one degree of freedom. Note that some of the weights λ_j in this linear combination appearing in (3.13) may be negative, since they are eigenvalues of the symmetric, but in general not positive definite matrix (3.16). Using modern computing power, (3.13) can equally easy be computed with Monte-Carlo integration techniques. Indeed, for a sufficiently high number of vectors $(W_1, \ldots, W_p)$ generated from a multivariate standard normal distribution, we check for every simulated vector whether the inequality in (3.13) holds for the given value of k. The probability in (3.13) is then being approximated as the corresponding empirical frequency.

For diagonal covariance matrices and $H^0 = H$, an expression of the TPM for QDA was presented by Houshmand (1993). Recently, McFarland and Richards (2002) considered the problem of computing exact misclassification probabilities in the normal case for finite samples. The expression for TPM in the setting of Linear Discriminant Analysis is much better known. In the normality case with equal covariances it is simply given $\mathrm{TPM}_{\mathrm{LDA}} = \Phi(\frac{-\Delta}{2})$, with $\Delta = \sqrt{(\mu_1 - \mu_2)^t \Sigma^{-1} (\mu_1 - \mu_2)}$ the Mahalanobis distance between the populations and Φ the c.d.f. of a standard normal. To study the effect of outliers on the total probability of misclassification, partial influence functions will be computed in the next section.

3.3 Partial influence functions

Influence functions have already been used for estimators that depend on more than one sample (e.g. Campbell, 1978; Fung, 1992, 1996b). We compute the influence of observations in the training sample on the TPM by using the formalism of partial influence functions (Pires and Branco, 2002). Partial influence functions (PIF) extend the traditional concept of influence functions to the multi-sample setting. The first PIF gives the influence on the classification error of an observation x being allocated to the first group of training data. The second PIF measures the influence on the TPM for training data being allocated to the second group. Formally,

$$\mathrm{PIF}_1(x; \mathrm{TPM}, H^0, H) = \lim_{\varepsilon \downarrow 0} \frac{\mathrm{TPM}\left((1-\varepsilon)H_1^0 + \varepsilon \Delta_x, H_2^0), H\right) - \mathrm{TPM}(H^0, H)}{\varepsilon}, \tag{3.18}$$

$$\mathrm{PIF}_2(x; \mathrm{TPM}, H^0, H) = \lim_{\varepsilon \downarrow 0} \frac{\mathrm{TPM}\left((H_1^0, (1-\varepsilon)H_2^0 + \varepsilon\Delta_x), H\right) - \mathrm{TPM}(H^0, H)}{\varepsilon}, \tag{3.19}$$

where Δ_x is a Dirac measure putting all its mass at x. One sees that for the first PIF contamination is only induced for H_1^0, the distribution generating the first group of training data, while the second distribution H_2^0 remains unaltered. Only contamination in the training sample is considered, the distribution H of the data to classify is not subject to contamination. When actually computing influence functions, we work at the model distribution $H^0 = (H_1, H_2)$. Indeed, when no contamination is present, one supposes that the data generating processes for the training data and for future data are the same. This model condition is natural and implicitly made in the classification literature. At the model, the notation $\mathrm{PIF}_s(x; \mathrm{TPM}, H) := \mathrm{PIF}_s(x; \mathrm{TPM}, (H_1, H_2), H)$, for $s = 1, 2$, can be used. For classical quadratic discriminant analysis the partial influence functions are written as $\mathrm{PIF}_s(x; \mathrm{TPM}_{Cl}, H^0, H)$, for $s = 1, 2$. When using robust plug-in estimates in the definition of Q, the notation $\mathrm{PIF}_s(x; \mathrm{TPM}_R, H^0, H)$ is used.

For linear discriminant analysis, the above influence functions have already been computed (e.g. Croux and Dehon, 2001). The result, when using standard population averages and covariances, is strikingly simple

$$\mathrm{PIF}_s(x; \mathrm{TPM}_{Cl}^{\mathrm{LDA}}, H^0, H) = (\pi_1 - \pi_2)\frac{\phi(\Delta/2)}{2\Delta}(L(x) - L(\mu_s)), \tag{3.20}$$

for $s = 1, 2$. Here ϕ is the density of a standard normal distribution and Δ as before the Mahalanobis distance between the 2 source populations. As Critchley and Vitiello (1991) noticed, the influence is determined by the factor $L(x) - L(\mu_s)$, which they consider as a residual. For QDA it seems very difficult to come up with an easily interpretable expression.

The next proposition shows how the partial influence functions of the TPM using the quadratic discriminant rule Q can be obtained.

Proposition 3.2. *Let H^0 be the distribution of the training data and $H = \pi_1 N_p(\mu_1, \Sigma_1) + \pi_2 N_p(\mu_2, \Sigma_2)$ the distribution of the data to classify. Suppose that*

(i) All eigenvalues of the matrix $\Sigma_1\Sigma_2^{-1}$ are distinct and different from one.

(ii) The partial influence function of the location functionals T_1 and T_2, and the scatter functionals C_1 and C_2 exist at H^0.

(iii) The model holds, i.e. $H^0 = (H_1, H_2)$.

The partial influence functions of the total probability of misclassification of a quadratic discriminant rule Q based on the location measures $T_1(H^0)$ and $T_2(H^0)$

and the scatter measures $C_1(H^0)$ and $C_2(H^0)$ is then given by

$$\mathrm{PIF}_s(x;\mathrm{TPM},H^0,H)=\pi_1\mathrm{PIF}_s(x;\Pi_{2|1},H^0,H)+\pi_2\mathrm{PIF}_s(x;\Pi_{1|2},H^0,H), \quad (3.21)$$

for $s=1,2$. Here

$$\begin{aligned}\mathrm{PIF}_s(x;\Pi_{2|1},H^0,H) &= \sum_{j=1}^{p}\frac{\partial\Pi_{2|1}(H^0,H)}{\partial\lambda_j}\cdot\mathrm{PIF}_s(x;\lambda_j,H^0,H)\\ &\quad+\sum_{j=1}^{p}\frac{\partial\Pi_{2|1}(H^0,H)}{\partial d_j^*}\cdot\mathrm{PIF}_s(x;d_j^*,H^0,H) \quad (3.22)\\ &\quad+\frac{\partial\Pi_{2|1}(H^0,H)}{\partial k}\cdot\mathrm{PIF}_s(x;k,H^0,H),\end{aligned}$$

where all the notations of Proposition 3.1 are used and for all $j=1,\ldots,p$ the notation $d_j^*(H^0,H)=v_j(H^0,H)^t d_{2|1}(H^0,H)$ is used. Furthermore

$$\mathrm{PIF}_s(x;\lambda_j,H^0,H) = v_j^t\Sigma_1^{1/2}\mathrm{PIF}_s(x;A,H^0)\Sigma_1^{1/2}v_j, \quad (3.23)$$

$$\begin{aligned}\mathrm{PIF}_s(x;d_j^*,H^0,H) &= \mathrm{PIF}_s(x;v_j,H^0,H)^t d_{2|1}(H^0,H) \quad (3.24)\\ &\quad+v_j^t\mathrm{PIF}_s(x;d_{2|1},H^0,H),\end{aligned}$$

$$\begin{aligned}\mathrm{PIF}_s(x;k,H^0) &= -b^tA^{-1}\mathrm{PIF}_s(x;A,H^0)A^{-1}b/4 \quad (3.25)\\ &\quad+b^tA^{-1}\mathrm{PIF}_s(x;b,H^0)/2-\mathrm{PIF}_s(x;c,H^0),\end{aligned}$$

while

$$\mathrm{PIF}_s(x;d_{2|1},H^0,H)=-\Sigma_1^{1/2}A^{-1}\left(\mathrm{PIF}_s(x;b,H^0)-\mathrm{PIF}_s(x;A,H^0)A^{-1}b\right)/2. \quad (3.26)$$

$$\mathrm{PIF}_s(x;v_j,H^0,H)=\sum_{k=1,k\neq j}^{p}\frac{v_k^t\Sigma_1^{1/2}\mathrm{PIF}_s(x;A,H^0)\Sigma_1^{1/2}v_j}{\lambda_j-\lambda_k}v_k, \quad (3.27)$$

for $j=1,\ldots,p$. The following shorthand notations $A=A(H^0)$, $b=b(H^0)$, $\lambda_j=\lambda_j(H^0,H)$ and $v_j=v_j(H^0,H)$ for $j=1,...,p$, are used. Furthermore,

$$\mathrm{PIF}_s(x;A,H^0) = (-1)^{s+1}\frac{1}{2}\left\{C_s^{-1}\mathrm{PIF}_s(x;C_s,H^0)C_s^{-1}\right\}, \quad (3.28)$$

$$\begin{aligned}\mathrm{PIF}_s(x;b,H^0) &= (-1)^{s+1}\{C_s^{-1}\mathrm{PIF}_s(x;T_s,H^0) \quad (3.29)\\ &\quad-C_s^{-1}\mathrm{PIF}_s(x;C_s,H^0)C_s^{-1}T_s\},\end{aligned}$$

$$\begin{aligned}\mathrm{PIF}_s(x;c,H^0) &= (-1)^{s+1}\frac{1}{2}\{T_s^tC_s^{-1}\mathrm{PIF}_s(x;C_s,H^0)C_s^{-1}T_s \quad (3.30)\\ &\quad-2T_s^tC_s^{-1}\mathrm{PIF}_s(x;T_s,H^0)-\mathrm{trace}(C_s^{-1}\mathrm{PIF}_s(x;C_s,H^0))\}.\end{aligned}$$

for $s=1,2$ using the short-hand notation T_s for $T_s(H^0)$ and C_s for $C_S(H^0)$. The partial derivatives $\frac{\partial\Pi_{2|1}(H^0,H)}{\partial\lambda_j}$, $\frac{\partial\Pi_{2|1}(H^0,H)}{\partial d_j^*}$ and $\frac{\partial\Pi_{2|1}(H^0,H)}{\partial k}$, for $j=1,...,p$,

do not depend on the argument x, neither on location and covariance functionals. Expressions for them are given in Lemma's 3.3, 3.4 and 3.5 in the Appendix. In order to compute $\mathrm{PIF}_s(x;\Pi_{1|2},H^0,H)$*, it suffices to replace* Σ_1 *by* Σ_2 *in the expressions* (3.23) *up to* (3.27) *and to interchange* $d_{2|1}$ *with* $d_{1|2}$*. The* λ_j *and* v_j *are then the eigenvectors and eigenvalues of the matrix* $\Sigma_2^{1/2}A(H^0)\Sigma_2^{1/2}$ *instead of the matrix* $\Sigma_1^{1/2}A(H^0)\Sigma_1^{1/2}$*.*

Computing the partial influence functions appearing in Proposition 3.2 is tedious, but straightforward. Building bricks are the expressions for the partial influence functions of the estimators of location and scatter. For the classical estimators it is immediate to check that

$$\begin{aligned}\mathrm{PIF}_s(x;C_s,H^0) &= (x-T_s(H^0))(x-T_s(H^0))^t - C_s(H^0)\\ \mathrm{PIF}_s(x;T_s,H^0) &= x - T_s(H^0),\end{aligned} \tag{3.31}$$

for $s=1,2$ while $\mathrm{PIF}_s(x;C_{s'},H^0) = \mathrm{PIF}_s(x;T_{s'},H^0) = 0$ for $s'\neq s$. From (3.31) all other auxiliary partial influence functions can be computed, resulting in $\mathrm{PIF}_1(x;\mathrm{TPM}_{Cl},H^0,H)$ and $\mathrm{PIF}_2(x;\mathrm{TPM}_{Cl},H^0,H)$.

Computation of the partial derivatives of $\Pi_{2|1}(H^0,H)$, appearing in (3.22), requires some care. These partial derivatives only depend on the population parameters, they do not depend on x, neither on the estimators used. Lemmas 3.3, 3.4, and 3.5 formulated in the Appendix express them in terms of integrals, which can be computed by numerical integration. Note that numerical integration is much more stable than numerical differentiation. Although the formulas for computing the PIF are cumbersome, there are no major computational difficulties. A Matlab program computing the partial influence functions is available from *www.econ.kuleuven.be/kristel.joossens*.

When deriving the expression for the PIF, the assumption " (i): All eigenvalues of the matrix $\Sigma_1\Sigma_2^{-1}$ are distinct and different from one" was needed. If the matrix $\Sigma_1\Sigma_2^{-1}$, or equivalently $\Sigma_2\Sigma_1^{-1}$, has eigenvalues close to 1, or close to each other, then it can be seen from (3.27) and Lemmas 3.3 and 3.4 in the Appendix that the influence function will tend to explode. If one is close to a setting where condition (i) is not valid, then the discriminant rule is very sensitive to single observations in the training data. One case where (i) is not valid is the equal covariance matrix case, where all eigenvalues of $\Sigma_1\Sigma_2^{-1}$ are equal to one. Hence, for reasons of local robustness, it is advised to use LDA whenever one is close to the equal covariance matrix case. Performing a test for equal covariance matrices (e.g. Bartlett 1937) before carrying out a QDA, as is common in applied research, can prevent construction of an unstable quadratic discriminant rule. However, there are other situations where condition (i) is not met, for example when Σ_1 and Σ_2 are both proportional to the identity matrix. The latter corresponds with a setting of two spherically symmetric data clouds. Here, alternative methods like regularised Gaussian discriminant analysis (Bensmail and Celeux, 1996) are preferable to keep the local sensitivity under control.

The eigenvalues of $\Sigma_1\Sigma_2^{-1}$ determine the nature of the quadratic form (3.2). For example, in the bivariate setting the eigenvalues determine whether the classification regions associated with the two groups are an ellipse and its complement or an hyperbola and its complement. When an eigenvalue passes from below to above one, the nature of the classification regions changes. Finally, note that interchanging two eigenvalues close to each other leads to a change in orientation of the quadratic form, which explains why the equal eigenvalue case is unstable as well (similar as in principal components analysis, see Critchley 1985).

Some pictures of partial influence functions in the univariate and bivariate case are represented. Figure 3.1 gives the first PIF for $H_1 = N(0,1)$ and $H_2 = N(1,\sigma^2)$, for $\sigma^2 = 0.6, 0.8, 1.2$ and 1.6, and equal prior probabilities for discriminant analysis based on Q_{Cl}. It is immediate to see that the influence functions have a quadratic shape and are unbounded. When the value of σ^2 approaches 1, the values for the PIF increase. For $\sigma^2 = 1.2$ the shape of the PIF is reversed: outliers for the first training data set tend to decrease the estimated error rate.

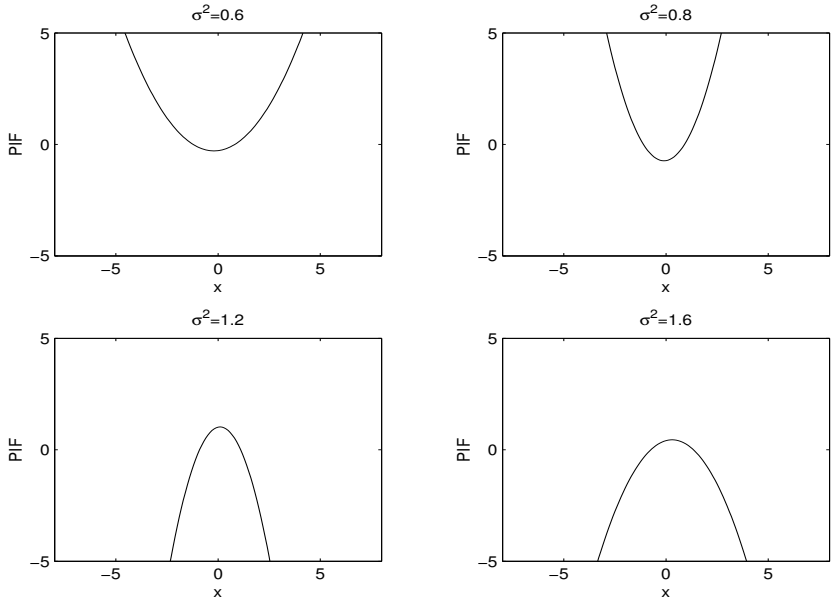

Figure 3.1: *First partial influence function* $\mathrm{PIF}_1(x; \mathrm{TPM}_{Cl}, H)$ *for* $H = 0.5N(0,1) + 0.5N(0,\sigma^2)$ *and for several values of* σ^2.

Of course, in practice one is interested in the higher dimensional case. The shape and sign of the PIF depend heavily on the parameter values and are difficult to predict, in contrast with the linear case. In Figure 3.2 the first partial influence function is shown for a bivariate distribution where $H_1 = N(0, I_2)$ and

$H_2 = N((1,1)^t, \text{diag}(0.3, 0.8))$. Notice again the quadratic shape of the influence surface, being quite flat in the central region here, but unbounded in the tails of the distribution.

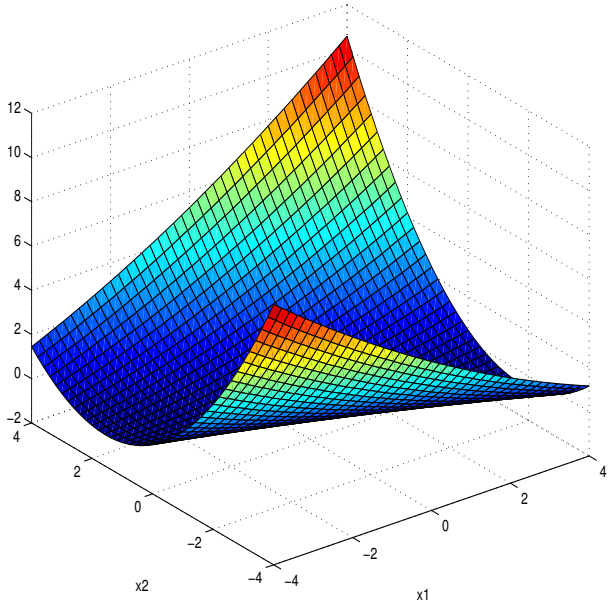

Figure 3.2: *First partial influence function* $\text{PIF}_1(x; \text{TPM}_{Cl}, H)$ *for* $H = 0.5N(0, I_2) + 0.5N((1,1)^t, \text{diag}(0.3, 0.8))$.

The expressions in Proposition 3.2 are not only valid for TPM_{Cl}, but they also apply when robust estimators are used for the parameters μ_1, μ_2, Σ_1 and Σ_2 in the discriminant rule Q. For example, Randles et al. (1978) proposed to use M-estimators. Since M-estimators loose robustness when the dimension p increases, we will use the highly robust Minimum Covariance Determinant (MCD) estimator (Rousseeuw and Van Driessen, 1999). The MCD-estimator is obtained by selecting the subsample of size h (we selected $h = 0.75n$) for which the determinant of the covariance matrix computed from that subsample is minimal, and computing afterwards the mean and the sample covariance matrix solely from this "optimal" subsample. The robustness of the MCD-estimator in the context of QDA has recently been shown by means of simulation studies (Joossens and Croux 2004; Hubert and Van Driessen, 2004). Now, using the results of Proposition 3.2, we are able to prove local robustness by means of partial influence functions. It is indeed immediate to see that $\text{PIF}_s(x; \text{TPM}, H^0, H)$ is bounded as soon as $\text{PIF}_s(x; \mu_s, H^0)$ and $\text{PIF}_s(x; \Sigma_s, H^0)$ are bounded. Influence functions for the MCD-estimator where computed by Butler, Davies and Jhun (1993) and Croux and Haesbroeck (1999) and were shown to be bounded at elliptical models.

Figure 3.3 shows the PIF for the same distributions as for Figure 3.1, but now using the robust MCD estimator to estimate the discriminant rule. The same scaling of the axes as in Figure 3.1 is used, and it is immediately observed how much lower the values for the PIF become. In the central part of the data, the PIF behaves like the PIF of the classical estimation procedure, but in the tails we observe a bounded influence. Hence far outliers receive a bounded, but non-zero, influence. Notice that for σ^2 close to 1, where condition C is not valid, the influence function also gets blown up, but to a much lesser degree. For σ^2 equal to one, the PIF will not exist either.

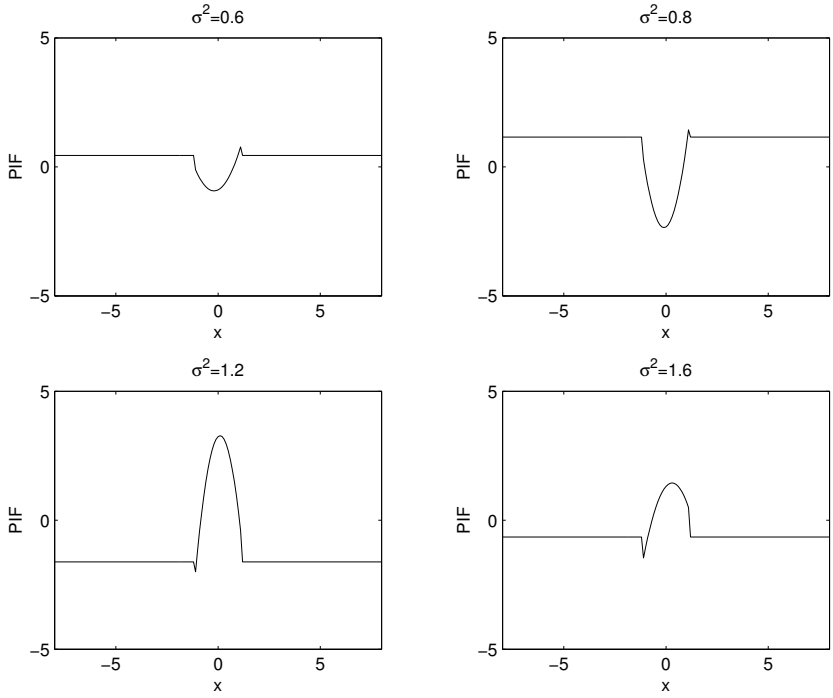

Figure 3.3: *First partial influence function* $\mathrm{PIF}_1(x; \mathrm{TPM}_R, H)$. *As in Figure 3.1, but now using the robust MCD-estimator for estimating the parameters in the discriminant rule* Q.

3.4 Robust diagnostic measures and examples

The heuristic interpretation of (partial) influence functions is that the estimated difference between the population TPM and its estimated value is approximatively given by the average of the values $\mathrm{PIF}(x_i; \mathrm{TPM}, H)$ for $i = 1, \ldots, n$ (cfr. Hampel et al., 1986; Pires and Branco, 2002). Hence the partial influence functions

evaluated at the sample points give the contribution of every observation in the training set to the misclassification rate. Large values for the PIF reveal points giving a large positive contribution to the TPM. We restrict ourselves to the detection of influential points for classical discriminant analysis. When a robust discriminant rule Q_R is used, it is less important to pinpoint the highly influential points, since the robust procedure has a bounded influence and is resistant to these observations.

Diagnostic measures are then computed using the first, respectively second, PIF for observations belonging to the first, respectively second, group of training data:

$$\begin{aligned} \mathrm{D}_{i,Cl}(\mu_1, \mu_2, \Sigma_1, \Sigma_2) &= |\mathrm{PIF}_1(x_i, \mathrm{TPM}_{Cl}, H)| \quad \text{for } i = 1, \ldots, n_1 \\ \mathrm{D}_{i,Cl}(\mu_1, \mu_2, \Sigma_1, \Sigma_2) &= |\mathrm{PIF}_2(x_i, \mathrm{TPM}_{Cl}, H)| \quad \text{for } i = n_1 + 1, \ldots, n. \end{aligned} \tag{3.32}$$

Plotting D_i with respect to the index i, or alternatively with respect to the value of $Q(x_i)$, then results in a diagnostic plot. The sign information in the PIF could be kept by dropping the absolute values in (3.32). To compute the diagnostics D_i, the parameters μ_1, μ_2, Σ_1 and Σ_2 need to be estimated. The prior probability π_1 can be estimated as the frequency of observation from the training sample belonging to the first group, and similarly for π_2.

The idea of using the influence function as a tool for sensitivity analysis has a long tradition in statistics. For applications in multivariate analysis see for example Critchley (1985), and Tanaka (1994). In the construction of the D_i the non-robust sample average and covariance matrix estimators could be used for estimating the population parameters. Though it is well-known that diagnostic measures based on non-robust estimators are subject to the masking effect. Outliers and atypical observations might shift the estimated means and blow up the dispersion matrices, resulting in a non reliable diagnostic measure. It might as well be possible that influential observations will not be detected anymore. To prevent this masking effect, it is proposed to estimate μ_1, μ_2, Σ_1 and Σ_2 using robust estimators, resulting in a robust diagnostic measure. A similar approach to robust diagnostics was taken by (Tanaka and Tarumi, 1996; Pison et al., 2003; and Boente et al., 2002) in different fields of multivariate statistics. In the construction of the robust diagnostic tool, the robust estimators are auxiliary and only serve to estimate the $\mathrm{D}_{i,Cl}(\mu_1, \mu_2, \Sigma_1, \Sigma_2)$ in a reliable way, not suffering from the masking effect. As such, the partial influence function of the non robust classical estimator is estimated in a robust way. The aim is to detect influential points when using Q_{Cl}. When no highly influential points are detected by the robust diagnostic, one could pass to a standard discriminant analysis, the latter one being more efficient at the normal model.

To illustrate the risk of masking when using non-robust diagnostics, consider the *Skull's data*, described in Flury and Riedwyl (1988, page 123-125). This well-known data set contains skull measurements (6 variables) on two species of

female voles: Microtus Californicus and Microtus Ochrogaster. The first group contains 41 observations, and the second 45. In Figure 3.4 diagnostic plots are made, once using the classical estimators, and once using robust plug-in estimators for $\mathrm{D}_{i,Cl}(\mu_1, \mu_2, \Sigma_1, \Sigma_2)$. The robust diagnostic measures, immediately reveal that there is a huge influential observation: number 73. The non-robust diagnostic measures suffer from the masking effect and cannot detect any influential observations anymore.

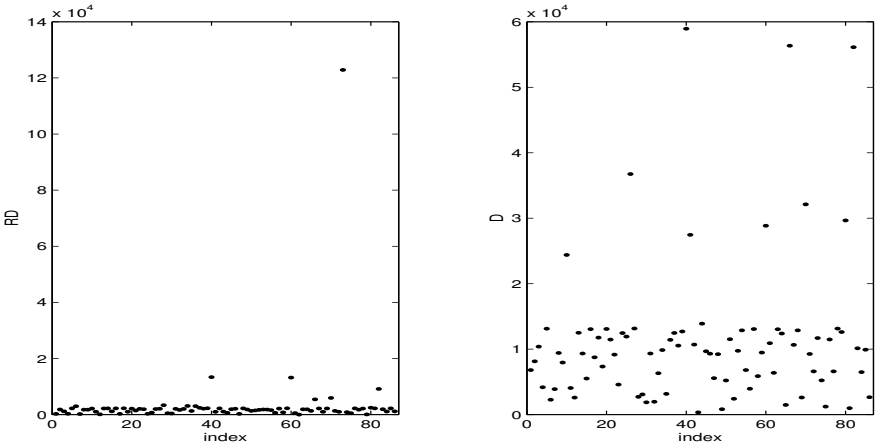

Figure 3.4: *Diagnostic plot for the Skull data using robust plug-in estimators (left figure) or using classical plug-in estimators (right figure) for* $D_{i,Cl}(\mu_1, \mu_2, \Sigma_1, \Sigma_2)$.

Several diagnostic measures for classical quadratic discriminant analysis have already been introduced by Fung (1996a). Influence is measured by looking at the effect of deleting an observation from the sample on the estimated probabilities of all other observations. Fung (1996a) proposed different variants, all based on the leave-one-out principle. One of them is the Relative Log-Odds SQuared influence for an observation i,

$$\mathrm{RLOSQ}_i = \frac{1}{n}\sum_{j=1}^{n}\left[\log\left\{\frac{\hat{p}_1(x_j)}{1-\hat{p}_1(x_j)}\right\} - \log\left\{\frac{\hat{p}_{1(i)}(x_j)}{1-\hat{p}_{1(i)}(x_j)}\right\}\right]^2,$$

where $\hat{p}_1(x)$ is the estimated probability that an observation x belongs to the first group,

$$\hat{p}_1(x) = \hat{f}_1(x)/[\hat{f}_1(x) + \hat{f}_2(x)],$$

with $\hat{f}_j$ the density of $N_p(\hat{\mu}_j, \hat{\Sigma}_j)$, for $j = 1, 2$. On the other hand, $\hat{p}_{1(i)}(x)$ estimates the same probability, but now using the sample where observation i has been removed.

Consider as a second example the *Biting flies* data, described in Johnson and Wichern (2002, page 373). Two species of flies, *Leptoconops cartei* and *Leptoconops torrens*, were thought for many years to be the same, because they are morphologically very similar. For each group a sample of 35 observations was drawn and seven measurements where taken. Figure 3.5 shows the comparison between the RLOSQ-diagnostic and the robust diagnostic based on the partial influence functions for the TPM_{Cl}. The robust diagnostic indicates only 36 as highly influential. The leave-one-out method suggests as well 2, 15 and 23. Further inspection of the data reveals that 2, 15 and 23 are outlying observations. Hence there is a risk that due to the presence of multiple outliers, the leave-one-out procedure becomes unreliable. Whether 2,3, and 15 are highly influential, or only outlying, is difficult to find out using the RLOSQ indices.

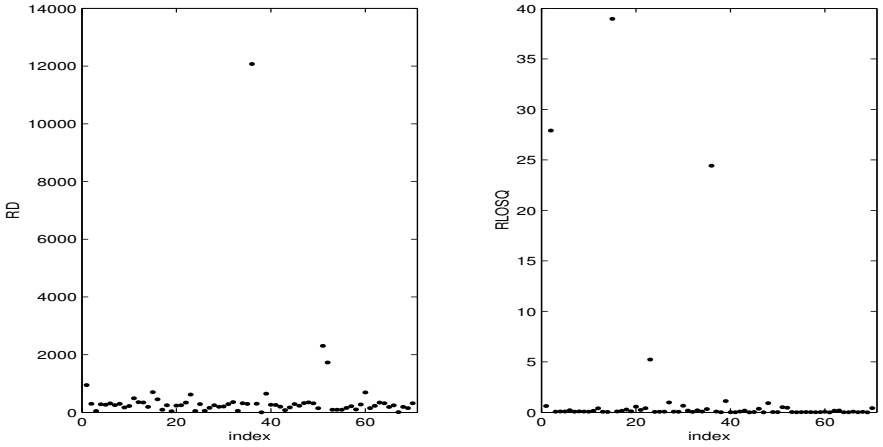

Figure 3.5: *Diagnostic plot for the Biting Flies data using robust diagnostics based on* TPM_{Cl} *(left figure) and using the leave-one out measure RLOSQ (right figure).*

3.5 Conclusions

This chapter is about computing the influence of observations in the training sample on the classification error of a discriminant rule. For linear discriminant analysis, answers have been given more than a decade ago, but quadratic discriminant analysis is a harder problem to tackle. Starting from an expression for the total probability of misclassification, in Section 2, and using the technology of Partial Influence Functions of Pires and Branco (2002), in Section 3, a computable expression for the partial influence function of the total probability of misclassification was found.

Not surprisingly, this influence function was showed to be quadratic and unbounded. Using robust plug-in estimators in the discriminant rule Q, however, yields bounded influence procedures. But it also turned out that whenever the matrix $\Sigma_1\Sigma_2^{-1}$ has eigenvalues close to each other or close to one, the QDA is unduly sensitive to small data perturbations. Focus was on the influence on the TPM, and not on the influence on the estimates of the parameters of the quadratic discriminant rule. The latter estimates are not of immediate interest in QDA. In some sense, one could think of $\mathrm{PIF}(x; \mathrm{TPM}, H)$ as an appropriate summary of the influences on the estimates of the $p(p+3)$ components of μ_1, μ_2, Σ_1 and Σ_2. Besides of theoretical interest, the PIF can also be used to construct a robust diagnostic tool for the detection of influential points in classical QDA.

Influence diagnostics in discriminant analysis for LDA, QDA, and for the multiple group case were proposed and studied in a sequence of papers by Fung (1995a, 1995b, 1996a, 1996b). In this chapter, a theoretical expression of an influence function is used as basis of the diagnostic measure being proposed, allowing to avoid case-wise deletion measures. A completely different approach is taken by Riani and Atkinson (2001), who proposed a forward search algorithm to avoid masking effects in detecting influential points. Their approach is a useful data-analytic tool for a robust sensitivity analysis of a discriminant analysis, and requires user-interactive analysis of the data.

Let us emphasise that we do not aim to develop a new kind of robust discriminant analysis. This chapter quantifies the influence of observations on the estimated error rate using plug-in estimates for the parameters of the quadratic discriminant rule. Robust high breakdown linear and quadratic discriminant analysis has been discussed in several papers, such as Hawkins and McLachlan (1997), He and Fung (2000), Croux and Dehon (2001), Joossens and Croux (2004) and Hubert and Van Driessen (2004). But most of them focus on computational aspects and simulation comparison. Programs for computing robust linear and quadratic discriminant analysis can be retrieved from *www.econ.kuleuven.be/kristel.joossens*.

Appendix

Proof of Proposition 3.1:

It is sufficient to prove (3.13). The quadratic discriminant function (3.7) can be rewritten as written as

$$Q(x; H^0) = (x - \tilde{d}(H^0))^t A(H^0)(x - \tilde{d}(H^0)) - k(H^0), \tag{3.33}$$

with $k = k(H^0)$ defined in (3.15), and $\tilde{d}(H^0) = -A(H^0)^{-1}b(H^0)/2$. Take now $X \sim H_1$, then $W = \Sigma_1^{-1/2}(X - \mu_1) \sim N(0, I_p)$, and definition (3.11) yields

$$\begin{aligned} \Pi_{2|1}(H^0, H) &= P_{H_1}((X - \tilde{d}(H^0))^t A(H^0)(X - \tilde{d}(H^0)) < k) \\ &= P_{N(0,I_p)}((W - d_{2|1})^t \bar{A}_{2|1}(H^0, H)(W - d_{2|1}) < k), \end{aligned}$$

where $d_{2|1} = d_{2|1}(H^0, H)$ is defined in (3.14). Since $\bar{A}_{2|1}(H^0, H)$ is a symmetric matrix, its eigenvalues λ_j are real and we can write

$$\bar{A}_{2|1}(H^0, H) = \sum_{j=1}^{p} \lambda_j v_j v_j^t,$$

where v_j are the corresponding eigenvectors. Moreover, the eigenvectors of $\bar{A}_{2|1}(H^0, H)$ are orthogonal implying that the variables $W_j = W^t v_j$, for $j = 1, \ldots, p$, are components of a multivariate standard normal distribution.

Proof of Proposition 3.2:

Equation (3.21) follows from the definition of TPM, and (3.22) results from a standard application of the chain rule. As a first step, the PIF for the estimates of the parameters of the quadratic discriminant rule Q are computed. The matrix derivation rules $\mathrm{PIF}_s(x; C_s^{-1}, H^0) = -C_s^{-1}(H^0)\mathrm{PIF}_s(x; \Sigma_s, H^0)C_s^{-1}(H^0)$ and $\mathrm{PIF}_s(x; \log|C_s|, H^0) = \mathrm{trace}\big(C_s^{-1}(H^0)\mathrm{PIF}_s(x; C_s, H^0)\big)$ for $s = 1, 2$ are used, cfr. Magnus and Neudecker (1999). Straightforward derivation from definitions (3.8), (3.9), (3.10) yields, then (3.28), (3.29), (3.30).

Since the functional k is a simple combination of the functionals A, b and c, equation (3.25) follows. Lemma 2.1 in Sibson (1979) or Lemma 3 in Croux and Haesbroeck (2000) give influence functions for the eigenvalues and eigenvectors of a symmetric matrix. Applying this result to $\bar{A}_{2|1}(H^0, H) = \Sigma_1^{1/2} A(H^0) \Sigma_1^{1/2}$ results in expressions (3.23) and (3.27). Note that by conditions (i) and (iii), and the fact $\Sigma_1^{1/2}\Sigma_2^{-1}\Sigma_1^{1/2} - I_p$ and $\Sigma_1\Sigma_2^{-1} - I_p$ have the same eigenvalues, division by zero in (3.27) is avoided. From (3.14), equation (3.26) follows and by the definition of d_j^*, equation (3.24) holds for $j = 1, \ldots, p$. Of course, similar arguments hold for deriving $\mathrm{PIF}_s(x; \Pi_{1|2}, H^0, H)$.

Computation of the partial derivatives of $\Pi_{2|1}(H^0, H)$ *w.r.t.* λ_j, d_j^* *and* k*:*

According to Proposition 3.1 and with $d_j^* = v_j^t d_{2|1}$, write

$$\Pi_{2|1}(H^0, H) = P\Big(\sum_{j=1}^{p} \mathrm{sign}(\lambda_j) X_j^2 < k\Big) \quad \text{where} \quad X_j \sim N_p(-d_j^* \sqrt{|\lambda_j|}, |\lambda_j|) \tag{3.34}$$

where the X_j are independent univariate normal variables, each having density

$$f_{X_j}(x_j) = \frac{1}{\sqrt{|\lambda_j|}} \varphi\Big(\frac{x_j}{\sqrt{|\lambda_j|}} + d_j^*\Big). \tag{3.35}$$

Now (3.34) can be written as the integral

$$\int f_{X_1}(x_1) \ldots f_{X_p}(x_p) I\Big(\sum_{j=1}^{n} \mathrm{sign}(\lambda_j) x_j^2 < k\Big) dx_1 \ldots dx_p.$$

By condition (iii) the eigenvalues λ_j of $\bar{A}_{2|1}$ are the same as those of $\Sigma_1\Sigma_2^{-1} - 1$ and by condition (i) they are different from zero.

Using the above notations, we get the following three lemmas.

Lemma 3.3. *The partial derivatives of* $\Pi_{2|1}(H^0, H)$ *with respect to* λ_j *are given by*

$$\frac{1}{2\lambda_j}\left\{-P(\Sigma_i\text{sign}(\lambda_i)X_i^2 < k) + E\Big[\frac{X_j(X_j + d_j^*\sqrt{|\lambda_j|})}{|\lambda_j|}I(\Sigma_i\text{sign}(\lambda_i)X_i^2 < k)\Big]\right\},$$

for $j = 1, \ldots, p$.

Proof. For each $1 \le j \le p$, it holds that $\frac{\partial}{\partial\lambda_j}\Pi_{2|1}(H^0, H)$ equals ($\diamond$ stands for $\varphi'(u) = -u\varphi(u)$)

$$\begin{aligned}
&\int \frac{\partial}{\partial\lambda_j}f_{X_j}(x_j)\prod_{m=1,m\neq j}^{p} f_{X_i}(x_i)I\Big(\sum_{i=1}^{p}\text{sign}(\lambda_i)x_i^2 < k\Big)\,dx_1\ldots dx_p \\
&= \int \text{sign}(\lambda_j)\frac{\partial}{\partial|\lambda_j|}f_{X_j}(x_j)\prod_{m=1,m\neq j}^{p} f_{X_m}(x_m)I\Big(\sum_{i=1}^{p}\text{sign}(\lambda_i)x_i^2 < k\Big)dx_1\ldots dx_p \\
&\overset{(3.35)}{=} \int \text{sign}(\lambda_j)\Big[-\frac{1}{2|\lambda_j|^{3/2}}\varphi\Big(\frac{x_j}{\sqrt{|\lambda_j|}} + d_j^*\Big) + \Big(\frac{x_j}{-2|\lambda_j|^2}\Big)\varphi'\Big(\frac{x_j}{\sqrt{|\lambda_j|}} + d_j^*\Big)\Big] \\
&\qquad \prod_{m=1,m\neq j}^{p} f_{X_m}(x_m)I\Big(\sum_{i=1}^{p}\text{sign}(\lambda_i)x_i^2 < k\Big)\,dx_1\ldots dx_p \\
&\overset{\diamond}{=} \int \text{sign}(\lambda_j)\frac{1}{2|\lambda_j|}\Big[-1 + \frac{x_j(x_j + d_j^*\sqrt{|\lambda_j|})}{|\lambda_j|}\Big] \\
&\qquad \prod_{m=1}^{p} f_{X_m}(x_m)I\Big(\sum_{i=1}^{p}\text{sign}(\lambda_i)x_i^2 < k\Big)\,dx_1\ldots dx_p,
\end{aligned}$$

from which the lemma follows directly. □

Lemma 3.4. *The partial derivatives of* $\Pi_{2|1}(H^0, H)$ *with respect to* d_j^* *are given by* ($\diamond$ *stands for* $\varphi'(u) = -u\varphi(u)$)

$$\frac{-1}{\sqrt{|\lambda_j|}}E[X_jI(\Sigma_i\text{sign}(\lambda_i)X_i^2 < k)] - d_j^*P(\Sigma_i\text{sign}(\lambda_i)X_i^2 < k),$$

for $j = 1, \ldots, p$.

Proof. For each $1 \le j \le p$, it holds that $\frac{\partial}{\partial d_j^*}\Pi_{2|1}(H^0, H)$ equals

$$\int \frac{\partial}{\partial d_j^*}f_{X_j}(x_j)\prod_{m=1,m\neq j}^{p} f_{X_m}(x_m)I\Big(\sum_{i=1}^{p}\text{sign}(\lambda_i)x_i^2 < k\Big)\,dx_1\ldots dx_p$$

$$\overset{(3.35)}{=} \int \frac{1}{\sqrt{|\lambda_j|}}\varphi'\Big(\frac{x_j}{\sqrt{|\lambda_j|}} + d_j^*\Big) \prod_{\substack{m=1\\ m\neq j}}^{p} f_{X_m}(x_m) I\Big(\sum_{i=1}^{p} \operatorname{sign}(\lambda_i) x_i^2 < k\Big) dx_1 \dots dx_p$$

$$\overset{\diamond}{=} \int \Big(-\frac{x_j + d_j^*\sqrt{|\lambda_j|}}{|\lambda_j|}\Big)\varphi\Big(\frac{x_j}{\sqrt{|\lambda_j|}} + d_j^*\Big)$$
$$\prod_{m=1, m\neq j}^{p} f_{X_m}(x_m) I\Big(\sum_{i=1}^{p} \operatorname{sign}(\lambda_i) x_i^2 < k\Big)\, dx_1 \dots dx_p$$
$$= \int \Big(-\frac{x_j}{\sqrt{|\lambda_j|}} - d_j^*\Big) \prod_{m=1}^{p} f_{X_m}(x_m) I\Big(\sum_{i=1}^{p} \operatorname{sign}(\lambda_i) x_i^2 < k\Big)\, dx_1 \dots dx_p,$$

from which the lemma follows directly. □

For the partial derivative with respect to k, we will reorder the components of X such that the corresponding eigenvalues satisfy

$$\lambda_{(1)} \geq \dots \geq \lambda_{(q)} > 0 > \lambda_{(q+1)} \geq \dots \geq \lambda_{(p)},$$

where q is the number of positive eigenvalues. Furthermore, let

$$S^+ = \sum_{j=1}^{q} X_{(j)}^2 \text{ and } S^- = \sum_{j=q+1}^{p} X_{(j)}^2$$

where empty sums are zero by convention. From (3.34) we have that $\Pi_{2|1}(H^0, H) = P(S^+ - S^- < k)$. Without loss of generality we will suppose that $k > 0$. For $k < 0$ one has

$$\frac{\partial \Pi_{2|1}(H^0, H)}{\partial k} = -\frac{\partial P(S^- - S^+ > |k|)}{\partial |k|} = \frac{\partial P(S^- - S^+ \leq |k|)}{\partial |k|}$$

and it suffices to interchange the roles of S^+ and S^- in the lemma below.

Lemma 3.5. *With this notations above, and for $k > 0$, the partial derivative of Π_{12} with respect to k is given by*

$$\begin{array}{lll} 0 & \text{if} & q = 0 \\ E\left[\left\{f_{X_{(1)}}(\sqrt{k+S^-}) + f_{X_{(1)}}(-\sqrt{k+S^-})\right\}/(2\sqrt{k+S^-})\right] & \text{if} & q = 1 \\ E\left[\pi^{q-1}(k+S^-)^{\frac{q-2}{2}} f_q(U\sqrt{k+S^-})\delta(\theta(U))\right] & \text{if} & q \geq 2 \end{array}$$

where f_q is joint density of $(X_{(1)}, \dots, X_{(q)})^t$ in polar coordinates, U is uniformly distributed on the periphery of the q dimensional unit sphere S^{q-1}, independently of S^-. Here $\delta(\theta(u)) = \sin^{q-2}\theta_1 \sin^{q-3}\theta_2 \dots \sin\theta_{q-2}$ for $q \geq 2$, with $\theta(u) = (\theta_1, \dots, \theta_q)$ the angles determining u.

Proof. The results is clear for $q = 0$ since it was supposed that $k > 0$. Now if $q = 1$ then

$$\begin{aligned}
\frac{\partial \Pi_{2|1}(H^0, H)}{\partial k} &= E\left[\frac{\partial}{\partial k} P(X^2_{(1)} \leq k + S^- | S^-)\right] \\
&= E\left[\frac{\partial}{\partial k} \int_0^{k+S^-} f_{X^2_{(1)}}(u) du\right] \\
&= E\left[f_{X^2_{(1)}}(k + S^-)\right] \\
&= E\left[\left\{f_{X_{(1)}}(\sqrt{k + S^-}) + f_{X_{(1)}}(-\sqrt{k + S^-})\right\} / (2\sqrt{k + S^-})\right].
\end{aligned}$$

For $q \geq 2$, a transformation $f_q(x_{(1)}, \ldots, x_{(q)}) := f_q(x^q) \to f_q(r, \theta)$ to polar coordinates will be carried out, where $r = \|x^q\|$ and $\theta \equiv (\theta_1, \ldots, \theta_{q-1})$, with $\theta_1, \ldots, \theta_{q-2} \in [0, \pi[$, $\theta_{q-1} \in [0, 2\pi[$ contains the corresponding angles. Let Θ be the space where the angles vary in, and let $\theta(u)$ be the set of angles associated with a unit vector. Then $\delta(\theta) = \sin^{q-2}\theta_1 \sin^{q-3}\theta_2 \ldots \sin\theta_{q-2}$ is the absolute value of the determinant of the Jacobian of this transformation. For every positive k one has

$$\begin{aligned}
& \frac{\partial}{\partial k} P(S^+ \leq k) \\
= \quad & \frac{\partial}{\partial k} \int f_q(x^q) I(\|x_q\|^2 < k) dx_q \\
= \quad & \frac{\partial}{\partial k} \int_0^{\sqrt{k}} \int_\Theta f_q(r, \theta) r^{q-1} \delta(\theta)\ d\theta\ dr \\
\overset{\text{Fubini}}{=} \quad & \int_\Theta \frac{\partial}{\partial k} \int_0^{\sqrt{k}} f_q(r, \theta) r^{q-1} \delta(\theta)\ d\theta\ dr \\
\overset{\text{Leibnitz}}{=} \quad & \int_\Theta \frac{1}{2\sqrt{k}} k^{\frac{q-1}{2}} f_q(\sqrt{k}, \theta) \delta(\theta) d\theta \\
= \quad & \frac{k^{\frac{q-2}{2}}}{2} \int_\Theta f(\sqrt{k}, \theta) \delta(\theta) d\theta, \\
= \quad & \frac{k^{\frac{q-2}{2}}}{2} 2\pi^{q-1}\ E_U[f_q(\sqrt{k}, U)\delta(\theta(U))],
\end{aligned}$$

where U is uniformly distributed over the q-dimensional unit sphere S^{q-1}. Then

$$\begin{aligned}
\frac{\partial}{\partial k} \Pi_{2|1}(H^0, H) &= E\left[\frac{\partial}{\partial k} P(S^+ \leq k + S^- | S^-)\right] \\
&= E\left[\pi^{q-1} k_q^{\frac{q-2}{2}} f_q(U\sqrt{k + S^-}) \delta(\theta(U))\right].
\end{aligned}$$

□

Finally, let us return to the proof of Proposition 3.2. It is easy to verify that the partial derivatives of $\Pi_{1|2}(H^0, H)$ with respect to λ_j, d_j^* and k are given by similar expressions as in Lemmas 3.3, 3.4 and 3.5. In Lemmas 3.3 and 3.4 the inequalities need to inverted, while the sign of the formula of Lemma 3.5 needs to be changed. □

Chapter 4

Logistic discrimination using robust estimators

Co-Author: C. Croux and G. Haesbroeck

Summary Logistic regression is frequently used for classifying observations into two groups. Unfortunately there are often outlying observations in a data set, who might affect the estimated model and the associated classification error rate. In this chapter, the effect of observations in the training sample on the error rate is studied by computing influence functions. It turns out that the usual influence function vanishes, and that the use of second order influence functions is appropriate. It is shown that using robust estimators in logistic discrimination strongly reduces the effect of outliers on the classification error rate. Furthermore, the second order influence function can be used as diagnostic tool to pinpoint outlying observations.

4.1 Introduction

In discriminant analysis one wants to classify multivariate observations into two different populations, using the outcome of a discriminant rule. The rule is constructed from a *training sample*, being observations for which it it known to which population they belong. The classical linear discriminant rule of Fisher is well-known and treated in every textbook on multivariate analysis. Many applied researchers, however, give preference to logistic regression as a tool for allocating observations to one out of two populations. It is a flexible method that can deal with different types of variables. Discriminant analysis resulting from an estimated logistic regression model is called logistic discrimination. Over the last decade, several more sophisticated classification methods like support vector machines and random forests have been proposed (see Friedman et al 2001), but

logistic discrimination remains a benchmark method performing well in many applications.

In this chapter the robustness of logistic discriminant analysis is studied. Focus is on the effect of observations in the training sample on the error rate of the associated classification rule. Influence functions measuring this effect will be computed for the normal discrimination model, where logistic discrimination achieves (asymptotically) the optimal error rate. It is shown that the usual influence function vanishes, and *second order influence functions* need to be computed. It turns out that the influence of outlying observations on the error rate can go beyond all bounds when estimating the logistic model by Maximum Likelihood (ML), but remains bounded when using an appropriate robust estimator.

For linear and quadratic discriminant analysis influence functions of the error rate were computed by Croux and Dehon (2001) and Croux and Joossens (2005). However, since they worked with non-optimal classification rules, they did not need to use second order influence functions. Up to our best knowledge, this chapter is one of the rare examples where the use of second order influence functions is natural and appropriate.

The non-robustness of the maximum likelihood estimator for logistic regression is well studied. Its influence function was computed in Künsch et al (1989), and breakdown point considerations were made in Christmann (1996) and Croux et al (2002). Tools for detecting influential observations in logistic regression analysis have been proposed in the literature (e.g. Pregibon 1981; Cook and Weisberg 1982, Chapter 5; Johnson 1985), but these diagnostics measure the influence relative to parameter estimates and predicted probabilities, and not the influence on the error rate. Moreover, they are all based on the classical ML-estimators computed from the sample with one or two observations deleted. In presence of multiple outliers, such case-wise deletion diagnostics suffer from the *masking effect*, meaning that influential points are not guaranteed to be detected due to bias in the diagnostic measure. It is hence recommended to rely on robust estimators.

Several proposals for robust logistic regression estimators have been made (e.g. Pregibon 1982 , Künsch et al. 1989, Carroll and Pederson 1993, Victoria-Feser 2002, Bondell 2005). Cox and Ferry (1991) considered a more robust version of logistic discrimination by adapting the logistic regression model and estimating it by maximum likelihood. In this chapter we stick to the traditional logistic regression model, although the theoretical results are valid for any robust estimator possessing an influence function.

The chapter is organised as follows: Section 4.2 reviews the normal logistic discrimination model and provides definitions of some robust estimators for logistic regression. An expression for the error rate is derived. The use of second order influence functions is motivated in Section 4.3, where the influence functions are derived and graphical presentations are given. Simulation results and an ap-

plication are presented in Section 4.4. In particular, a robust diagnostic tool is proposed to detect influential points for the error rate. Finally, some conclusions are given in Section 4.5.

4.2 Logistic discrimination and error rate

4.2.1 The normal discrimination model

Theoretical results will be derived at the normal discrimination model (e.g. Efron 1975). Suppose there are two p-dimensional source populations, both normally distributed with different means but the same covariance matrix. The variable X can arise from one of these populations:

$$X \sim \begin{cases} H_1 = N_p(\mu_1, \Sigma) & \text{with probability} \quad \pi_1, \\ H_0 = N_p(\mu_0, \Sigma) & \text{with probability} \quad \pi_0, \end{cases} \tag{4.1}$$

where $\pi_0 + \pi_1 = 1$. Let the variable Y indicate the source population of the corresponding X, then

$$Y = \begin{cases} 1 & \text{with probability} \quad \pi_1, \\ 0 & \text{with probability} \quad \pi_0 = 1 - \pi_1, \end{cases} \tag{4.2}$$

and

$$X \mid Y = y \sim N_p(\mu_y, \Sigma). \tag{4.3}$$

The joint distribution of (X, Y) is from now on denoted by H_m. It easily follows now, using Bayes' rule, that

$$P_{H_m}(Y = 1 \mid X = x) = F(\alpha + x^t\beta), \tag{4.4}$$

where $F(u) = 1/(1 + \exp(-u))$ is the logit cumulative distribution function,

$$\beta = \Sigma^{-1}(\mu_1 - \mu_0) \qquad \text{and} \qquad \alpha = \log(\pi_1/\pi_0) - \beta^t(\mu_0 + \mu_1)/2. \tag{4.5}$$

The discriminant rule is then as follows: an observation x is assigned to population 1 if $\alpha + x^t\beta > 0$ and to population 0 otherwise.

Given a random sample $\{(y_1, x_1), (y_2, x_2), \ldots, (y_n, x_n)\}$ drawn from the model distribution H_m, one can estimate the discriminant rule via estimation of the unknown parameters α and β. In a logistic discrimination procedure, these parameters are directly estimated via the logit model (4.4). This is in contrast with linear discriminant analysis (Fisher's rule) where the parameters μ_1, μ_2 and Σ are estimated, from which an estimated discriminant rule is obtained via (4.5) (see also Sapra 1991). The advantage of logistic discrimination is that one only relies on the specification (4.4) of the conditional distribution $Y|X$, while the normality

assumption is not used. This makes logistic regression more "robust" with respect to model misspecification. On the other hand, if the normal discrimination model perfectly holds, then the linear method is more efficient since it uses the full maximum likelihood estimators of the joint distribution.

4.2.2 Logistic regression estimators

In this section we introduce the logistic regression estimators that are used in this chapter, in particular the estimator of Bianco and Yohai (BY, 1996) and a weighted maximum likelihood estimator. Let $\gamma = (\alpha, \beta^t)^t$ and $z_i = (1, x_i^t)^t$ for all $1 \leq i \leq n$. An estimator for γ computed from the sample $S_n = \{(y_1, x_1), \ldots, (y_n, x_n)\}$ is denoted by $\hat{\gamma}_n$. The maximum likelihood (ML) estimator $\hat{\gamma}_n^{\mathrm{ML}}$ is given by

$$\hat{\gamma}_n^{\mathrm{ML}} = \underset{\gamma}{\operatorname{argmax}} \log L(\gamma; S_n) = \underset{\gamma}{\operatorname{argmin}} \sum_{i=1}^{n} d(z_i^t \gamma; y_i), \tag{4.6}$$

where $\log L(\gamma; S_n)$ is the conditional log-likelihood function and $d(\cdot; y_i)$ is the deviance function $d(s; y_i) = -y_i \log F(s) - (1 - y_i) \log(1 - F(s))$. Definition (4.6) can be generalised to

$$\hat{\gamma}_n = \underset{\gamma}{\operatorname{argmin}} \sum_{i=1}^{n} \varphi(z_i^t \gamma; y_i), \tag{4.7}$$

where $\varphi(s, y_i)$ is a positive and almost everywhere differentiable function in s, with the property $\varphi(s; 0) = \varphi(-s; 1)$ for any s. Bianco and Yohai (1996) showed that by selecting an appropriate φ function, a consistent, asymptotically normal, and resistant estimation procedure is obtained. In this chapter we will work with the φ function proposed by Croux and Haesbroeck (2003), having the property that the corresponding estimator exists whenever the ML-estimator exists. These authors also provided a fast and stable algorithm for its computation and showed in a simulation study the good performance of this estimator with respect to other proposals.

To reduce the influence of outlying observations in the covariate space, weights can be added to control for leverage points (e.g. Carroll and Pederson 1993). The weighted version of the Bianco and Yohai estimator is then defined as

$$\hat{\gamma}_n = \underset{\gamma}{\operatorname{argmin}} \sum_{i=1}^{n} w_i \varphi(z_i^t \gamma; y_i),$$

where the weights depend on the *Robust Distance* of the observation x_i. This robust distance RD_i is equal to the Mahalanobis distance of x_i to the center of the data cloud in the covariate space, with the center and covariance-matrix robustly estimated. For the latter, S-estimators of multivariate location and covariance

(Davies 1987, Rousseeuw and Leroy 1987, p. 174) are used. The weights are generated as $w_i = W(\mathrm{RD}_i)$, with weight function

$$W(t) = I(t^2 \leq \chi^2_{p,0.975}),$$

and the resulting estimator is called the Weighted Bianco and Yohai (WBY) estimator. Similarly, by taking $\varphi_{\mathrm{ML}}(s, y) = d(s, y)$, the *Weighted Maximum Likelihood estimator* (WML) is obtained (see also Rousseeuw and Christmann 2003).

In the sequel of the chapter, the functional representation of the estimators $\hat{\gamma}_n = (\hat{\alpha}_n, \hat{\beta}_n^t)^t$ of the parameters of the logistic regression model is used. Let S_n be a sample from a distribution H, and denote H_n the associated empirical distribution function. The statistical functionals $A(H)$ and $B(H)$ corresponding to the intercept and slope estimators verify $\hat{\alpha}_n = A(H_n)$ and $\hat{\beta}_n = B(H_n)$. If the estimators are consistent at the distribution H, then $A(H)$ and $B(H)$ are the limit values of $\hat{\alpha}_n$ and $\hat{\beta}_n$. At the model distribution $H = H_m$, it holds that $A(H_m) = \alpha$ and $B(H_m) = \beta$ for all functionals corresponding to consistent estimators at the logistic regression model.

4.2.3 Error rate

The classification performance of the logistic discrimination procedure is quantified by its error rate. Denote by Π_{01} the probability that an observation of population 1 is misclassified (so classified as an observation coming from population 0) and Π_{10} the probability that an observation of population 0 is misclassified. The data to classify are supposed to come from the model distribution H_m. The data used to estimate the logistic discriminant rule, i.e. the *training data*, come from a distribution H. In ideal circumstances $H = H_m$, but it might be that the training data are contaminated and contain outliers. The error rate (ER) is defined as

$$\mathrm{ER}(H) = \pi_1 \Pi_{01}(H) + (1 - \pi_1)\Pi_{10}(H),$$

with $\pi_1 = P_{H_m}(Y = 1)$. Using the previously defined functionals A and B, the probability of misclassifying an observation of population 1 can be written as

$$\begin{aligned}
\Pi_{01}(H) &= P(X^t B(H) + A(H) < 0 \mid X \sim N(\mu_1, \Sigma)) \\
&= P(X^t B(H) < -A(H) \mid X \sim N(\mu_1, \Sigma)) \\
&= P\Big(Z \leq \frac{-A(H) - \mu_1^t B(H)}{\sqrt{B^t(H)\Sigma B(H)}} \mid Z \sim N(0,1)\Big) \\
&= \Phi\Big(\frac{-A(H) - \mu_1^t B(H)}{\sqrt{B^t(H)\Sigma B(H)}}\Big), \qquad (4.8)
\end{aligned}$$

with Φ the cumulative distribution function of a univariate standard normal. In the same way, the probability of misclassifying an observation of population 0 is given by

$$\begin{aligned}\Pi_{10}(H) &= P(X^t B(H) + A(H) > 0 \mid X \sim N(\mu_0, \Sigma)) \\ &= \Phi\Big(\frac{A(H) + \mu_0^t B(H)}{\sqrt{B^t(H)\Sigma B(H)}}\Big). \end{aligned} \tag{4.9}$$

Using (4.8) and (4.9), the error rate using training data coming from a distribution H is given by

$$\mathrm{ER}(H) = \pi_1 \Phi\Big(\frac{-A(H) - \mu_1^t B(H)}{\sqrt{B^t(H)\Sigma B(H)}}\Big) + (1-\pi_1)\Phi\Big(\frac{A(H) + \mu_0^t B(H)}{\sqrt{B^t(H)\Sigma B(H)}}\Big). \tag{4.10}$$

At the model distribution $H = H_m$, where $A(H_m) = \alpha$ and $B(H_m) = \beta$, one gets

$$\mathrm{ER}(H_m) = \pi_1 \Phi\Big(\frac{-\alpha - \mu_1^t \beta}{\sqrt{\beta^t \Sigma \beta}}\Big) + (1-\pi_1)\Phi\Big(\frac{\alpha + \mu_0^t \beta}{\sqrt{\beta^t \Sigma \beta}}\Big).$$

4.3 Influence function

4.3.1 Second order influence functions

Expression (4.10) for the error rate defines a statistical functional $H \to \mathrm{ER}(H)$, of which the influence function (see Hampel et al (1986)) is defined as

$$\begin{aligned}\mathrm{IF}((x,y); \mathrm{ER}, H) &= \lim_{\varepsilon \downarrow 0} \frac{\mathrm{ER}((1-\varepsilon)H + \varepsilon \Delta_{(x,y)}) - \mathrm{ER}(H)}{\varepsilon} \\ &= \frac{\partial}{\partial \varepsilon} \mathrm{ER}((1-\varepsilon)H + \varepsilon \Delta_{(x,y)})\Big|_{\varepsilon\,=\,0} \end{aligned} \tag{4.11}$$

in those (x,y) where the limit exists. The notation $\Delta_{(x,y)}$ is used for a Dirac measure putting all its mass at (x,y). The heuristic interpretation of the influence function is that it measures the influence of an observation x in the training sample, being assigned to population y (where $y = 0$ or 1), on the error rate of the discriminant analysis procedure.

In this chapter we also need the *second order influence function*, defined here as

$$\mathrm{IF2}((x,y); \mathrm{ER}, H) = \frac{\partial^2}{\partial \varepsilon^2} \mathrm{ER}((1-\varepsilon)H + \varepsilon \Delta_{(x,y)})\Big|_{\varepsilon\,=\,0}.$$

If there is a (small) amount of contamination ε in the training data, due to the presence of a possible outlier (x,y), then the error rate of the discriminant procedure will be affected and can be approximated by the following Taylor expansion:

$$\mathrm{ER}(H_\varepsilon) \approx \mathrm{ER}(H_m) + \varepsilon \mathrm{IF}((x,y); \mathrm{ER}, H_m) + \frac{1}{2}\varepsilon^2 \,\mathrm{IF2}((x,y); \mathrm{ER}, H_m). \tag{4.12}$$

In Figure 4.1, we picture $\mathrm{ER}(H_\varepsilon)$ as a function of ε. The Fisher discriminant rule is optimal at the model distribution H_m, and therefore we denote $\mathrm{ER}(H_m) = \mathrm{ER}_{\mathrm{opt}}$. This implies that any other discriminant rule, in particular the one based on a contaminated training sample, can never have an error rate smaller than $\mathrm{ER}_{\mathrm{opt}}$. Hence, negative values of the influence function are excluded. From the well known property that $E[\mathrm{IF}((x,y);\mathrm{ER},H_m)] = 0$, (Hampel et al 1986, page 84), it follows that

$$\mathrm{IF}((x,y);\mathrm{ER},H_m) \equiv 0$$

almost surely. The behaviour of the error rate under small amounts of contamination is then characterised by the *second order influence function* IF2. Note that this second order influence function should be non-negative everywhere.

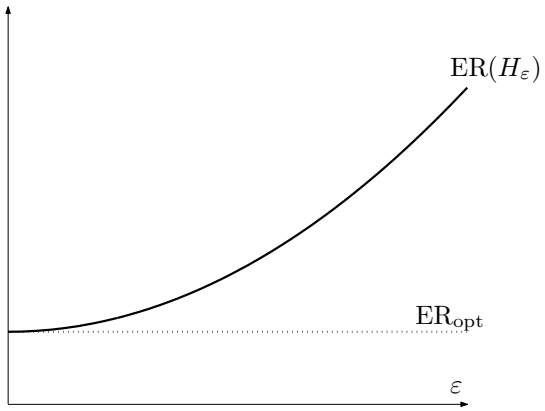

Figure 4.1: *Error rate of a discriminant rule based on a contaminated model distribution as a function of the amount of contamination ε.*

In the next proposition the second order influence functions of the error rate at the normal discrimination model is given. The obtained expression depends on the log odds ratio

$$\theta = \log \frac{\pi_1}{1-\pi_1}$$

and on the (squared) Mahalanobis distance between the centers of the two populations

$$\Delta^2 = (\mu_1 - \mu_0)^t \Sigma^{-1} (\mu_1 - \mu_0) = \beta^t \Sigma \beta.$$

Proposition 4.1. *Using the above notations, the influence function on the error rate of logistic discriminant analysis at the normal discriminant model H_m is zero and the second order influence function is given by*

$$\begin{aligned}\mathrm{IF2}((x,y);\mathrm{ER},H_m) = \pi_1\phi\Big(-\frac{\theta}{\Delta}-\frac{\Delta}{2}\Big)\Delta\Big[\Big(\frac{\mathrm{IF}((x,y);A,H_m)}{\Delta} \\ -\frac{\theta}{\Delta^3}(\mu_1-\mu_0)^t\mathrm{IF}((x,y);B,H_m)+\Big(\frac{\mu_1+\mu_0}{2}\Big)^t\frac{\mathrm{IF}((x,y);B,H_m)}{\Delta}\Big)^2 \\ +\frac{\mathrm{IF}((x,y);B,H_m)^t}{\Delta}\Big(\Sigma-\Big(\frac{\mu_1-\mu_0}{\Delta}\Big)\Big(\frac{\mu_1-\mu_0}{\Delta}\Big)^t\Big)\frac{\mathrm{IF}((x,y);B,H_m)}{\Delta}\Big]\end{aligned} \tag{4.13}$$

where $\mathrm{IF}((x,y);A,H_m)$ *and* $\mathrm{IF}((x,y);B,H_m)$ *are the influence functions of the estimators of the intercept and slope parameter of the logistic regression model, and* ϕ *is the standard normal density function.*

The proof is in the appendix. For different estimators of the parameters α and β in (4.4), different expressions for IF2 are obtained. In particular, one sees that bounded influence for the error rate is attained as soon as the IF of the functionals A and B are bounded. In the next subsection, plots of the second order influence functions will be presented.

4.3.2 Graphical representations

In this subsection, IF2 will be visualised for the ML and Bianco and Yohai estimators, as well as for their weigthed versions. Expressions for $\mathrm{IF}((x,y);A,H_m)$ and $\mathrm{IF}((x,y);B,H_m)$, needed to evaluate the second order influence function for the error rate in (4.13), are given in Croux and Haesbroeck (2003). Since all these estimators are equivariant with respect to an affine transformation of the vector of explanatory variables, without loss of generality, it may be assumed that $\mu_1 = -\mu_0 = (\Delta/2, 0, \ldots, 0)^t$, and $\Sigma = I_p$, yielding a *Canonical Model* H_m.

In Figure 4.2, $\mathrm{IF2}((x,y);\mathrm{ER},H_m)$ is pictured at the canonical model with $p = 1$, $\Delta = 2$ and $\theta = \log(2)$. The latter implies unequal group probabilities: $\pi_1 = 2/3$ and $\pi_2 = 1/3$. In this univariate setting, IF2 is plotted as a function of x with the value of y kept fixed, yielding one curve for $y = 1$ and another for $y = 0$. The curve for $y = 1$ gives then the influence that an observation in the training data, being allocated to the group with label $y = 1$, has on the error rate of the discriminant procedure. From Figure 4.2 one can see that, for one single covariate, the BY discriminant procedure has a bounded influence, while this does not hold for the ML-based method. For example, the IF2 goes beyond all bounds when the x-value of an observation corresponding to the population $N(\Delta/2, 1)$ tends to $-\infty$. Such observations are called bad leverage points, since they are both misclassified and leverage points in the covariate space. For the BY-procedure the bad leverage points only have a bounded effect, and the IF redescends to zero for extreme leverage points. The weighted estimators even give zero weight to

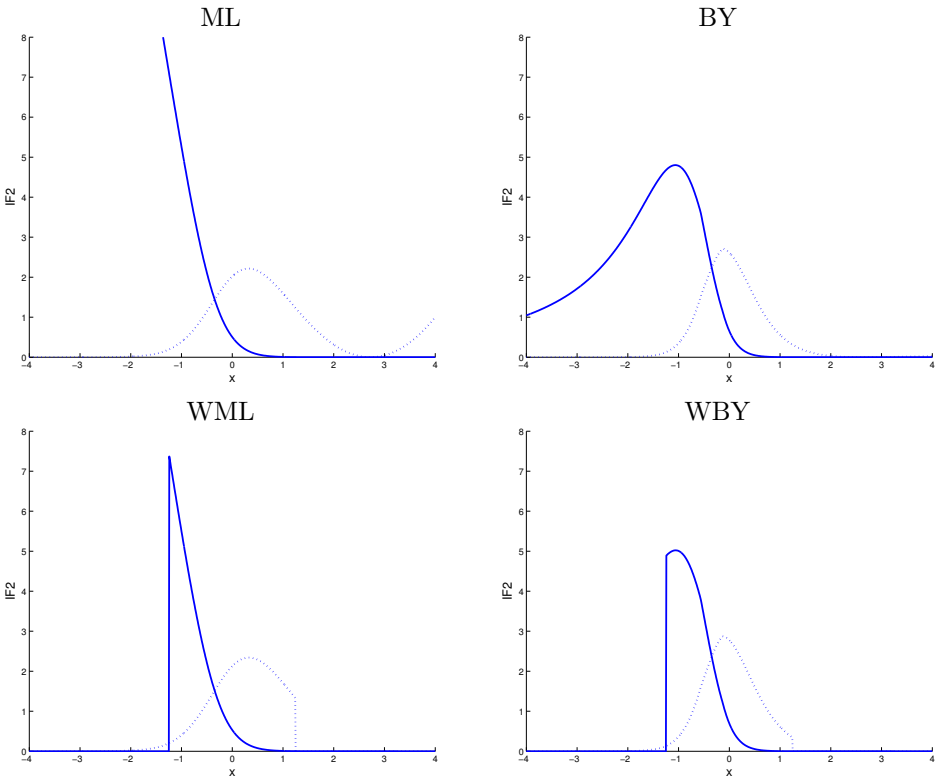

Figure 4.2: *Second order influence function* $\mathrm{IF2}((x,y);\mathrm{ER},H_m)$ *at the canonical model* H_m*, with* $p=1$*,* $\Delta=2$ *and* $\theta=\log(2)$ *for logistic discrimination based on the ML-estimator (left), on the Bianco and Yohai estimator (right), as well as their weighted versions (lower). We distinguish between* $y=1$ *(solid lines) and* $y=0$ *(dashed lines).*

high leverage points, as is reflected in their IF2. Except for the leverage points, the general shape of all second order influence functions is pretty similar. For all 4 considered discriminant procedures one sees that (i) good leverage points, i.e. correctly classified observations being outlying in the covariate space, have almost no influence on the error rate; (ii) incorrectly classified observations have a higher influence on the Error Rate; (iii) observations in the training sample being allocated to the group with the largest prior probability have more influence on the error rate.

Figure 4.3 represents $\mathrm{IF2}((x,1);\mathrm{ER},H_m)$ for $p=2$, $\Delta=2$ and $\theta=0$, corresponding to training data coming from a bivariate normal with mean $(1,0)^t$. The hyperplane separating the two groups of data has equation $x_1=0$. Similar conclusions as in the univariate case can be made, but there is a remarkable difference. For the BY estimator we observe that an observation, lying close to the discriminating hyperplane, while having a large value for the covariate variable, can have a value of the IF2 going beyond all bounds. These highly influential observations for the error rate of BY are neither good or bad leverage points. Therefore, as soon as the dimension of the covariate space is larger than one, a weighting step needs to be added to BY to get a fully bounded influence discriminant rule. Also note that the magnitude of the influence of a bad leverage point at x on the error rate depends heavily on the position in the covariate space. For the ML, for example, the IF2 is much smaller for observations being closer to the line connecting the two population centers.

We conclude that the BY discriminant procedure has no bounded influence on the error rate, and that weighting is recommended. Comparing the plots of WML and WBY, Figure 4.3 shows that their influence behaviour (on the error rate) is very similar. Taking into account the fact that WML is easier to compute than WBY, we favour this WML in the numerical applications we present in the next section.

4.4 Numerical results

4.4.1 Simulation study for the error rate

By means of a simulation experiment, we compare the finite sample error rate of robust (using the WML-estimator) and classical logistic discriminant analysis. Moreover, we also compare with Fisher's linear discriminant analysis, and a robustified version of it using S-estimators (as in He and Fung, 2000, or Croux and Dehon, 2001). Several sampling schemes are considered, for $p=3$ and $n=200$. For every sampling scheme we generated $m=1000$ training data sets of size n, and computed the associated error rate. This error rate is obtained by evaluating the discriminant rule estimated from the training data on a test data set of size 10^5 generated from the model distribution. Average error rates over the m simulations are then reported in Table 4.1.

In the first three sampling scheme, training samples are generated according

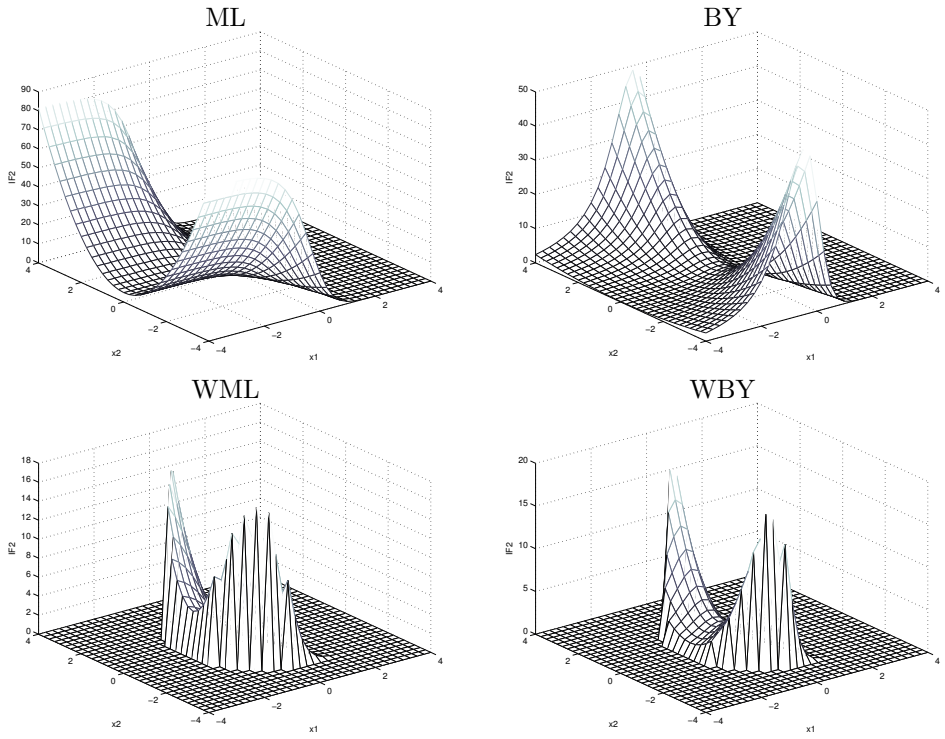

Figure 4.3: *Second order influence function* $\mathrm{IF2}((x,1);\mathrm{ER},H_m)$ *at the canonical model* H_m*, with* $p=2$*,* $\Delta=2$ *and* $\theta=0$ *for logistic discrimination based on the ML-estimator (left), on the Bianco and Yohai estimator (right), as well as their weighted versions (lower).*

to a canonical normal discrimination model H_m, with $\mu_1 = -\mu_0 = (\Delta/2, 0, 0)^t$, and $\Sigma = I_p$. In the first simulation experiment we take $\Delta = 1$ and $\theta = 0$, afterwards $\Delta = 1$ and $\theta = \log(2)$, and in the third setting $\Delta = 3$ and $\theta = 0$. The 2 other sampling schemes take $\Delta = 1$ and 2, respectively, and $\theta = 0$, but they do not follow the normal discrimination model discussed in Section 2.1. In the fourth scheme the data are simulated from normal distributions with unequal covariance matrices: $H_1 = N(\mu_1, I_p)$ and $H_0 = N(\mu_0, 0.25I_p)$, while in a last simulation setting a exponential transformation is applied to the explicative variables, creating asymmetric distributions for the two source populations.

To investigate the robustness of the procedures, we add 10 leverage points to the training data, inducing 5% of contamination. These leverage points are all attributed to the group $y = 1$, and distributed according to $\lambda\Delta N(-(\lambda, 1, 1)^t, (0.01) * I_p)$. Intermediate outliers correspond then with $\lambda = 2$, and extreme outliers with $\lambda = 5$.

In Table 4.1 simulated error rates are given, where the standard error around the reported results ranges from about 0.02% (for the cases where not outliers are present) up to 0.1%. Let us first investigate the effect of the outliers on the error rates. We see that outliers may have a disastrous effect on the classification performance of the classical procedures. In presence of the extreme outliers (type 2), the classical procedures can even have unacceptably high error rates around 50%, which happens for schemes (1) and (4). When the contamination in the training data is of the first type, and closer to the data clouds of the clean observations, the error rate of the classical procedure is still significantly driven upwards, but we also note that the robust discriminant procedures are much more vulnerable to these intermediate than to extreme outliers. The reason is that the robust estimators involved are redescending, and by giving a zero weight, the extreme outliers become harmless. For the second sampling scheme, with $\theta = \log(2)$, the effect of outliers is less pronounced than in the first case. The reason is that the contamination level, expressed as a percentage of the number of group "$y = 1$" observations, is smaller than for the first sampling scheme. For sampling scheme (3), similar conclusions as before can be made, but all error rates are smaller now since the two source populations are easier to discriminate here.

Table 4.1 also allows to compare standard linear and logistic discrimination. When no outliers are present, working at the normal discrimination model (the first three cases), linear discriminant analysis has slightly smaller error rates for $n = 200$, the reason being that Fisher's method is based on the full maximum likelihood estimators here. Logistic discrimination, however, is not losing much in error rate, since it is also consistently estimating the optimal discriminant boundary. For the last two sampling schemes, Fisher's linear discriminant analysis is no longer optimal. In the simulation experiment with unequal covariances, it still results in slightly better error rates, but at the asymmetric lognormal distributions logistic discrimination outperforms Fisher's method.

Table 4.1: *Simulated average error rates for logistic and linear discriminant analysis with classical and robust estimators, for five different sampling schemes, and in presence of intermediate outliers (type I outliers), and extreme outliers (type II outliers).*

	no outliers		type I outliers		type II outliers	
	Classic	*Robust*	*Classic*	*Robust*	*Classic*	*Robust*
(1) $\Delta = 1,\ \theta = 0$						
Logistic	31.52	31.56	36.64	34.57	49.39	31.55
Linear	31.52	31.82	36.59	35.30	49.01	31.91
(2) $\Delta = 1,\ \theta = \log(2)$						
Logistic	27.58	27.65	30.83	28.64	33.91	27.60
Linear	27.57	27.88	30.79	29.60	33.88	28.01
(3) $\Delta = 3,\ \theta = 0$						
Logistic	7.03	7.09	19.80	7.06	36.02	7.07
Linear	6.89	7.09	19.76	7.01	35.97	7.07
(4) Unequal covariances						
Logistic	24.62	24.70	34.15	30.35	47.92	24.83
Linear	24.10	24.46	33.73	31.21	47.58	25.27
(5) Log-normal, $\Delta = 2$						
Logistic	17.33	16.89	28.94	26.72	43.08	17.01
Linear	25.54	23.10	31.79	28.72	43.68	24.04

Comparing the performance of robust logistic and robust linear discriminant analysis turns out to be favourable for robust logistic discrimination. In most cases the differences in simulated error rate between both robust procedures is very small, but for the lognormal distributions there is a clear advantage for the logistic approach. A conclusion from this simulation experiment is that robust logistic discrimination leads only to a very small loss in classification performance when no outliers are present. On the other hand, the effect of outliers, both extreme and intermediate, in the training sample on the error rate remains within bounds, while this does not hold for the classical procedures. Finally, robust logistic discrimination can compete with robust versions of Fisher's linear discriminant analysis.

4.4.2 A diagnostic measure for detecting influential observations

Consider the well-known Vaso Constriction data set of Finney (1947), see also Pregibon (1981). The binary outcomes (presence or absence of vaso constriction of the skin of the digits after air inspiration) are explained by two continuous variables: x_1 the volume of air inspired and x_2 the inspiration rate, both log-transformed. Figure 4.4 gives the scatter plot of the 40 observations in the covariate space, together with the y-values. To asses the effect of contamination on

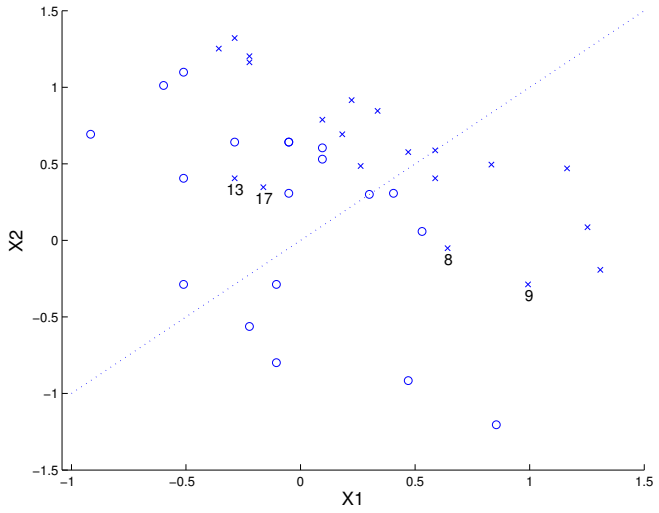

Figure 4.4: *The Vaso Constriction data set. The circles represent the group in absence of vaso constriction ($y = 0$) and the crosses the group in presence of vaso constriction ($y = 1$).*

the ML-estimator and on the robust WML-estimator, an observation is added to the population with $y = 0$ at position $(x_1, x_2) = (s, s)$. In Figure 4.4 the dotted line represents the line along which this extra observation moves. For negative values of s, the added observation will be correctly classified and therefore it is a good leverage point. For large values of s, we get a bad leverage point. To study the effect of adding this extra observation we compute the apparent error rate from the 40 observations, where s varies from -1 to 10. From Figure 4.5, it is confirmed that the robust WML estimator limits the influence of outliers. On the other hand, the error rate of the classical ML estimator can increase to about 50% when adding only one outlier.

In the same spirit as in Boente et al (2002) or Pison et al (2003), the influence functions can be used to detect influential points in the training data set. The value of IF2 evaluated at the sample points indicates the contribution of each particular observation in the training set to the error rate. Aim is to detect influential observations for the ML-estimator, being most vulnerable to outliers. The diagnostic measures are defined as

$$D_i = \mathrm{IF2}((x_i, y_i)); \mathrm{ER}, H_m)/c_{y_i}, \tag{4.14}$$

for $1 \leq i \leq n$. In (4.14), the constant c_j corresponds to the 95% quantile of the distribution of $\mathrm{IF2}((X, j); \mathrm{ER}, H_m)$, with $X \sim H_j$, for $j = 0, 1$. For more information on critical values for influence function diagnostics, we refer to Pison

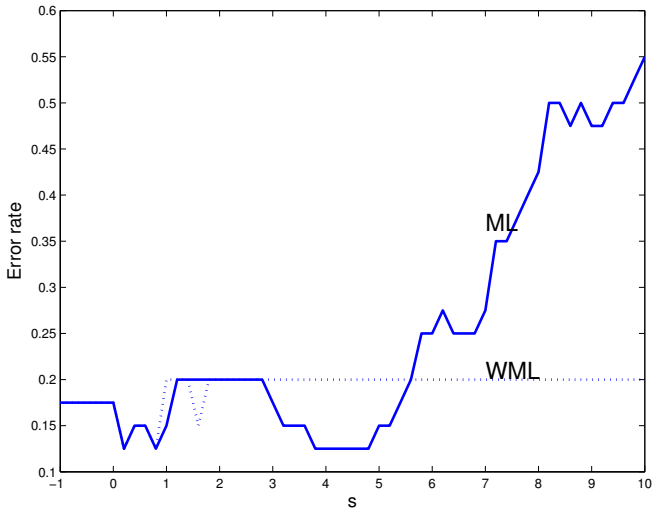

Figure 4.5: *Misclassification rate for the ML-estimator (solid line) and for the WML-estimator (dotted line) after adding observation* $(s, s, 0)$*, where* s *varies from -1 to 10.*

and Van Aelst (2004). This allows to flag an observation as being significantly influential as soon as $D_i > 1$. Note that the unknown parameters in H_m need to be estimated robustly to avoid the masking effect, hereby yielding a robust diagnostic measure.

A plot of the diagnostic measures D_i with respect to the index of the observation gives a graphical diagnostic tool to detect influential observations. The diagnostic measures were computed for the Vaso Constriction data, and also for the contaminated data sets where the 21-st observation is the added observation $(s, s, 0)$, for respectively $s = 4, 7, 10$. Figure 4.6 presents the 4 corresponding plots. From the upper left plot, it is seen that there are a few influential points: observations 8 and 9, and to a lesser extent observations 13 and 17. These observations, as can be seen from Figure 4.4, are incorrectly classified, and somehow at the border of the data cloud for $y = 1$. Although these observations are quite influential on the ML-estimator, they are by no means heavy outliers. From the other plots of Figure 4.6, it is seen that the values of D_i, with the exception of the added observation, remain quite stable. This illustrates the robustness of the diagnostics. Regarding the added observation, it is seen from Figure 4.6 that it only becomes highly influential for $s = 7$ and $s = 10$. This confirms Figure 4.5, where the contamination for $s = 4$ is not yet affecting the error rate of the ML-procedure. It is worth noting that $s = 4$ corresponds to a huge outlier in the covariate space, but even more extreme values of s are needed to become influential. The reason is that the added outliers are close to a line through the center

and orthogonal to the separating hyperplane, where the influence on the error rate is smallest, as can be seen from Figure 4.3.

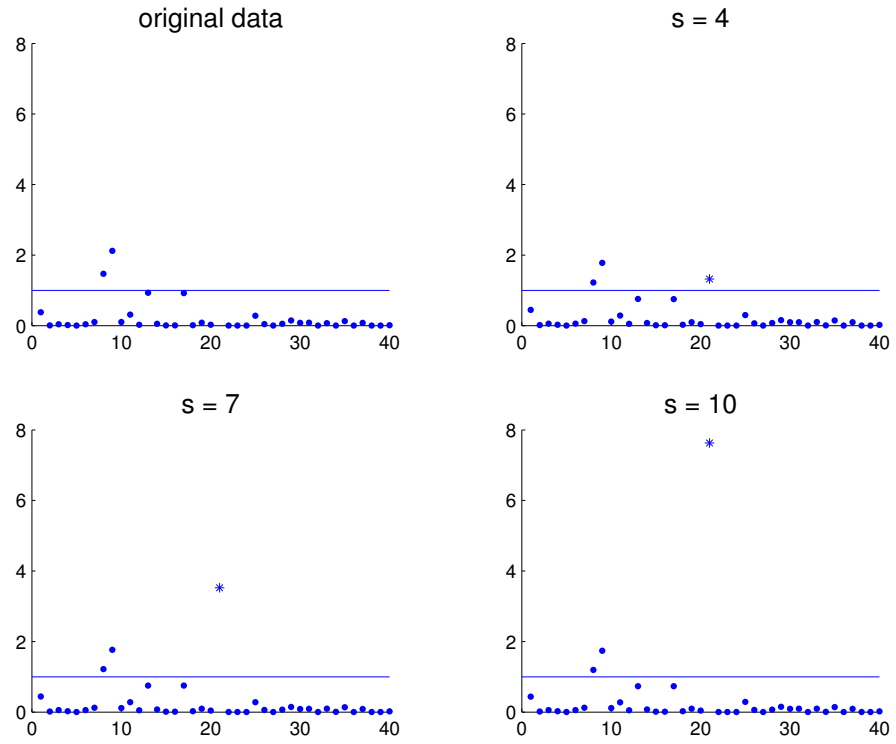

Figure 4.6: *Diagnostic plots for the Vaso Constriction data set (upper left) and for the data set with an added observation* $(s, s, 0)$ *with index 21, for* $s = 4, 7$ *and 10.*

4.5 Conclusions

In this chapter second order influence functions for the error rate have been computed. Due to the optimality of logistic discrimination at the normal discrimination model the use of the second order influence functions is natural and appropriate, as motivated in Section 4.3. The expressions obtained are not only valid for the classical maximum likelihood estimator, but also for robust estimators. While influence analysis for estimators of the parameters of the logistic regression model has already been carried out before, this is not the case for the corresponding error rate. Besides of theoretical interest, it has also been shown how an empirical version of the second order influence function can be used as a robust

diagnostic tool.

Logistic discrimination is easy to carry out, since the Maximum Likelihood estimator for the logistic regression model is implemented in all statistical software packages. Unfortunately the ML-estimator is not robust: although outliers cannot occur in the dependent variable (taking only the values 0 or 1), outliers in the space of the explicative variables, i.e. leverage points, can ruin the ML-procedure. Indeed, as shown in this chapter, outliers may have an unlimited influence on the error rate corresponding to the ML-based procedure. Using the weighted ML-estimator instead, an alternative robust procedure for logistic discrimination is obtained.

Appendix

Before starting the proof of Proposition 4.1, we first need the two following Lemmas.

Lemma 4.2. *Set* $D_1 = -\theta/\Delta - \Delta/2$ *and* $D_0 = \theta/\Delta - \Delta/2$. *Then*

(i) $\mathrm{ER}(H_m) = \pi_1\Phi(D_1) + \pi_0\Phi(D_0)$

(ii) $\pi_1\phi(D_1) = \pi_0\phi(D_0)$

Proof. (i) This is straightforward from (4.10). For example

$$\frac{\alpha + \beta^t\mu_0}{\sqrt{\beta^t\Sigma\beta}} = \frac{\theta + \beta^t\frac{\mu_0-\mu_1}{2}}{\Delta} = \frac{\theta - 1/2(\mu_1-\mu_0)^t\Sigma^{-1}(\mu_1-\mu_0)}{\Delta} = \frac{\theta}{\Delta} - \frac{\Delta}{2}.$$

(ii) It is sufficient to note that $\log(\phi(D_0)/\phi(D_1)) = D_1^2/2 - D_0^2/2 = \theta = \log(\pi_1/\pi_0)$. □

Lemma 4.3. *Consider the two functionals* $E(H) = A(H)/\sqrt{B^t(H)\Sigma B(H)}$ *and* $F(H) = B(H)/\sqrt{B^t(H)\Sigma B(H)}$. *Then*

(i) $\mathrm{IF}((x,y); E, H_m) = \mathrm{IF}((x,y); A, H_m)/\Delta - \alpha\beta^t\Sigma\mathrm{IF}((x,y); B, H_m)/\Delta^3$.

(ii) $\mathrm{IF}((x,y); F, H_m) = \mathrm{IF}((x,y); B, H_m)/\Delta - \beta\beta^t\Sigma\mathrm{IF}((x,y); B, H_m)/\Delta^3$.

(iii) $\mathrm{IF}((x,y); F, H_m)^t(\mu_1 - \mu_0) = 0$.

(iv) $\mathrm{IF2}((x,y); F, H_m)^t(\mu_1-\mu_0) = -\Delta\frac{\mathrm{IF}((x,y);B,H_m)^t}{\Delta}\left\{\Sigma - \left(\frac{\mu_1-\mu_0}{\Delta}\right)\left(\frac{\mu_1-\mu_0}{\Delta}\right)^t\right\}\frac{\mathrm{IF}((x,y);B,H_m)}{\Delta}$.

Proof. (i) and (ii) can be obtained via straightforward derivation. For a given fixed (x,y), we set $H_\varepsilon = (1-\varepsilon)H_m + \varepsilon\Delta_{(x,y)}$. Now by definition of F, we have

$F(H)^t\Sigma F(H) = 1$ for any H, and in particular $F(H_\varepsilon)^t\Sigma F(H_\varepsilon) = 1$. From the latter it follows that

$$\Big(\frac{\partial}{\partial\varepsilon}F(H_\varepsilon)\Big)^t\Sigma F(H_\varepsilon) = 0, \tag{4.15}$$

for any $\varepsilon > 0$. Evaluating (4.15) at $\varepsilon = 0$ and noting that $F(H_m) = \beta/\Delta = \Sigma^{-1}(\mu_1 - \mu_0)/\Delta$ yields (iii). Deriving (4.15) ones more w.r.t. ε and evaluating at $\varepsilon = 0$ results in

$$\mathrm{IF2}((x,y);F,H_m)^t\Sigma F(H_m) + \mathrm{IF}((x,y);F,H_m)^t\Sigma\mathrm{IF}((x,y);F,H_m) = 0,$$

from which it follows that

$$\mathrm{IF2}((x,y);F,H_m)^t(\mu_1-\mu_0) = -\Delta\mathrm{IF}((x,y);F,H_m)^t\Sigma\mathrm{IF}((x,y);F,H_m). \tag{4.16}$$

Denote now

$$P = I - \left(\frac{\Sigma^{-1/2}(\mu_1-\mu_0)}{\Delta}\right)\left(\frac{\Sigma^{-1/2}(\mu_1-\mu_0)}{\Delta}\right)^t$$

a projection matrix such that $P^tP = P$ and $P = P^t$. Then we can rewrite (ii) as

$$\mathrm{IF}((x,y);F,H_m) = \Sigma^{-1/2}P\Sigma^{1/2}\mathrm{IF}((x,y);B,H_m)/\Delta.$$

From the above, it follows immediately from (4.16) that

$$\mathrm{IF2}((x,y);F,H_m)^t(\mu_1-\mu_0) = -\Delta\frac{\mathrm{IF}((x,y);F,H_m)^t}{\Delta}\Sigma^{1/2}P\Sigma^{1/2}\frac{\mathrm{IF}((x,y);F,H_m)}{\Delta},$$

implying (iv). □

Proof of Proposition 4.1: At the contaminated distribution H_ε, it follows from (4.10) that

$$\mathrm{ER}(H_\varepsilon) = \pi_1\Phi(-E(H_\varepsilon) - F(H_\varepsilon)^t\mu_1) + \pi_0\Phi(E(H_\varepsilon) + F(H_\varepsilon)^t\mu_0) \tag{4.17}$$

Standard derivations results in

$$\begin{aligned}\mathrm{IF}((x,y);\mathrm{ER},H_m) =&(-\pi_1\phi(D_1) + \pi_0\phi(D_0))\mathrm{IF}((x,y);E,H_m)\\ &- \pi_1\phi(D_1)\mathrm{IF}((x,y);F,H_m)^t(\mu_1-\mu_0),\end{aligned} \tag{4.18}$$

using the notations of Lemma 4.2. The first term of (4.18) cancels due to Lemma 4.2(ii) and the second term due to Lemma 4.3(iii), showing already that $\mathrm{IF}((x,y);\mathrm{ER},H_m) = 0$.

Computing the second derivative of (4.17) results in

$$\begin{aligned}\mathrm{IF2}((x,y);\mathrm{ER},H_m) &= \pi_1\phi'(D_1)[\mathrm{IF}((x,y);E,H_m) + \mu_1^t\mathrm{IF}((x,y);F,H_m)]^2\\ &+ \pi_0\phi'(D_0)[\mathrm{IF}((x,y);E,H_m) + \mu_0^t\mathrm{IF}((x,y);F,H_m)]^2\\ &- \pi_1\phi(D_1)[\mathrm{IF2}((x,y);E,H_m) + \mu_1^t\mathrm{IF2}((x,y);F,H_m)]\\ &+ \pi_0\phi(D_0)[\mathrm{IF2}((x,y);E,H_m) + \mu_0^t\mathrm{IF2}((x,y);F,H_m)]\end{aligned}$$

Using $\phi'(u) = -u\phi(u)$, $D_0 + D_1 = -\Delta$, Lemma 4.3(iii) and Lemma 4.2(ii), the above expression reduces to

$$\begin{aligned}\mathrm{IF2}((x,y);\mathrm{ER}H_m) =& \pi_1\Delta\phi(D_1)[\mathrm{IF}((x,y);E,H_m) + \mu_1^t\mathrm{IF}((x,y);F,H_m)]^2 \\ &- \pi_1\phi(D_1)\mathrm{IF2}((x,y);F,H_m)^t(\mu_1-\mu_0).\end{aligned} \tag{4.19}$$

From Lemma 4.3(i) and 4.3(ii) it follows after some calculations that the term $\mathrm{IF}((x,y);E,H_m) + \mu_1^t\mathrm{IF}((x,y);F,H_m)$ is equal to

$$\frac{\mathrm{IF}((x,y);A,H_m)}{\Delta} + \Big[\Big(\frac{\mu_1+\mu_0}{2}\Big) - \frac{\theta(\mu_1-\mu_0)}{\Delta^2}\Big]^t\frac{\mathrm{IF}((x,y);B,H_m)}{\Delta},$$

where it was used that $\alpha = \theta - \beta^t\frac{\mu_1+\mu_0}{2}$ and $\beta = \Sigma^{-1}(\mu_1-\mu_0)$. From (4.19), the above equation and Lemma 4.3(iv), the expression for $\mathrm{IF2}((x,y);\mathrm{ER},H_m)$ can be obtained immediately. □

Chapter 5

Robust linear discriminant analysis for multiple groups: Influence and classification efficiencies

Co-Author: C. Croux and P. Filzmoser

Summary Linear discriminant analysis for multiple groups is typically carried out using Fisher's method. This method relies on the sample averages and covariance matrices computed from the different groups constituting the training sample. Since sample averages and covariance matrices are not robust, it is proposed to use robust estimators of location and covariance instead, yielding a robust version of Fisher's method.

In this chapter expressions are derived for the influence that an observation in the training set has on the error rate of the Fisher method for multiple linear discriminant analysis. These influence functions on the error rate turn out to be unbounded for the classical rule, but bounded when using a robust approach. Using these influence functions, we compute relative classification efficiencies of the robust procedures with respect to the classical method. It is shown that, by using an appropriate robust estimator, the loss in classification efficiency at the normal model remains limited. These findings are confirmed by finite sample simulations.

5.1 Introduction

In discriminant analysis one observes several groups of multivariate observations, forming together the *training sample*. For the data in this training sample, it is known to which group they belong. Discriminant functions, aimed at separating the different groups, are constructed on the basis of the training sample. These discriminant functions are then used to classify new observations into one of the groups. A popular discrimination method is Fisher's linear discriminant analysis, introduced for two populations by Fisher (1938) and generalised to multiple populations by Rao (1948). Over the last decade several more sophisticated classification methods, like support vector machines and random forests, have been proposed (see Friedman et al 2001). But Fisher's method is still often used and performs well in many applications. Also, the Fisher discriminant functions are linear combinations of the measured variables, making them easy to interpret.

At the population level, the Fisher discriminant functions are obtained as follows. Consider g populations in a p-dimensional space, being distributed with centers $\mu_1, \ldots, \mu_g$ and covariance matrices $\Sigma_1, \ldots, \Sigma_g$. The probability that an observation to classify belongs to group j is denoted by π_j, for $j = 1, \ldots, g$, with $\sum_j \pi_j = 1$. Then the *between groups covariance matrix* $\mathcal{B}$ is defined as

$$\mathcal{B} = \sum_{j=1}^{g} \pi_j (\mu_j - \bar{\mu})(\mu_j - \bar{\mu})^t, \tag{5.1}$$

with $\bar{\mu} = \sum_j \pi_j \mu_j$ the weighted average of the population centers. The *within groups covariance matrix* $\mathcal{W}$ is given by the pooled version of the different scatter matrices

$$\mathcal{W} = \sum_{j=1}^{g} \pi_j \Sigma_j. \tag{5.2}$$

The aim of Fisher's method is to project the data onto a lower dimensional subspace of dimension s by maximising the between groups variance of the projected data, while keeping the within groups variance constant. Moreover, the within groups covariance matrix of the projected data should be the unity matrix. This leads to an eigenvalue analysis of the matrix

$$\mathcal{W}^{-1}\mathcal{B}. \tag{5.3}$$

For details and proofs we refer to Johnson and Wichern (1998). Denote now the eigenvectors corresponding to the largest s strictly positive eigenvalues of (5.3) by $v_1, \ldots, v_s$, and scale them such that $v_j^t \mathcal{W} v_j = 1$, for $1 \leq j \leq s$. If x is an observation to classify, then the linear combinations $v_1^t x, \ldots, v_s^t x$ are the values of, respectively, the first,..., s-th *Fisher linear discriminant functions*. Note that the value of s is at most equal to the number of strictly positive eigenvalues of

$\mathcal{W}^{-1}\mathcal{B}$, so $s \leq \min(g-1, p)$. With the aim of dimension reduction and visualisation (e.g. Cook and Yin 2001), s may be taken smaller than $\min(g-1, p)$.

The observation to classify is assigned to that group for which the "distance" between the projected observation and the group center is smallest. Formally, x is assigned to population k for which

$$D_k(x) = \min_{j=1,\ldots,g} D_j(x),$$

where

$$D_j^2(x) = [V^t(x-\mu_j)]^t[V^t(x-\mu_j)] - 2\log\pi_j \tag{5.4}$$

and $V = (v_1, \ldots, v_s)$ is the matrix having the eigenvectors in its columns. Note that the squared distances, also called the Fisher discriminant scores, in (5.4) are penalised by the term $-2\log\pi_j$, so that an observation is less likely to be assigned to groups with smaller prior probabilities. A prior probability π_j is unknown, but can be estimated by the empirical frequency of observations in the training data belonging to group j, for $1 \leq j \leq g$. By adding the penalty term in (5.4), the Fisher discriminant rule is optimal (in the sense of having a minimal total probability of misclassification), for source populations being normally distributed with equal covariance matrix and for s equal to the maximum number of strictly positive eigenvalues of $\mathcal{W}^{-1}\mathcal{B}$ (see Johnson and Wichern 1998, page 685).

At the sample level, the centers μ_j and covariance matrices Σ_j of each group need to be estimated, which is typically done using sample averages and sample covariance matrices. But sample averages and covariance matrices are not robust, and outliers in the training sample may have an unduly large influence on the classical Fisher discriminant rule. Hence it has been proposed to use robust estimators of location and covariance instead and plugging them into (5.1) and (5.2), yielding a robust version of Fisher's method. Such a straightforward plug-in approach for obtaining a robust discriminant analysis procedure was already taken by Randles et al (1978), using M-estimators, and afterwards by Chork and Rousseeuw (1992), Hawkins and McLachlan (1997) and Hubert and Van Driessen (2004) using Minimum Covariance Determinant estimators, and by He and Fung (2000) and Croux and Dehon (2001) using S-estimators. In most of these papers the good performance of the robust discriminant procedures was shown by means of simulations and examples, but we would like to obtain some theoretical results concerning robustness and efficiency of the discrimination method. The performance of the discriminant rules will be measured by their *error rate*, being the total probability of misclassification.

The contribution here is twofold. First of all, influence functions, measuring the effect of an observation in the training sample on the error rate, are computed theoretical. In robustness it is standard to compute an influence function for estimators, but here the focus is on the error rate of a classification rule. Computation

of such a theoretical influence function for the error rate is difficult, and results are presented for a model where the different populations are normally distributed, with equal covariance matrices, and collinear centers. In this case the Fisher discriminant rule is optimal, and it turns out that one needs to compute a *second order influence function*, since the usual first order influence function equals zero. It is shown that the Fisher rule, using the sample averages and sample covariance matrices of each group, yields unbounded influence functions for the error rate, while using robust estimates instead gives bounded influence procedures.

A second contribution of this chapter is that *asymptotic relative classification efficiencies* are computed, using the second order influence functions. As such, one can measure how much increase of the error rate is expected when a robust instead of the classical procedure is used in case when no outliers are present. Classification efficiencies were introduced by Efron (1975), who compared the performance of logistic discrimination with linear discrimination for two-group discriminant analysis. These results were then extended to multi-group settings by Bull and Donner (1987) and Campbell and Donner (1989). Also these authors made the assumption of collinear population centers, to keep the calculations feasible. Note that for two-group discrimination, the population centers are always collinear. Up to our best knowledge, asymptotic relative classification efficiencies for *robust* discriminant procedures have never been computed before.

This chapter is organised as follows. In Section 5.2, an expression for the error rate of Fisher's multiple discriminant analysis at the model distribution is given. Section 5.3 defines the influence of an observation on the error rate and derives expressions for the second order influence function. Asymptotic relative classification efficiencies are then given in Section 5.4, followed by a simulation study, presented in Section 5.5, and conclusions, made in Section 5.6.

5.2 Error rate

Let X be a p-variate stochastic variable containing the predictor variables, and Y be the variable indicating the group membership, so $Y \in \{1, \ldots, g\}$. The training sample $(X_1, Y_1), \ldots, (X_n, Y_n)$ is a random sample from the distribution H. In this section we will define the Error Rate (ER) as a function of the distribution H, yielding a statistical functional $H \to \mathrm{ER}(H)$, which allows to compute influence functions in Section 5.3.

Denote $T_j(H)$ and $C_j(H)$ the location and scatter of the conditional distribution $X|Y = j$, for $j = 1, \ldots, g$. The location and scatter functionals may correspond to the expected value and the covariance matrix, but any other affine equivariant location and scatter measure is allowed. The functional representa-

tions of the between and within groups covariance matrices are then

$$B(H) = \sum_{j=1}^{g} \pi_j(H)(T_j(H) - \bar{T}(H))(T_j(H) - \bar{T}(H))^t$$

and

$$W(H) = \sum_{j=1}^{g} \pi_j(H) C_j(H),$$

with $\bar{T}(H) = \sum_j \pi_j(H) T_j(H)$ and $\pi_j(H) = P_H(Y = j)$, for $j = 1, \ldots, g$. The first s eigenvectors of $W^{-1}(H)B(H)$, with $s \leq \min(g-1, p)$, are then collected in the matrix $V(H)$, allowing us to compute the Fisher discriminant scores

$$D_j^2(x, H) = (x - T_j(H))^t V(H) V(H)^t (x - T_j(H)) - 2 \log \pi_j(H), \tag{5.5}$$

for $j = 1 \ldots, g$. A new observation x will be assigned to population k for which the discriminant score is minimal. In the above formula, the prior group probabilities $\pi_j(H)$ are estimated from the training data. So we have a *prospective* sampling scheme in mind, meaning that the group proportions of the data to classify are the same as for the training data [1].

Let us denote by H_m the distribution of the data to classify. Then, with $\pi_j = P_{H_m}(Y = j)$, for $j = 1, \ldots, g$, the error rate is given by

$$\mathrm{ER}(H) = \sum_{j=1}^{g} \pi_j \; P_{H_m}\Big(D_j(X, H) > \min_{\substack{k \neq j \\ k=1,\ldots,g}} D_k(X, H) \mid Y = j\Big). \tag{5.6}$$

In ideal circumstances we have that the data to classify are generated from the same distribution as the training data set, so $H = H_m$. When computing the influence function, however, we need to take for H a contaminated version of H_m.

Expression (5.6) is difficult to evaluate. To make theoretical results possible, we restrict to normal distributions with identical covariance matrices and collinear centers. Note that for discriminating $g = 2$ groups, the collinearity condition is automatically verified. Formally, we require the model distribution H_m to verify

(M) At the model distribution H_m, $X|Y = j$ follows a normal distribution $N(\mu_j, \Sigma)$ for $j = 1, \ldots, g$. The centers μ_j are different and collinear, and the matrix Σ is non-singular. Furthermore, every $\pi_j = P_{H_m}(Y = j)$ is strictly positive.

Since we will only work with location and scatter functionals being consistent at normal distributions, we have $(T_j(H_m), C_j(H_m)) = (\mu_j, \Sigma)$ for $1 \leq j \leq g$.

[1] Results for a retrospective sampling scheme, where the prior probabilities differ from the sampling proportions in the training set, can be obtained in a completely analogous way.

Furthermore, since $B(H_m) = \mathcal{B}$ has rank 1, we only can have one strictly positive eigenvalue of $\mathcal{W}^{-1}\mathcal{B}$, implying $s = 1$. The matrix $V(H_m)$ reduces then to the vector

$$v_1 = \Sigma^{-1}\frac{\mu_j - \mu_{j+1}}{\Delta_j}, \tag{5.7}$$

with

$$\Delta_j = \sqrt{(\mu_j - \mu_{j+1})^t\Sigma^{-1}(\mu_j - \mu_{j+1})}, \tag{5.8}$$

for $j = 1, \ldots, g-1$.

Taking H_m as distribution of the data to classify (with $s = 1$), expression (5.6) becomes tractable. Let H be any distribution of the training data. We will reorder the labels of the groups such that $V^t(H)T_1(H) < V^t(H)T_2(H) < \ldots < V^t(H)T_{g'}(H)$, with $g' \leq g$, and such that observations belonging to groups with a label $j > g'$ are misclassified with probability one. In the Appendix, a procedure for doing this relabelling is outlined. The following result holds. Throughout the chapter, we use the notation Φ for the cumulative distribution function of a univariate standard normal, and ϕ for its density.

Proposition 5.1. *If the observations to classify are distributed according to a model H_m verifying* **(M)**, *the error rate of the Fisher discriminant rule (with $s = 1$) is given by*

$$\begin{aligned}\mathrm{ER}(H) &= \sum_{j=1}^{g'-1}\Big\{\pi_j\Phi\Big(\frac{A_j(H) + B_j^t(H)\mu_j}{\sqrt{B_j^t(H)\Sigma B_j(H)}}\Big) + \pi_{j+1}\Phi\Big(\frac{-A_j(H) - B_j^t(H)\mu_{j+1}}{\sqrt{B_j^t(H)\Sigma B_j(H)}}\Big)\Big\} \\ &\quad + \sum_{j=g'+1}^{g}\pi_j \end{aligned} \tag{5.9}$$

with

$$B_j(H) = V(H)V(H)^t(T_{j+1}(H) - T_j(H)) \tag{5.10}$$

$$A_j(H) = \log(\pi_{j+1}(H)/\pi_j(H)) - B_j(H)^t(T_j(H) + T_{j+1}(H))/2 \tag{5.11}$$

for $1 \leq j \leq g$ and H the distribution of the training sample.

For $H = H_m$ formula (5.9) reduces further to

$$\mathrm{ER}(H_m) = \sum_{j=1}^{g'-1}\Big\{\pi_j\Phi\Big(\frac{\theta_j}{\Delta_j} - \frac{\Delta_j}{2}\Big) + \pi_{j+1}\Phi\Big(-\frac{\theta_j}{\Delta_j} - \frac{\Delta_j}{2}\Big)\Big\} + \sum_{j=g'+1}^{g}\pi_j, \tag{5.12}$$

where $\theta_j = \log(\pi_{j+1}/\pi_j)$ and Δ_j is defined in (5.8) for $j = 1, \ldots, g-1$.

5.3 Influence functions

To study the effect of an observation on a statistical functional it is common in the robustness literature to use influence functions (see Hampel et al 1986). As such, the influence function of the error rate at the model H_m is defined as

$$\begin{aligned} \mathrm{IF}((x,y);\mathrm{ER},H_m) &= \lim_{\varepsilon\to 0}\frac{\mathrm{ER}\left((1-\varepsilon)H_m+\varepsilon\Delta_{(x,y)}\right)-\mathrm{ER}(H_m)}{\varepsilon} \\ &= \frac{\partial}{\partial\varepsilon}\mathrm{ER}(1-\varepsilon)H_m+\varepsilon\Delta_{(x,y)})\Big|_{\varepsilon\,=\,0}, \end{aligned}$$

with $\Delta_{(x,y)}$ the Dirac measure putting all its mass in (x,y). Recall that x is a p-variate observation, and y indicates the group membership. More generally, we define[2] the k-th order influence function as

$$\mathrm{IFk}((x,y);T,H)=\frac{\partial^k}{\partial\varepsilon^k}\mathrm{ER}((1-\varepsilon)H_m+\varepsilon\Delta_{(x,y)})\Big|_{\varepsilon\,=\,0}. \tag{5.13}$$

If there is a (small) amount of contamination in the training data, due to the presence of a possible outlier (x,y), the error rate of the discriminant procedure based on $H_\varepsilon=(1-\varepsilon)H_m+\varepsilon\Delta_{(x,y)}$ can be approximated by the following Taylor expansion:

$$\mathrm{ER}(H_\varepsilon)\approx\mathrm{ER}(H_m)+\varepsilon\mathrm{IF}((x,y);\mathrm{ER},H_m)+\frac{1}{2}\varepsilon^2\,\mathrm{IF2}((x,y);\mathrm{ER},H_m). \tag{5.14}$$

In Figure 5.1, we picture $\mathrm{ER}(H_\varepsilon)$ as a function of ε. The Fisher discriminant rule is optimal at the model distribution H_m, and therefore we denote $\mathrm{ER}(H_m)=\mathrm{ER}_{\mathrm{opt}}$ throughout the text. This implies that any other discriminant rule, in particular the one based on a contaminated training sample, can never have an error rate smaller than $\mathrm{ER}_{\mathrm{opt}}$. Hence, negative values of the influence function are excluded. From the well known property that $E[\mathrm{IF}((x,y);\mathrm{ER},H_m)]=0$ (Hampel et al 1986, page 84), it follows that

$$\mathrm{IF}((x,y);\mathrm{ER},H_m)\equiv 0$$

almost surely. According to (5.14), the behaviour of the error rate under small amounts of contamination is then characterised by the *second order influence function* IF2. Note that this second order influence function should be non-negative everywhere.

In the next proposition, we derive the second order influence function for the error rate. The obtained expression is quite complex, and depends on populations quantities of the model H_m, and on the influence functions of the location and scatter functionals used. At a p-dimensional distribution F, these influence

[2] Note that our definition of higher order influence function differs from the one used in Gatto and Ronchetti (1996).

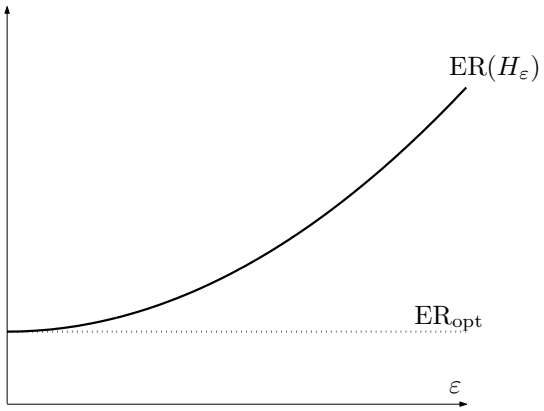

Figure 5.1: *Error rate of a discriminant rule based on a contaminated model distribution as a function of the amount of contamination ε.*

functions are denoted by $\mathrm{IF}(x;T,F)$ and $\mathrm{IF}(x;C,F)$. We will need to evaluate them at the normal distributions $H_j \sim N(\mu_j,\Sigma)$. For the functionals associated with sample averages and covariances we have $\mathrm{IF}(x;T,H_j) = x-\mu_j$ and $\mathrm{IF}(x;C,H_j) = (x-\mu_j)(x-\mu_j)^t - \Sigma$. Influence functions for several robust location and scatter functionals have been computed in the literature: we will use the expressions of Croux and Haesbroeck (1999) for the Minimum Covariance Determinant (MCD) estimator, and of Lopuhaä (1989) for S-estimators. For definitions of these estimators, we refer to Rousseeuw (1985) for the MCD, and to Davies (1987) for multivariate S-estimators. In this chapter, we use the 25% breakdown point versions of these estimators, with a Tukey Biweight loss function for the S-estimator.

Proposition 5.2. *At the model distribution H_m verifying* **(M)**, *the influence function of the error rate of the Fisher discriminant rule (with $s=1$) is zero, and* $\mathrm{IF2}((x,y);\mathrm{ER},H_m)$ *equals*

$$\sum_{j=1}^{g'-1} \pi_j \phi\left(\frac{\theta_j}{\Delta_j} - \frac{\Delta_j}{2}\right)\Delta_j \tag{5.15}$$

$$\Big\{\Big[\frac{\mathrm{IF}((x,y);A_j,H_m)}{\Delta_j} + \left(\frac{\mu_j+\mu_{j+1}}{2} - \frac{\theta_j(\mu_{j+1}-\mu_j)}{\Delta_j^2}\right)^t \frac{\mathrm{IF}((x,y);B_j,H_m)}{\Delta_j}\Big]^2$$

$$+\frac{\mathrm{IF}((x,y);B_j,H_m)^t}{\Delta_j}\left[\Sigma - \left(\frac{\mu_{j+1}-\mu_j}{\Delta_j}\right)\left(\frac{\mu_{j+1}-\mu_j}{\Delta_j}\right)^t\right]\frac{\mathrm{IF}((x,y);B_j,H_m)}{\Delta_j}\Big\}$$

with A_j and B_j the functionals defined in (5.10) *and* (5.11), *Δ_j is defined in* (5.8), *and $\theta_j = \log(\pi_{j+1}/\pi_j)$ for $j=1,\ldots,g'-1$.*

The influence functions of the functionals A_j and B_j are easy to compute and given by

$$\mathrm{IF}((x,y);B_j,H_m) = \mathrm{IF}((x,y);VV^t,H_m)(\mu_{j+1}-\mu_j)+\frac{\delta_{y,j+1}-\delta_{y,j}}{\pi_y}v_1v_1^t\mathrm{IF}(x;T,H_y) \tag{5.16}$$

and

$$\begin{aligned}\mathrm{IF}((x,y);A_j,H_m) = &-\mathrm{IF}((x,y);B_j,H_m)^t\frac{\mu_j+\mu_{j+1}}{2} \\ &-\frac{1}{2\pi_y}(\delta_{y,j}+\delta_{y,j+1})(\mu_{j+1}-\mu_j)^t\Sigma^{-1}\mathrm{IF}(x;T,H_y)+\frac{\delta_{y,j+1}-\delta_{y,j}}{\pi_y}\end{aligned} \tag{5.17}$$

for $1 \le j \le g'$, and with $\delta_{y,j}$ the Kronecker symbol (so $\delta_{y,j} = 1$ for $y = j$ and zero for $y \neq j$). Furthermore, $\mathrm{IF}((x,y);VV^t,H_m) = \mathrm{IF}((x,y);V,H_m)v_1^t + v_1\mathrm{IF}((x,y);V,H_m)^t$. Finally, it is shown in the appendix that $\mathrm{IF}((x,y);V,H_m)$ equals

$$c_y(\Sigma^{-1}-v_1v_1^t)\mathrm{IF}(x;T,H_y)-\Sigma^{-1}\mathrm{IF}(x;C,H_y)v_1+\frac{v_1^tIF(x;C,H_y)v_1}{2}v_1, \tag{5.18}$$

with $c_y = (\mu_y-\bar{\mu})^tV/(V^tBV)$.

From the expressions above for the second order influence function of the error rate, one can see that the effect of an observation is bounded as soon as the IF of the location and scatter functionals are bounded. The MCD- and S-estimators have bounded influence functions, yielding a bounded $\mathrm{IF2}(\cdot;\mathrm{ER},H_m)$. The structure of the obtained expression becomes more apparent by considering the case $p = 1$. In this univariate setting, $s = 1 = \min(g-1,p)$, and the Fisher discriminant rule becomes affine equivariant. Hence we may assume, without loss of generality, that $\Sigma = 1$. The corollary below writes $\mathrm{IF2}((x,y);\mathrm{ER};H_m)$ as an explicit function of the IF of the location/scatter measures.

Corollary 5.3. *For $p = 1$ and $\Sigma = 1$, we have that $\mathrm{IF2}((x,y);\mathrm{ER};H_m)$ is given by*

$$\begin{aligned}\sum_{j=1}^{g'-1}\frac{\pi_j}{\Delta_j}\phi\left(\frac{\theta_j}{\Delta_j}-\frac{\Delta_j}{2}\right) \quad &\left\{\theta_j\mathrm{IF}(x;C,H_y)+\frac{\delta_{y,j+1}-\delta_{y,j}}{\pi_y}\right. \\ &\left.+\left[\delta_{y,j}\left(\frac{\theta_j}{\Delta_j}-\frac{\Delta_j}{2}\right)+\delta_{y,j+1}\left(-\frac{\theta_j}{\Delta_j}-\frac{\Delta_j}{2}\right)\right]\frac{\mathrm{IF}(x;T,H_y)}{\pi_y}\right\}^2.\end{aligned} \tag{5.19}$$

In Figure 5.2, we plot the IF2 in (5.19) as a function of x, and this for every possible value of y separately. The plots in the left column of the panel correspond to two groups with $\mu_1 = -0.5$, $\mu_2 = 0.5$ and $\pi_1 = \pi_2 = 0.5$, and the right column to three groups with $\mu_1 = -1, \mu_2 = 0, \mu_3 = 1$, and $\pi_1 = \pi_2 = \pi_3 = 1/3$. The first

row corresponds to Fisher discriminant analysis using the classical estimators, the second to the MCD, and the third row to the S-estimator. Note that the second order influence function is non-negative everywhere, since contamination in the training sample may only increase the error rate, given that we work with an optimal classification rule at the model.

From Figure 5.2, we see that outlying observations may have an unbounded influence on the error rate of the classical procedure. The MCD yields a bounded IF2, but we see that it is more vulnerable to inliers, as is perceived by the high peaks quite near the population centers. The S-based discriminant procedure is doing much better in this respect, having a much smaller value for the maximum influence (the so-called "gross-error sensitivity"). Moreover, its IF2 is smooth and has no jumps. Notice that extreme outliers still have a positive bounded influence on the error rate of the robust methods, even though we know that both MCD and S location and scatter estimators have a redescending influence function. This is caused by the fact that an extreme outlier in the training sample will still have an effect on the estimates of the prior probabilities estimates in (5.5). These above findings hold for both two and three groups. In the three groups case we also see that outliers being allocated to the second group (indicated by the dotted line), have, in general, a higher value for the influence function. An explanation is that the observations in the centrally located group will affect misclassification probabilities in all groups, while observations in a more outwards located group will basically only have influence on the misclassification probabilities of two groups. In the next section we will use IF2 to compute classification efficiencies.

5.4 Asymptotic relative classification efficiencies

At finite samples, discrimination rules are estimated from a training sample, resulting in an error rate ER_n. This error rate depends on the sample, and gives the total probability of misclassification when working with the estimated discriminant functions. When sampling training data from the model H_m, the expected loss in classification performance is

$$\mathrm{Loss}_n = E_{H_m}[\mathrm{ER}_n - \mathrm{ER}_{\mathrm{opt}}]. \tag{5.20}$$

This is a measure of our expected regret, in terms of increased error rate, when using some estimated discrimination procedure (see Efron 1975). The larger the size of the training sample, the more information available for accurate discrimination, and the closer the error rate will be to the optimal one. Efron (1975, Theorem 1) showed that the expected loss decreases to zero at a rate of $1/n$. Campbell and Donner (1989, Theorem 1) extended Efron's result to multiple groups to compute the classification efficiency of multinomial w.r.t. ordinal logistic regression. O'Neill (1980) discusses the large-sample distribution of the error rate of an arbitrary estimator of the optimal classification rule. These authors did not use

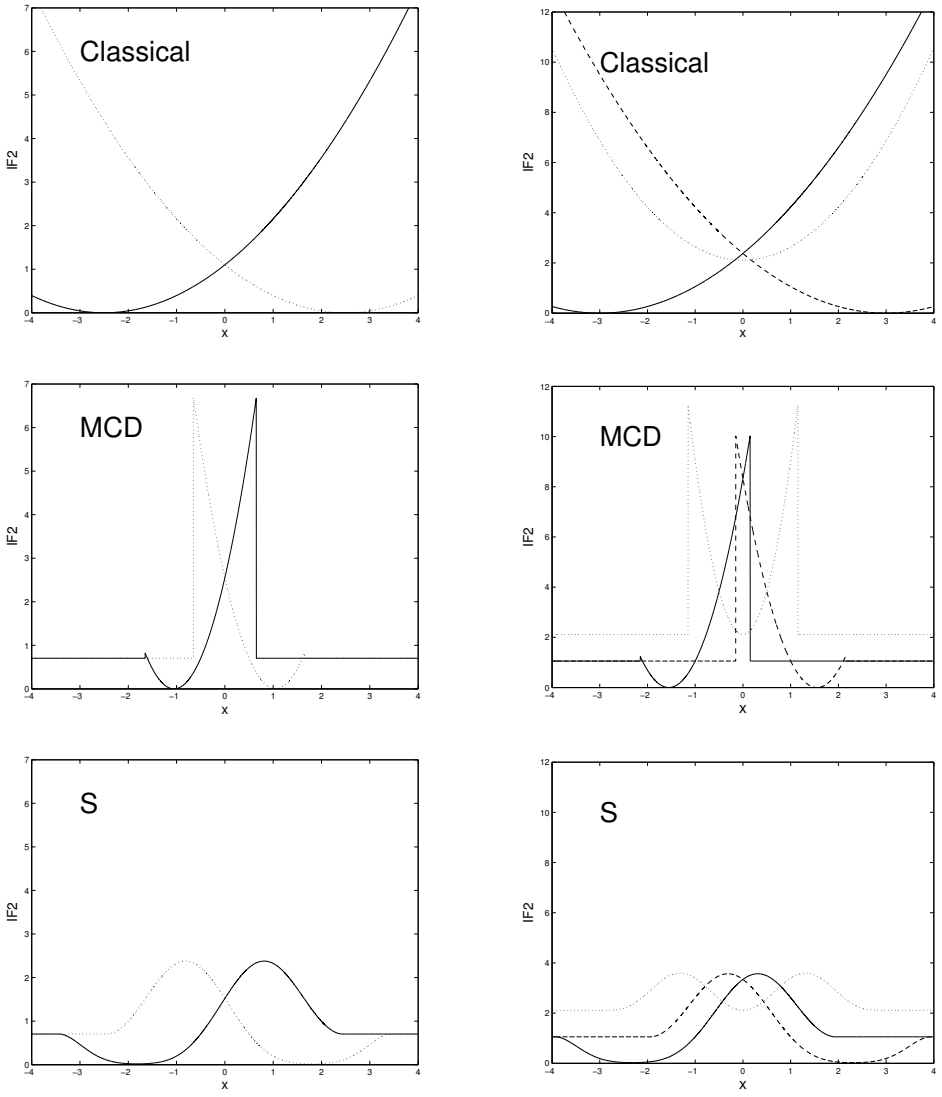

Figure 5.2: *Second order influence functions for $p = 1$ and $\Sigma = 1$, for multiple group discriminant analysis using the classical estimators (top), the MCD (middle), and S-estimators (bottom). Figures on the left correspond to two groups with $\pi_1 = \pi_2$, and on the right to three groups with $\pi_1 = \pi_2 = \pi_3$. The solid curve gives* IF2 *for an observation with $y = 1$, the dotted line for $y = 2$, and the dashed line for $y = 3$.*

influence functions, and in the following proposition we show how their results may be reformulated in terms of the expected value of the second order influence function. Some standard regularity conditions on the location/scatter estimators are needed and stated at the beginning of the proof in the Appendix.

Proposition 5.4. *At the model distribution H_m verifying* **(M)**, *we have that the expected loss in error rate of an estimated optimal discriminant rule verifies*

$$\text{Loss}_n = \frac{1}{2n} E_{H_m}[\text{IF2}((X,Y);\text{ER},H_m)] + o_p(n^{-1}). \tag{5.21}$$

The above expression (5.21) corresponds to (5.14) with $\varepsilon = 1/\sqrt{n}$, and allows to define an *Asymptotic Loss* as

$$\text{A-Loss} = \lim_{n\to\infty} n\text{Loss}_n = \frac{1}{2} E[\text{IF2}((X,Y);\text{ER},H_m)].$$

Efron (1975) proposed then to compare the classification performance of two estimators by computing *Asymptotic Relative Classification Efficiencies* (ARCE). Here, we would like to compare the loss in expected error rate using the classical procedure, Loss(Cl), with the loss of the robust Fisher's discriminant analysis, Loss(Robust). The ARCE of the robust with respect to classical Fisher's discriminant analysis is then

$$\text{ARCE(Robust, Cl)} = \frac{\text{A-Loss(Cl)}}{\text{A-Loss(Robust)}}. \tag{5.22}$$

At the model (**M**), where the different populations are normally distributed, the classical procedure uses the Maximum Likelihood estimates, and we have $0 \leq$ ARCE(Robust,Cl) ≤ 1.

In the case of $g = 2$ groups, an explicit expression for the ARCE can be obtained. For $g = 2$, we have that $s = 1 = \min(g-1, p)$ and the discriminant procedure is affine equivariant. Without loss of generality, we may assume that $\mu_1 = (-\Delta/2, \ldots, 0)^t$, $\mu_2 = -\mu_1$ and $\Sigma = I_p$. Then the following proposition holds.

Proposition 5.5. *The asymptotic loss of Fisher's discriminant analysis based on the location and scatter measures T and C, for $g = 2$ groups being normally distributed with equal covariance matrices, is given by*

$$\begin{aligned}\text{A-Loss} \;=\; & \frac{\phi(\theta/\Delta - \Delta/2)}{2\pi_2\Delta}\Big\{(p-1+\frac{\Delta^2}{4}+\frac{\theta^2}{\Delta^2}+(\pi_1-\pi_2)\theta)\,ASV(T_1) \\ & +(p-1)\Delta^2\;\pi_1\pi_2\,ASV(C_{12})+\theta^2\pi_1\pi_2\,ASV(C_{11})+1\Big\}, \end{aligned} \tag{5.23}$$

with $\Delta = \mu_2 - \mu_1$ and $\theta = \log(\pi_2/\pi_1)$. Here, $ASV(T_1)$, $ASV(C_{12})$, and $ASV(C_{11})$ stands for the asymptotic variance of, respectively, a component of T, an off-diagonal element of C, and a diagonal element of C, all evaluated at $N(0, I_p)$.

Evaluating expression (5.23), for both the robust and the classical procedure, immediately gives the asymptotic relative classification efficiencies in (5.22). We will compute the ARCE for S-estimators and for the Reweighted MCD-estimator (RMCD), both with 25% breakdown point. Note that it is common to perform a reweighing step for the MCD, in order to improve its efficiency. Asymptotic variances for the S- and RMCD-estimator are reported in Croux and Haesbroeck (1999), using results of Lopuhaä (1989, 1999). From Figure 5.4, we see how the ARCE of both estimators varies with Δ and with the log-odds ratio θ, for $p = 5$ (other values of p give similar results). First we note that the classification efficiency of both robust procedures is quite high, where the S-based method is the more efficient. Both robust discriminant rules lose some classification efficiency when the distance between the population centers increases, and this loss is more pronounced for the RMCD-estimator. On the other hand, the effect of θ on the ARCE is very limited; changing the group proportions has almost no effect on the relative performance of the different discriminant methods we considered.

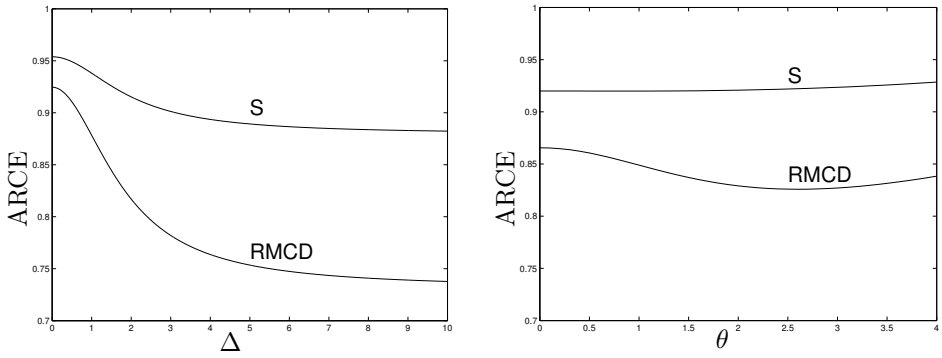

Figure 5.3: *The asymptotic relative classification efficiency of Fisher's discriminant analysis based on RMCD and S w.r.t. the classical method, for $p = 2$, as a function of Δ (left figure, for $\theta = 0$) and as a function of θ (right figure, for $\Delta = 1$).*

5.5 Simulations

The results of the previous section were derived at the population level. In a first simulation experiment we show that the derived asymptotic classification efficiencies of Section 5.4 are confirmed by finite sample results. Afterwards, we present simulation experiments where we generate training samples from models not satisfying condition **(M)**: one where the population centers are not collinear, and one where outliers were induced in the training sample. We will compare three different versions of Fisher's discrimination method: the classical method,

where sample averages and covariance matrices are used in (5.1) and (5.2), and the methods using RMCD and S-estimators. We compute them using the fast algorithms of Rousseeuw and Van Driessen (1999) for the RMCD, and Salibian-Barrera and Yohai (2005) for the S-estimator.

In a first simulation setting we generate $m = 10000$ training samples of size n according to a mixture of two normal distributions. We set $\pi_1 = \pi_2 = 0.5$, $\mu_2 = (\frac{1}{2}, 0, \ldots, 0) = -\mu_1$, and $\Sigma = I_2$. For every training sample, we compute the discriminant rule and denote the associated error rate by ER_n^k, for $k = 1, \ldots, m$. Since we know the true distribution of the data to classify, ER_n^k can be estimated without any significant error by generating a test sample from the model distribution of size 100000, and computing the empirical frequency of misclassified observations over this test sample. Since the model distribution satisfies condition **(M)**, it is possible to compute the optimal error rate according to formula (5.12). Then we can approximate the expected loss in error rate by the Monte Carlo average

$$\overline{\mathrm{Loss}}_n = \frac{1}{m}\sum_{k=1}^{m} \mathrm{ER}_n^k - \mathrm{ER}_{\mathrm{opt}} = \overline{\mathrm{ER}}_n - \mathrm{ER}_{\mathrm{opt}}. \tag{5.24}$$

The *finite sample relative classification efficiency* of the robust method with respect to the classical procedure is then given by

$$\mathrm{RCE}_n(\mathrm{Robust}, \mathrm{Cl}) = \frac{\mathrm{Loss}_n(\mathrm{Cl})}{\mathrm{Loss}_n(\mathrm{Robust})}, \tag{5.25}$$

and is estimated via Monte Carlo techniques by

$$\widehat{\mathrm{RCE}}_n(\mathrm{Robust}, \mathrm{Cl}) = \frac{\overline{\mathrm{Loss}}_n(\mathrm{Cl})}{\overline{\mathrm{Loss}}_n(\mathrm{Robust})}. \tag{5.26}$$

In Table 5.1 these efficiencies are reported for different training sample sizes[3] for dimensions $p = 2$ and $p = 5$, and for using the RMCD- and the S-estimator as robust estimators. We also added the asymptotic classification efficiency, using formula (5.23), in the row "$n = \infty$". We see from Table 1 that the finite sample results are very close to the asymptotic efficiency; only for the RMCD the convergence is somehow slower for $p = 5$. Note that the finite sample efficiencies of both robust procedures are very high. The average classification errors are reported as well. Standard errors around the reported results have been computed and are small.[4] Table 5.1 shows that for $n = 50$ there is still a gap of a few percentages

[3] The training sample size needs to be large enough to ensure that the robust high breakdown estimators can still be computed in each group. For larger dimensions, those require a large enough sample size to be computable.

[4] More precisely, for $p = 2$ standard errors around the reported average error rates are about $0.06, 0.03, 0.01\%$ for $n = 50, 100, 200$ and for $p = 5$ about $0.05, 0.02\%$ for $n = 100, 200$. The standard error around the finite sample relative classification efficiencies are for $p = 2$ about 0.007 and for $p = 5$ about 0.004

between the optimal error rate and the finite sample error rate. For $n = 200$ we are already getting very close to the optimal error rate, illustrating the fast (order n^{-1}) convergence to $\mathrm{ER}_{\mathrm{opt}}$.

In a second simulation experiment, we simulate according to a normal model H_m^* with $\mu_1 = (1, 0, \ldots, 0)^t$, $\mu_2 = (-\frac{1}{2}, \frac{\sqrt{3}}{2}, 0, \ldots, 0)^t$, $\mu_3 = (-\frac{1}{2}, -\frac{\sqrt{3}}{2}, 0, \ldots, 0)^t$, $\Sigma = I_p$, and $\pi_1 = \pi_2 = \pi_3$. This distribution does not obey condition **(M)**, since the population centers are not collinear. The centers are at equal distance $\Delta = \sqrt{3}$ from each other, which makes it possible to derive an explicit expression for the optimal error rate. It is not difficult to verify that

$$\mathrm{ER}(H_m^*) = 1 + \Phi\left(\frac{\Delta}{\sqrt{3}}\right) - 2\int_{-\Delta/\sqrt{3}}^{\infty} \Phi(\sqrt{3}z + \Delta) d\Phi(z).$$

If we select $s = \min(g - 1, p) = 2$ discriminant functions, then $\mathrm{ER}(H_m^*) = \mathrm{ER}_{\mathrm{opt}}$, and we can compute finite sample relative classification efficiencies using (5.26). We do not have an expression for the A-loss if $s = 2$, hence asymptotic efficiencies are not available. From Table 5.2 we see that the error rates converge quite quickly to $\mathrm{ER}_{\mathrm{opt}}$, for the three considered methods. Clearly, the loss in error rate is more important for the higher dimensions. Due to the choice of the sampling scheme, there is no loss in discrimination power by projecting the sample onto the two-dimensional subspace spanned by the first two basis vectors. Clearly, estimating this subspace is somehow harder in a higher dimensional space. By looking at the values of the RCE_n, the very high efficiency of the S-based procedure is revealed, while the RMCD also performs well. We also see that the finite sample efficiencies are quite stable over the different sample sizes.

In Table 5.3 the results are reported by using only one discriminant function. Such an approach has the advantage of dimension reduction, but at the model $\mathrm{ER}(H_m^*)$ this leads to a loss of discrimination power. Again, we see that the error rates ER_n are quite stable over the different sample sizes, and are converging quickly to the asymptotic error rate (this convergence is a bit slower for $p = 5$) for all estimators considered. The latter error rate will be suboptimal, leading to an increased probability of misclassification of about 14% (compared to $\mathrm{ER}_{\mathrm{opt}}$) in this example. Hence the discriminant rule is not "consistent", in the sense of not being asymptotically optimal, and one cannot compute asymptotic relative efficiencies. This is comparable to the asymptotic efficiency of an estimator, which can only be compared among consistent estimators.

Finally, we illustrate the robustness of the RMCD- and S-based discriminant procedure by introducing outliers in the training sample. We generate 10% of the data according to a contaminated model H_c, being identical to model H_m^*, but with population centers being shifted to $-9 * \mu_j$, for $j = 1, \ldots, 3$. Empirical error rates are computed for $s = 2$ and $s = 1$ and need to be compared with the results from Tables 5.2 and 5.3. Table 5.4 clearly shows that the error rates of the robust procedure are only slightly affected by the outliers. The classical

Table 5.1: *Finite sample relative classification efficiencies, together with average error rates in percentages, for RMCD- and S-based discriminant analysis, for several values of n and for $p = 2, 5$. Results for $g = 2$ groups, and $\Delta = 1$.*

		Relative Efficiencies $\mathrm{RCE}_n(\mathrm{Cl},\cdot)$		Error rates $\overline{\mathrm{ER}}_n(\cdot)$		
	n	RMCD	S	Cl	RMCD	S
p=2	50	0.8732	0.9828	32.66	32.92	32.69
	100	0.8813	0.9772	31.77	31.89	31.79
	200	0.9204	0.9788	31.28	31.32	31.29
	∞	0.8783	0.9381	30.85	30.85	30.85
p=5	50	0.7977	0.9983	33.01	33.55	33.01
	100	0.8320	0.9894	31.93	32.15	31.94
	200	0.8872	0.9936	31.39	31.45	31.39
	∞	0.9219	0.9783	30.85	30.85	30.85

procedure, however, is completely misled by the outliers, and gives unacceptable high misclassification probabilities of around 64%. (Note that in the three group case, random guessing would already give an error rate of 66.67%.

5.6 Conclusions

This chapter studies classification efficiencies and robustness properties of Fisher's linear discriminant analysis. The centers and covariances appearing in the population discriminant rule can be estimated by their sample counterparts, but the theory also allows for plugging in robust estimates instead, yielding a robust discriminant procedure. Influence functions and asymptotic relative classification efficiencies were computed at a model where all groups are normally distributed with equal covariance and collinear group means. At this model, the Fisher discriminant rule is optimal. In Section 5.3 it is shown that for optimal classification rules the influence function vanishes, and that the second order influence function is the appropriate tool to use. Taking the expected value of the second order influence function allows then to compute asymptotic relative classification efficiencies. This efficiency measures the loss in classification performance (at the model) when using a robust instead of the classical procedure. It was shown that this loss remains very limited, if one uses efficient robust estimators of location and scatter like RMCD- and S-estimators. If outliers are present, the robust method completely outperforms the Fisher rule based on sample averages and covariances.

Table 5.2: *Finite sample relative classification efficiencies, together with average error rates in percentages, for RMCD- and S-based discriminant analysis, for several values of n and for $p = 2, 5$. Results for a setting with $g = 3$ groups, and $s = 2$.*

		Relative Efficiencies $\mathrm{RCE}_n(\mathrm{Cl},\cdot)$		Error rates $\overline{\mathrm{ER}}_n(\cdot)$		
	n	RMCD	S	Cl	RMCD	S
p=2	50	0.8790	0.9995	32.48	32.77	32.48
	100	0.8633	0.9897	31.41	31.58	31.42
	200	0.8898	0.9864	30.90	30.96	30.90
	∞			30.35	30.35	30.35
p=5	100	0.8757	0.9689	35.53	36.27	35.70
	200	0.8614	0.9650	33.88	34.45	34.01
	∞			30.35	30.35	30.25

Table 5.3: *Finite sample average error rates in percentages, for the same sampling scheme as in Table 5.2, but with $s = 1$.*

		Error rates		
	n	$\overline{\mathrm{ER}}_n(\mathrm{Cl})$	$\overline{\mathrm{ER}}_n(\mathrm{RMCD})$	$\overline{\mathrm{ER}}_n(\mathrm{S})$
p=2	50	47.19	47.23	47.25
	100	46.63	46.64	46.65
	200	46.28	46.22	46.28
	∞	44.33	44.33	44.33
p=5	100	49.08	49.29	49.20
	200	47.99	48.27	48.09
	∞	44.33	44.33	44.33

Table 5.4: *Finite sample average error rates in percentages, for the same sampling scheme as in Table 5.2 and 5.3, with* $p = 2$*, but with* 10% *of outliers introduced in the training sample. Results are given for* $s = 2$ *and* $s = 1$*.*

		Error rates		
	n	$\overline{\text{ER}}_n$(Cl)	$\overline{\text{ER}}_n$(RMCD)	$\overline{\text{ER}}_n$(S)
s=2	50	62.94	34.87	39.42
	100	64.45	31.55	34.82
	200	64.97	30.89	31.71
s=1	50	62.31	46.90	47.40
	100	63.91	46.68	46.97
	200	64.78	46.24	46.59

For the two-group case, influence functions for the error rate of linear discriminant analysis were already computed by Croux and Dehon (2001) and for quadratic discriminant analysis by Croux and Joossens (2005). However, they used a non-optimal classification rule, by omitting the penalty term in (5.4), leading to essentially different expressions for the influence function (in particular, the first order IF will not vanish); they also did not consider classification efficiencies. A next challenge would be to compute asymptotic classification efficiencies for the multiple group case with non-collinear centers. However, in the general setting, no tractable expression for the error rate is available. One might fear that it will not be possible to obtain theoretical results here, and that only simulations and numerical experiments (as those reported in Section 5.5) are possible.

Appendix

Description of the procedure for ordering the group labels: We will drop the dependency on H in the notation. Since $s = 1$, it follows from (5.5) that $D_j^2(x) - cx_1^2 = b_j x_1 + a_j$, with $b_j = -2T_j^t V$, $a_j = (V^t T_j)^2 - 2\log\pi_j$, for $j = 1, \ldots, g$, and with $x_1 = V^t x$. The minimum of the discriminant scores can thus be found by minimising a set of g linear functions in x_1. The resulting minimum, denoted here by $f(x_1)$, will be piecewise linear. Let now $s_1 = -\infty < s_2 < \ldots < s_{g'} < s_{g'+1} = \infty$ such that f is linear on every interval $]s_j, s_{j+1}[$ for $1 \leq j \leq g'$. We will relabel now the groups in such a way that $D_j^2(x) \equiv f(x_1)$ on the intervals $]s_j, s_{j+1}[$. Moreover, it is not difficult to see that $s_j < s_{j+1}$ implies $b_j > b_{j+1}$, for $j = 1, \ldots, g' - 1$. It is then clear that $R_j = \{x \in \mathbb{R}^p \mid \min_k D_k^2(x) = D_j^2(x)\}$, for $1 \leq j \leq g'$. If a function $b_j x_1 + a_j$ is not corresponding to any of the intervals on which f in linear, then the label j needs to be set larger than g', and $R_j = \emptyset$.

To conclude, we will order the groups with respect to decreasing values of b_j, or increasing values of $V^t T_j$, and remove the indices j corresponding to empty regions R_j. □

Proof of Proposition 5.1: We will use the notation of the above description of the procedure to order the group labels. Let $(X, Y) \sim H_m$. First note that if $Y = j$, with $j > g'$, then $R_j = \emptyset$ and the observation will always be misclassified. This explains the presence of the last term in (5.9). Now for $1 \leq j \leq g'$, denote $\Pi_j^R = P(V^t X > s_{j+1} | Y = j)$ and $\Pi_j^L = P(V^t X < s_j | Y = j)$. Then the probability that an observation coming from one of the first g' groups is misclassified is given by

$$\sum_{j=1}^{g'-1} \pi_j \Pi_j^R + \sum_{j=2}^{g} \pi_j \Pi_j^L.$$

Now for $1 \leq j \leq g' - 1$, we have

$$\begin{aligned}
\Pi_j^R &= P_{H_m}\big(b_j(V^t X) + a_j > b_{j+1}(V^t X) + a_{j+1} | Y = j\big) \\
&= P_{H_j}(-2(T_j - T_{j+1})^t V V^t [X - (T_j + T_{j+1})/2] > 2\log(\pi_j/\pi_{j+1}) \mid Y = j) \\
&= P_{H_j}(-B_j^t X < A_j) \\
&= P\Big(Z < \frac{A_j + B_j^t \mu_j}{\sqrt{B_j^t \Sigma B_j}} \mid Z \sim N_p(0, I_p)\Big) = \Phi\Big(\frac{A_j + B_j^t \mu_j}{\sqrt{B_j^t \Sigma B_j}}\Big),
\end{aligned}$$

where $A_j = A_j(H)$ and $B_j = B_j(H)$ are defined in (5.10) and (5.11). Similarly

$$\Pi_j^L = \Phi\Big(\frac{-A_{j-1}(H) - B_{j-1}^t(H)\mu_j}{\sqrt{B_{j-1}^t(H)\Sigma B_{j-1}(H)}}\Big).$$

Collecting terms yields the result. □

Proof of Proposition 5.2: We fix (x, y) and denote $H_\varepsilon = (1-\varepsilon)H_m + \varepsilon\Delta_{(x,y)}$. To compute IF and IF2, we need to compute the first and second order derivative of $\mathrm{ER}(H_\varepsilon)$. Expression (5.9) can be structured as

$$\mathrm{ER}(H_\varepsilon) = \sum_{j=1}^{g'-1} [\pi_j \Pi_j^R(H_\varepsilon) + \pi_{j+1} \Pi_{j+1}^L(H_\varepsilon)] + \sum_{j=g'+1}^{g} \pi_j. \tag{5.27}$$

Since the last term in the above expression is constant, it will not infer in the expression for the influence function. We will also use the functionals $E_j = A_j(B_j^t \Sigma B_j)^{-1/2}$ and $F_j = B_j(B_j^t \Sigma B_j)^{-1/2}$, where we drop the dependency on H.

Throughout this proof, we also use that at the model, that is for $\varepsilon = 0$, the following identities hold: $\beta_j := B_j(H_m) = \Sigma^{-1}(\mu_{j+1} - \mu_j) = v_1 \Delta_j$ and $\alpha_j :=$

$A_j(H_m) = \theta_j - \beta_j^t(\mu_j + \mu_{j+1})/2$. Furthermore $\beta_j^t\Sigma\beta_j = \Delta_j^2$, $E_j(H_m) = \alpha_j/\Delta_j$ and $F_j(H) = \beta_j/\Delta_j$ such that, for $j = 1, \ldots, g-1$

$$\Pi_j^R(H_m) = \Phi(\frac{\theta_j}{\Delta_j} - \frac{\Delta_j}{2}) \qquad \text{and} \qquad \Pi_{j+1}^L(H_m) = \Phi(-\frac{\theta_{j+1}}{\Delta_{j+1}} - \frac{\Delta_{j+1}}{2}).$$

Before continuing we need the following lemmas. We use the shorthand notation $\mathrm{IF}(\cdot) = \mathrm{IF}((x,y); \cdot, H_m)$

Lemma:

(i) $\mathrm{IF}(E_j) = \mathrm{IF}(A_j)/\Delta_j - \alpha_j\beta_j^t\Sigma\ \mathrm{IF}(B_j)/\Delta_j^3$.

(ii) $\mathrm{IF}(F_j) = (I_p - \beta_j\beta_j^t\Sigma/\Delta_j^2)\ \mathrm{IF}(B_j)/\Delta_j$.

(iii) $\mathrm{IF}(F_j)^t(\mu_{j+1} - \mu_j) = 0$.

(iv) $\mathrm{IF2}(F_j)^t(\mu_{j+1} - \mu_j) = -\Delta_j \frac{\mathrm{IF}(B_j)^t}{\Delta_j}\{\Sigma - \left(\frac{\mu_{j+1}-\mu_j}{\Delta_j}\right)\left(\frac{\mu_{j+1}-\mu_j}{\Delta_j}\right)^t\}\frac{\mathrm{IF}(B_j)}{\Delta_j}$.

(v) $\pi_j\phi(\theta_j/\Delta_j - \Delta_j/2) = \pi_{j+1}\phi(-\theta_j/\Delta_j - \Delta_j/2)$.

Proof:
(i) and (ii) can be obtained via straightforward derivation. By definition of F_j, we have $F_j^t(H_\varepsilon)\Sigma F_j(H_\varepsilon) = 1$ for all H_ε. From the latter it follows that

$$\Big(\frac{\partial}{\partial\varepsilon}F_j^t(H_\varepsilon)\Big)\Sigma F_j(H_\varepsilon) = 0, \tag{5.28}$$

for any $\varepsilon > 0$. Evaluating (5.28) at $\varepsilon = 0$ results in (iii). Deriving (5.28) once more w.r.t. ε and evaluating at $\varepsilon = 0$ results in

$$\mathrm{IF2}(F_j)^t(\mu_{j+1} - \mu_j)/\Delta_j = -\mathrm{IF}(F_j)^t\Sigma\ \mathrm{IF}(F_j). \tag{5.29}$$

Since

$$\Big(I_p - \frac{\beta_j\beta_j^t\Sigma}{\Delta_j^2}\Big)^t\Sigma\Big(I_p - \frac{\beta_j\beta_j^t\Sigma}{\Delta_j^2}\Big) = \Sigma - \frac{\Sigma\beta_j\beta_j^t\Sigma}{\Delta_j^2} = \Sigma - \Big(\frac{\mu_{j+1}-\mu_j}{\Delta_j}\Big)\Big(\frac{\mu_{j+1}-\mu_j}{\Delta_j}\Big)^t,$$

(iv) follows. Finally (v) follows from

$$\log\frac{\phi(\theta_j/\Delta_j - \Delta_j/2)}{\phi(-\theta_j/\Delta_j - \Delta_j/2)} = \Big(-\frac{\theta_j}{\Delta_j} - \frac{\Delta_j}{2}\Big)^2/2 - \Big(\frac{\theta_j}{\Delta_j} - \frac{\Delta_j}{2}\Big)^2/2 = \theta_j = \log\frac{\pi_{j+1}}{\pi_j}$$

which ends the proof of the Lemma. □

The first order derivative of $\pi_j \Pi_j^R(H_\varepsilon) + \pi_{j+1}\Pi_{j+1}^L(H_\varepsilon)$ equals now, for $1 \le j \le g' - 1$,

$$\begin{aligned}
&\pi_j \phi\Big(\frac{\theta_j}{\Delta_j} - \frac{\Delta_j}{2}\Big) \frac{\partial}{\partial \varepsilon} \left[E_j(H_\varepsilon) + F_j^t(H_\varepsilon)\mu_j\right]\Big|_{\varepsilon\,=\,0} \\
&+\pi_{j+1}\phi\Big(-\frac{\theta_j}{\Delta_j} - \frac{\Delta_j}{2}\Big) \frac{\partial}{\partial \varepsilon} \left[-E_j(H_\varepsilon) - F_j^t(H_\varepsilon)\mu_{j+1}\right]\Big|_{\varepsilon\,=\,0} \\
&= -\pi_j \phi\Big(\frac{\theta_j}{\Delta_j} - \frac{\Delta_j}{2}\Big) \mathrm{IF}(F_j)^t(\mu_{j+1} - \mu_j) \\
&= 0
\end{aligned}$$

using lemma (iii) and (v). This implies that $\mathrm{IF}((x,y); \mathrm{ER}, H_m) = 0$. The second order derivative of $\pi_j \Pi_j^R(H_\varepsilon) + \pi_{j+1}\Pi_{j+1}^L(H_\varepsilon)$ at $\varepsilon = 0$ equals

$$\begin{aligned}
&\pi_j \phi'\Big(\frac{\theta_j}{\Delta_j} - \frac{\Delta_j}{2}\Big) \Big[\frac{\partial}{\partial \varepsilon}(E_j(H_\varepsilon) + F_j^t(H_\varepsilon)\mu_j)\Big|_{\varepsilon\,=\,0}\Big]^2 \\
&+\pi_{j+1}\phi'\Big(-\frac{\theta_j}{\Delta_j} - \frac{\Delta_j}{2}\Big) \Big[\frac{\partial}{\partial \varepsilon}(-E_j(H_\varepsilon) - F_j^t(H_\varepsilon)\mu_{j+1}\Big|_{\varepsilon\,=\,0}\Big]^2 \\
&+\pi_j \phi\Big(\frac{\theta_j}{\Delta_j} - \frac{\Delta_j}{2}\Big) \frac{\partial^2}{\partial \varepsilon^2}(E_j(H_\varepsilon) + F_j^t(H_\varepsilon)\mu_j)\Big|_{\varepsilon\,=\,0} \\
&+\pi_{j+1}\phi\Big(-\frac{\theta_j}{\Delta_j} - \frac{\Delta_j}{2}\Big) \frac{\partial^2}{\partial \varepsilon^2}(-E_j(H_\varepsilon) - F_j^t(H_\varepsilon)\mu_{j+1})\Big|_{\varepsilon\,=\,0}.
\end{aligned}$$

Using $\phi'(u) = -u\phi(u)$ this can be written as

$$\begin{aligned}
&-\pi_j \Big(\frac{\theta_j}{\Delta_j} - \frac{\Delta_j}{2}\Big) \phi\Big(\frac{\theta_j}{\Delta_j} - \frac{\Delta_j}{2}\Big) [\mathrm{IF}(E_j) + \mathrm{IF}(F_j)^t\mu_j]^2 \\
&+\pi_{j+1}\Big(\frac{\theta_j}{\Delta_j} + \frac{\Delta_j}{2}\Big) \phi\Big(-\frac{\theta_j}{\Delta_j} - \frac{\Delta_j}{2}\Big) [\mathrm{IF}(E_j) + \mathrm{IF}(F_j)^t\mu_{j+1}]^2 \\
&+\pi_j \phi\Big(\frac{\theta_j}{\Delta_j} - \frac{\Delta_j}{2}\Big) [\mathrm{IF2}(E_j) + \mathrm{IF2}(F_j)^t\mu_j] \\
&+\pi_{j-1}\phi\Big(-\frac{\theta_j}{\Delta_j} - \frac{\Delta_j}{2}\Big) [-\mathrm{IF2}(E_j) - \mathrm{IF2}(F_j)^t\mu_{j+1}].
\end{aligned}$$

Using lemmas (iii) and (v) the above equation reduces to $\pi_j \phi(\theta_j/\Delta_j - \Delta_j/2)$ times

$$\begin{aligned}
&-\Big(\frac{\theta_j}{\Delta_j} - \frac{\Delta_j}{2}\Big) [\mathrm{IF}(E_j) + \mathrm{IF}(F_j)^t\mu_j]^2 + \Big(\frac{\theta_j}{\Delta_j} + \frac{\Delta_j}{2}\Big) [\mathrm{IF}(E_j) + \mathrm{IF}(F_j)^t\mu_{j+1}]^2 \\
&-\mathrm{IF2}(F_j)^t(\mu_{j+1} - \mu_j) \\
&= \Delta_j[\mathrm{IF}(E_j) + \mathrm{IF}(F_j)^t(\mu_j + \mu_{j+1})/2]^2 - \Delta_j[\mathrm{IF2}(F_j)^t(\mu_j - \mu_{j+1})/\Delta_j].
\end{aligned}$$

The above expression together with (5.27) results in (5.15). □

Proof of equation (5.18) **for IF**$((x,y);V,H_m)$**:** At the model H_m, let λ_1 be the largest eigenvalue of the matrix $\mathcal{W}^{-1}\mathcal{B}$ and denote $v_2,\ldots,v_p$ for the eigenvectors corresponding to the null eigenvalues. The influence function of the functional V_1, being the first eigenvector of the matrix $W^{-1}B$ is

$$\text{IF}((x,y);V_1,H_m)=\frac{1}{\lambda_1}\sum_{k=2}^{p}(v_k^t\text{IF}((x,y);W^{-1}B,H_m)v_1)v_k-\frac{1}{2}(v_1^t\text{IF}(x;C,H_y)v_1)v_1. \tag{5.30}$$

(See Lemma 3 of Croux and Dehon, 2002, for the influence function of the eigenvectors of a non-symmetric matrix.) Using the fact that

$$\text{IF}((x,y);W^{-1}B,H_m)=\mathcal{W}^{-1}\text{IF}((x,y);B,H_m)-\mathcal{W}^{-1}\text{IF}((x,y);W,H_m)\mathcal{W}^{-1}\mathcal{B},$$

it is easy to see that the $\text{IF}((x,y);V_1,H_m)$ can be written as

$$\frac{1}{\lambda_1}\sum_{k=2}^{p}v_k^t(\text{IF}((x,y);B,H_m)-\lambda_1\text{IF}(x;C,H_y)v_1)v_k-\frac{1}{2}v_1^t\text{IF}(x,C,H_y)v_1v_1. \tag{5.31}$$

Now, it is not difficult to verify that $\text{IF}((x,y);B,H_m)$ equals

$$(\mu_y-\bar{\mu})(\mu_y-\bar{\mu})^t-\mathcal{B}+\text{IF}(x;T,H_y)(\mu_y-\bar{\mu})^t+(\mu_y-\bar{\mu})\text{IF}(x;T,H_y)^t.$$

Since the eigenvectors $v_2,\ldots,v_k$ are perpendicular to $\mu_y-\bar{\mu}$, (5.31) simplifies to

$$c_y\sum_{k=2}^{p}(v_k^t\text{IF}(x;T,H_y))v_k-\sum_{k=2}^{p}(v_k^t\text{IF}(x;C,H_y)c_1)v_k-(v_1^t\text{IF}(x;C,H_y)v_1)v_1/2,$$

with $c_y=(\mu_y-\bar{\mu})^tv_1/\lambda_1$. The nice property that $\Sigma^{-1}=\sum_{k=1}^{p}v_kv_k^t$ and the fact that $v_1^tBv_1=\lambda_1$ yields the equations (5.18). □

Proof of Proposition 5.4: Collect the estimates of location and scatter being used to construct the discriminant rule in a vector $\hat{\theta}_n$ and denote Θ the corresponding functional. Suppose that $\text{IF}((X,Y);\Theta,H_m)$ exists and that $\hat{\theta}_n$ is consistent and asymptotically normal with

$$\lim_{n\to\infty}n\text{Cov}(\hat{\theta})=\text{ASV}(\hat{\theta}_n)=E_{H_m}[\text{IF}((X,Y);\Theta,H_m)\text{IF}((X,Y);\Theta,H_m)^t]. \tag{5.32}$$

Evaluating (5.9) at the empirical distribution function $H=H_n$, gives $\text{ER}_n=\text{ER}(H_n)=g(\hat{\theta}_n)$, for a certain (complicated) function g. Denote θ_0 the true parameter, for which $g(\theta_0)=\text{ER}_{\text{opt}}$. Since θ_0 corresponds to a minimum of g, the derivative of g evaluated at θ_0 equals zero. A Taylor expansion of g around θ_0 yields then

$$\text{ER}_n=\text{ER}_{\text{opt}}+\frac{1}{2}(\hat{\theta}_n-\theta_0)^tH_g(\hat{\theta}_n-\theta_0)+o_p(\|\hat{\theta}_n-\theta_0\|^2),$$

with H_g the Hessian matrix of g at θ_0. It follows that

$$\begin{aligned} nE[\mathrm{ER}_n - \mathrm{ER}_{\mathrm{opt}}] &= \frac{1}{2}E\Big[\Big(n^{1/2}(\hat{\theta}_n - \theta_0)\Big)^t H_g\Big(n^{1/2}(\hat{\theta}_n - \theta_0)\Big)\Big] + o_p(1) \\ &= \frac{1}{2}H_g\,\mathrm{trace}E\Big[\Big(n^{1/2}(\hat{\theta}_n - \theta_0)\Big)\Big(n^{1/2}(\hat{\theta}_n - \theta_0)\Big)^t\Big] + o_p(1) \\ &= \frac{1}{2n}H_g\,\mathrm{trace}\ \mathrm{ASV}(\hat{\theta}_n) + o_p(1). \end{aligned}$$

From (5.32) and definition (5.24) we have then

$$\mathrm{Loss}_n = \frac{1}{2n}H_g\mathrm{trace}\left(E_{H_m}[\mathrm{IF}((X,Y);\Theta,H_m)\mathrm{IF}((X,Y);\Theta,H_m)^t]\right) + o_p(n^{-1}). \tag{5.33}$$

On the other hand, at the level of the functional it holds that $\mathrm{ER} \equiv g(\Theta)$, and definition (5.13) and the chain rule imply

$$\mathrm{IF2}((x,y);\mathrm{ER},H_m) = \mathrm{IF}((x,y);\Theta,H_m)^t H_g \mathrm{IF}((x,y);\Theta,H_m),$$

since $\Theta(H_m) = \theta_0$ and the derivative of g at θ_0 vanishes. Using trace properties, we get

$$E[\mathrm{IF2}((x,y);\mathrm{ER},H_m)] = H_g\mathrm{trace}\left(E_{H_m}[\mathrm{IF}((X,Y);\Theta,H_m)\mathrm{IF}((X,Y);\Theta,H_m)^t]\right). \tag{5.34}$$

Combining (5.33) and (5.34) yields the result (5.21) of proposition 5.4. □

Proof of Proposition 5.5 Without loss of generality, for the case of 2 groups, take a model H_m with $\mu_1 = -\frac{\Delta}{2}e_1$, $e_1 = (1,0,\ldots,0)^t$, $\mu_2 = \frac{\Delta}{2}e_1$ and $\Sigma = I_p$. Denote $e_2,\ldots,e_p$ the other basis vectors. The second order influence function of the error rate in (5.15) simplifies then to $\pi_1\Delta\phi(\theta/\Delta - \Delta/2)$ times

$$\Big[\frac{\mathrm{IF}((x,y);A,H_m)}{\Delta} - \frac{\theta}{\Delta}\frac{e_1^t\mathrm{IF}((x,y);B,H_m)}{\Delta}\Big]^2 + \sum_{k=2}^{p}\Big[\frac{e_k^t\mathrm{IF}((x,y);B,H_m)}{\Delta}\Big]^2.$$

Using obvious notations, we have $ASV(A) = E[\mathrm{IF}(A)^2]$, for $k = 1,\ldots,p$ $ASV(B_k) = e_k^t E[\mathrm{IF}(B)\mathrm{IF}(B)^t]e_k$, and $ASV(A,B_1) = e_1^t[\mathrm{IF}(B)\mathrm{IF}(A)]$. By a symmetry argument, $ASV(B_2) = \ldots = ASV(B_p)$. Taking the expected value of the above gives then

$$\begin{aligned} \text{A-loss} &= (\pi_1/\Delta)\phi(\theta/\Delta - \Delta/2)\{ASV(A) - (2\theta/\Delta)\,ASC(A,B_1) \\ &\quad + (\theta^2/\Delta^2)\,ASV(B_1) + (p-1)\,ASV(B_2)\}. \end{aligned} \tag{5.35}$$

At our model H_m, equations (5.16) and (5.17) become

$$\mathrm{IF}((x,y);A,H_m) = -\Delta e_1^t\mathrm{IF}(x;T,H_y)/(2\pi_y) + (\delta_{y,2} - \delta_{y,1})/\pi_y$$

and

$$\mathrm{IF}((x,y);B,H_m) = (\delta_{y,2}-\delta_{y,1})\mathrm{IF}(x;T,H_y)/\pi_y - \Delta\mathrm{IF}(x;C,H_y)e_1,$$

from which it follows

$$\begin{aligned} ASV(A) &= ((\Delta/2)^2\, ASV(T_1)+1)/(\pi_1\pi_2) \\ ASV(B_1) &= ASV(T_1)/(\pi_1\pi_2) + \Delta^2\, ASV(C_{11}) \\ ASV(A,B_1) &= -\Delta(\pi_1-\pi_2)\, ASV(T_1)/(2\pi_1\pi_2) \\ ASV(B_2) &= \Delta^2\, ASV(C_{12}) + ASV(T_1)/(\pi_1\pi_2). \end{aligned}$$

Inserting the above equations in (5.35) results in (5.23), and ends the proof. □

Chapter 6

Robust estimation of the vector autoregressive model by a least trimmed squares procedure

Co-Author: C. Croux

Summary The vector autoregressive model is very popular for modelling multiple time series. Estimation of its parameters is done by a least squares procedure. However, this estimation method is unreliable when outliers are present in the data, and therefore we propose to estimate the vector autoregressive model by using a least trimmed squares estimator. We show how the order of the autoregressive model can be determined in a robust way, and how confidence bounds around the robustly estimated impulse response functions can be constructed. The robust procedure is illustrated on two real data sets.

6.1 Introduction

The use of autoregressive models for predicting and modelling univariate time series is standard and well known. In many applications, one does not observe a single time series, but several series, possibly interacting with each other. For these multiple time series the vector autoregressive model became very popular, and is described in standard textbooks on time series (e.g. Brockwell and Davis 2003, Chapter 7; Stock and Watson 2003, Chapter 14). In this chapter we propose a robust procedure to estimate vector autoregressive models, to select their order, and to construct confidence bounds around the impulse response functions.

Let $\{y_t \mid t \in \mathbb{Z}\}$ be a p-dimensional stationary time series. The vector autoregressive model of order k, denoted by VAR(k), is given by

$$y_t = \mathcal{B}_0' + \mathcal{B}_1' y_{t-1} + \ldots + \mathcal{B}_k' y_{t-k} + \varepsilon_t, \tag{6.1}$$

with y_t a p-dimensional vector, the intercept parameter $\mathcal{B}_0'$ a vector in $\mathbb{R}^p$ and the slope parameters $\mathcal{B}_1, \ldots, \mathcal{B}_k$ being matrices in $\mathbb{R}^{p\times p}$. Throughout the chapter M' will stand for the transpose of a matrix M. The p-dimensional error terms ε_t are supposed to be independently and identically distributed with a density of the form

$$f_{\varepsilon_t}(u) = \frac{g(u'\Sigma^{-1}u)}{(\det\Sigma)^{1/2}}, \tag{6.2}$$

with Σ a positive definite matrix, called the *scatter matrix* and g a positive function. If the second moment of ε_t exists, Σ will be (proportional to) the covariance matrix of the error terms. Existence of a second moment, however, will not be required for the robust estimator. We focus on the unrestricted VAR(k) model, where no restrictions are put on the parameters $\mathcal{B}_0, \mathcal{B}_1, \ldots, \mathcal{B}_k$.

Suppose that the multivariate time series y_t is observed for $t = 1, \ldots, T$. The vector autoregressive model (6.1) can be rewritten as a multivariate regression model

$$y_t = \mathcal{B}' x_t + \varepsilon_t, \tag{6.3}$$

for $t = k+1, \ldots, T$ and with $x_t = (1, y_{t-1}', \ldots, y_{t-k}')' \in \mathbb{R}^q$, where $q = pk+1$. The matrix $\mathcal{B} = (\mathcal{B}_0', \mathcal{B}_1', \ldots, \mathcal{B}_k')' \in \mathbb{R}^{q\times p}$ contains all unknown regression coefficients. In the language of regression, $X = (x_{k+1}, \ldots, x_T)' \in \mathbb{R}^{n\times q}$ is the matrix containing the values of the explanatory variables and $Y = (y_{k+1}, \ldots, y_T)' \in \mathbb{R}^{n\times p}$ the matrix of responses, where $n = T - k$. The classical least squares estimator for the regression parameter $\mathcal{B}$ in (6.3) is given by the well known formula

$$\hat{\mathcal{B}}_{\mathrm{OLS}} = (X'X)^{-1}X'Y,$$

and the scatter matrix Σ is unbiasedly estimated by

$$\hat{\Sigma}_{\mathrm{OLS}} = \frac{1}{n-p}(Y - X\hat{\mathcal{B}}_{\mathrm{OLS}})'(Y - X\hat{\mathcal{B}}_{\mathrm{OLS}}). \tag{6.4}$$

In applied time series research, one is aware of the fact that outliers can seriously affect parameter estimates, model specification and forecasts based on the selected model. Outliers in time series can be of different natures (Fox 1972), the most well known types being additive outliers and innovational outliers. With respect to the autoregressive model (6.1), an observation y_t is an additive outlier if only its own value has been affected by contamination. On the other hand, an outlier is said to be innovational if the error term ε_t in (6.1) is contaminated.

Innovational outliers will therefore have an effect on the next observations as well, due to the dynamic structure in the series. Additive outliers have an isolated effect on the time series, but they still may seriously affect the parameter estimates.

Several procedures to detect different types of outliers for univariate time series have been proposed, e.g. Chang, Tiao and Chen (1988), and Gerlach, Carter and Kohn (1999). Bianco, García Ben, Martínez and Yohai (2001) and Riani (2004) proposed diagnostics based on robust estimators. Other robust estimators for univariate ARMA models have been proposed by Bustos and Yohai (1986) and De Luna and Genton (2001). For further references and a detailed treatment of robust univariate time series analysis we refer to Maronna, Martin and Yohai (2006, Chapter 8). While all of the above studies focus on a single series, this chapter deals with robust analysis of multivariate time series.

A common practice for handling outliers in a multivariate process is to first apply univariate techniques to the component series in order to remove the outlier, followed by treating the adjusted series as outlier-free and model them jointly. But this procedure encounters several difficulties. First, in a multivariate process, contamination in one component may be caused by an outlier in the other components. Secondly, a multivariate outlier cannot always be detected by looking at the component series separately, since it can be an outlier for the correlation structure only. Therefore it is better to cope with outliers in a multivariate framework. Tsay, Peña and Pankratz (2000) discuss the problem of multivariate outliers in detail.

The aim of this chapter is to propose a robust estimation procedure for the vector autoregressive model, the most popular model for multiple time series analysis. Not much work has been done for the robust estimation of multivariate time series. Franses, Kloek and Lucas (1999) used Generalized M-estimators, which are known to have low robustness in higher dimensions. Another approach was taken by García Ben, Martínez and Yohai (1999), using so-called *Residual Autocovariance* (RA)-estimators, being an affine equivariant version of the estimators of Li and Hui (1989). García Ben et al (1999) showed, by means of a simulation study, that the RA-estimators are resistant to outliers. Using an appropriate starting value, the RA-estimators are iteratively computed as solutions of certain estimating equations.

Our proposal for obtaining a resistant estimator for the VAR model is to replace the multivariate least squares estimator for (6.3) by a highly robust estimator. We will use the *Multivariate Least Trimmed Squares* (MLTS) estimator, introduced by Agulló, Croux and Van Aelst (2002). This estimator is defined by minimising a trimmed sum of squared Mahalanobis distances, and can be computed by a fast algorithm. The procedure also provides a natural estimator for the scatter matrix of the residuals, which can then be used for model selection criteria. This estimator is presented in Section 6.2. The robustness of the estimator is studied by means of several simulation experiments in Section 6.3, where also

a comparison with the RA-estimators is made. In Section 6.4 it is explained how to select the autoregressive order of the model in a robust way. In Section 6.5 confidence bounds around the impulse response functions obtained from the robust estimates are constructed. The robust VAR methodology is applied on real data sets in Section 6.6, while Section 6.7 contains some conclusions.

6.2 The multivariate least trimmed squares estimator

The unknown parameters of the VAR(k) will be estimated via the multivariate regression model (6.3). For this the Multivariate Least Trimmed Squares estimator (MLTS), based on the idea of the Minimum Covariance Determinant estimator (Rousseeuw and Van Driessen 1999), is used. The MLTS selects the subset of h observations having the property that the determinant of the covariance matrix of its residuals from a least squares fit, solely based on this subset, is minimal.

Consider the data set $Z = \{(x_t, y_t),\ t = k+1, \ldots, T\} \subset \mathbb{R}^{p+q}$. Let $\mathcal{H} = \{H \subset \{k+1, \ldots, T\} \mid \#H = h\}$ be the collection of all subsets of size h. For any subset $H \in \mathcal{H}$, let $\hat{\mathcal{B}}_{\text{OLS}}(H)$ be the classical least squares fit based on the observations of the subset:

$$\hat{\mathcal{B}}_{\text{OLS}}(H) = (X_H' X_H)^{-1} X_H' Y_H,$$

where X_H and Y_H are submatrices of X and Y, consisting of the rows of X, respectively Y, having an index in H. The corresponding scatter matrix estimator computed from this subset is then

$$\hat{\Sigma}_{\text{OLS}}(H) = \frac{1}{h-p}(Y_H - X_H \hat{\mathcal{B}}_{\text{OLS}}(H))'(Y_H - X_H \hat{\mathcal{B}}_{\text{OLS}}(H)).$$

The MLTS estimator is now defined as

$$\hat{\mathcal{B}}_{\text{MLTS}}(Z) = \hat{\mathcal{B}}_{\text{OLS}}(\hat{H}) \quad \text{where} \quad \hat{H} = \underset{H \in \mathcal{H}}{\operatorname{argmin}} \det \hat{\Sigma}_{\text{OLS}}(H), \tag{6.5}$$

and the associated estimator of the scatter matrix of the error terms is given by

$$\hat{\Sigma}_{\text{MLTS}}(H) = c_\alpha \hat{\Sigma}_{\text{OLS}}(\hat{H}). \tag{6.6}$$

In definition (6.6), c_α is a correction factor to obtain consistent estimation of Σ at the model distribution (6.2) of the error terms, and α the trimming proportion for the MLTS estimator, i.e. $\alpha \approx 1 - h/n$. In the case of multivariate normal error terms it has been shown (e.g. Croux and Haesbroeck 1999) that $c_\alpha = (1-\alpha)/F_{\chi^2_{p+2}}(q_\alpha)$. Here $F_{\chi^2_q}$ is the cumulative distribution function of a χ^2 distribution with q degrees of freedom, and $q_\alpha = \chi^2_{q,1-\alpha}$ is the upper α-quantile of this distribution.

Equivalent characterisations of the MLTS estimator were given by Agulló, Croux and Van Aelst (2002). They proved that any $\tilde{\mathcal{B}} \in \mathbb{R}^{p\times q}$ minimising the sum of the h smallest squared Mahalanobis distances of its residuals (subject to $\det\Sigma = 1$) is a solution of (6.5). In mathematical terms,

$$\hat{\mathcal{B}}_{\mathrm{MLTS}} = \underset{\mathcal{B},\ \Sigma;\ |\Sigma|=1}{\operatorname{argmin}} \sum_{s=1}^{h} d^2_{s:n}(\mathcal{B},\Sigma).$$

Here $d_{1:n}(\mathcal{B},\Sigma) \le \ldots \le d_{n:n}(\mathcal{B},\Sigma)$ is the ordered sequence of the residual Mahalanobis distances

$$d_s(\mathcal{B},\Sigma) = \left((y_t - \mathcal{B}'x_t)'\Sigma^{-1}(y_t - \mathcal{B}'x_t)\right)^{1/2}. \tag{6.7}$$

for $\mathcal{B} \in \mathbb{R}^{p\times q}$. Therefore, we see that the MLTS-estimator minimises the sum of the h smallest squared distances of its residuals, and is therefore the multivariate extension of the Least Trimmed Squares (LTS) estimator of Rousseeuw (1984).

Since the efficiency of the MLTS estimator is rather low, the reweighted version is used in this chapter, to improve the performance of MLTS. The Reweighted Multivariate Least Trimmed Squares (RMLTS) estimates are defined as

$$\hat{\mathcal{B}}_{\mathrm{RMLTS}} = \hat{\mathcal{B}}_{\mathrm{OLS}}(J) \quad \text{and} \quad \hat{\Sigma}_{\mathrm{RMLTS}} = c_\delta \hat{\Sigma}_{\mathrm{OLS}}(J), \tag{6.8}$$

where $J = \{j \in \{1,\ldots,n\} \mid d^2_j(\hat{\mathcal{B}}_{\mathrm{MLTS}}, \hat{\Sigma}_{\mathrm{MLTS}}) \le q_\delta\}$ and $q_\delta = \chi^2_{q,1-\delta}$. The idea is that outliers have large residuals with respect to the initial robust MLTS estimator, resulting in a large residual Mahalanobis distance $d^2_j(\hat{\mathcal{B}}_{\mathrm{MLTS}}, \hat{\Sigma}_{\mathrm{MLTS}})$. If the latter is above the critical value q_δ, then the observation is flagged as an outlier. The final RMLTS is then based on those observations not having been detected as outliers. In this chapter, we set $\delta = 0.01$ and take as trimming proportion for the initial MLTS estimator $\alpha = 25\%$.

6.3 Simulation experiments

In order to study the robustness of the estimators, we perform a simulation study comparing the OLS estimator with the robust RMLTS and the RA estimators. As in García Ben et al (1999), RA estimators are computed as iteratively reweighted maximum likelihood estimates, with a Tukey Biweight weight function (tuned to have a 95% relative asymptotic efficiency for Gaussian innovations). Since this weight function is redescending, it is important to use a robust starting value to ensure convergence to the "right" solution. In our implementation, the RMLTS was used as starting value.

We generate bivariate time series according to the VAR(2) model

$$\begin{pmatrix} y_{1,t} \\ y_{2,t} \end{pmatrix} = \begin{pmatrix} .10 \\ .02 \end{pmatrix} + \begin{pmatrix} .40 & .03 \\ .04 & .20 \end{pmatrix} \begin{pmatrix} y_{1,t-1} \\ y_{2,t-1} \end{pmatrix} + \begin{pmatrix} .100 & .005 \\ .010 & .080 \end{pmatrix} \begin{pmatrix} y_{1,t-2} \\ y_{2,t-2} \end{pmatrix} + \begin{pmatrix} \varepsilon_{1,t} \\ \varepsilon_{2,t} \end{pmatrix}, \tag{6.9}$$

where $\varepsilon_t \sim N_2(0, \Sigma)$ with

$$\Sigma = \begin{pmatrix} 1 & .2 \\ .2 & 1 \end{pmatrix}. \tag{6.10}$$

The aim is to look at the effect of the outliers on the parameter estimates. There are 10 parameters to be estimated, and we simulate total bias and total Mean Squared Error (MSE) as performance measures. The former is computed as

$$\text{Bias} = \sqrt{\sum_{i=1}^{q}\sum_{j=1}^{p}\left(\frac{1}{\text{nsim}}\sum_{s=1}^{\text{nsim}}\hat{\mathcal{B}}_{ij}^{s} - \mathcal{B}_{ij}\right)^2},$$

where $\hat{\mathcal{B}}^s$, for $s = 1, ..., \text{nsim}$, is the estimate obtained from the s-th generated series, $\mathcal{B}$ is the true parameter value and nsim= 1000 the number of simulations. The MSE is given by

$$\text{MSE} = \sum_{i=1}^{q}\sum_{j=1}^{p}\left[\frac{1}{\text{nsim}}\sum_{s=1}^{\text{nsim}}(\hat{\mathcal{B}}_{ij}^{s} - \mathcal{B}_{ij})^2\right].$$

After generating series of length T =500, according to model (6.9), m outliers will be introduced. The classical and robust estimators are used to estimate this VAR(2) model for the uncontaminated series ($m = 0$), and for the contaminated ones ($m > 0$), where several types of outliers are considered. Below we look at the effect of additive, innovational, and correlation outliers on the different estimators. Note that other types of contamination do exist, like level shifts and patches of outliers.

Additive outliers are introduced by randomly selecting m bivariate observations, and contaminate them by adding the value 10 to all the components of the selected observations. We consider different contamination levels, ranging from one single outlier up to 5% of additive outliers, i.e. $m = 25$. The Bias and MSE for the OLS, RA and RMLTS estimator are given in Table 6.1, as a function of the number m of additive outliers.

Both Bias and MSE grow for an increasing number of outliers, the increase being much faster for the non robust OLS. Using the robust estimators instead of OLS leads to a very small loss in efficiency at the model when no outliers are present. When even only one outlier is present, the RA and RMLTS are already more efficient, and this gain in MSE becomes very substantial for the larger amounts of outliers. Comparing the robust procedures, RMLTS performs slightly better as RA in this simulation setting.

Innovational outliers are generated by first randomly selecting m innovation terms ε_t in (6.9). Then add the value 10 to the first component of the innovations, yielding the contaminated innovations series ε_t^C. Bivariate series are then simulated

Table 6.1: *Simulated Bias and Mean Squared Error for the OLS, and the robust RA and RMLTS estimator of a bivariate VAR(2) model, in presence of m additive outliers in a series of length 500.*

	OLS		RA		RMLTS	
m	Bias	MSE	Bias	MSE	Bias	MSE
0	0.00	0.020	0.00	0.022	0.00	0.022
1	0.08	0.030	0.02	0.023	0.02	0.023
2	0.14	0.045	0.03	0.025	0.03	0.024
3	0.18	0.063	0.05	0.028	0.04	0.026
4	0.22	0.079	0.06	0.031	0.04	0.027
5	0.25	0.096	0.07	0.035	0.05	0.029
10	0.38	0.193	0.14	0.061	0.07	0.039
15	0.51	0.319	0.21	0.086	0.11	0.057
20	0.64	0.478	0.25	0.101	0.17	0.080
25	0.76	0.659	0.29	0.115	0.25	0.104

according to (6.9), but with ε_t replaced by ε_t^C. The Bias and MSE when estimating the uncontaminated ($m = 0$) and contaminated series are given in Table 6.2, for the classical as well as the robust estimation procedures.

The Bias and MSE for OLS grow for an increasing number of outliers, although at a smaller rate than for contamination with additive outliers. For the robust estimator we see a small decrease of the Bias and MSE, implying that the robust procedure becomes more efficient in presence of innovational outliers. This is due to the fact that an innovational outlier in the time series results in a single vertical outlier, but also in several good leverage points when estimating the autoregressive model. The robust method can cope with the vertical outlier and takes profit of the good leverage points to decrease the MSE. The OLS estimator gets biased due to the vertical outliers, but the presence of the good leverage points explains why the effect of innovational outliers is less strong than for additive outliers. Note that when no outliers are present, the RMLTS is almost as efficient as the OLS, the loss in MSE being marginal. Finally, note that the difference between the two robust approaches is not significant here, showing again that RMLTS and RA perform very similarly. Hence, the RA method does neither improves, neither deteriorates the initial MLTS estimate.

Correlation outliers are generated as innovational outliers, but instead of (6.10), we take

$$\Sigma = \begin{pmatrix} 1 & .9 \\ .9 & 1 \end{pmatrix} \tag{6.11}$$

Table 6.2: *Simulated Bias and Mean Squared error for the OLS, and the robust RA and RMLTS estimator of a bivariate VAR(2) model, in presence of m innovational outliers in a series of length 500.*

	OLS		RA		RMLTS	
m	Bias	MSE	Bias	MSE	Bias	MSE
0	0.00	0.021	0.00	0.022	0.00	0.022
1	0.02	0.022	0.00	0.021	0.00	0.021
2	0.04	0.023	0.01	0.020	0.01	0.020
3	0.06	0.025	0.01	0.019	0.01	0.019
4	0.08	0.029	0.01	0.018	0.01	0.018
5	0.10	0.033	0.01	0.018	0.01	0.018
10	0.20	0.068	0.01	0.017	0.01	0.017
15	0.30	0.123	0.01	0.016	0.01	0.016
20	0.40	0.198	0.01	0.016	0.01	0.016
25	0.49	0.289	0.01	0.017	0.01	0.016

and place the innovation outliers all at the same position $(2, -2)'$. By placing the outliers in this way, they are only outlying for the correlation structure, and not with respect to the marginal distributions of the innovations. This type of outliers strongly influences results of a (robust) univariate analysis. To illustrate this, we will estimate the VAR model (6.9) equation by equation, applying twice a univariate reweighted least trimmed squares estimator (RLTS) instead of the RMLTS. Bias and MSE when estimating the uncontaminated and contaminated series by OLS, the univariate RLTS and the multivariate RMLTS, are given in Table 6.3.

When no outliers are present, there is hardly any difference between the different estimation procedures: the robust procedures show only a marginal loss in MSE. From Table 6.3 one can see that the univariate RLTS yields comparable Bias as for OLS, growing for an increasing number of correlation outliers. On the other hand, the multivariate RMLTS approach offers protection against the correlation outliers, remaining almost without bias. As for the previous simulation scheme, the MSE tends to decrease with the number of outliers (because the latter introduce good leverage points). We conclude from this simulation experiment that a fully multivariate robust approach is necessary when estimating a VAR model.

Table 6.3: *Simulated Bias and Mean Squared error for the OLS, robust univariate (RLTS) and multivariate (RMLTS) estimators of a bivariate VAR(2) model in presence of m correlation outliers in a series of length 500.*

	OLS		RLTS		RMLTS	
m	Bias	MSE	Bias	MSE	Bias	MSE
0	0.01	0.084	0.01	0.098	0.01	0.093
1	0.01	0.074	0.01	0.088	0.01	0.083
2	0.02	0.069	0.02	0.083	0.01	0.076
3	0.02	0.056	0.02	0.074	0.01	0.069
4	0.02	0.054	0.03	0.067	0.01	0.062
5	0.03	0.046	0.03	0.065	0.01	0.059
10	0.06	0.046	0.06	0.054	0.01	0.044
15	0.08	0.043	0.08	0.049	0.01	0.037
20	0.11	0.044	0.11	0.048	0.01	0.032
25	0.14	0.049	0.14	0.053	0.01	0.030

6.4 Determining the autoregressive order

To select the order k of a vector autoregressive model, information criteria are computed for several values of k and an optimal order is selected by minimising the criterion. Most information criteria are in terms of the value of the log likelihood l_k of the VAR(k) model. Using the model assumption (6.2) for the distribution of the error terms, we get

$$l_k = \sum_{t=k+1}^{T} g(\varepsilon_t' \Sigma^{-1} \varepsilon_t) - \frac{n}{2} \log\det\Sigma,$$

with $n = T - k$. When error terms are multivariate normal the above leads to

$$l_k = -\frac{n}{2} \log\det\Sigma - \frac{np}{2}\log(2\pi) - \frac{1}{2}\sum_{t=k+1}^{T} \varepsilon_t' \Sigma^{-1} \varepsilon_t. \tag{6.12}$$

The log likelihood will depend on the autoregressive order via the estimate of the covariance matrix of the residuals. For the ordinary least squares estimator we have

$$\hat{\Sigma}_{\mathrm{OLS}} = \frac{1}{n-p} \sum_{t=k+1}^{T} \hat{\varepsilon}_t(k) \hat{\varepsilon}_t'(k),$$

where the $\hat{\varepsilon}_t(k)$ are the residuals corresponding with the estimated VAR(k) model. Using trace properties, the last term in (6.12) equals the constant $-(n-p)p/2$ for the OLS estimator. To prevent that outliers might affect the optimal selection of the information criteria, we estimate Σ by the RMLTS estimator:

$$\hat{\Sigma}_{\text{RMLTS}} = \frac{c_\delta}{m(k)-p} \sum_{t \in J(k)} \hat{\varepsilon}_t(k)\hat{\varepsilon}_t'(k),$$

with $J(k)$ as in (6.8) and $m(k)$ the number of elements in $J(k)$. The last term in (6.12) equals now $-(m(k)-p)p/(2c_\delta)$.

The most popular information criteria to select the order of the autoregressive model are of the form

$$\frac{-2}{n} l_k + h(n)\frac{(kp+1)p}{n},$$

where $(kp+1)p$ is the number of unknown parameters, which penalises for model complexity, and where $h(n)$ can take different forms. We will consider the following three criteria: the popular Akaike info criterion (Akaike 1973), corresponding to $h(n) = 2$, the Hannan-Quinn criterion (Hannan & Quinn 1979), corresponding to $h(n) = 2\log(\log(n))$ and the Schwarz criterion (Schwarz 1978, also called the Bayesian Information Criterium), for which $H(n) = \log(n)$. Other choices of $h(n)$ have been proposed in the literature (Smith and Spiegelhalter,1 1980), but we stick to the most important ones.

6.5 Impulse response function

After selecting and estimating the VAR model it is common, in particular in economics and business, to look at the Impulse Response Function (IRF), permitting to quantify variable responses to shocks on different horizon lengths (e.g. Hamilton 1994, chapter 11, or Enders 2004, Chapter 5). To define the Impulse Response Functions, let $\mathcal{B}(L)$ be the autoregressive polynomial $I_p - \mathcal{B}_1'L \ldots - \mathcal{B}_k'L^k$, where L is the lag operator defined as $Ly_t = y_{t-1}$. The VAR(k) model (6.1) can then be written as $\mathcal{B}(L)y_t = \mathcal{B}_0' + \varepsilon_t$, or $y_t = \mathcal{B}(L)^{-1}\mathcal{B}_0' + \mathcal{B}(L)^{-1}\varepsilon_t$, yielding the infinite moving average representation

$$y_t = a + \varepsilon_t + \mathcal{A}_1\varepsilon_{t-1} + \ldots + \mathcal{A}_l\varepsilon_{t-l} + \ldots$$

where a is a vector and $\mathcal{A}_1,\ \mathcal{A}_2,\ \ldots, \mathcal{A}_l,\ \ldots$ are $p \times p$ matrices, depending on the parameters in $\mathcal{B}$. The function mapping l on $(\mathcal{A}_l)_{ij}$ for $l = 0, 1, 2, 3, \ldots$, with $1 \leq i, j \leq p$, is called an impulse response function. It measures the response of component i of y_t to an impulse of one unit in component j of ε_{t-l}. In total, p^2 possible impulse response functions can be constructed. In practice, the IRF meeds to be estimated, via estimation of the underlying parameter $\mathcal{B}$. Using a

robust estimate of $\mathcal{B}$ yields a robustly estimated IRF. In the sequel, we consider two methods to construct confidence bounds around the estimated IRFs.

To construct analytic confidence bounds, we start from the asymptotic normality of the estimators in the multivariate regression model (6.3):

$$\sqrt{T}(\text{vec}\hat{\mathcal{B}} - \text{vec}\mathcal{B}) \to N_{p(pk+1)}(0, d_p\Sigma \otimes Q^{-1}). \tag{6.13}$$

Here $Q = E[x_t x_t']$, with x_t as in Section 6.1, "vec" is the operator which vectorizes a matrix and $\otimes$ stands for the Kronecker product. The constant d_p depends on the chosen estimator. For the OLS estimator we have $d_p = 1$, and for the RMLTS the value of d_p will be larger than 1 and can be retrieved from the asymptotic variance of the RMLTS estimator for the multivariate regression model (Agulló, Croux and Van Aelst 2002). The constant d_p will not only depend on the dimension, but also on the trimming fraction α of the initial MLTS estimator and on the value of δ used in the reweighting step:

$$d_p = d_p(\alpha, \delta) = \frac{c^2}{F_\alpha} + \frac{F_\delta}{(1-\delta)^2} + \frac{c}{1-\delta}\frac{F_\delta}{F_\alpha},$$

with $c = 1 - F_\delta/(1-\delta)$, $F_\delta = F_{\chi^2_{p+2}}(\chi^2_{p,1-\delta})$ and $F_\alpha = F_{\chi^2_{p+2}}(\chi^2_{p,1-\alpha})$.

By using the Delta method, as in Hamilton (1994, page. 186), we get from (6.13) that

$$\sqrt{T}(\text{vec}\mathcal{A}_l(\hat{\mathcal{B}}) - \text{vec}\mathcal{A}_l(\mathcal{B})) \to N_{p^2}(0, d_p G_l(\Sigma \otimes Q^{-1})G_l'),$$

where

$$G_l = \frac{\partial \text{vec}A_l(\mathcal{B})}{\partial(\text{vec}\mathcal{B})'}.$$

Standard errors around the values $(\mathcal{A}_l)_{ij}$ of the IRFs are then obtained as the square roots of the diagonal elements of $d_p\hat{G}_l(\hat{\Sigma} \otimes \hat{Q}^{-1})\hat{G}_l'/T$, for $l = 0, 1, 2, \ldots$ Here we take $\hat{\Sigma}$ as in (6.8), and $\hat{Q} =$ average $\{x_t x_t'; t \in J\}$ with J the set of indices used in the definition of the RMLTS estimator. Furthermore, G_l can be calculated recursively via the formula

$$G_l = \sum_{s=1}^{l} \left[\mathcal{A}_{s-1} \otimes (O_{n1}\ \mathcal{A}_{l-s}\ \mathcal{A}_{l-s-1}\ \ldots \mathcal{A}_{l-s-k-1})\right],$$

where O_{n1} is a zero matrix of size $n \times 1$ (e.g. Hamilton 1994, p. 337). The estimate $\hat{G}_l$ is then simply obtained by replacing the matrices $\mathcal{A}_l$ in the above expression by $\mathcal{A}_l(\hat{\mathcal{B}})$.

We can also obtain Monte Carlo confidence bounds using a parametric bootstrap procedure. We first estimate the model from the original data. Then we

generate 1000 series according to the estimated VAR(k) model, with errors following a multivariate normal distribution with mean zero and covariance matrix $\hat{\Sigma}$. From these 1000 generated series, 1000 impulse response values can be computed. By sorting these values and taking the 2.5% and 97.5% quantile, 95% confidence bounds can be constructed for the impulse response functions.

6.6 Examples

As a first example, we consider the bivariate time series of *maturity rates* (Tsay 2002, p. 324–325). The first series "GS1" is the 1-year Treasury constant maturity rate, and the second series "N3" is the 3-year Treasury constant maturity rate. The data are monthly and sampled from April 1953 to January 2001. As in the book of Tsay (2002), we work with the log-transformed version of both series. From the plot of the series (Figure 6.1), it can be seen that there might be some outliers around the years 1954 and 1958.

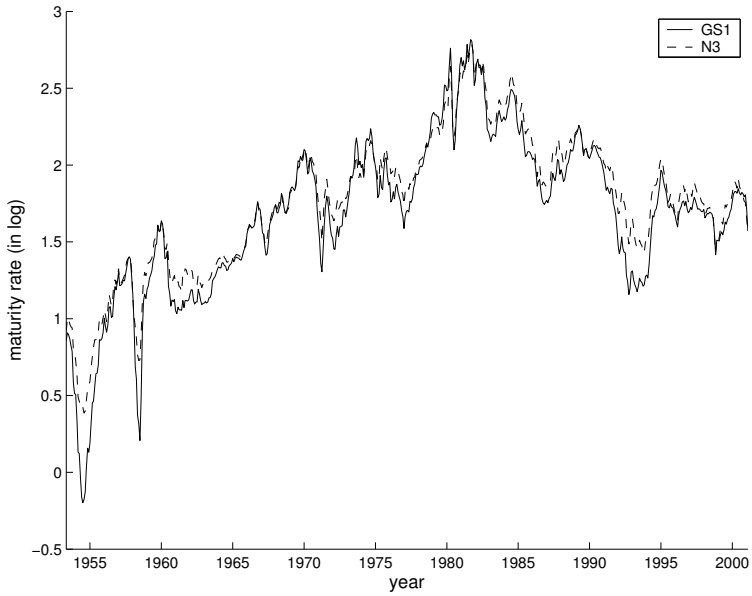

Figure 6.1: *Time plot of the "maturity rate" series. The solid line represents the 1-Year Treasury constant maturity rate and the dashed line the 3-Year Treasury constant maturity rate, both in logs.*

In Table 6.4 different lag length criteria, as discussed in Section 6.4, are presented, once based on the OLS estimator, and once based on the RMLTS. The information criteria clearly depend on the chosen estimator. For example, when using the AIC the classical method suggests a VAR(8) model while the robust indicates a VAR(6) model. On the other hand the Schwarz criterion selects an optimal order 3 for both estimators. Since the latter criterion yields a consistent estimate of the optimal order (Hannan 1980) we continue the analysis with $k = 3$.

Table 6.4: *Lag length criteria using the OLS and RMLTS estimator for the "maturity rate" series.*

k	1	2	3	4	5	6	7	8
Based on OLS estimation								
AIC	-7.3552	-7.5823	-7.6179	-7.6261	-7.6149	-7.6078	-7.6268	**-7.6276**
HQ	-7.3374	-7.5526	**-7.5763**	-7.5726	-7.5494	-7.5302	-7.5372	-7.5258
SC	-7.3096	-7.5062	**-7.5113**	-7.4889	-7.4470	-7.4090	-7.3972	-7.3669
Based on RMLTS estimation								
AIC	-7.4386	-7.6282	-7.6795	-7.6997	-7.6943	**-7.7490**	-7.6961	-7.7118
HQ	-7.4208	-7.5985	-7.6380	-7.6461	-7.6288	**-7.6714**	-7.6065	-7.6101
SC	-7.3930	-7.5522	**-7.5730**	-7.5624	-7.5264	-7.5502	-7.4665	-7.4512

After estimating the VAR(3) model with the robust RMLTS estimator, the corresponding robust residual distances $d_t(\hat{\mathcal{B}}_{\text{RMLTS}}, \hat{\Sigma}_{\text{RMLTS}})$ are computed as in (6.7), for $t = k+1, \ldots, T$. Figure 6.2 displays these distances with respect to the time index, and high residual distances indicate outlying observations. It is important to compute these distances based on the robust RMLTS, in order to avoid the well-known masking effect. Furthermore, it is common to compare these distances with a critical value from the chi-square distribution with p degrees of freedom, and we took $\chi_{p,0.99}$ (Rousseeuw and Van Zomeren 1990). Figure 6.2 reveals that several suspectable high residuals are detected, in particular around the years 1954 and 1958. But there are also a couple of other, less extreme outliers, which are more difficult to retrieve from the time series plot in Figure 6.1. Due to the presence of outliers, it is appropriate to make use of robust methods for further analysis of this data set.

The robustly estimated impuls respons functions for the maturity rate series, together with their confidence bounds, are presented in Figure 6.3. Both analytic (Figure 6.3a) and Monte Carlo confidence intervals (Figure 6.3b) around the IRFs are computed for this example. We see that the impulse response functions for the analytic method and the Monte Carlo method are very similar. The Monte Carlo based 95% confidence bounds are somehow larger and less smooth in comparison with the analytic confidence bounds. Of course, the Monte Carlo bounds require more computing time. It is seen from Figure 6.3 that the effect of a unit shock at the innovation of GS1 on the response of GS1 is significant (meaning significantly

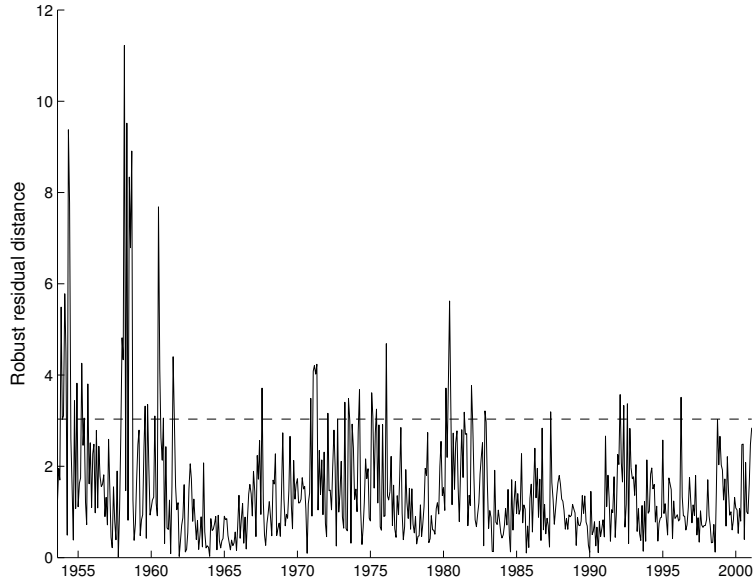

Figure 6.2: *Robust residual distances for the "maturity rate" series, based on RMLTS estimator of a VAR(3) model. The dashed line represents the critical value at the 1% level.*

different from zero) up to 11 months. On the other hand the variable N3 is non-responsive to such a shock. The effects of a unit shock in the innovation series driving N3 are more important: the response of GS1 to such a shock is significant up to about 2 years, and the effect on the response of N3 even remains significant for lags up to 40 months.

As a second example we consider the *housing data* (Diebold 2001, p. 109). Housing is a bivariate series of monthly data of housing starts and housing completions from January 1968 until June 1996. We plot the series in Figure 6.4, and notice immediately the presence of two huge outliers around 1971 and 1977. For this reasons, we proceed with a robust analysis of the bivariate time series. To find the optimal order of the vector autoregressive model, the robust lag length criteria are computed and each of them suggests a VAR(6) model. After estimating this model, robust residuals distances are computed and plotted in Figure 6.5. The two severe outliers are detected, together with some other less extreme observations.

Impulse response functions resulting from the estimated VAR(6) model using RMLTS are presented in Figure 6.6, together with 95% analytic confidence bounds (Monte carlo confidence bounds give almost identical results). A unit shock in the innovation for the variable "housing starts" gives rise to a large response

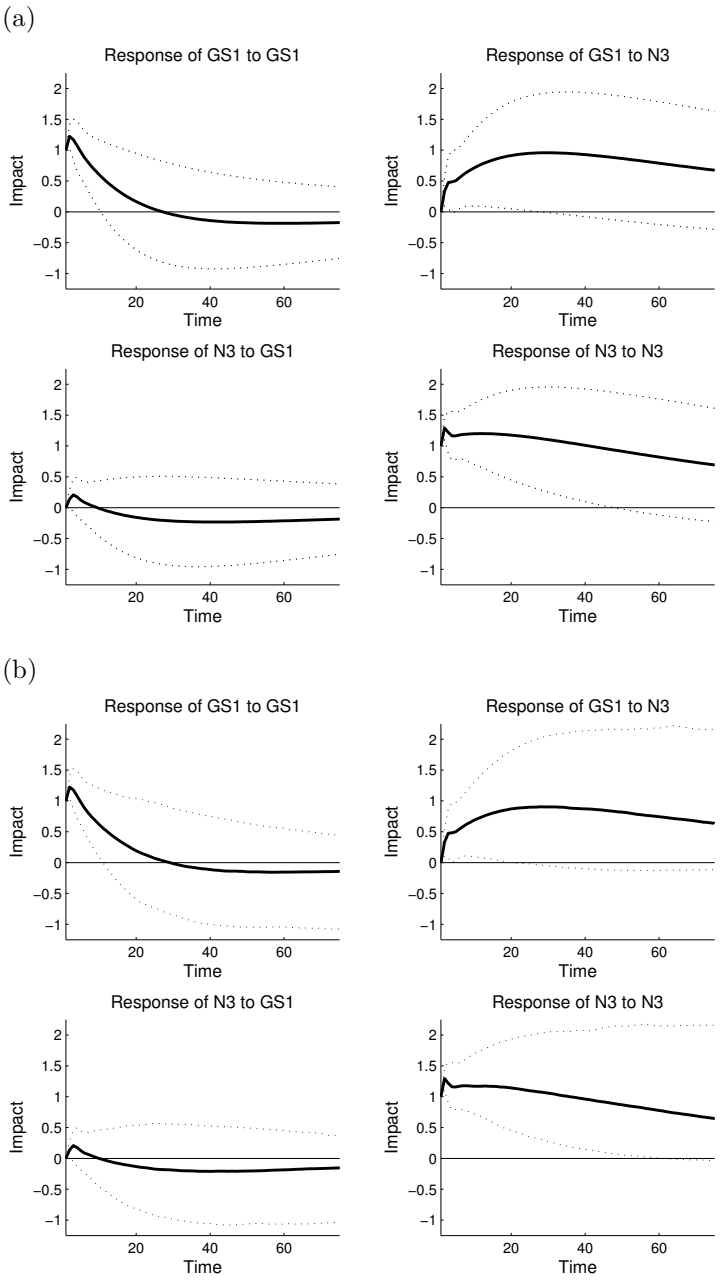

Figure 6.3: *The impulse response functions (solid lines) for the "maturity rate" series estimated by the robust RMLTS estimator, together with the (a) analytic; (b) Monte Carlo confidence bounds (dotted lines).*

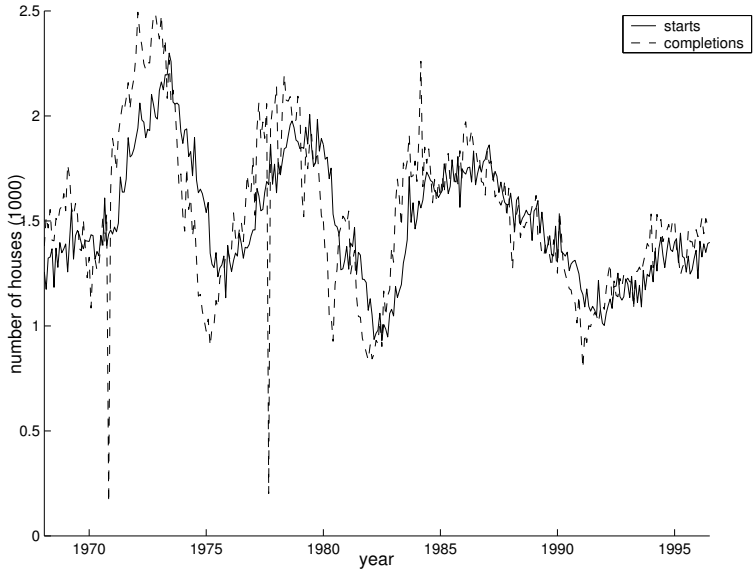

Figure 6.4: *Time plot of the "housing data" series. The solid line represents housing starts and the dashed line housing completions (in thousands).*

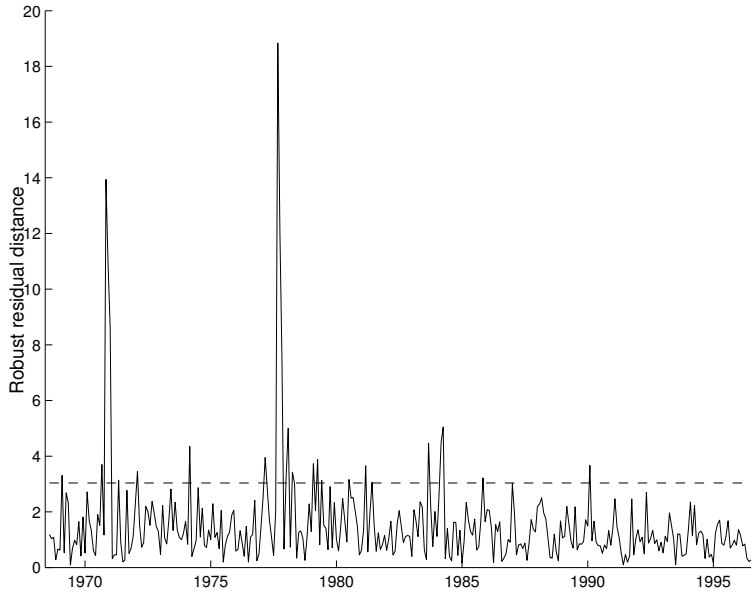

Figure 6.5: *Robust residual distances for the "housing data" series. The dashed line represents the critical value at the 1% level.*

on the series itself, showing that there is quite some persistency is this series. Without much surprise, "housing completions" is also strongly affected by unit shocks in "housing starts", with a maximum effect after 16 months, and remaining significant until even more than 2 years. On the other hand, a unit shock in the innovations of the completions series has only a limited impact on both series. In particular, the response of housing starts on completions turns out to be non-significant, since the horizontal line at zero in included within the confidence bounds

Note that we do not report the results of the classical analysis here. Indeed, if no (or harmless) outliers are present, then both methods of analysis produce very similar results. On the other hand, if there are outliers, then the robust procedure is more reliable. For example, the IRFs based on the OLS method for the "housing data" give much less significant responses. Since outliers are present in the data, more confidence should be given to the results of the robust analysis.

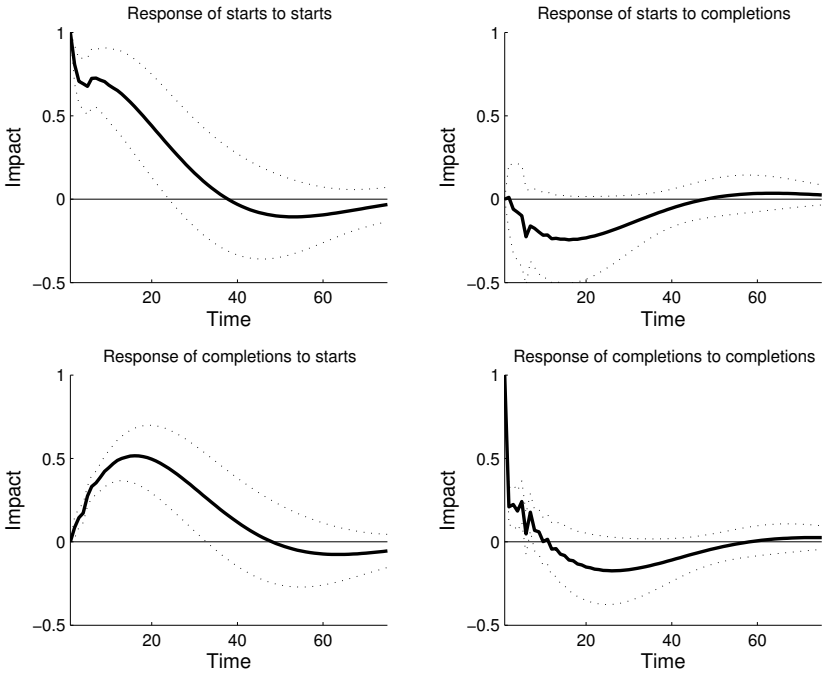

Figure 6.6: *The impulse response functions (solid lines) for the "housing data" series, together with analytic confidence bounds (dotted lines).*

6.7 Conclusions

For multivariate time series correlation outliers can be present, which are not necessarily visible in plots of the single univariate series. Development of robust procedures for multiple time series analysis is therefore even more important than for univariate time series analysis.

In this chapter we have shown how robust multivariate regression estimators can be used to estimate Vector Autoregressive models. We use the reweighted multivariate least trimmed squares estimator, but other robust multivariate regression estimators could be used as well (e.g. the MM estimators of Tatsuoka and Tyler 2000, the robust covariance based estimators of Rousseeuw et al 2004, or the τ-estimators of García Ben, Martínez and Yohai, 2006). Software to robustly estimate the VAR model is available from *www.econ.kuleuven.be/christophe.croux*. This software computes different robust lag-length selection criteria, the robustly estimated impulse response functions, together with their confidence bounds, and provides robust residual distances as a tool for outlier detection. It was used to analyse the real data sets in Section 6.6

The estimation of VAR models as multivariate regression models has one major disadvantage. A fraction ε of outliers in the original series can produce up to $k\varepsilon$ outliers for the regression model (6.1), due to the fact that k delayed versions of the time series are used as explanatory variables. Hence, if a robust regression estimator has a breakdown point of, for example, $1/2$, this reduces to $1/(2k)$ when estimating the VAR(k) model. To solve this problem of propagation of outliers, it has been proposed to first filter the series with a robust filter, and then to apply a robust estimator on the robustly filtered data (see Bianco et al 2001, Maronna et al 2006). Other types of robust filters were proposed by Davies et al (2004) and Fried et al (2006). However, while robust filters are available for univariate series, multivariate versions have not been developed yet, up to our best knowledge, and we leave this for future research.

In the simulation experiments the RMLTS estimators have been compared with the residual autocovariance (RA) estimators of García Ben et al (1999). The RA estimates are computed iteratively, and we propose to use the RMLTS as a starting value for computing the RA estimators. It turned out that both robust estimators behave then similarly. If there are no outliers in the data set present, the robust estimators performs almost as good as the classical estimator. But if there are outliers, bias and MSE only remain under control when using the robust estimator.

List of figures

List of tables

Bibliography

Agulló, J.; Croux, C. and Van Aelst, S. (2002), "The multivariate least trimmed squares estimator," *Research report*, Dept. of Applied Economics, K.U.Leuven, Belgium.

Akaike, H. (1973), "Information theory and an extension of the maximum likelihood principle," in *2nd International Symposium on Information Theory*, eds. Petrov, B. N. and Csáki, F., Académiai Kiadó: Budapest, pp. 267–281.

Bartlett, M. S. (1937), "Properties of sufficiency and statistical tests," *Proceedings of the Royal Statistical Society, Series A*, 160, 318–335.

Bensmail, H. and Celuex, G. (1996), "Regularized Gaussian discriminant analysis through eigenvalue decomposition," *Journal of the American Statistical Association*, 91, 1743–1748.

Bianco, A. M.; García Ben, M.; Martínez, E. J. and Yohai, V. J. (2001), "Outlier detection in regression models with ARIMA errors using robust estimates," *Journal of Forecasting*, 20, 565–579.

Bianco, A. M. and Yohai, V. J. (1996), "Robust estimation in the logistic regression model," in *Robust Statistics, Data Analysis and Computer Intensive Methods*, ed. Reider, H., Springer Verlag: New York, pp. 17–34.

Boente, G.; Pires, A. M. and Rodrigues, I. M. (2002), "Influence functions and outlier detection under the common principal components model: A robust approach," *Biometrika*, 89, 861–875.

Bondell, H. D. (2005), "Minimum distance estimation for the logistic regression model," *Biometrika*, 92, 724–731.

Brockwell, P. J. and Davis, R. A. (2003), *Introduction to Time Series and Forecasting*, Wiley: New York, 2nd ed.

Bull, S. B. and Donner, A. (1987), "The efficiency of multinomial logistic regression compared with multiple group discriminant analysis," *Journal of the American Statistical Association*, 82, 1118–1122.

Bustos, H. and Yohai, V. J. (1986), "Robust estimates of ARMA models," *Journal of the American Statistical Association*, 81, 155–159.

Butler, R. W.; Davies, P. L. and Jhun, M. (1993), "Asymptotics for the Minimum Covariance Determinant estimator," *Annals of Statistics*, 21, 1385–1400.

Campbell, M. K. and Donner, A. (1989), "Classification efficiency of multinomial logistic regression relative to ordinal logistic regression," *Journal of the American Statistical Association*, 84, 587–591.

Campbell, N. A. (1978), "The influence as an aid in outlier detection in discriminant analysis," *Applied Statistics*, 27, 251–258.

Carroll, R. J. and Pederson, S. (1993), "On robust estimation in the logistic regression model," *Journal of the Royal Statistical Society, Series B*, 55, 693–706.

Chang, I.; Tiao, G. C. and Chen, C. (1988), "Estimation of time series parameters in the presence of outliers," *Technometrics*, 30, 193–204.

Chatterjee, S.; Price, B. and Hadi, A. (2000), *Regression Analysis by Example*, Wiley: New York.

Chen, X.; Ender, P.; Mitchell, M. and Wells, C. (2003), *Regressions with STATA*, Stata Web Books, http://www.ats.ucla.edu/stat/sas/webbooks/reg/.

Chork, C. Y. and Rousseeuw, P. J. (1992), "Integrating a high-breakdown option into discriminant analysis in exploration geochemistry," *Journal of Geochemical Exploration*, 43, 191–203.

Christmann, A. (1996), "High breakdown point estimators in logistic regression," in *Robust Statistics, Data Analysis and Computer Intensive Methods*, ed. Reider, H., Springer Verlag: New York, pp. 79–89.

Cook, R. D. and Weisberg, S. (1982), *Residuals and Influence in Regression*, Chapman and Hall: London.

Cook, R. D. and Weisberg, S. (1999), *Applied Regression Including Computing and Graphics*, Wiley: New York.

Cook, R. D. and Yin, Z. (2001), "Dimension reduction and visualization in discriminant analysis," *Australian and New Zealand Journal of Statistics*, 43, 147–199.

Cox, T. F. and Ferry, G. (1991), "Robust logistic discrimination," *Biometrika*, 78, 841–849.

Critchley, F. (1985), "Influence in principal components analysis," *Biometrika*, 72, 627–636.

Critchley, F. and Vitiello, C. (1991), "The influence of observations on misclassification probability estimates in linear discriminant analysis," *Biometrika*, 78, 677–690.

Croux, C. and Dehon, C. (2001), "Robust linear discriminant analysis using S-estimators," *The Canadian Journal of Statistics*, 29, 473–492.

Croux, C. and Dehon, C. (2002), "Analyse canonique basée sur des estimateurs robustes de la matrice de covariance," *La Revue de Statistique Appliqueé*, 2, 5–26.

Croux, C.; Filzmoser, P.; Pison, G. and Rousseeuw, P. J. (2003), "Fitting multiplicative models by robust alternating regressions," *Statistics and Computing*, 13, 23–36.

Croux, C.; Flandre, C. and Haesbroeck, G. (2002), "The breakdown behavior of the maximum likelihood estimator in the logistoc regression model," *Statistics and Probability Letters*, 60, 377–386.

Croux, C. and Haesbroeck, G. (1999), "Influence function and efficiency of the Minumum Covariance Determinant scatter matrix estimator," *Journal of Multivariate Analysis*, 71, 161–190.

Croux, C. and Haesbroeck, G. (2000), "Principal component analysis based on robust estimators of the covariance or correlation matrix: influence functions and efficiencies," *Biometrika*, 87, 603–618.

Croux, C. and Haesbroeck, G. (2002), "Maxbias curves of location estimators based on subranges," *Journal of Nonparametric Statistics*, 14, 295–306.

Croux, C. and Haesbroeck, G. (2003), "Implementing the Bianco and Yohai estimator for logistic regression," *Computational Statistics and Data Analysis*, 44, 273–295.

Croux, C. and Joossens, K. (2005), "Influence of observations on the misclassification probability in quadratic discriminant analysis," *Journal of Multivariate Analysis*, 96, 384–403.

Davies, P. L. (1987), "Asymptotic behavior of S-estimators of multivariate location parameters and dispersion matrices," *Annals of Statistics*, 15, 1269–1292.

Davies, P. L.; Fried, R. and Gather, U. (2004), "Robust signal extraction for on-line monitoring data," *Journal of Statistical Planning and Inference*, 122, 65–78.

De Leval, D. (2001), "Etude comptable et actuarielle des plans de pension: Confrontation des normes IAS/US GAAP et introduction de tables de mortalité prospectives," *Mémoire de fin d'études sous la supervision de D. Justens, Département de gestion, Université de Liège.*

De Luna, X. and Genton, M. G. (2001), "Robust simulation-based estimation of ARMA models," *Journal of Computational and Graphical Statistics*, 10, 370–387.

Diebold, F. X. (2001), *Elements of Forecasting*, South-Western College Publishing, 2nd ed.

Draper, N. and Smith, H. (1998), *Applied Regression Analysis*, Wiley: New York.

Efron, B. (1975), "The efficiency of logistic regression compared to normal discriminant analysis," *Journal of the American Statistical Association*, 70, 892–898.

Enders, W. (2004), *Applied Econometric Time Series*, Wiley: New York.

Fama, E. and French, K. (1992), "The cross-section of expected stocks returns," *Journal of Finance*, 47, 427–465.

Finney, D. J. (1947), "The estimation from individual records of the relationship between dose and quantal response," *Biometrika*, 34, 320–334.

Fisher, R. A. (1938), "The statistical utilization of multiple measurements," *Annals of Eugenics*, 8, 376–386.

Flury, B. and Riedwyl, H. (1988), *Multivariate Statistics: a Practical Approach*, Chapman and Hall: London.

Fox, A. (1972), "Outliers in time series," *Journal of the Royal Statistical Society, Series B*, 34, 350–363.

Franses, H. P.; Kloek, T. and Lucas, A. (1999), "Outlier robust analysis of long-run marketing effects for weekly scanning data," *Journal of Econometrics*, 89, 293–315.

Friedman, J.; Hastie, T. and Tibshirani, R. (2001), *The Elements of Statistical Learning: Data Mining, Inference and Prediction*, Springer Verlag: New York.

Fung, W. K. (1992), "Some diagnostic measures in discriminant analysis," *Statistics and Probability Letters*, 13, 279–285.

Fung, W. K. (1995a), "Diagnostics in linear discriminant analysis," *Journal of the American Statistical Association*, 90, 952–956.

Fung, W. K. (1995b), "Detecting influential observations for estimated probabilities in multiple discriminant analysis," *Computational Statistics and Data Analysis*, 20, 557–568.

Fung, W. K. (1996a), "Diagnosing influential observations in quadratic discriminant analysis," *Biometrics*, 52, 1235–1241.

Fung, W. K. (1996b), "The influence of an observation on the misclassification probability in multiple discriminant analysis," *Communications in Statistics - Theory and Methods*, 25, 1917–1930.

García Ben, M.; Martínez, E. J. and Yohai, V. J. (1999), "Robust estimation in vector autoregressive moving average models," *Journal of Time Series Analysis*, 20, 381–399.

García Ben, M.; Martínez, E. J. and Yohai, V. J. (2006), "Robust estimation for the multivariate linear model based on a τ-scale," *Journal of Multivariate Analysis*, forthcoming.

Gatto, R. and Ronchetti, E. (1996), "General saddlepoint approximations of marginal densities and tail probabilities," *Journal of the American Statistical Association*, 91, 666–673.

Gerlach, R.; Carter, C. and Kohn, R. (1999), "Diagnostics for time series analysis," *Journal of Time Series Analysis*, 20, 309–330.

Hamilton, J. D. (1994), *Time Series Analysis*, Princeton University Press.

Hampel, F. R. (1971), "A general qualitative definition of robustness," *Annals of Mathematical Statistics*, 42, 1887–1896.

Hampel, F. R.; Ronchetti, E. M.; Rousseeuw, P. J. and Stahel, W. A. (1986), *Robust Statistics: The Approach Based on Influence Functions*, Wiley: New York.

Hannan, E. J. (1980), "The estimation of the order of an ARMA process," *Annals of Statistics*, 8, 1071–1081.

Hannan, E. J. and Quinn, B. G. (1979), "The determination of the order of an autoregression," *Journal of the Royal Statistical Society, Series B*, 41, 190–195.

Hawkins, D. M. and McLachlan, G. J. (1997), "High-breakdown linear discriminant analysis," *Journal of the American Statistical Association*, 92, 136–143.

He, X. M. and Fung, W. K. (2000), "High breakdown estimation for multiple populations with applications to discriminant analysis," *Journal of Multivariate Analysis*, 72, 151–162.

Hoaglin, D. A.; Mosteller, F. and Tukey, J. W. (1982), *Understanding Robust and Exploratory Data Analysis*, Wiley: New York.

Houshmand, A. A. (1993), "Misclassification probabilities for the quadratic discriminant function," *Communications in Statistics, Series B*, 81–98.

Huber, P. J. (1964), "Robust estimation of a location parameter," *Annals of Mathematical Statistics*, 35, 73–101.

Huber, P. J. (1981), *Robust Statistics*, Wiley: New York.

Hubert, M. and Van Driessen, K. (2004), "Fast and robust discriminant analysis," *Computational Statistics and Data Analysis*, 45, 301–320.

Johnson, R. A. and Wichern, D. W. (1998), *Applied Multivariate Statistical Analysis*, Prentic Hall: New York, 4th ed.

Johnson, R. A. and Wichern, D. W. (2002), *Applied Multivariate Statistical Analysis*, Prentic Hall: London, 4th ed.

Johnson, W. (1985), "Influence measures for logistic regression: Another point of view," *Biometrika*, 72, 59–65.

Joossens, K. and Croux, C. (2004), "Empirical comparison of the classification performance of robust linear and quadratic discriminant analysis," in *Statistics for Industry and Technology*, eds. Hubert, M.; Pison, G.; Struyf, A. and Van Aelst, S., Birkhäuser Verlag: Basel-Switzerland, pp. 131–140.

Jureckova, J. and Sen, P. K. (1996), *Robust Statistical Procedures: Asymptotics and Interrelations*, Wiley: New York.

KBC Bank en Verzekeringen (2001), "Economisch profiel van de Europese Unie," *Economische Financile Berichten*, 8.

Knez, P. J. and Ready, M. J. (1997), "On the robustness of size and book-to-market in cross-sectional regressions," *Journal of Finance*, 52, 1355–1382.

Künsch, H. R.; Stefanski, L. A. and Carroll, R. J. (1989), "Conditionally unbiased bounded influence estimation in general regression models, with applications to generalized linear models," *Journal of the American Statistical Association*, 84, 460–466.

Lachenbruch, P. A. (1979), "Note on initial misclassification effects on the quadratic discriminant function," *Technometrics*, 21, 129–132.

Li, W. K. and Hui, Y. V. (1989), "Robust multiple time series modelling," *Biometrika*, 76, 309–315.

Lopuhaä, H. P. (1989), "On the relation between S-estimators and M-estimators of multivariate location and covariance," *Annals of Statistics*, 17, 1662–1683.

Lopuhaä, H. P. (1999), "Asymptotics of reweighted estimators of multivariate location and scatter," *Annals of Statistics*, 27, 1638–1665.

Maddala, G. S. and Rao, C. R. (1997), *Handbook of Statistics 15: Robust Inference*, Elsevier: Amsterdam.

Magnus, J. R. and Neudecker, H. (1999), *Matrix differential calculus with applications in statistics and econometrics*, John Wiley: New York, 2nd ed.

Marazzi, A. (1993), *Algorithms, Routines, and S-functions for Robust Statistics*, Champan and Hall: New York.

Maronna, R. and Yohai, V. (1998), "Robust estimation of multivariate location and scatter," in *Encyclopedia of Statistical Sciences Update Volume 2*, eds. Kotz, S.; Read, C.; and Banks, D., John Wiley: New York, pp. 589–596.

Maronna, R. A.; Martin, R. D. and Yohai, V. Y. (2006), *Robust Statistics: Theory and Practice*, Wiley: New York, forthcoming.

McFarland, H. R. and Richards, D. S. P. (2002), "Exact misclassification porbabilities for plug-in normal quadratic discriminant functions II. The heterogeneous case," *Journal of Multivariate Analysis*, 82, 299–330.

McKean, J. W. and Hettmansperger, T. P. (1998), *Robust Nonparametric Statistical Methods*, Arnold: London.

O'neil, T. J. (1980), "The general distribution of the error rate of a classification procedure with application to logistic regression discrimination," *Journal of the American Statistical Association*, 75, 154–160.

Pires, A. M. and Branco, J. A. (2002), "Partial influence functions," *Journal of Multivariate Analysis*, 83, 451–468.

Pison, G.; Rousseeuw, P. J.; Filzmoser, P. and Croux, C. (2003), "Robust factor analysis," *Journal of Multvariate Analysis*, 84, 145–172.

Pison, G. and Van Aelst, S. (2004), "Diagnostic plots for robust multivariate methods," *Journal of Computational and Graphical Statistics*, 13, 310–329.

Pregibon, D. (1981), "Logistic regression diagnostics," *Annals of Statistics*, 9, 705–724.

Pregibon, D. (1982), "Resistant fits for some commonly used logistic models with medical applications," *Biometrics*, 38, 485–498.

Randles, R. H.; Broffitt, J. D.; Ramber, J. S. and Hogg, R. V. (1978a), "Generalized linear and quadratic discriminant functions using robust estimates," *Journal of the American Statistical Association*, 73, 564–568.

Randles, R. H.; Brofitt, J. D.; Ramberg, J. S. and Hogg, R. V. (1978b), "Generalized linear and Quadratic Discriminant Functions Using Robust Estimates," *Journal of the American Statistical Association*, 73, 564–568.

Rao, C. R. (1948), "The utilization of multiple measurements in problems of biological classification," *Journal of the Royal Statistical Society, Series B*, 10, 159–203.

Riani, M. (2004), "Extensions of the forward search to time series," *Studies in Non Linear Dynamics and Econometrics*, 8, Article 2.

Riani, M. and Atkinson, A. C. (2000), *Robust Diagnostic Regression Analysis*, Spinger Verlag: New York.

Riani, M. and Atkinson, A. C. (2001), "A unified approach to outliers, influence and transformations in discriminant analysis," *Journal of Computational and Graphical Statistics*, 10, 513–544.

Rieder, H. (1994), *Robust Asymptotic Statistics*, Springer Verlag: New York.

Rousseeuw, P. J. (1984), "Least median of squares regression," *Journal of the American Statistical Association*, 79, 871–880.

Rousseeuw, P. J. (1985), "Multivariate estimation with high breakdown point," in *Mathematical Statistics and applications*, eds. Grossman, W.; Pflug, G.; Vincze, I. and Wertz, W., Reidel, Dordrecht, vol. B, pp. 283–297.

Rousseeuw, P. J. and Christmann, A. (2003), "Robustness against separation and outliers in logistic regression," *Computational Statistics and Data Analysis*, 43, 315–332.

Rousseeuw, P. J. and Leroy, A. M. (1987), *Robust Regression and Outlier Detection*, Wiley: New York.

Rousseeuw, P. J. and Van Driessen, K. (1999), "A fast algorithm for the minimum covariance determinant estimator," *Technometrics*, 41, 212–223.

Rousseeuw, P. J.; Van Driessen, K.; Van Aelst, S. and Agulló, J. (2004), "Robust multivariate regression," *Technometrics*, 46, 293–305.

Rousseeuw, P. J. and Van Zomeren, B. C. (1990), "Unmasking multivariate outliers and leverage points," *Journal of the American Statistical Association*, 85, 633–639.

Ruppert, D. (1992), "Computing S-estimators for regression and multivariate location/dispersion," *Journal of Computational and Graphical Statistics*, 1, 253–270.

Salibian-Barrera, M. and Yohai, V. J. (2005), "A fast algorithm for S-regression estimates," *Journal of Computational and Graphical Statistics*, forthcoming.

SAS OnlineDoc (2002), "IML: Robust regression," *Sas Institute, Cary, NC, http://v8doc.sas.com/sashtml*.

Sapra, S. K. (1991), "A connection between the logit model, normal discriminant analysis, and multivariate normal mixtures," *The American Statistician*, 45, 265–268.

Schwarz, G. (1978), "Estimating the dimension of a model," *Annals of Statistics*, 6, 461–464.

Sibson, R. (1979), "Studies in the robustness of multidimensional scaling: perturbational analysis of classical scaling," *Journal of the Royal Statistical Society, Series B*, 41, 217–229.

Smith, A. F. M. and Spiegelhalter, D. J. (1980), "Bayes factors and choice criteria for linear models," *Journal of the Royal Statistical Society, Series B*, 42, 213–220.

Staudte, R. G. and Seather, S. J. (1990), *Robust Estimation and Testing*, John Wiley and Sons: New York.

Stock, J. H. and Watson, M. W. (2003), *Introduction to Econometrics*, Addison Wesley.

Tanaka, Y. (1994), "Recent advance in sensitivity analysis in multivariate statistical methods," *Journal of the Japanese Society of Computational Statistics*, 7, 1–25.

Tanaka, Y. and Tarumi, T. (1996), "Sensitivity analysis in multivariate methods: General procedure based on influence functions and its robust version," in *Compstat: Proceedings in Computational Statistics*, ed. Prat, A., Physica-Verlag: Heidelberg, pp. 185–186.

Tatsuoka, K. S. and Tyler, D. E. (2000), "On the uniqueness of the S-functionals and the M-functionals under nonelliptical distributions," *The Annals of Statistics*, 28, 1219–1243.

Tsay, R. S. (2002), *Analysis of Financial Time Series*, John Wiley and Sons: New York.

Tsay, R. S.; Peña, D. and Pankratz, A. E. (2000), "Outliers in multivariate time series," *Biometrika*, 87, 789–804.

Victoria-Feser, M. P. (2002), "Robust inference with binary data," *Psychometrica*, 67, 21–32.

Wilcox, R. (1997), *Introduction to Robust Estimation and Hypothesis Testing*, Academic Press: San Diego.

Zaman, A.; Rousseeuw, P. J. and Orhan, M. (2001), "Econometric applications of high-breakdown robust regression techniques," *Economics Letters*, 71, 1–8.

Doctoral dissertations from the Faculty of Economic and Applied Economic Sciences

From August 1, 1971.

1. GEPTS Stefaan (1971)
   Stability and efficiency of resource allocation processes in discrete commodity spaces. Leuven, K. U. Leuven, 1971. 86 pp.

2. PEETERS Theo (1971)
   Determinanten van de internationale handel in fabrikaten. Leuven, Acco, 1971. 290 pp.

3. VAN LOOY Wim (1971)
   Personeelsopleiding: een onderzoek naar investeringsaspekten van opleiding. Hasselt, Vereniging voor wetenschappelijk onderzoek in Limburg, 1971. VII, 238 pp.

4. THARAKAN Mathew (1972)
   Indian exports to the European community: problems and prospects. Leuven, Faculty of Economics and Applied Economics, 1972. X, 343 pp.

5. HERROELEN Willy (1972)
   Heuristische programmatie: methodologische benadering en praktische toepassing op complexe combinatorische problemen. Leuven, Aurelia scientifica, 1972. X, 367 pp.

6. VANDENBULCKE Jacques (1973)
   De studie en de evaluatie van data-organisatiemethodes en data-zoekmethodes. Leuven, s.n., 1973. 3 V.

7. PENNYCUICK Roy A. (1973)
   The economics of the ecological syndrome. Leuven, Acco, 1973. XII, 177 pp.

8. KAWATA T. Bualum (1973)
   Formation du capital d'origine belge, dette publique et stratégie du développement au Zaire. Leuven, K. U. Leuven, 1973. V, 342 pp.

9. DONCKELS Rik (1974)
   Doelmatige oriëntering van de sectorale subsidiepolitiek in België: een theoretisch onderzoek met empirische toetsing. Leuven, K. U. Leuven, 1974. VII, 156 pp.

10. VERHELST Maurice (1974)
    Contribution to the analysis of organizational information systems and their financial benefits. Leuven, K. U. Leuven, 1974. 2 V.

11. CLEMEUR Hugo (1974)
    Enkele verzekeringstechnische vraagstukken in het licht van de nutstheorie. Leuven, Aurelia scientifica, 1974. 193 pp.

12. HEYVAERT Edward (1975)
    De ontwikkeling van de moderne bank- en krediettechniek tijdens de zestiende en zeventiende eeuw in Europa en te Amsterdam in het bijzonder. Leuven, K. U. Leuven, 1975. 186 pp.

13. VERTONGHEN Robert (1975)
    Investeringscriteria voor publieke investeringen: het uitwerken van een operationele theorie met een toepassing op de verkeersinfrastructuur. Leuven, Acco, 1975. 254 pp.

14. Niet toegekend.

15. VANOVERBEKE Lieven (1975)
    Microeconomisch onderzoek van de sectoriële arbeidsmobiliteit. Leuven, Acco, 1975. 205 pp.

16. DAEMS Herman (1975)
    The holding company: essays on financial intermediation, concentration and capital market imperfections in the Belgian economy. Leuven, K. U. Leuven, 1975. XII, 268 pp.

17. VAN ROMPUY Eric (1975)
    Groot-Brittannië en de Europese monetaire integratie: een onderzoek naar de gevolgen van de Britse toetreding op de geplande Europese monetaire unie. Leuven, Acco, 1975. XIII, 222 pp.

18. MOESEN Wim (1975)
    Het beheer van de staatsschuld en de termijnstructuur van de intrestvoeten met een toepassing voor België. Leuven, Vander, 1975. XVI, 250 pp.

19. LAMBRECHT Marc (1976)
Capacity constrained multi-facility dynamic lot-size problem. Leuven, K. U. Leuven, 1976. 165 pp.

20. RAYMAECKERS Erik (1976)
De mens in de onderneming en de theorie van het producenten-gedrag: een bijdrage tot transdisciplinaire analyse. Leuven, Acco, 1976. XIII, 538 pp.

21. TEJANO Albert (1976)
Econometric and input-output models in development planning: the case of the Philippines. Leuven, K. U. Leuven, 1976. XX, 297 pp.

22. MARTENS Bernard (1977)
Prijsbeleid en inflatie met een toepassing op België. Leuven, K. U. Leuven, 1977. IV, 253 pp.

23. VERHEIRSTRAETEN Albert (1977)
Geld, krediet en intrest in de Belgische financiële sector. Leuven, Acco, 1977. XXII, 377 pp.

24. GHEYSSENS Lieven (1977)
International diversification through the government bond market: a risk-return analysis. Leuven, s.n., 1977. 188 pp.

25. LEFEBVRE Chris (1977)
Boekhoudkundige verwerking en financiële verslaggeving van huurkooptransacties en verkopen op afbetaling bij ondernemingen die aan consumenten verkopen. Leuven, K. U. Leuven, 1977. 228 pp.

26. KESENNE Stefan (1978)
Tijdsallocatie en vrijetijdsbesteding: een econometrisch onderzoek. Leuven, s.n., 1978. 163 pp.

27. VAN HERCK Gustaaf (1978)
Aspecten van optimaal bedrijfsbeleid volgens het marktwaardecriterium: een risico-rendements-analyse. Leuven, K. U. Leuven, 1978. IV, 163 pp.

28. VAN POECK Andre (1979)
World price trends and price and wage development in Belgium: an investigation into the relevance of the Scandinavian model of inflation for Belgium. Leuven, s.n., 1979. XIV, 260 pp.

29. VOS Herman (1978)
De industriële technologieverwerving in Brazilië: een analyse. Leuven, s.n., 1978. onregelmatig gepagineerd.

30. DOMBRECHT Michel (1979)
Financial markets, employment and prices in open economies. Leuven, K. U. Leuven, 1979. 182 pp.

31. DE PRIL Nelson (1979)
Bijdrage tot de actuariële studie van het bonus-malussysteem. Brussel, OAB, 1979. 112 pp.

32. CARRIN Guy (1979)
Economic aspects of social security: a public economics approach. Leuven, K. U. Leuven, 1979. onregelmatig gepagineerd

33. REGIDOR Baldomero (1979)
An empirical investigation of the distribution of stock-market prices and weak-form efficiency of the Brussels stock exchange. Leuven, K. U. Leuven, 1979. 214 pp.

34. DE GROOT Roger (1979)
Ongelijkheden voor stop loss premies gebaseerd op E.T. systemen in het kader van de veralgemeende convexe analyse. Leuven, K. U. Leuven, 1979. 155 pp.

35. CEYSSENS Martin (1979)
On the peak load problem in the presence of rationizing by waiting. Leuven, K. U. Leuven, 1979. IX, 217 pp.

36. ABDUL RAZK Abdul (1979)
Mixed enterprise in Malaysia: the case study of joint venture between Malysian public corporations and foreign enterprises. Leuven, K. U. Leuven, 1979. 324 pp.

37. DE BRUYNE Guido (1980)
Coordination of economic policy: a game-theoretic approach. Leuven, K. U. Leuven, 1980. 106 pp.

38. KELLES Gerard (1980)
Demand, supply, price change and trading volume on financial markets of the matching-order type. Vraag, aanbod, koersontwikkeling en omzet op financiële markten van het Europese type. Leuven, K. U. Leuven, 1980. 222 pp.

39. VAN EECKHOUDT Marc (1980)
De invloed van de looptijd, de coupon en de verwachte inflatie op het opbrengstverloop van vastrentende finaciële activa. Leuven, K. U. Leuven, 1980. 294 pp.

40. SERCU Piet (1981)
Mean-variance asset pricing with deviations from purchasing power parity. Leuven, s.n., 1981. XIV, 273 pp.

41. DEQUAE Marie-Gemma (1981)
Inflatie, belastingsysteem en waarde van de onderneming. Leuven, K. U. Leuven, 1981. 436 pp.

42. BRENNAN John (1982)
An empirical investigation of Belgian price regulation by prior notification: 1975 - 1979 - 1982. Leuven, K. U. Leuven, 1982. XIII, 386 pp.

43. COLLA Annie (1982)
Een econometrische analyse van ziekenhuiszorgen. Leuven, K. U. Leuven, 1982. 319 pp.

44. Niet toegekend.

45. SCHOKKAERT Eric (1982)
Modelling consumer preference formation. Leuven, K. U. Leuven, 1982. VIII, 287 pp.

46. DEGADT Jan (1982)
Specificatie van een econometrisch model voor vervuilingsproblemen met proeven van toepassing op de waterverontreiniging in België. Leuven, s.n., 1982. 2 V.

47. LANJONG Mohammad Nasir (1983)
A study of market efficiency and risk-return relationships in the Malaysian capital market. s.l., s.n., 1983. XVI, 287 pp.

48. PROOST Stef (1983)
De allocatie van lokale publieke goederen in een economie met een centrale overheid en lokale overheden. Leuven, s.n., 1983. onregelmatig gepagineerd.

49. VAN HULLE Cynthia (1983)
Shareholders' unanimity and optimal corporate decision making in imperfect capital markets. s.l., s.n., 1983. 147 pp. + appendix.

50. VAN WOUWE Martine (2/12/83)
Ordening van risico's met toepassing op de berekening van ultieme ruïnekansen. Leuven, s.n., 1983. 109 pp.

51. D'ALCANTARA Gonzague (15/12/83)
SERENA: a macroeconomic sectoral regional and national account econometric model for the Belgian economy. Leuven, K. U. Leuven, 1983. 595 pp.

52. D'HAVE Piet (24/02/84)
De vraag naar geld in België. Leuven, K. U. Leuven, 1984. XI, 318 pp.

53. MAES Ivo (16/03/84)
The contribution of J.R. Hicks to macro-economic and monetary theory. Leuven, K. U. Leuven, 1984. V, 224 pp.

54. SUBIANTO Bambang (13/09/84)
A study of the effects of specific taxes and subsidies on a firms' R&D investment plan. s.l., s.n., 1984. V, 284 pp.

55. SLEUWAEGEN Leo (26/10/84)
Location and investment decisions by multinational enterprises in Belgium and Europe. Leuven, K. U. Leuven, 1984. XII, 247 pp.

56. GEYSKENS Erik (27/03/85)
Produktietheorie en dualiteit. Leuven, s.n., 1985. VII, 392 pp.

57. COLE Frank (26/06/85)
Some algorithms for geometric programming. Leuven, K. U. Leuven, 1985. 166 pp.

58. STANDAERT Stan (26/09/86)
A study in the economics of repressed consumption. Leuven, K. U. Leuven, 1986. X, 380 pp.

59. DELBEKE Jos (03/11/86)
Trendperioden in de geldhoeveelheid van België 1877-1983: een theoretische en empirische analyse van de "Banking school" hypothese. Leuven, K. U. Leuven, 1986. XII, 430 pp.

60. VANTHIENEN Jan (08/12/86)
Automatiseringsaspecten van de specificatie, constructie en manipulatie van beslissingstabellen. Leuven, s.n., 1986. XIV, 378 pp.

61. LUYTEN Robert (30/04/87)
A systems-based approach for multi-echelon production/inventory systems. s.l., s.n., 1987. 3V.

62. MERCKEN Roger (27/04/87)
De invloed van de data base benadering op de interne controle. Leuven, s.n., 1987. XIII, 346 pp.

63. VAN CAYSEELE Patrick (20/05/87)
Regulation and international innovative activities in the pharmaceutical industry. s.l., s.n., 1987. XI, 169 pp.

64. FRANCOIS Pierre (21/09/87)
De empirische relevantie van de independence from irrelevant alternatives. Assumptie indiscrete keuzemodellen. Leuven, s.n., 1987. IX, 379 pp.

65. DECOSTER André (23/09/88)
Family size, welfare and public policy. Leuven, K. U. Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 1988. XIII, 444 pp.

66. HEIJNEN Bart (09/09/88)
Risicowijziging onder invloed van vrijstellingen en herverzekeringen: een theoretische analyse van optimaliteit en premiebepaling. Leuven, K. U. Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 1988. onregelmatig gepagineerd.

67. GEEROMS Hans (14/10/88)
Belastingvermijding. Theoretische analyse van de determinanten van de belastingontduiking en de belastingontwijking met empirische verificaties. Leuven, s.n., 1988. XIII, 409, 5 pp.

68. PUT Ferdi (19/12/88)
Introducing dynamic and temporal aspects in a conceptual (database) schema. Leuven, K. U. Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 1988. XVIII, 415 pp.

69. VAN ROMPUY Guido (13/01/89)
A supply-side approach to tax reform programs. Theory and empirical evidence for Belgium. Leuven, K. U. Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 1989. XVI, 189, 6 pp.

70. PEETERS Ludo (19/06/89)
Een ruimtelijk evenwichtsmodel van de graanmarkten in de E.G.: empirische specificatie en beleidstoepassingen. Leuven, K. U. Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 1989. XVI, 412 pp.

71. PACOLET Jozef (10/11/89) Marktstructuur en operationele efficiëntie in de Belgische financiële sector. Leuven, K. U. Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 1989. XXII, 547 pp.

72. VANDEBROEK Martina (13/12/89)
Optimalisatie van verzekeringscontracten en premieberekeningsprincipes. Leuven, K. U. Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 1989. 95 pp.

73. WILLEKENS Francois (1990)
Determinance of government growth in industrialized countries with applications to Belgium. Leuven, K. U. Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 1990. VI, 332 pp.

74. VEUGELERS Reinhilde (02/04/90)
Scope decisions of multinational enterprises. Leuven, K. U. Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 1990. V, 221 pp.

75. KESTELOOT Katrien (18/06/90)
Essays on performance diagnosis and tacit cooperation in international oligopolies. Leuven, K. U. Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 1990. 227 pp.

76. WU Changqi (23/10/90)
Strategic aspects of oligopolistic vertical integration. Leuven, K. U. Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 1990. VIII, 222 pp.

77. ZHANG Zhaoyong (08/07/91)
A disequilibrium model of China's foreign trade behaviour. Leuven, K. U. Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 1991. XII, 256 pp.

78. DHAENE Jan (25/11/91)
Verdelingsfuncties, benaderingen en foutengrenzen van stochastische grootheden geassocieerd aan verzekeringspolissen en -portefeuilles. Leuven, K. U. Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 1991. 146 pp.

79. BAUWELINCKX Thierry (07/01/92)
Hierarchical credibility techniques. Leuven, K. U. Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 1992. 130 pp.

80. DEMEULEMEESTER Erik (23/3/92)
Optimal algorithms for various classes of multiple resource-constrained project scheduling problems. Leuven, K. U. Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 1992. 180 pp.

81. STEENACKERS Anna (1/10/92)
Risk analysis with the classical actuarial risk model: theoretical extensions and applications to Reinsurance. Leuven, K. U. Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 1992. 139 pp.

82. COCKX Bart (24/09/92)
The minimum income guarantee. Some views from a dynamic perspective. Leuven, K. U. Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 1992. XVII, 401 pp.

83. MEYERMANS Eric (06/11/92)
Econometric allocation systems for the foreign exchange market: Specification, estimation and testing of transmission mechanisms under currency substitution. Leuven, K. U. Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 1992. XVIII, 343 pp.

84. CHEN Guoqing (04/12/92)
Design of fuzzy relational databases based on fuzzy functional dependency. Leuven, K. U. Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 1992. 176 pp.

85. CLAEYS Christel (18/02/93)
Vertical and horizontal category structures in consumer decision making: The nature of product hierarchies and the effect of brand typicality. Leuven, K. U. Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 1993. 348 pp.

86. CHEN Shaoxiang (25/03/93)
The optimal monitoring policies for some stochastic and dynamic production processes. Leuven, K. U. Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 1993. 170 pp.

87. OVERWEG Dirk (23/04/93)
Approximate parametric analysis and study of cost capacity management of computer configurations. Leuven, K. U. Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 1993. 270 pp.

88. DEWACHTER Hans (22/06/93)
Nonlinearities in speculative prices: The existence and persistence of nonlinearity in foreign exchange rates. Leuven, K. U. Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 1993. 151 pp.

89. LIN Liangqi (05/07/93)
Economic determinants of voluntary accounting choices for R & D expenditures in Belgium. Leuven, K. U. Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 1993. 192 pp.

90. DHAENE Geert (09/07/93)
Encompassing: formulation, properties and testing. Leuven, K. U. Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 1993. 117 pp.

91. LAGAE Wim (20/09/93)
Marktconforme verlichting van soevereine buitenlandse schuld door private crediteuren: een neo-institutionele analyse. Leuven, K. U. Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 1993. 241 pp.

92. VAN DE GAER Dirk (27/09/93)
Equality of opportunity and investment in human capital. Leuven, K. U. Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 1993. 172 pp.

93. SCHROYEN Alfred (28/02/94)
Essays on redistributive taxation when monitoring is costly. Leuven, K. U. Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 1994. 203 pp. + V.

94. STEURS Geert (15/07/94)
Spillovers and cooperation in research and development. Leuven, K. U. Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 1994. 266 pp.

95. BARAS Johan (15/09/94)
The small sample distribution of the Wald, Lagrange multiplier and likelihood ratio tests for homogeneity and symmetry in demand analysis: a Monte Carlo study. Leuven, K. U. Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 1994. 169 pp.

96. GAEREMYNCK Ann (08/09/94)
The use of depreciation in accounting as a signalling device. Leuven, K. U. Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 1994. 232 pp.

97. BETTENDORF Leon (22/09/94)
A dynamic applied general equilibrium model for a small open economy. Leuven, K. U. Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 1994. 149 pp.

98. TEUNEN Marleen (10/11/94)
Evaluation of interest randomness in actuarial quantities. Leuven, K. U. Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 1994. 214 pp.

99. VAN OOTEGEM Luc (17/01/95)
An economic theory of private donations. Leuven. K. U. Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 1995. 236 pp.

100. DE SCHEPPER Ann (20/03/95)
Stochastic interest rates and the probabilistic behaviour of actuarial functions. Leuven, K. U. Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 1995. 211 pp.

101. LAUWERS Luc (13/06/95)
Social choice with infinite populations. Leuven, K. U. Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 1995. 79 pp.

102. WU Guang (27/06/95)
A systematic approach to object-oriented business modeling. Leuven, K. U. Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 1995. 248 pp.

103. WU Xueping (21/08/95)
Term structures in the Belgian market: model estimation and pricing error analysis. Leuven, K. U. Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 1995. 133 pp.

104. PEPERMANS Guido (30/08/95)
Four essays on retirement from the labor force. Leuven, K. U. Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 1995. 128 pp.

105. ALGOED Koen (11/09/95)
Essays on insurance: a view from a dynamic perspective. Leuven, K. U. Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 1995. 136 pp.

106. DEGRYSE Hans (10/10/95)
Essays on financial intermediation, product differentiation, and market structure. Leuven, K. U. Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 1995. 218 pp.

107. MEIR Jos (05/12/95)
Het strategisch groepsconcept toegepast op de Belgische financiële sector. Leuven, K. U. Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 1995. 257 pp.

108. WIJAYA Miryam Lilian (08/01/96)
Voluntary reciprocity as an informal social insurance mechanism: a game theoretic approach. Leuven, K. U. Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 1996. 124 pp.

109. VANDAELE Nico (12/02/96)
The impact of lot sizing on queueing delays: multi product, multi machine models. Leuven, K. U. Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 1996. 243 pp.

110. GIELENS Geert (27/02/96)
Some essays on discrete time target zones and their tails. Leuven, K. U.

Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 1996. 131 pp.

111. GUILLAUME Dominique (20/03/96)
Chaos, randomness and order in the foreign exchange markets. Essays on the modelling of the markets. Leuven, K. U. Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 1996. 171 pp.

112. DEWIT Gerda (03/06/96)
Essays on export insurance subsidization. Leuven, K. U. Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 1996. 186 pp.

113. VAN DEN ACKER Carine (08/07/96)
Belief-function theory and its application to the modeling of uncertainty in financial statement auditing. Leuven, K. U. Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 1996. 147 pp.

114. IMAM Mahmood Osman (31/07/96)
Choice of IPO Flotation Methods in Belgium in an Asymmetric Information Framework and Pricing of IPO's in the Long-Run. Leuven, K. U. Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 1996. 221 pp.

115. NICAISE Ides (06/09/96)
Poverty and Human Capital. Leuven, K. U. Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 1996. 209 pp.

116. EYCKMANS Johan (18/09/97)
On the Incentives of Nations to Join International Environmental Agreements. Leuven, K. U. Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 1997. XV + 348 pp.

117. CRISOLOGO-MENDOZA Lorelei (16/10/97)
Essays on Decision Making in Rural Households: a study of three villages in the Cordillera Region of the Philippines. Leuven, K. U. Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 1997. 256 pp.

118. DE REYCK Bert (26/01/98)
Scheduling Projects with Generalized Precedence Relations: Exact and Heuristic Procedures. Leuven, K. U. Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 1998. XXIV + 337 pp.

119. VANDEMAELE Sigrid (30/04/98)
Determinants of Issue Procedure Choice within the Context of the French IPO Market: Analysis within an Asymmetric Information Framework. Leuven, K. U. Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 1998. 241 pp.

120. VERGAUWEN Filip (30/04/98)
Firm Efficiency and Compensation Schemes for the Management of Innovative Activities and Knowledge Transfers. Leuven, K. U. Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 1998. VIII + 175 pp.

121. LEEMANS Herlinde (29/05/98)
The Two-Class Two-Server Queueing Model with Nonpreemptive Heterogeneous Priority Structures. Leuven, K. U. Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 1998. 211 pp.

122. GEYSKENS Inge (4/09/98)
Trust, Satisfaction, and Equity in Marketing Channel Relationships. Leuven, K. U. Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 1998. 202 pp.

123. SWEENEY John (19/10/98)
Why Hold a Job ? The Labour Market Choice of the Low-Skilled. Leuven, K. U. Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 1998. 278 pp.

124. GOEDHUYS Micheline (17/03/99)
Industrial Organisation in Developing Countries, Evidence from Côte d'Ivoire. Leuven, K. U. Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 1999. 251 pp.

125. POELS Geert (16/04/99)
On the Formal Aspects of the Measurement of Object-Oriented Software Specifications. Leuven, K. U. Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 1999. 507 pp.

126. MAYERES Inge (25/05/99)
The Control of Transport Externalities: A General Equilibrium Analysis. Leuven, K. U. Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 1999. XIV + 294 pp.

127. LEMAHIEU Wilfried (5/07/99)
Improved Navigation and Maintenance through an Object-Oriented Approach to Hypermedia Modelling. Leuven, K. U. Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 1999. 284 pp.

128. VAN PUYENBROECK Tom (8/07/99)
Informational Aspects of Fiscal Federalism. Leuven, K. U. Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 1999. 192 pp.

129. VAN DEN POEL Dirk (5/08/99)
Response Modeling for Database Marketing Using Binary Classification. Leuven, K. U. Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 1999. 342 pp.

130. GIELENS Katrijn (27/08/99)
International Entry Decisions in the Retailing Industry: Antecedents and Performance Consequences. Leuven, K. U. Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 1999. 336 pp.

131. PEETERS Anneleen (16/12/99)
Labour Turnover Costs, Employment and Temporary Work. Leuven, K. U. Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 1999. 207 pp.

132. VANHOENACKER Jurgen (17/12/99)
Formalizing a Knowledge Management Architecture Meta-Model for Integrated Business Process Management. Leuven, K. U. Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 1999. 252 pp.

133. NUNES Paulo (20/03/2000)
Contingent Valuation of the Benefits of Natural Areas and its Warmglow Component. Leuven, K. U. Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 2000. XXI + 282 pp.

134. VAN DEN CRUYCE Bart (7/04/2000)
Statistische discriminatie van allochtonen op jobmarkten met rigide lonen. Leuven, K. U. Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 2000. XXIII + 441 pp.

135. REPKINE Alexandre (15/03/2000)
Industrial restructuring in countries of Central and Eastern Europe: Combining branch-, firm- and product-level data for a better understanding of Enterprises' behaviour during transition towards market economy. Leuven, K. U. Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 2000. VI + 147 pp.

136. AKSOY, Yunus (21/06/2000)
Essays on international price rigidities and exchange rates. Leuven, K. U. Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 2000. IX + 236 pp.

137. RIYANTO, Yohanes Eko (22/06/2000)
Essays on the internal and external delegation of authority in firms. Leuven, K. U. Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 2000. VIII + 280 pp.

138. HUYGHEBAERT, Nancy (20/12/2000)
The Capital Structure of Business Start-ups. Leuven, K. U. Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 2000. VIII + 332 pp.

139. FRANCKX Laurent (22/01/2001)
Ambient Inspections and Commitment in Environmental Enforcement. Leuven, K. U. Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 2001. VIII + 286 pp.

140. VANDILLE Guy (16/02/2001)
Essays on the Impact of Income Redistribution on Trade. Leuven, K. U. Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 2001. VIII + 176 pp.

141. MARQUERING Wessel (27/04/2001)
Modeling and Forecasting Stock Market Returns and Volatility. Leuven, K. U. Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 2001. V + 267 pp.

142. FAGGIO Giulia (07/05/2001)
Labor Market Adjustment and Enterprise Behavior in Transition. Leuven, K. U. Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 2001. 150 pp.

143. GOOS Peter (30/05/2001)
The Optimal Design of Blocked and Split-plot experiments. Leuven, K. U. Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 2001. X + 224 pp.

144. LABRO Eva (01/06/2001)
Total Cost of Ownership Supplier Selection based on Activity Based Costing and Mathematical Programming. Leuven, K. U. Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 2001. 217 pp.

145. VANHOUCKE Mario (07/06/2001)
Exact Algorithms for various Types of Project Scheduling Problems. Non-regular Objectives and time/cost Trade-offs. Leuven, K. U. Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 2001. 316 pp.

146. BILSEN Valentijn (28/08/2001)
Entrepreneurship and Private Sector Development in Central European Transition Countries. Leuven, K. U. Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 2001. XVI + 188 pp.

147. NIJS Vincent (10/08/2001)
Essays on the dynamic Category-level Impact of Price promotions. Leuven, K. U. Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 2001.

148. CHERCHYE Laurens (24/09/2001)
Topics in Non-parametric Production and Efficiency Analysis. Leuven, K. U. Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 2001. VII + 169 pp.

149. VAN DENDER Kurt (15/10/2001)
Aspects of Congestion Pricing for Urban Transport. Leuven, K. U. Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 2001. VII + 203 pp.

150. CAPEAU Bart (26/10/2001)
In defence of the excess demand approach to poor peasants' economic behaviour. Theory and Empirics of non-recursive agricultural household modelling. Leuven, K. U. Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 2001. XIII + 286 pp.

151. CALTHROP Edward (09/11/2001)
Essays in urban transport economics. Leuven, K. U. Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 2001.

152. VANDER BAUWHEDE Heidi (03/12/2001)
Earnings management in an Non-Anglo-Saxon environment. Leuven, K. U. Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 2001. 408 pp.

153. DE BACKER Koenraad (22/01/2002)
Multinational firms and industry dynamics in host countries: the case of Belgium. Leuven, K. U. Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 2002. VII + 165 pp.

154. BOUWEN Jan (08/02/2002)
Transactive memory in operational workgroups. Concept elaboration and case study. Leuven, K. U. Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 2002. 319 pp. + appendix 102 pp.

155. VAN DEN BRANDE Inge (13/03/2002)
The psychological contract between employer and employee: a survey among Flemish employees. Leuven, K. U. Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 2002. VIII + 470 pp.

156. VEESTRAETEN Dirk (19/04/2002)
Asset Price Dynamics under Announced Policy Switching. Leuven, K. U. Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 2002. 176 pp.

157. PEETERS Marc (16/05/2002)
One Dimensional Cutting and Packing: New Problems and Algorithms. Leuven, K. U. Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 2002. IX + 247 pp.

158. SKUDELNY Frauke (21/05/2002)
Essays on The Economic Consequences of the European Monetary Union. Leuven, K. U. Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 2002.

159. DE WEERDT Joachim (07/06/2002)
Social Networks, Transfers and Insurance in Developing countries. Leuven, K. U. Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 2002. VI + 129 pp.

160. TACK Lieven (25/06/2002)
Optimal Run Orders in Design of Experiments. Leuven, K. U. Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 2002. XXXI + 344 pp.

161. POELMANS Stephan (10/07/2002)
Making Workflow Systems work. An investigation into the Importance of Task-appropriation fit, End-user Support and other Technological Characteristics. Leuven, K. U. Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 2002. 237 pp.

162. JANS Raf (26/09/2002)
Capacitated Lot Sizing Problems : New Applications, Formulations and Algorithms. Leuven, K. U. Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 2002.

163. VIAENE Stijn (25/10/2002)
Learning to Detect Fraud from enriched Insurance Claims Data (Context, Theory and Applications). Leuven, K. U. Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 2002. 315 pp.

164. AYALEW Tekabe (08/11/2002)
Inequality and Capital Investment in a Subsistence Economy. Leuven, K. U. Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 2002. V + 148 pp.

165. MUES Christophe (12/11/2002)
On the Use of Decision Tables and Diagrams in Knowledge Modeling and Verification. Leuven, K. U. Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 2002. 222 pp.

166. BROCK Ellen (13/03/2003)
The Impact of International Trade on European Labour Markets. Leuven, K. U. Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 2002.

167. VERMEULEN Frederic (29/11/2002)
Essays on the collective Approach to Household Labour Supply. Leuven, K. U. Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 2002. XIV + 203 pp.

168. CLUDTS Stephan (11/12/2002)
Combining participation in decision-making with financial participation: theoretical and empirical perspectives. Leuven, K. U. Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 2002. XIV + 247 pp.

169. WARZYNSKI Frederic (09/01/2003)
The dynamic effect of competition on price cost margins and innovation. Leuven, K. U. Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 2003.

170. VERWIMP Philip (14/01/2003)
Development and genocide in Rwanda ; a political economy analysis of peasants and power under the Habyarimana regime. Leuven, K. U. Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 2003.

171. BIGANO Andrea (25/02/2003)
Environmental regulation of the electricity sector in a European Market Framework. Leuven, K. U. Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 2003. XX + 310 pp.

172. MAES Konstantijn (24/03/2003)
Modeling the Term Structure of Interest Rates Across Countries. Leuven, K. U. Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 2003. V+246 pp.

173. VINAIMONT Tom (26/02/2003)
The performance of One- versus Two-Factor Models of the Term Structure of Interest Rates. Leuven, K. U. Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen. 2003.

174. OOGHE Erwin (15/04/2003)
Essays in multi-dimensional social choice. Leuven, K. U. Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 2003. VIII+108 pp.

175. FORRIER Anneleen (25/04/2003)
Temporary employment, employability and training. Leuven, K. U. Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 2003.

176. CARDINAELS Eddy (28/04/2003)
The role of cost system accuracy in managerial decision making. Leuven, K. U. Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 2003. 144 pp.

177. DE GOEIJ Peter (02/07/2003)
Modeling Time-Varying Volatility and Interest Rates. Leuven, K. U. Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 2003. VII+225 pp.

178. LEUS Roel (19/09/2003)
The generation of stable project plans. Complexity and exact algorithms. Leuven, K. U. Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 2003.

179. MARINHEIRO Carlos (23/09/2003)
EMU and fiscal stabilisation policy: the case of small countries. Leuven, K. U. Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 2003.

180. BAESSENS Bart (24/09/2003)
Developing intelligent systems for credit scoring using machine learning techniques. Leuven, K. U. Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 2003.

181. KOCZY Laszlo (18/09/2003)
Solution concepts and outsider behaviour in coalition formation games. Leuven, K. U. Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 2003.

182. ALTOMONTE Carlo (25/09/2003)
Essays on Foreign Direct Investment in transition countries: learning from the evidence. Leuven, K. U. Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 2003.

183. DRIES Liesbeth (10/11/2003)
Transition, Globalisation and Sectoral Restructuring: Theory and Evidence from the Polish Agri-Food Sector. Leuven, K. U. Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 2003.

184. DEVOOGHT Kurt (18/11/2003)
Essays On Responsibility-Sensitive Egalitarianism and the Measurement of Income Inequality. Leuven, K. U. Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 2003.

185. DELEERSNYDER Barbara (28/11/2003)
Marketing in Turbulent Times. Leuven, K. U. Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 2003.

186. ALI Daniel (19/12/2003)
Essays on Household Consumption and Production Decisions under Uncertainty in Rural Ethiopia. Leuven, K. U. Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 2003.

187. WILLEMS Bert (14/01/2004)
Electricity networks and generation market power. Leuven, K. U. Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 2004.

188. JANSSENS Gust (30/01/2004)
Advanced Modelling of Conditional Volatility and Correlation in Financial Markets. Leuven, K. U. Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 2004.

189. THOEN Vincent (19/01/2004)
On the valuation and disclosure practices implemented by venture capital providers. Leuven, K. U. Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 2004.

190. MARTENS Jurgen (16/02/2004)
A fuzzy set and stochastic system theoretic technique to validate simulation models. Leuven, K. U. Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 2004.

191. ALTAVILLA Carlo (21/05/2004)
Monetary policy implementation and transmission mechanisms in the Euro area. Leuven, K. U. Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 2004.

192. DE BRUYNE Karolien (07/06/2004)
Essays in the location of economic activity. Leuven, K. U. Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 2004.

193. ADEM Jan (25/06/2004)
Mathematical programming approaches for the supervised classification problem. Leuven, K. U. Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 2004.

194. LEROUGE Davy (08/07/2004)
Predicting Product Preferences: the effect of internal and external cues. Leuven, K. U. Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 2004.

195. VANDENBROECK Katleen (16/07/2004)
Essays on output growth, social learning and land allocation in agriculture : micro-evidence from Ethiopia and Tanzania. Leuven, K. U. Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 2004.

196. GRIMALDI Maria (03/09/2004)
The exchange rate, heterogeneity of agents and bounded rationality. Leuven, K. U. Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 2004.

197. SMEDTS Kristien (26/10/2004)
Financial integration in EMU in the framework of the no-arbitrage theory. Leuven, K. U. Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 2004.

198. KOEVOETS Wim (12/11/2004)
Essays on Unions, Wages and Employment. Leuven, K. U. Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 2004.

199. CALLENS Marc (22/11/2004)
Essays on multilevel logistic Regression. Leuven, K. U. Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 2004.

200. RUGGOO Arvind (13/12/2004)
Two stage designs robust to model uncertainty. Leuven, K. U. Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 2004.

201. HOORELBEKE Dirk (28/01/2005)
Bootstrap and Pivoting Techniques for Testing Multiple Hypotheses. Leuven, K. U. Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 2005.

202. ROUSSEAU Sandra (17/02/2005)
Selecting Environmental Policy Instruments in the Presence of Incomplete Compiance. Leuven, K. U. Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 2005.

203. VAN DER MEULEN Sofie (17/02/2005)
Quality of Financial Statements: Impact of the external auditor and applied accounting standards. Leuven, K. U. Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 2005.

204. DIMOVA Ralitza (21/02/2005)
Winners and Losers during Structural Reform and Crisis: the Bulgarian Labour Market Perspective. Leuven, K. U. Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 2005.

205. DARKIEWICZ Grzegorz (28/02/2005)
Value-at-risk in Insurance and Finance: the Comonotonicity Approach. Leuven, K. U. Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 2005.

206. DE MOOR Lieven (20/05/2005)
The Structure of International Stock Returns: Size, Country and Sector Effects in Capital Asset Pricing. Leuven, K. U. Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 2005.

207. EVERAERT Greetje (27/06/2005)
Soft Budget Constraints and Trade Policies: The Role of Institutional and External Constraints. Leuven, K. U. Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 2005.

208. SIMON Steven (06/07/2005)
The Modeling and Valuation of complex Derivatives: The Impact of the Choice of the term structure model. Leuven, K. U. Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 2005.

209. MOONEN Linda (23/09/2005)
Algorithms for some Graph-Theoretical Optimization Problems. Leuven, K. U. Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 2005.

210. COUCKE Kristien (21/09/2005)
Firm and industry adjustment under de-industrialisation and globalization of the Belgian economy. Leuven, K. U. Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 2005.

211. DECAMPS Marc (21/10/2005)
Some actuarial and financial applications of generalized diffusion processes. Leuven, K. U. Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 2005.

212. KIM Helena (29/11/2005)
Escalation games: an instrument to analyze conflicts. The strategic approach to the bargaining problem. Leuven, K. U. Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 2005.

213. GERMENJI Etleva (06/01/2006)
Essays on the Economics of Emigration from Albania. Leuven, K. U. Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 2006.

214. BELIEN Jeroen (18/01/2006)
Exact and Heuristic Methodologies for Scheduling in Hospitals: Problems, Formulations and Algorithms. Leuven, K. U. Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 2006.

215. JOOSSENS Kristel (10/02/2006)
Robust discriminant analysis. Leuven, K. U. Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 2006.