FACULTY OF ECONOMICS AND BUSINESS ADMINISTRATION

CLASSIFYING YIELD SPREAD MOVEMENTS IN SPARSE DATA THROUGH TRIPLOTS

Carel Johannes van der Merwe

Supervisors: Prof. Dr. Tertius de Wet (Stellenbosch University) and Prof. Dr. Koen Inghelbrecht (Ghent University)

A dissertation submitted to Ghent University in fulfilment of the requirements for the degree of Doctor of Economics. This dissertation has also been presented at Stellenbosch University in terms of a joint-degree agreement for the degree of Doctor of Philosophy in Mathematical Statistics.

Academic year: 2019 - 2020

GHENT UNIVERSITY

UNIVERSITEIT iYUNIVESITHI STELLENBOSCH UNIVERSITY

By submitting this dissertation electronically, I declare that the entirety of the work contained therein is my own, original work, that I am the sole author thereof (save to the extent explicitly otherwise stated), that reproduction and publication thereof by Ghent University will not infringe any third party rights and that I have not previously in its entirety or in part submitted it for obtaining any qualification. This dissertation has also been presented at Stellenbosch University in terms of a joint-degree agreement.

To everyone by whom I was ever taught and
whom I had the honour to teach.

EXAMINATION BOARD

Prof. Dr. Patrick Van Kenhove,
Dean of the Faculty of Economics and Business Administration, Ghent University, Belgium

Prof. Dr. Koen Inghelbrecht,
Supervisor,
Department of Economics, Ghent University, Belgium

Prof. Dr. Tertius de Wet,
Supervisor,
Department of Statistics and Actuarial Science, Stellenbosch University, South Africa

Prof. Dr. Michèle Vanmaele,
Co-Supervisor,
Department of Applied Mathematics, Computer Science and Statistics, Ghent University, Belgium

Prof. Dr. Willie Conradie,
Co-Supervisor,
Department of Statistics and Actuarial Science, Stellenbosch University, South Africa

Prof. Dr. Frank de Jonghe,
Department of Economics, Ghent University, Belgium

Prof. Dr. Jan Annaert,
Department of Accountancy and Finance, University of Antwerp, Belgium

Prof. Dr. Riaan de Jongh,
Centre for Business Mathematics and Informatics, North West University, South Africa

Prof. Dr. Sugnet Lubbe,
Department of Statistics and Actuarial Science, Stellenbosch University, South Africa

ACKNOWLEDGEMENTS

It is almost an unfathomable task to give thanks to every person who contributed to the success of this thesis. These contributions came in many different forms and I would like to use this opportunity to acknowledge them.

Firstly, I would like to thank my examiners, Jan, Frank, Riaan, and Sugnet, for their time and effort to read through and comment on this thesis. Your comments were invaluable to improving the readability, presentation, and understanding of this manuscript, and also helped to identify a wide variety of further research opportunities. I am truly lucky to have had such a highly respected group of individuals on my panel.

Vir Prof de Wet wat die afgelope vier jaar soveel tyd aan my afgestaan het - baie dankie vir al die advies en leiding wat Prof vir my gegee het tydens my PhD. Ek kan net hoop om eendag in Prof se voetspore te kan volg!

Aan Koen, Michèle en Dries - bedankt voor de tijd en begeleiding die jullie me ook tijdens mijn PhD hebben gegeven. Jullie accepteerden me zonder aarzelen als een van je PhD-studenten en maakten mijn joint-PhD-reis zeer aangenaam.

I can imagine that doing a joint-PhD without the necessary institutional support would probably be as difficult as completing the PhD itself. I was fortunate enough not to face such a dilemma as I received amazing support throughout my PhD.

I would therefore like to thank the following people who assisted me during the journey, thereby ensuring that the joint-PhD was a success: Frank for introducing me to the Department of Economics at Ghent University; Rudi for agreeing to take me in as a joint-PhD student; Dorothy and her counterpart at UGent for setting up all the relevant agreements; Lidia, the Stellenbosch University International office, and their counterparts at UGent for providing me with the funding to visit Ghent University for a total of 6 months; Jaco and the Faculty of Economic and Management Sciences at Stellenbosch University for providing much required lecturer replacement funds; Paul and Prof Conradie for accommodating my PhD by lightening my

lecturing load and also providing me with funds in order to complete it and to attend conferences; lastly the support staff at both universities who I could always count on - Elizna, Heleen, and Sabine - thank you very much. A special note of thanks as well to Ann, Sandy, and the Faculty of Economics and Business Administration at UGent, for organising and administrating my private and public defenses.

I would also like to give a specific word of thanks to Retief and Chantelle for generously opening up their home for me during parts of my stay in Ghent, and also for Mieke for the time she took to proofread this thesis.

Finally, this journey would also not have been possible without the support of my family and friends - new and old. Aan my ouers, baie dankie vir julle ondersteuning en liefde my hele lewe lank - niemand kan vir beter ouers as julle vra nie. Aan my susters, en uitgebreide familie, dankie ook vir julle ondersteuning. Aan al my vriende wat oor die afgelope paar jaar my ondersteun het - elke keer as julle gevra het hoe dit gaan het dit soveel beteken - baie dankie hiervoor. To all my fellow PhD students at Ghent University and new friends that I have made during my stay in Ghent - bedankt dat jullie van mijn tijd in Gent een onvergetelijke ervaring hebben gemaakt.

The past four years have comprised a truly unforgettable journey - not only in my academic career, but in my personal life as well. I am truly grateful for the positive impact each and every person that crossed paths with me has had in my life.

Carel Johannes van der Merwe

January 2020

The computational resources (Stevin Supercomputer Infrastructure) and services used in this work were provided by the VSC (Flemish Supercomputer Center), funded by Ghent University, FWO and the Flemish Government department EWI.

CONTENTS

SAMENVATTING

In veel ontwikkelingslanden, waaronder Zuid-Afrika, zijn niet alle gegevens beschikbaar die nodig zijn om de reële waarde van financiële instrumenten te berekenen. Bovendien zijn bedrijven die niet over de nodige kwantitatieve vaardigheden beschikken in sommige gevallen terughoudend om de juiste reële waardering op te nemen door verkeerde technieken te gebruiken. Dit probleem is het meest opvallend met betrekking tot niet-genoteerde schuldinstrumenten.

Er zijn twee belangrijke inputs met betrekking tot de waardering van niet-beursgenoteerde schuldinstrumenten, met name de risicovrije curve en de renteverschuiving. Onderzoek naar deze twee componenten vormt de basis van dit proefschrift. Eerst wordt er een analyse uitgevoerd en een methode ontwikkeld om de risicovrije curves te benaderen zelfs wanneer data schaars is. Daarna wordt onderzocht of er voldoende aanwijzingen zijn voor een significante wijziging in de rendementsspreads van niet-genoteerde schuldinstrumenten. Om deze veranderingen te bepalen, werd een nieuwe methode ontwikkeld – triplot-classificatie met polybags – die zowel de visualisatie als classificatie van gegevens mogelijk maakt. Deze nieuwe classificatietechniek laat ook toe om misclassificatietarieven te beperken.

In het eerste artikel wordt een proxy voor de uitgebreide nulcurve berekend obv andere waarneembare inputs. Hiervoor wordt een simulatiebenadering gebruikt waarbij twee nieuwe technieken, gepermuteerde integer multiple lineaire regressie en geaggregeerde gestandaardiseerde modelscoring, worden geïntegreerd. Een Nelson Siegel-fit op een gereduceerde dataset, met een mix van één jaar forwardtarieven als proxy voor het nulpunt op de lange termijn presteerde relatief goed in de trainings- en testdatasets. Deze nieuwe methode maakt de benadering van risicovrije curven mogelijk wanneer er geen lange-termijnpunten beschikbaar zijn, en laat ook toe om de determinanten van de vorm van de rentecurve te berekenen door andere beschikbare gegevens te overwegen. De veranderingen in deze vormbepalende parameters worden in het laatste artikel gebruikt als determinanten voor veranderingen in de opbrengstspreads.

Voor het tweede artikel is een nieuwe classificatietechniek ontwikkeld die in het laatste artikel wordt gebruikt. Classificatietechnieken zijn moeilijk visueel te interpreteren en laten niet toe om de foutnegatieve en foutpositieve percentages te beperken. Voor sommige onderzoeksgebieden en praktische toepassingen zijn deze tekortkomingen van belang. In dit artikel worden classificatietechnieken gecombineerd met biplots, waardoor gelijktijdige visuele weergave en classificatie van de gegevens mogelijk is, wat resulteert in de zogenaamde triplot. Door verder polybags te integreren, wordt ook het vermogen om fouten van het misclassificatie-type te beperken geïntroduceerd. Een simulatieonderzoek en een toepassing tonen aan dat de resultaten van deze methode vergelijkbaar zijn met bestaande methoden, maar met toegevoegde visualisatievoordelen. Het artikel richt zich puur op het ontwikkelen van een statistische techniek die in elk veld kan worden toegepast. Zo betreft de toepassing in het artikel, bijvoorbeeld, een set van medische gegevens. In het laatste artikel wordt de techniek gebruikt om veranderingen in renteverschillen te meten.

Het derde artikel beschouwt veranderingen in rendementsspreads. Deze rendementspreads werden geanalyseerd via verschillende covariaten om te bepalen of er significante dalingen of stijgingen zouden hebben plaatsgevonden voor niet-genoteerde schuldinstrumenten. De methodologie bepaalt niet de nieuwe spread, maar geeft aan of de aanvankelijke impliciete spread dezelfde kan blijven of dat er een nieuwe spread moet worden bepaald. Deze renteverschuivingsbewegingen worden geclassificeerd met behulp van verschillende aandelen, rentetarieven, financiële ratio's en economische covariaten op een visueel interpreteerbare manier. Dit geeft ook een beter inzicht in hoe verschillende factoren de veranderingen in de rendementspreads beïnvloeden.

Ten slotte werd als aanvulling op elke paper een webgebaseerde applicatie gebouwd waarmee de lezer kan communiceren met alle gegevens en eigenschappen van de besproken methoden. De volgende links verschaffen toegang tot deze drie applicaties:

- Artikel 1: `https://carelvdmerwe.shinyapps.io/ProxyCurve/`
- Artikel 2: `https://carelvdmerwe.shinyapps.io/TriplotSimulation/`
- Artikel 3: `https://carelvdmerwe.shinyapps.io/SpreadsTriplot/`

SUMMARY

In many developing countries, including South Africa, all data that are required to calculate the fair values of financial instruments are not always readily available. Additionally, in some instances, companies who do not have the necessary quantitative skills are reluctant to incorporate the correct fair valuation by failing to employ the appropriate techniques. This problem is most notable with regards to unlisted debt instruments.

There are two main inputs with regards to the valuation of unlisted debt instruments, namely the the risk-free curve and the the yield spread. Investigation into these two components forms the basis of this thesis. Firstly, an analysis is carried out to derive approximations of risk-free curves in areas where data is sparse. Thereafter it is investigated whether there is sufficient evidence of a significant change in yield spreads of unlisted debt instruments. In order to determine these changes, however, a new method that allows for simultaneous visualisation and classification of data was developed - termed triplot classification with polybags. This new classification technique also has the ability to limit misclassification rates.

In the first paper, a proxy for the extended zero curve, calculated from other observable inputs, is found through a simulation approach by incorporating two new techniques, namely permuted integer multiple linear regression and aggregate standardised model scoring. It was found that a Nelson Siegel fit, with a mixture of one year forward rates as proxies for the long term zero point, and some discarding of initial data points, performs relatively well in the training and testing data sets. This new method allows for the approximation of risk-free curves where no long term points are available, and further allows for the determinants of the yield curve shape by considering other available data. The changes in these shape determining parameters are used in the final paper as determinants for changes in yield spreads.

For the second paper, a new classification technique is developed that was used in the final paper. Classification techniques do not easily allow for visual interpretation, nor do they usually allow for the limitation of the false negative and positive

error rates. For some areas of research and practical applications these shortcomings are important to address. In this paper, classification techniques are combined with biplots, allowing for simultaneous visual representation and classification of the data, resulting in the so-called triplot. By further incorporating polybags, the ability of limiting misclassification type errors is also introduced. A simulation study as well as an application is provided showing that the method provides similar results compared to existing methods, but with added visualisation benefits. The paper focuses purely on developing a statistical technique that can be applied to any field. The application that is provided, for example, is on a medical data set. In the final paper the technique is applied to changes in yield spreads.

The third paper considered changes in yield spreads which were analysed through various covariates to determine whether significant decreases or increases would have been observed for unlisted debt instruments. The methodology does not specifically determine the new spread, but gives evidence on whether the initial implied spread could be left the same, or whether a new spread should be determined. These yield spread movements are classified using various share, interest rate, financial ratio, and economic type covariates in a visually interpretive manner. This also allows for a better understanding of how various factors drive the changes in yield spreads.

Finally, as supplement to each paper, a web-based application was built allowing the reader to interact with all the data and properties of the methodologies discussed. The following links can be used to access these three applications:

- Paper 1: `https://carelvdmerwe.shinyapps.io/ProxyCurve/`
- Paper 2: `https://carelvdmerwe.shinyapps.io/TriplotSimulation/`
- Paper 3: `https://carelvdmerwe.shinyapps.io/SpreadsTriplot/`

OPSOMMING

In baie ontwikkelende lande, insluitend Suid-Afrika, is al die inligting wat benodig word om die billike waardes van finansiële instrumente te bereken, nie altyd geredelik beskikbaar nie. In sommige gevalle is ondernemings, wat nie oor die nodige kwantitatiewe vaardighede beskik nie, teësinnig om die regte billike waardasie te bereken deur nie-toepaslike tegnieke te gebruik. Hierdie probleem is veral opvallend ten opsigte van ongenoteerde skuldinstrumente.

Daar is twee hoof insette met betrekking tot die waardasie van ongenoteerde skuldinstrumente, naamlik die risiko-vrye kromme en die opbrengskoersspreiding. Die ondersoek na hierdie twee komponente vorm die basis van hierdie tesis. Eerstens word ’n analise uitgevoer om benaderings vir die risiko-vrye kurwes af te lei in areas waar die data skaars is. Daarna word ondersoek gedoen om vas te stel of daar voldoende bewyse is van betekenisvolle veranderinge in die opbrengskoersspreiding van ongenoteerde skuldinstrumente. Ten einde hierdie veranderinge te bepaal, is ’n nuwe metode wat gelyktydige visualisering en klassifikasie van data moontlik maak, ontwikkel - genaamd tri-stipping-klassifisering met poli-sakke. Hierdie nuwe klassifikasietegniek het ook die vermoë om wanklassifikasiekoerse te beperk.

In die eerste artikel word ’n benadering vir die uitgebreide nul-kromme bereken uit ander waarneembare insette. Dit word gevind deur middel van ’n simulasiebenadering deur twee nuwe tegnieke, naamlik gepermuteerde heelgetal meervoudige liniêre regressie en totale gestandaardiseerde model-telling, te gebruik. Dit is gevind dat ’n Nelson Siegel-passing, met ’n kombinasie van een jaar vooruitkoerse as benaderings vir die langtermyn nulpunt, en ’n mate van weglating van die aanvanklike datapunte, relatief goed in die leer en toetsing van datastelle presteer. Hierdie nuwe metode maak voorsiening vir die benadering van risiko-vrye krommes waar geen langtermynpunte beskikbaar is nie. Dit maak ook voorsiening vir die komponente van die opbrengskrommevorm deur ander beskikbare data in ag te neem. Die veranderinge in hierdie vormbepalingsparameters word in die finale artikel as komponente vir veranderinge in opbrengskoersspreidings gebruik.

In die tweede artikel word ’n nuwe klassifikasietegniek ontwikkel wat in die finale artikel gebruik word. Klassifikasietegnieke laat nie maklik visuele interpretasie toe nie, en maak gewoonlik ook nie die beperking van die vals negatiewe en positiewe foutkoerse moontlik nie. Hierdie tekortkominge is belangrik vir sommige navorsings- en praktiese toepassingsareas. In hierdie artikel word klassifikasietegnieke gekombineer met bi-stippings, waardeur die data gelyktydig visueel voorgestel en geklassifiseer word, wat die sogenaamde tri-stipping tot gevolg het. Deur poli-sakke in te bring, word die vermoë om foute in die wanklassifikasie te beperk geïnkorporeer. ’n Simulasie-studie sowel as ’n toepassing word word geïllustreer. Dit toon aan dat die metode soortgelyke resultate lewer in vergelyking met die bestaande metodes, maar met ekstra visualiseringsvoordele. Die artikel fokus slegs op die ontwikkeling van ’n statistiese tegniek wat op enige veld toegepas kan word. Die toepassing wat byvoorbeeld verskaf is, was op ’n mediese datastel. In die finale artikel word die tegniek op veranderinge in opbrengskoersspreidings toegepas.

In die derde artikel word veranderinge in opbrengskoersspreidings ondersoek en word dit deur middel van verskillende ko-variate ontleed om te bepaal of betekenisvolle daling of stygings by ongenoteerde skuldinstrumente waargeneem word. Die metodologie bepaal nie die nuwe spreiding spesifiek nie, maar lewer ’n bewys of die aanvanklike geïmpliseerde spreiding dieselfde gelaat kan word, of dat ’n nuwe spreiding bepaal moet word. Hierdie opbrengskoersspreidingbewegings word op ’n visueel interpretatiewe wyse geklassifiseer met behulp van verskillende aandeel-, rentekoers-, finansiële verhouding- en ekonomiese tipe ko-variate. Dit gee ook ’n beter begrip van hoe verskillende faktore die veranderinge in opbrengskoerse beïnvloed.

Ten slotte, aanvullend tot elke artikel, is ’n webtoepassing gebou wat die leser in staat stel om met al die data en eienskappe van die metodologieë wat bespreek is, te eksperimenteer. Die volgende skakels kan gebruik word om toegang tot hierdie drie toepassings te verkry:

- Artikel 1: `https://carelvdmerwe.shinyapps.io/ProxyCurve/`
- Artikel 2: `https://carelvdmerwe.shinyapps.io/TriplotSimulation/`
- Artikel 3: `https://carelvdmerwe.shinyapps.io/SpreadsTriplot/`

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

APPROXIMATING RISK-FREE CURVES IN SPARSE DATA ENVIRONMENTS

A shortened version of this paper was published in Finance Research Letters 26 (2018) pp. 112–118. The published paper can be found at http://doi.org/10.1016/j.frl.2017.12.016. This paper was co-authored with supervisors (at the time of publication) Dries Heyman and Tertius de Wet.

Abstract

Accounting standards require one to minimise the use of unobservable inputs when calculating fair values of financial assets and liabilities. In emerging markets and less developed countries, zero curves are not as readily observable over the longer term, as data are often more sparse than in developed countries. A proxy for the extended zero curve, calculated from other observable inputs, is found through a simulation approach by incorporating two new techniques, namely permuted integer multiple linear regression and aggregate standardized model scoring. A Nelson Siegel fit, with a mixture of one year forward rates as proxies for the long term zero point, and some discarding of initial data points, was found to perform relatively well in the training and testing data sets.

1 Introduction

The International Accounting Standards Board defines, in the International Financial Reporting Standards (IFRS) 13 *Fair Value Measurement*, the fair value for financial instruments as the price that would be received from selling an asset or paid to transfer a liability in an orderly transaction between market participants at the measurement date, i.e. an exit price. The definition of fair value is similar to

that of the Financial Accounting Standards Board's Accounting Standards Codification (ASC) Topic 810 (formerly, Statement of Financial Accounting Standards (SFAS) 157) *Fair Value Measurement*. These fair value accounting standards are relatively similar, with the most significant differences between them being the recognition of day one gains and losses, accounting for alternative investments, some quantitative sensitivity disclosure requirements, and disclosure exemptions (RMS US, 2012).

Further guidance is provided in the standards in terms of considerations when deciding the most appropriate method and inputs to determine the fair value of a financial instrument. Considerations include the condition, location, restriction of sale, the principal or most advantageous market, assumptions market participants would use, as well as maximising the use of observable inputs in the fair value calculation. These inputs are classified into three levels, with the first level being broadly directly observable inputs, the second level inputs that are derived via models from other observable inputs, and lastly unobservable inputs.

Some emerging markets and less developed countries, however, lack these observable inputs. While there exists a vast amount of research on parameterising the interest rate curve, extrapolating it, and forecasting it as well, very little research has been and is being done on extending it in sparse data environments. The approach followed in this research was to simulate such sparse environments from data rich environments and find methods that perform well in extrapolating curves under these conditions.

This paper focuses specifically on zero coupon risk-free curves, and finds a proxy that can be used to obtain the extended curve through other observable inputs and specified models, allowing them to be considered as second level, rather than third level inputs under IFRS 13. The approximations are found through a phased simulation[1] approach which incorporates two new techniques, namely permuted integer multiple linear regression (PIMLR) and aggregate standardised model scoring (ASMS).

The approach followed deviates from the classical econometric approaches and attempts to utilise the power of high performance computing to solve the research question. The PIMLR and ASMS are techniques which were designed to be used in

[1] When referring to simulation in the text it refers to repetitive evaluation of models over a set of predefined models and input values.

conjunction with high performance computing - they are discussed in more detail later.

The research provides a method for obtaining a possible extrapolation of the zero curve. It can easily be used within the valuation of financial instruments, or could possibly be used as a starting value for the pricing of them as well.

2 Background and literature overview

Kumarasiri and Fisher (2011) summarised Pacter's (2007) concerns regarding the application of fair value measurement in developing countries. The first being that inactive markets cause unreliable fair value estimates due to infrequent transactions, large bid-ask spreads, and market prices only being influenced by a few market participants or transactions. Secondly, there exists a trade-off between cost and benefit of implementing sophisticated fair valuation techniques. Furthermore, there are significant skill shortages in these countries - not only in-house, but externally as well.

Further, various transactions are entered into with related parties, hence it might be struck at non-market prices causing mismatches between market implied prices. Market prices could also be influenced by government, and therefore might not reflect normal market interactions. Also, a weak regulatory environment in some developing countries, could see low compliance with financial reporting standards. Lastly, lack of valuation standards and guidance on how to determine fair value raises additional concerns.

In addition to the above, some other studies have been performed with regards to the appropriateness of fair value accounting. Palea and Maino (2013) investigated whether the application of IFRS 13 for private equity valuation actually does contribute to the enhancing of transparency and comparability in financial statements (stated as one of the objectives in the EU Regulation 1606/2002). This also relates to the IASB's Conceptual Framework for Financial Reporting, where one of the fundamental qualitative characteristics of useful (financial statement) information is to provide a faithful representation of the underlying events and transactions, which includes completeness.

Another relates to the enhancing qualitative characteristics with regards to the comparability of financial statements. Benston (2008) found that fair values other than those that are directly observable in the market, could be manipulated easily and often are difficult to verify. Laux and Leuz (2009) discussed the different views regarding fair value accounting, and pointed to further research. Barth (2004) investigated the impact of the volatility of estimates on financial statements due to fair value, Penman (2007) discussed the benefits and disadvantages of fair value over historical cost estimates for various assets, while Ryan (2008) criticised the definition and measurement of fair value during a financial crisis.

Kumarasiri and Fisher (2011) surveyed 156 Sri Lankan practitioners with regards to their perception of fair value accounting. They identified various areas for future research, one of which includes the extension of their study to other developing countries due to the concerns raised regarding the credibility of financial statements prepared on the basis of fair value accounting in developing markets. They further stated that there is a perceived lack of technical guidance for preparers and auditors regarding fair valuation in developing countries, and that further research should consider the optimal nature, form, and source of such guidance.

As an example, some issued application guidance does exist with regards to credit value adjustments, such as, EY (2014), Deloitte (2013), PwC (2013), and KPMG (2015). While these all tend to agree on the standard guidance as per Gregory (2012), there is some degree of divergence on the approximation approaches. The issued guidance in most cases merely states what has been observed from market participants, and often lacks the underlying (published) scientific research supporting the methods.

Therefore, in summary it can be seen that even in developed markets, after the crisis there is a concern about fair value in certain asset classes, and in the developing markets, this problem is worse and structural.

3 Research methodology

It is clear from the above discussion that there is a certain expectation from the fair value accounting standards that an appropriate fair value be calculated, or

at a minimum that the relevant risks be identified and incorporated in the valuation. Paragraph 11 in IFRS 13 also states that 'an entity shall take into account the characteristics of the asset or liability if market participants would take those characteristics into account when pricing the asset or liability at the measurement date', which indicates that the notional of 'relevant' is not absolute but relative to market practices for a given use.

The focus of this paper will be the approximation of the zero coupon risk-free curves that can be used in valuation in sparse data environments. In markets where risk-free curves are sparse and it is required to value unlisted financial instruments, an estimation of the relevant curve needs to be obtained.

A simulation study was designed and performed in order to find the overall average optimal proxy for the risk-free rates where there are no data in the longer end of the curve. This is done through fitting various different models to an artificially created sparse environment and comparing it to the actual observed rates. The models were then scored to determine the average best performing model across various data sets. Out-of-sample data sets were used to test for consistency of the overall results. The observed data points were interpolated up to the relevant sparse environment, after which it is extrapolated using various techniques discussed later.

Figure 1.1 provides a graphical representation of the approach followed.

In the figure assume that the data represented by the line AD are known, and an artificially sparse environment is created by discarding CD. This allows for various curves ($A'D'$) to be fitted to the remaining AC data points. In order to improve a fit over the area of interest CD, some initial data (AB) is truncated and a data point E is added. The goodness of fit of the newly fitted curve $C'D'$, based on BC and E, is then measured by considering the squared differences between CD and $C'D'$. The single additional data point was chosen in order to 'pull' the longer term estimated curve towards the true values, something which would be difficult to accomplish by only considering the sparse data. The original curve on a specific date t is denoted by $Z_t(\tau) = R_t(0, \tau)$, where τ indicates the relevant term.

The simulation study was split into two phases, the estimation methods for the additional data point forms the basis of Phase I, while simulating the different variations of the fitted curve was done in Phase II. That is, the first defines a modelling frame-

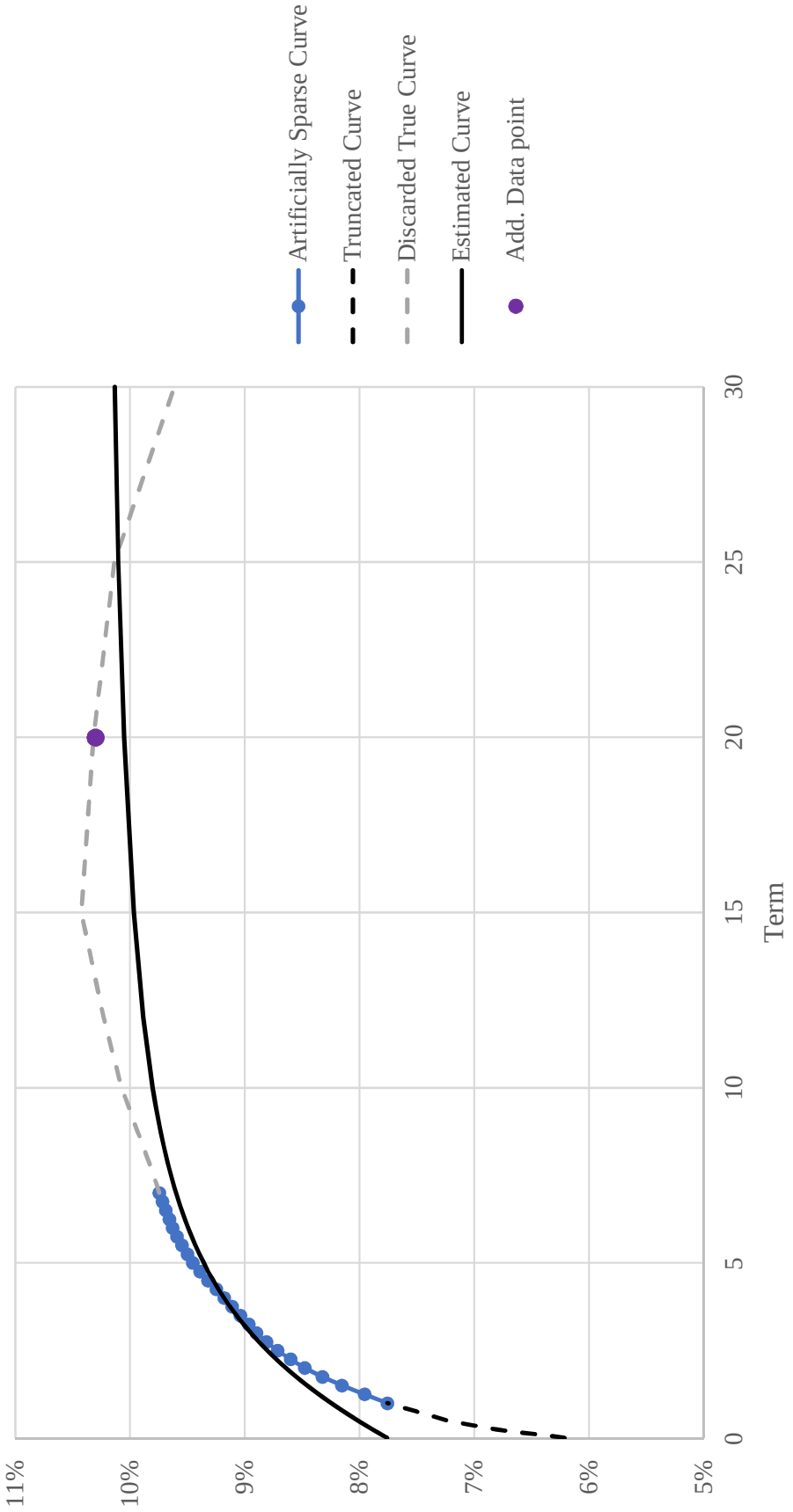


Figure 1.1: *Example of the various input data used in Phase II, where the Additional Data point is an output from Phase I.*

work for the additional data point and the second sets up the simulation design and carries out the simulation.

The framework for the additional data point followed an approach, whereby a training data set was used to obtain a number of additional data point models from presumed observable data through permuted integer multiple linear regression (PIMLR). These models were then scored using the aggregate standardised model scoring (ASMS) technique in order to obtain the better performing regression output across the training set.

Once these models were obtained, they were incorporated into the simulation design. The simulation fitted a number of different Nelson Siegel and Svensson parameterisations to artificially created sparse data environments, and found the better performing models again using the ASMS technique across the training data set. The process was repeated on a test data set (using the same modelling framework from Phase I), in order to test robustness and consistency of results.

In section 4 the various modelling techniques that were used are discussed. They include the Nelson-Siegel method used in estimating the additional data point as well as the curve fitting, together with the PIMRL and ASMS techniques. Thereafter, in section 5, the data that was used is discussed. This is followed, in section 6, with the practical implementation of the modelling framework for the additional data point, or rather the first phase, together with the results and a discussion thereof. The paper is concluded with the simulation framework for the second phase in section 7, the presentation of the results in section 8, a discussion thereof and some areas for future research in section 10.

4 Modeling framework

Three core theoretical concepts were incorporated into the research. The first being the well known Nelson Siegel parameterisation of the interest rate curve, together with the Svensson extension. The other two are new techniques, namely the PIMLR and ASMS. These three methods are theoretically discussed in the following sections, and are referred to in various other sections of the paper.

4.1 Nelson-Siegel family type parametric curve fitting

Nelson and Siegel (1987) introduced a simplified approach to the modelling of the term structure of interest rates. This comes after Friedman, in his 1977 paper, identified the need for one to be modelled with fewer parameters.

The approach was built on the notion that, if the term structure of interest rates can be generated by a differential equation, then the solution to that differential equation can be regarded as the forward rates.

Their formulation, after some further simplifying assumptions, is the following:

$$r(\tau) = \beta_0 + \beta_1 e^{-\lambda\tau} + \beta_2 (\lambda\tau) e^{-\lambda\tau}$$

with $r(\tau)$ the instantaneous forward rate. From this the zero rates can be obtained as:

$$R(0,\tau) \equiv \frac{1}{\tau}\int_0^\tau r(x)\,\mathrm{d}x = \beta_0 + (\beta_1+\beta_2)\left(\frac{1-e^{-\lambda\tau}}{\lambda\tau}\right) - \beta_2 e^{-\lambda\tau}.$$

Diebold and Li (2006) rewrote the above $R(0,\tau)$ in the form of level (L), slope (S), and curvature (C) coefficients. In fact, they only parametrised it with β's, while Diebold et al. (2006) renamed them L, S, and C. Additionally, they also made these parameters time varying.

This gave the following generalised form obtained for the interest rate curve (whether zero or yield):

$$R_t(0,\tau) = L_t + S_t\left(\frac{1-e^{-\lambda\tau}}{\lambda\tau}\right) + C_t\left(\frac{1-e^{-\lambda\tau}}{\lambda\tau} - e^{-\lambda\tau}\right) \tag{1.1}$$

They further noted that, approximations for the various factor coefficients can be estimated as follows:

$$\begin{aligned} L_t &\approx R_t(0,\infty) \approx R_t(0,10) \\ S_t &\approx -(R_t(0,\infty) - R_t(0,0)) \approx -(R_t(0,10) - R_t(0,0.25)) \\ C_t &\approx 2 \times R_t(0,2) - R_t(0,0.25) - R_t(0,10) \end{aligned} \tag{1.2}$$

After fitting the coefficients over various data sets, they found that little accuracy is lost if λ is fixed. The value of λ determines the maturity at which the loading on the medium-term (or curvature) factor achieves its maximum. They fixed λ at 0.0609 (if τ is measured in months, therefore it should be $12 \times 0.0609 = 0.7308$ when τ is the number of years), which maximizes the loading on the medium-term factor. This allowed them to compute the loading factors, and find the estimates of the factor coefficients through least squares.

Another well known extension of the Nelson Siegel parametrisation of the interest rate curve is that of Svensson (1994), which added an additional curvature term to (1.1), namely

$$C_t' \left(\frac{1 - e^{-\lambda' \tau}}{\lambda' \tau} - e^{-\lambda' \tau} \right).$$

The above parameterisations are conveniently parsimonious, however they do not fit the actual term structure on a specific time t, which would open up the possibility of arbitrage opportunities. They do, however, provide the user with intuitive latent factor loadings and coefficients.

One of the key concepts that will be utilised in going forward is that some of the coefficients could be proxied through the linear use of the current term structure as per (1.2).

In the next subsection, a new regression methodology is discussed, which allows one to obtain similar approximations.

4.2 Permuted Integer Multiple Linear Regression (PIMLR)

PIMLR is a regression method which allows one to limit the number of variables as well as placing certain restrictions on the values of the coefficients, resulting in a more intuitive interpretation of the result, allowing for similar results as in (1.2). While restricting coefficients to predetermined values and limiting the number of covariates will certainly reduce the accuracy of the regression, it is traded for the benefit of simplicity. In essence it allows for the use of a range of predefined regression coefficients and covariates to be evaluated through high performance computing.

More formally, let Y be a dependent variable (response), depending on the independent variables (covariates) $X_1, X_2, \ldots, X_{n_I}$. A linear model is assumed in their relationship, i.e.

$$Y = c_1 X_1 + c_2 X_2 + \ldots + c_{n_I} X_{n_I} + \epsilon,$$

with ϵ denoting the usual error term.

Furthermore, a sequence of predetermined coefficients $\{c_1, \ldots, c_{n_C}\}$ is chosen, with n_C arbitrary. In the application, these coefficients will be chosen as integers, i.e. $c_i \in \mathbb{Z}$, $i = 1, 2, \ldots, n_C$.

Now, consider size $n_P \leq n_I$ subsets of combinations of covariates $\{X'_1, X'_2, \ldots, X'_{n_P}\} \subseteq \{X_1, \ldots, X_{n_I}\}$, with repetitions not allowed, and of coefficients $\{c'_1, c'_2, \ldots, c'_{n_P}\} \subseteq \{c_1, \ldots, c_{n_C}\}$, with repetitions allowed. Combining each subset of covariates with each subset of coefficients, results in a total number of

$$n_M = \frac{n_I!}{(n_I - n_P)! n_P!} (n_C)^{n_P} \tag{1.3}$$

possible pairings. Given a certain choice for the coefficients, there could be duplicates, hence only the unique pairings need to be considered. For each of these pairings, the response can be estimated as

$$\hat{Y} = c'_1 X'_1 + c'_2 X'_2 + \ldots + c'_{n_P} X'_{n_P}. \tag{1.4}$$

Denote the estimated response based on a specific pairing, or model, M, by $\hat{Y}^{[M]}$, $M = 1, 2, \ldots, n_M$. Repeat this process for each of the n_t observations (in data set, D) of Y with the covariates $X_1, \ldots, X_{n_I}$. For the i^{th} observation y_i and the specified model M, denote the corresponding estimate, obtained from (1.4), by $\hat{y}_i^{[M]}$, $i = 1, 2, \ldots, n_t$, $M = 1, 2, \ldots, n_M$.

Define the root mean squared error (RMSE) as

$$\theta_{D,M} = \sqrt{\frac{1}{n_t}\sum_{i=1}^{n_t}\left(y_i - \hat{y}_i^{[M]}\right)^2}, \tag{1.5}$$

for model $M = 1, 2, \ldots, n_M$, and n_t the number of observations in data set $D = 1, 2, \ldots, n_D$.

In addition, the coefficients can be normalised through the transformation

$$c_j^* = \frac{c_j'}{\sum_{j=1}^{n_P} c_j'},$$

with $j = 1, 2, \ldots, n_P$. Here again, $\{c_1', c_2', \ldots, c_{n_P}'\} \subseteq \{c_1, \ldots, c_{n_C}\}$, however $c_k \in \mathbb{Z}_+ \cup \{0\}$ and $\sum_{j=1}^{n_P} c_j' > 0$. This is referred to as the fraction variation of the PIMLR.

In the following section a scoring methodology will be introduced, which allows one to identify the top performing model from a pool (of size n_M) of predetermined models, over n_D data sets. The methodology utilises the RMSE (similar to (1.5)), in the scoring process. It will be denoted by $\theta_{D,M}$, with D referring to a specific data set $D = 1, \ldots, n_D$, and M relating to a specific model $M = 1, \ldots, n_M$.

4.3 Aggregate standardised model scoring (ASMS)

If n_M models are fitted across n_D data sets, it becomes difficult to assess the RMSE-based goodness of fit. The ASMS technique helps to overcome this problem.

In choosing an average top performing model, the following requirements were incorporated:

1. *Standardisation:* The RMSEs could differ in size across the various data sets since the magnitude of the values in the data sets could differ. A standardised measure was therefore used.
2. *Transformation:* Models were ranked in terms of RMSE, with models having values close to each other not penalised, but assigned a relatively similar standardised measure while still maintaining ranking.
3. *Penalisation:* A model that performed well across most data sets, but extremely poorly in a single one, was significantly penalised.
4. *Removal of outliers:* Due to the nature of a simulation study, outliers could be present. Such values were removed since they could distort the results.

The following methodology was followed to address these requirements.

As defined above, let $\theta_{D,M}$ denote the RMSE for a specific model (M) and data set (D), and let $\theta_D^{(r)}$ denote the r^{th} ranked RMSE for data set D across the models.

Tukey (1977) defined an outlier detection rule with which to detect outliers based on the three quartiles of a probability distribution together with its inter-quartile range ($IQR = \theta^{(Q3)} - \theta^{(Q1)}$). Tukey's rule states that any observation above $\theta^{(Q3)} + 1.5 \times IQR$ can be considered an outlier. Now, applying Tukey's rule, the scoring methodology for model M, across the various data sets, is defined as:

$$\theta''_M = \left(\prod_{D=1}^{n_D} \theta'_{D,M} \right)^{\frac{1}{n_D}}, \tag{1.6}$$

with

$$\theta'_{D,M} = \begin{cases} 1 - \frac{\theta_{D,M} - \theta_D^{(1)}}{\theta_D^{(U)} - \theta_D^{(1)}} & : \theta_{D,M} < \theta_D^{(U)} \\ 0 & : \theta_{D,M} \geq \theta_D^{(U)} \end{cases} \tag{1.7}$$

and

$$\theta_D^U = \theta_D^{(Q3)} + 1.5 \left(\theta_D^{(Q3)} - \theta_D^{(Q1)} \right), \tag{1.8}$$

with $Q1 = 0.25 \cdot n_M$ and $Q3 = 0.75 \cdot n_M$, and $\theta_D^{(U)}$ the largest ranked RMSE smaller than θ_D^U.

Equations (1.7) and (1.8) allow for the removal of outliers. Furthermore, (1.7) standardises and transforms the resulting RSMEs such that an individual score between 0 and 1 is obtained. Finally, using the geometric average of these scores, in (1.6), the relevant models are penalised if they perform poorly. The resulting θ''_M, with $0 \leq \theta''_M \leq 1$, are then ranked in order to investigate how the models performed on average across the various data sets.

The simulation study incorporated various different variables that were required to complete the modelling framework. These will be discussed in the next section.

5 Description of data

The various response variables and covariates are discussed in general in this section, and in each case the specific values and data used will be stated. Interest rate data were sourced from Reuters using the unique identification codes for the various term points[2], while the non-term dependent data points were sourced from a combination of Quantec and Reuters.

All variables were obtained across a training set of currencies, namely ZAR, GBP, JPY, USD, and AUD. The training set is used in the first phase in order to define the full modelling framework, after which both the training and test data sets are used in the second phase. The test data sets were obtained from KES, HKD, CAD, NZD, and EUR. The lengths of the sparse environments, T_S, were chosen as 2, 6, and 9 years. Here, T_S is equivalent to point C in figure 1.1.

Table 1.1 provides a more detailed description of the data. The data were chosen such that only dates with observed terms longer than 10 years were included, and where outliers were observed (in the case of KES), they were removed. For this reason, it can be observed from the table that not all currencies had complete data

[2]Note that, while different bootstrapping techniques could result in different zero curves, it is not expected that these will significantly influence the final result due to the nature of the PIMLR technique using predefined coefficients.

sets over the full sampling period from the start of 2000. The standard zero coupon swap curve as defined by Reuters was used for each currency.[3]

Table 1.1: *Description of training (in-sample) and testing (out-of-sample) data used.*

Training data			**Testing data**		
Currency	**Timespan**	**Data points**	**Currency**	**Timespan**	**Data points**
AUD	2005/05/16 - 2016/10/31	2894	NZD	2010/05/31 - 2016/10/31	1601
GBP	2001/01/02 - 2016/10/31	4035	EUR	2000/01/03 - 2016/10/31	4341
JPY	2000/08/29 - 2016/10/31	4034	HKD	2009/09/17 - 2016/10/31	1745
USD	2000/01/03 - 2016/10/31	4282	CAD	2002/11/19 - 2016/10/31	3539
ZAR	2005/06/09 - 2016/10/31	2867	KES	2010/10/21 - 2016/10/31	814
Total training data points		18 112	Total testing data points		12 040

5.1 Response variables

Four models were employed to obtain the additional data point at certain terms. These included predicting the zero rate, the bullet forward rate (defined later), and utilising the Nelson Siegel parameterisation with and without a proxy for the level parameter. For each of these models, a number of parameters, or response variables, needed to be estimated in order to obtain the additional data point. These four models together with their parameters are discussed in the following subsections. In each of the various models, the term τ was chosen as 10, 20, and 30 years.

[3] Further research could be done on how the choice of time window influences the results. The choice of time window for this research was done on the basis of trade-off on the availability of the various currencies' data points.

Zero coupon term structure, $Z_t(\tau)$

The first dependent variable was chosen to be the actual observed zero rates for certain terms, τ, over certain times (or dates), t. It is denoted by $R_t(0,\tau)$, $\tau \in [0,T]$, and $t = 1, 2, \ldots, n_t$. In some instances the zero rate for a certain term is referred to as $Z_t(\tau)$, with $\tau \in [0,T]$ and $t = 1, 2, \ldots, n_t$. Here, T denotes the last observable point on the actual term structure, and T_S the last observable point on the sparse term structure.

Bullet forward rates, $B_t^{[\tau_B]}(\tau)$

Next, the concept of bullet forward rates is introduced. These bullet rates were chosen such that, for any zero rate of term τ_B, an accompanying forward rate to a predetermined tenor (τ) could be obtained.

That is, given an arbitrary term τ_B, the bullet rate, $B_t^{[\tau_B]}$, was defined such that

$$B_t^{[\tau_B]}(\tau) = \frac{\tau R_t(0,\tau) - \tau_B R_t(0,\tau_B)}{\tau - \tau_B},$$

with $\tau_B \in [0,\tau)$, $\tau \leq T$, and $t = 1, 2, \ldots, n_t$.

Note that, these bullet rates can be used to calculate the zero rate through

$$Z_t(\tau) = \frac{\tau_B R_t(0,\tau_B) + (\tau - \tau_B) B_t^{[\tau_B]}(\tau)}{\tau}.$$

The value of τ_B was chosen as the last observable point in the sparse data environments (T_S), such that $\tau_B = T_S < \tau$.

Nelson Siegel parameters (without level proxy)

As described in section 4.1, through the application of the NS methodology on the fully observed curve, the three parameters, Level (L), Slope (S), and Curve (C), can be obtained.

They can be estimated over a single daily observation of the zero curve (considered here), as well as over a number of days. The estimated parameters are therefore denoted by L_t^{NS}, S_t^{NS}, and C_t^{NS}, with $t \in \{1, 2, \ldots, n_t\}$.

Note that these estimated parameters can be used to estimate the zero rate, by applying (1.1) to obtain

$$Z_t(\tau) = L_t^{NS} + S_t^{NS}\left(\frac{1-e^{-\lambda\tau}}{\lambda\tau}\right) + C_t^{NS}\left(\frac{1-e^{-\lambda\tau}}{\lambda\tau} - e^{-\lambda\tau}\right)$$

with some predetermined estimate of λ, $\tau \geq 0$, and $t \in \{1, 2, \ldots, n_t\}$.

The λ was fixed at 0.7308 (as discussed in section 4.1) after being tested for robustness at arbitrary different values of 0.375 and 1.125 which delivered similar results for the overall analysis. The chosen λ corresponds to a maximum of the curvature factor at 2.5 years, with the robustness λ's to years 4.8 and 1.6. The parameters were not calculated on an average basis, but rather on a daily basis.

Nelson Siegel parameters (with level proxy)

Here the same procedure was performed as in the previous case, except that the level parameter was not estimated but rather proxied separately. That is, L_t^{EST}, was chosen as some predetermined proxied value, after which the optimal fits for the slope and curve parameters were found, denoted by S_t^{NSL} and C_t^{NSL}. Using these, the zero rates were estimated as:

$$Z_t(\tau) = L_t^{EST} + S_t^{NSL}\left(\frac{1-e^{-\lambda\tau}}{\lambda\tau}\right) + C_t^{NSL}\left(\frac{1-e^{-\lambda\tau}}{\lambda\tau} - e^{-\lambda\tau}\right) \tag{1.9}$$

with some predetermined estimate of λ, $\tau \geq 0$, and $t \in \{1, 2, \ldots, n_t\}$.

Again, λ was fixed at 0.7308, and additionally L_t^{EST} was calculated as the average medium term one month forward rate, taken over

$$\underline{\tau}_A = \begin{cases} \{T\} & : T < 3 \\ \left\{3, 3\frac{3}{12}, 3\frac{6}{12}, \ldots, \min(T, 6)\right\} & : T \geq 3 \end{cases} \tag{1.10}$$

with T the last observable point on the actual term structure. The parameters were also calculated on a daily basis as before, and T in (1.10) was replaced with T_S in the second phase.

5.2 Covariates

The previous section described the various models together with their parameters (response variables obtained from the true data) that will be estimated with some covariate sets. Table 1.2 provides a summary of the three additional data points at $\tau = 10, 20$ and 30, as discussed in the previous section.

Table 1.2: *List of all methods and parameters estimated in Phase I for the additional data point.*

Model	**Additional data point:** $Z_t(\tau), \tau = 10, 20, 30$
Zero coupon term structure	$\hat{Z}_t(\tau) = \hat{R}_t(0, \tau)$
Bullet forward rates	$\hat{Z}_t(\tau) = \left(\tau_B R_t(0, \tau_B) + (\tau - \tau_B)\hat{B}_t^{[\tau_B]}(\tau)\right)/\tau,\ \tau_B = T_S < \tau$
Nelson Siegel (without level proxy)	$\hat{Z}_t(\tau) = \hat{L}_t^{NS} + \hat{S}_t^{NS}\left(\frac{1-e^{-\lambda\tau}}{\lambda\tau}\right) + \hat{C}_t^{NS}\left(\frac{1-e^{-\lambda\tau}}{\lambda\tau} - e^{-\lambda\tau}\right)$
Nelson Siegel (with level proxy)	$\hat{Z}_t(\tau) = L_t^{EST} + \hat{S}_t^{NSL}\left(\frac{1-e^{-\lambda\tau}}{\lambda\tau}\right) + \hat{C}_t^{NSL}\left(\frac{1-e^{-\lambda\tau}}{\lambda\tau} - e^{-\lambda\tau}\right)$

The covariates used to estimate the response variables were chosen in a manner as to represent observed variables within sparse data environments. They were broadly split into three categories, with one or more from each category used in the final analysis. These three categories are single term structure points, average term structure points, and other non-term dependent data points. When used as covariates, they are grouped together, and in no instance used in combination with each other. For the sparse data sets, the last observable term on the sparse term structure was chosen as $T_S \in \{2, 6, 9\}$, with $T_S \leq T$. This is done so that the last observable points in the covariates are always observed before the response variables. Therefore, from each of the 5 sparse training data sets, three sparse data environments were constructed. These data sets were then used to obtain the covariates.

Single term structure points

The first group of covariates comprises single points in the term structure. These take the form of $R(\tau_1, \tau_2)$ with $0 \leq \tau_1 < \tau_2 \leq T$ and T the term of the last observable point on the curve.

If $\tau_1 = 0$, then the single term structure point takes the form of a zero rate, simply denoted as $Z_t(\tau_2)$. If $\tau_1 > 0$, then the single term structure point takes the form of a forward rate from τ_1 for a term $\tau_F = \tau_2 - \tau_1$, denoted by $F_t^{[\tau_F]}(\tau_1)$, with $\tau_2 \leq T_S \leq T$.[4]

Given the details above, it is clear that the number of different variables will differ for a given sparse data environment. A standardised approach to obtain variables from the full term structure was therefore needed. To overcome this problem, equally spaced term points were taken from the available terms, i.e. for $R(\tau_1, \tau_2)$, with $0 \leq \tau_1 < \tau_2 \leq T_S$.

Fixing $\tau_F = \tau_2 - \tau_1$, implies that $0 \leq \tau_1 \leq T_S - \tau_F$. Now, by choosing n_B blocks of length $\tau_B = (T_S - \tau_F)/n_B$, results in $\tau_1 \in \{i \cdot \tau_B;\ i = 0, \ldots, n_B\}$. The resulting variables are therefore, $F_t^{[\tau_F]}(\tau_1)$, with the special case of $\tau_F = \tau_1 = 0$, and $0 \leq \tau_2 \leq T_S$, resulting in $Z_t(\tau_2)$.

For the data used in Phase I, three single term structure data sets were used, namely the zero coupon rates, $Z_t(\tau_2)$, the one-month ($1M$) forward rates, $F_t^{[1M]}(\tau_1)$, and one-year ($1Y$) forward rates, $F_t^{[1Y]}(\tau_1)$. The number of points, or data blocks, for each of these data sets was chosen as $n_B = 10$, resulting in 11 covariates for each sparse data set. That is $Z^{[i]}$, $F^{[1M,i]}$, and $F^{[1Y,i]}$ with $i = 0, 0.1, 0.2, \ldots, 1$ indicating the equally spaced term points corresponding to the data block.[5]

Average term structure points

The next group of covariates follows from the single term structure points as it takes the average of a set of points over a certain term. That is,

[4] Note that the last covariate forward point will always end at T_S.

[5] These can also be denoted by a block number, in combination with the amount of blocks, as per table 1.3.

$$\begin{aligned}\bar{F}_t^{[\underline{\tau},\tau_F]} &= \frac{1}{n_\tau \tau_F}\sum_{i=1}^{n_\tau}(\tau_i R_t(0,\tau_i) - (\tau_i - \tau_F) R_t(0,\tau_i - \tau_F)) \\ &= \frac{1}{n_\tau}\sum_{i=1}^{n_\tau} F_t^{[\tau_F]}(\tau_i - \tau_F)\end{aligned}$$

where n_τ denotes the number of term points over which is averaged and $\underline{\tau}$ is a vector containing the specific terms being averaged over, i.e. $\tau_i \in \underline{\tau}$, $i = 1, \ldots, n_\tau$. The averaging was only employed for forward rates, but could easily be adjusted to incorporate the average zero rates as:

$$\bar{Z}_t^{\tau} = \frac{1}{n_\tau}\sum_{i=1}^{n_\tau} R_t(0,\tau_i),$$

where $\{\tau_i\}$ indicates the respective zero term structure points that are incorporated.

Similar to the single term structure points, if a sparse data environment is present, then $\tau_{n_\tau} \leq T_S \leq T$.

The averaging employed consists of an average of either short (S), medium (M), or long (L) terms, i.e.

$$\begin{aligned}\underline{\tau}_S &= \left\{\tau_{S,1}, \ldots, \tau_{S,n_A^S}\right\} &&\in [\tau_S, \min(\tau_M, T_S)] \\ \underline{\tau}_M &= \left\{\tau_{M,1}, \ldots, \tau_{M,n_A^M}\right\} &&\in (\min(\tau_M, T_S), \min(\tau_L, T_S)] \\ \underline{\tau}_L &= \left\{\tau_{L,1}, \ldots, \tau_{L,n_A^L}\right\} &&\in (\min(\tau_L, T_S), T_S]\end{aligned}$$

The various covariates therefore take the form of $\bar{F}^{[X,\tau_F]}$, where $X = \underline{\tau}_S, \underline{\tau}_M, \underline{\tau}_L$ indicates the underlying terms used in the averaging, as described above.

For the purpose of the analysis, τ_F was taken as one month ($1M$), 6 months ($6M$), and 1 year ($1Y$). Additionally, $\tau_S = 1/365 + \tau_F$, $\tau_M = 3$, and $\tau_L = 6$. The possible values that τ could take were, $\underline{\tau} = \{1/365,\ 2/365,\ 7/365,\ 14/365,\ 1/12,\ 2/12,\ 0.25,\ 0.5,\ 0.75,\ 1,\ 1.25,\ 1.5,\ \ldots,\ 9.75,\ 10,\ 11,12,\ldots,50\}$. More explicitly, with writing $\{a_i\} + c$ as shorthand for $\{a_1 + c, a_2 + c, \ldots\}$,

$$\underset{\sim}{\tau}_S = \left\{ \frac{1}{365}, \frac{2}{365}, \frac{7}{365}, \frac{14}{365}, \frac{1}{12}, \frac{2}{12}, \frac{3}{12}, \frac{6}{12}, \frac{9}{12}, \ldots, \min(T_S, 3) - \tau_F \right\} + \tau_F \quad (1.11)$$

$$\underset{\sim}{\tau}_M = \begin{cases} \{T_S\} & : T_S \leq 3 \\ \{3\frac{3}{12}, 3\frac{6}{12}, 3\frac{9}{12}, \ldots, \min(T_S, 6)\} & : T_S > 3 \end{cases} \quad (1.12)$$

$$\underset{\sim}{\tau}_L = \begin{cases} \{T_S\} & : T_S \leq 6 \\ \{6\frac{3}{12}, 6\frac{6}{12}, 6\frac{9}{12}, \ldots, 10, 11, 12, \ldots, T_S\} & : T_S > 6 \end{cases} \quad (1.13)$$

Therefore, the total number of variables in the average term structure points cohort, is $\bar{F}^{[X,Y]}$, with $X \in \{\underset{\sim}{\tau}_S, \underset{\sim}{\tau}_M, \underset{\sim}{\tau}_L\}$ and $Y \in \{1/12, 6/12, 1\}$, i.e. 9 in total.

Non-term dependent data points

Finally, non-term dependent data points were incorporated in the analysis. These are denoted by $E_t^{[1]}, E_t^{[2]}, \ldots, E_t^{[n_{IE}]}$, with n_{IE} the total number of covariates. These variables are usually unrelated to each other. This, however, does not imply that they are uncorrelated.

In this study, the economic variables used were deposit rate, lending rate, money market rate, central bank policy rate, inflation (lagged by 4 months), real GDP growth rate (lagged by 4 months), and nominal GDP growth (lagged by 4 months). These covariates are denoted by $E^{[D]}$, $E^{[L]}$, $E^{[M]}$, $E^{[C]}$, $E^{[I]}$, $E^{[R]}$, and $E^{[N]}$. A lag of 4 months was chosen in order to be more conservative with regards to the availability of data in sparse data environments. As mentioned, these non-term dependent data points were sourced from a combination of Quantec and Reuters.

Summary

Table 1.3 provides a summary of all the covariates discussed in this section.

Table 1.3: *List of covariates incorporated in Phase I.*

Covariate group	**Description**	**List of covariates**
Single term structure points	Zero Rates	$Z_t^{[i]} = R_t(0, i \cdot (T_S/n_B))$, with $i = 0, \ldots, n_B$
	1 Month forward rates	$F_t^{[1M,i]} = R_t\left(i \cdot \frac{T_S - 1/12}{n_B}, i \cdot \frac{T_S - 1/12}{n_B} + 1/12\right)$, with $i = 0, \ldots, n_B$
	1 Year forward rates	$F_t^{[1Y,i]} = R_t\left(i \cdot \frac{T_S - 1}{n_B}, i \cdot \frac{T_S - 1}{n_B} + 1\right)$, with $i = 0, \ldots, n_B$
Average term structure points	Groups of average forward rates	$\bar{F}_t^{[\underline{\tau}, \tau_F]} = \frac{1}{n_\tau \tau_F} \sum_{i=1}^{n_\tau} (\tau_i R_t(0, \tau_i) - (\tau_i - \tau_F) R_t(0, \tau_i - \tau_F))$
Non-term dependent data points	Economic variables	Deposit rate (D), Lending rate (L), money market rate (M), central bank policy rate (C), inflation (lagged by 4 months) (I), real GDP growth rate (lagged by 4 months) (R), and nominal GDP growth (lagged by 4 months) - denoted with $E^{[\cdot]}$

6 Phase I: Calculation and results of additional data point methods

The different parameters from the additional data point models were estimated with the various covariate groups through the PIMLR. The RMSEs obtained for each combination of response variable and covariate sets were then scored according to the ASMS methodology across 5 training currency data sets and 3 artificially created sparse data environments, i.e. 15 data sets. This provided the top performing model that was used to estimate the additional data point in the simulation framework discussed in the following section. Given that estimates for each of Z_{10}, Z_{20}, Z_{30}, B_{10}, B_{20}, B_{30}, L^{NS}, S^{NS}, C^{NS}, S^{NSL}, and S^{NSL} were needed, and 5 covariate data sets ($Z^{[\cdot]}$, $F^{[1M,\cdot]}$, $F^{[1Y,\cdot]}$, $\bar{F}^{[\cdot,\cdot]}$, and $E^{[\cdot]}$) were used, a total number of 55 top performing estimation methods, across the 15 data sets were obtained. Figure A1.1 in the appendix provides a diagram detailing the approach followed in this section.

In the PIMLR process, the coefficients, $\{c_i\}$, were chosen as $\{-1, 0, 1\}$ for the integer variation, and $\{0, 1, 2\}$ for the fraction variation. The number of covariates

used in the regression was limited to $n_P = 3$ per covariate set.[6] All other response and covariate data were used as described in the previous sections. Table 1.4 provides a summary of the number of possible pairings for each of the covariate sets that would be linked to the response variables, as per (1.3), per integer/fraction variation:

Table 1.4: *Number of simulated variations per covariate set for both variations of the PIMLR and response variable.*

Simulation summary	$\boldsymbol{Z}^{[\cdot]}$	$\boldsymbol{F}^{[\mathbf{1M},\cdot]}$	$\boldsymbol{F}^{[\mathbf{1Y},\cdot]}$	$\bar{\boldsymbol{F}}^{[\cdot,\cdot]}$	$\boldsymbol{E}^{[\cdot]}$
Covariates (n_I)	11	11	11	9	7
Coefficients per method (n_C)	3	3	3	3	3
Covariates used (n_P)	3	3	3	3	3
Total variations	8910	8910	8910	4536	1890
Unique variations (n_M)	2883	2883	2883	1531	687

Given the above, it can therefore be seen that a total number of 10 867 variations were used across the 11 response variables and 3 sparse data sets. Therefore 358 611 model variations were fitted on daily data across 5 currency training data. Given that there were on average 3622 data points for each of the 5 currencies, it equated to a total number of 6.494 billion fits.

Only a single result will be described in detail, while the other results are given in the appendix. The result given was used as an example due to its importance in the results of the next phase of the research, and relates to the one year forward rates used to predict the zero rate.

Table 1.5 provides an example of the top combination of a response and covariate pairing for all the combinations of sparse environments and currencies. The specific response was the Z_{30}, with the covariate set as $F^{[1Y,\cdot]}$. Recall from table 1.4 that there were 2883 unique variations of this covariate, the top estimation took the form

$$\hat{Z}_{30} = -F^{[1Y,0.7]} + F^{[1Y,0.8]} + F^{[1Y,1]}, \tag{1.14}$$

[6] The values of these covariates and coefficients can be extended, however for this research they were chosen as the three smallest whole numbers for each of the PIMLR variations.

for which the resulting RMSEs, or $\theta_{D,M}$, are given. Through the use of (1.6) to (1.8), the resulting θ''_M was obtained. These were then compared to the other θ''_M in the response/covariate set combination in order to choose the highest performing model over the data sets.

The process followed in table 1.5 was repeated for the 10 867 unique model combinations for each of the 11 response variables, such that a top performing model from each covariate set was obtained for each response variable. For the response variable Z_{30}, the following top performing models were obtained from each of the covariate sets:

$$\hat{Z}_{30} = \begin{cases} -Z^{[0.5]} + Z^{[0.9]} + Z^{[1]} \\ -F^{[1M,0.5]} + F^{[1M,0.6]} + F^{[1M,0.9]} \\ -F^{[1Y,0.7]} + F^{[1Y,0.8]} + F^{[1Y,1]} \\ -\bar{F}^{[\underline{\tau}_S,1/12]} + \bar{F}^{[\underline{\tau}_S,6/12]} + \bar{F}^{[\underline{\tau}_L,1]} \\ 0.67E^{[L]} + 0.33E^{[R]} \end{cases}$$

The results for the other parameters are provided in tables A1.1 to A1.5 in the appendix, and were all used to obtain additional data points on the zero curve at 10, 20, and 30 years.

7 Phase II: Simulation design

Within the second phase, repetitive refits of variations of the Nelson Siegel and Svensson extensions are performed using the sparse data environments together with the additional data point. These fits are then used to forecast the zero rates into the so-called 'empty' part of the curve. These different fits' RMSEs are then evaluated using the ASMS.

That is, once the 55 models for the parameters of the additional data points from the first phase were obtained, the models could be incorporated in the second phase of the simulation, to obtain 60 additional zero points by using the models in table 1.2. That is, term points at 10, 20, and 30 years with the four different meth-

Table 1.5: *ASMS calculation for top performing unique combination of response Z_{30} and covariate set $F^{[1Y,\cdot]}$, as per (1.14).*

Data set		$\boldsymbol{\theta_{D,M}}$	$\boldsymbol{\theta_D^{(1)}}$	$\boldsymbol{\theta_D^{(Q1)}}$	$\boldsymbol{\theta_D^{(Q3)}}$	$\boldsymbol{\theta_D^{U}}$	$\boldsymbol{\theta_D^{(U)}}$	$\boldsymbol{\theta'_{D,M}}$**%**
AUD	2	1.14	1.07	1.25	7.73	17.47	15.43	99.49%
	6	0.53	0.45	0.68	8.00	18.98	18.45	99.51%
	9	0.6	0.35	0.55	8.30	19.92	19.34	98.70%
GBP	2	1.29	1.12	1.55	7.37	16.11	15.16	98.78%
	6	0.47	0.46	0.87	7.72	18.01	16.78	99.94%
	9	0.56	0.29	0.61	7.87	18.76	17.33	98.42%
JPY	2	1.70	1.14	1.77	2.53	3.67	3.67	77.89%
	6	0.72	0.50	1.25	2.37	4.06	4.06	93.82%
	9	0.82	0.39	0.96	2.64	5.16	5.16	91.09%
USD	2	1.99	1.66	2.26	7.43	15.19	14.71	97.46%
	6	0.27	0.27	1.04	7.52	17.24	17.21	100.00%
	9	0.45	0.15	0.63	7.97	18.96	18.94	98.39%
ZAR	2	1.45	1.19	1.55	14.2	33.16	29.46	99.08%
	6	1.41	0.82	1.16	15.04	35.85	33.48	98.20%
	9	1.39	0.87	1.14	15.35	36.67	34.33	98.43%
							$\boldsymbol{\theta''_M}$**%**	96.44%

ods, across the 5 covariate data sets. These models, together with numerous other variations (discussed below) were then applied to both the training and testing data sets in order to obtain the better overall performing parameterisation procedure in the sparse data environments. Here, only the ASMS technique is applied in order to score the different variations.

In order to approximate data points that are not observed in sparse data environments, a simulation of various parametrisation variations was performed. These variations were tested, and subsequently scored using the ASMS methodology. The variations in the models and data sets are described below.

The first variation, which was discussed in length in the previous section, relates to the additional data point obtained from some observable data in the sparse environments. This is used together with the artificially sparse data environments. It should further be noted that the 55 models that were obtained, were estimated from a training data set, and were kept the same when applied on the testing data

set. The combination of term, method, and covariates results in 60 variations of additional data points.

The second variation relates to the sparse data environments. These were chosen such that they represent various sparse environments observed in the market. In order to prevent overlaps in the simulated models, these sparse data environments were limited to terms shorter than the first additional data point generated (as per Phase I). That is, the sparse data environments could only be smaller than 10 years.

A further variation, also relating to the data points used in the parameterisation, is with regards to the discarding of initial term points. This was done in order to provide a possible better approximation of the long term points, when the sparse data environment is used together with the additional data point.

Finally, two types of parametrisation methods were applied to the data points described above, namely the Nelson Siegel and Svensson methods. The lambda values for these methods were chosen in such a way that it minimises the least square fit for each data set across λ. More specifically, for the Nelson Siegel, λ was limited to values that maximised the C factor loading between 0 and T_S. The Svensson extension limited the choice of λ to values that maximised C between 0 and the median term, and λ' to values that maximised C' between the median term and T_S.

Table 1.6 summarises the total number of combinations utilised in the simulation study. Furthermore, figure 1.1, together with figure A1.2 in the appendix, provides a graphical representation of the approach described above. Note that the 'error terms' form the input to the RMSE calculation which is then utilised in the ASMS procedure in order to obtain the better overall performing model. Furthermore, the 'estimated zero curve' relates to either the Nelson Siegel or the Svensson parameterisation of the artificially sparse zero curve (excluding the truncated zero curve), together with the additional data point.

The variations performed in the simulation described above, included 2 curve parameterisation methods, 3 initial data truncations, 60 additional data points, 10 currencies (average 3622 in-sample and 2408 out-of-sample), and three sparse environments, which resulted in 32.562 million fits. The 360 models' RMSEs across the 15 data sets in the training and testing data sets were then scored according to the ASMS and the results are summarised in the next section.

Table 1.6: *List of all simulation variations incorporated in Phase II.*

Simulation Variation	**Variations description**	**Number of variations**
Additional data points	Term points: Z_{10}, Z_{20}, Z_{30}	3
	Methods: Z, B, NS, NSL	4
	Covariate sets: $Z^{[\cdot]}$, $F^{[1M,\cdot]}$, $F^{[1Y,\cdot]}$, $\bar{F}^{[\cdot,\cdot]}$, $E^{[\cdot]}$	5
Initial data discarded	0, 0.5, 1.5 years	3
Model parameterisation	Nelson Siegel (NS), Svensson (SV)	2
Total model variations		360

8 Results and Interpretation

In order to obtain the better performing parameterisation of the zero curves in sparse environments, the different variations' RMSEs were scored according to the ASMS. The result of the top 10% performing variations, or models, for the in- and out-of-sample data is provided in table 1.7. The notation of table 1.6 is used.

Considering the results, the second model will be formalised below as a possible method to approximate the extension of the curve in a sparse data environment.[7]

Now, given a sparse zero curve environment, with $R(0,\tau)$, $\tau \in [0, T_S]$. Then $R(0,\tau)$, with $\tau > T_S$, can be approximated through obtaining the least squares parameter estimates of the Nelson Siegel method.

That is, through obtaining the $\hat{L}$, $\hat{S}$, $\hat{C}$, and $\hat{\lambda}$, that minimise the following sum of squared residuals (SSR)[8]:

$$SSR = \sum_{\forall \tau} \left(R'(0,\tau) - \hat{R}(0,\tau)\right)^2,$$

[7] An Excel implemented model can be obtained from the authors upon request.
[8] Here, λ is limited to values that maximise the curvature loading factor between 0 and 30.

Table 1.7: *Top 10% best in- and out-of-sample performing parameterisation variations. The additional data point description takes the form of Term point - Method - Covariate set*

In-sample Ranking ($n = 360$)	Model Parameter-isation	Initial data dis-carded	Additional data point	Score	Out-of-sample Ranking
12	NS	0.5	Z_{30} - NS - $F^{[1Y,\cdot]}$	79.7%	22
14	NS	0.5	Z_{30} - Z - $F^{[1Y,\cdot]}$	79.2%	10
15	NS	0.5	Z_{30} - NS - $\bar{F}^{[\cdot,\cdot]}$	79.2%	6
17	NS	1.5	Z_{30} - NS - $F^{[1M,\cdot]}$	79.0%	19
18	NS	1.5	Z_{30} - B - $F^{[1M,\cdot]}$	78.5%	21
19	NS	1.5	Z_{30} - Z - $F^{[1M,\cdot]}$	78.4%	23
20	NS	1.5	Z_{30} - NS - $F^{[1Y,\cdot]}$	78.4%	4
21	NS	0.5	Z_{20} - NS - $\bar{F}^{[\cdot,\cdot]}$	78.3%	9
27	NS	0.5	Z_{30} - Z - $\bar{F}^{[\cdot,\cdot]}$	78.0%	7
29	NS	1.5	Z_{30} - NS - $\bar{F}^{[\cdot,\cdot]}$	77.9%	1
30	NS	1.5	Z_{30} - Z - $F^{[1Y,\cdot]}$	77.7%	3
32	NS	1.5	Z_{30} - Z - $\bar{F}^{[\cdot,\cdot]}$	77.4%	2
33	NS	0	Z_{30} - NS - $\bar{F}^{[\cdot,\cdot]}$	76.8%	32
35	NS	1.5	Z_{30} - B - $F^{[1Y,\cdot]}$	76.8%	8
36	NS	0.5	Z_{30} - B - $\bar{F}^{[\cdot,\cdot]}$	76.7%	27

with

$$\hat{R}(0,\tau) = \hat{L} + \hat{S}\left(\frac{1-e^{-\hat{\lambda}\tau}}{\hat{\lambda}\tau}\right) + \hat{C}\left(\frac{1-e^{-\hat{\lambda}\tau}}{\hat{\lambda}\tau} - e^{-\hat{\lambda}\tau}\right),$$

over $\tau \in \{0.5 \leq \tau \leq T_S\} \cup \{30\}$, such that $R'(0,\tau) \in \{R(0,\tau) : 0.5 \leq \tau \leq T_S\} \cup \{\hat{Z}_{30}\}$, with

$$\hat{Z}_{30} = -F^{[1Y,7]} + F^{[1Y,8]} + F^{[1Y,10]},$$
$$F^{[1Y,i]} = R\left(i \cdot \frac{T_S - 1}{10}, i \cdot \frac{T_S - 1}{10} + 1\right).$$

Given that the other variations also performed relatively well out-of-sample, they can also be considered as some of the better performing models.

Various methods can be employed to connect the actual sparse curve to the extrapolated curve in order to make it continuous between the observed and fitted parts, these could range from simple interpolation to more advanced methods such as the Smith Wilson extrapolation technique (Smith and Wilson, 2001).

9 Web-based application

A web-based application (see Van der Merwe, 2019) was also built to show how the methods discussed in this paper are implemented. The user has access to all the various currencies discussed, change the sparsity of the data, change the determinants of the additional data point, change the initial discarded data, and choose between the various fits of the curve. A screenshot of the application can be seen in figure 1.2.

10 Discussion and Concluding Remarks

In the literature review and background in section 2, it was noted that there is a clear need for more guidance on fair valuation in developing countries. One of the most prevalent shortcomings of developing countries is their sparsity of data, one of which is the risk-free zero curves needed for fair valuation calculations.

In the absence of data, management requires an estimate of such values in order to estimate reliable fair values. IFRS 13 suggests that inputs estimated with models that use observable inputs (level 2) are preferred to unobservable inputs (level 3). The results presented in this paper provide a method to estimate unobserved zero rates from other observable data, resulting in level 2 inputs, rather than level 3.

The results allow not only increased transparency in the extrapolation of the zero coupon risk-free curve through the use of observable points, but also incorporate consistency between market participants. Additionally, this also provides the auditors of financial statements a much simpler task of assessing the reasonableness of the firm's estimate through the use of a standard model, rather than a management estimate. Having a level 2 instead of level 3 input will also reduce the additional disclosures required.

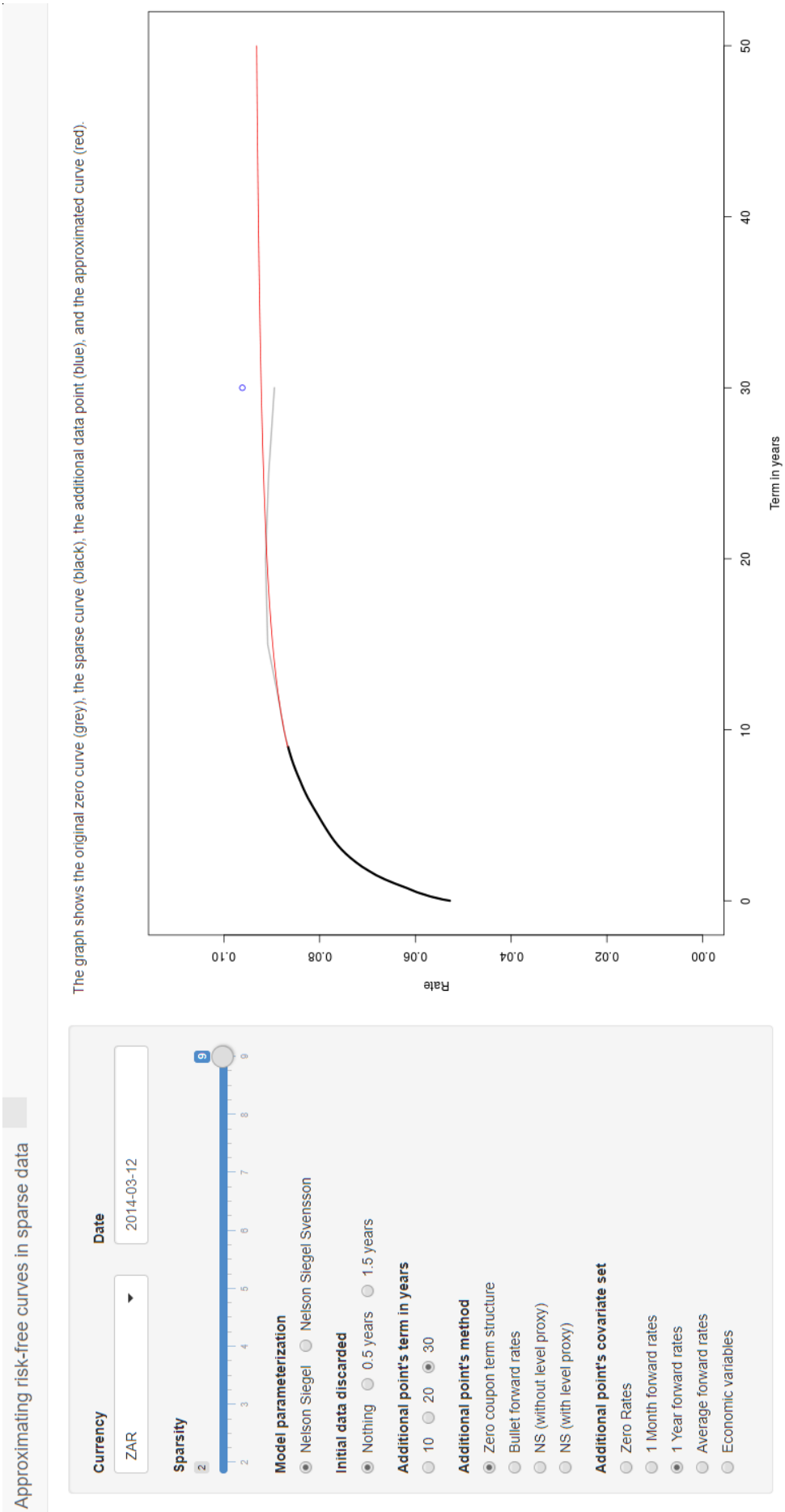


Figure 1.2: *A screenshot of the web-based application that was built as supplementary data for this paper. The link to the application can be found at: `https://doi.org/10.5281/zenodo.3355465`.*

The total wall time of the simulation study amounted to 105 days, a task that could not have been performed without the help of high performance computing. A limitation of this research still remains, however, that the results are limited to the models incorporated. Further research is planned to increase the number of models in order to find a more refined method.

Areas of further research include the application of other techniques to further smooth the extended curve from the observed points and alternative variations of the parameterisation procedure. Lastly, the new PIMLR and ASMS techniques can be applied easily to other simulation studies in sparse environments, such as for example credit spreads.

REFERENCES

Barth, M.E., 2004. Fair values and financial statement volatility, in: Claudio Borio, William C. Hunter, G.G.K., Tsatsaronis, K. (Eds.), Market Discipline Across Countries and Industries. MIT press, Oxford, pp. 323–333.

Benston, G.J., 2008. The shortcomings of fair-value accounting described in SFAS 157. Journal of Accounting and Public Policy 27, 101–114. URL: https://doi.org/10.1016/j.jaccpubpol.2008.01.001.

Deloitte, 2013. Fair value measurement of financial instruments under IFRS 13. a closer look. (accessed 29.06.17). URL: http://www.iasplus.com/en/publications/uk/closer-look/2013/april-2013.

Diebold, F.X., Li, C., 2006. Forecasting the term structure of government bond yields. Journal of econometrics 130, 337–364. URL: https://doi.org/10.1016/j.jeconom.2005.03.005.

Diebold, F.X., Rudebusch, G.D., Aruoba, S.B., 2006. The macroeconomy and the yield curve: a dynamic latent factor approach. Journal of econometrics 131, 309–338. URL: https://doi.org/10.1016/j.jeconom.2005.01.011.

EY, 2014. Credit valuation adjustments for derivative contracts. (accessed 29.06.17). URL: http://www.ey.com/Publication/vwLUAssets/EY-credit-valuation-adjustments-for-derivative-contracts/$FILE/EY-Applying-FV-April-2014.pdf.

Friedman, M., 1977. Time perspective in demand for money. The Scandinavian Journal of Economics 79, 397–416. URL: http://dx.doi.org/10.2307/3439699.

Gregory, J., 2012. Counterparty credit risk and credit value adjustment. volume 171. Wiley. URL: http://dx.doi.org/10.1002/9781118673638.

KPMG, 2015. Fair value measurement: Questions and answers. (accessed 29.06.17). URL: https://home.kpmg.com/content/dam/kpmg/pdf/2015/12/fair-value-qa-2015.pdf.

Kumarasiri, J., Fisher, R., 2011. Auditors' perceptions of fair-value accounting: Developing country evidence. International Journal of Auditing 15, 66–87. URL: http://dx.doi.org/10.1111/j.1099-1123.2010.00423.x.

Laux, C., Leuz, C., 2009. The crisis of fair-value accounting: Making sense of the recent debate. Accounting, organizations and society 34, 826–834. URL: https://doi.org/10.1016/j.aos.2009.04.003.

Nelson, C.R., Siegel, A.F., 1987. Parsimonious modeling of yield curves. Journal of business , 473–489URL: http://dx.doi.org/10.1086/296409.

Pacter, P., 2007. Fair value under ifrs: Issues for developing countries and smes. The Routledge Companion to Fair Value Financial Reporting. London: Routledge .

Palea, V., Maino, R., 2013. Private equity fair value measurement: a critical perspective on IFRS 13. Australian Accounting Review 23, 264–278. URL: http://dx.doi.org/10.1111/auar.12018.

Penman, S.H., 2007. Financial reporting quality: is fair value a plus or a minus? Accounting and business research 37, 33–44. URL: http://dx.doi.org/10.1080/00014788.2007.9730083.

PwC, 2013. Fair value measurements. (accessed 29.06.17). URL: http://www.pwc.com/en_US/us/cfodirect/assets/pdf/accounting-guides/pwc-fair-value-measurement-2015.pdf.

RMS US, 2012. U.S. GAAP vs. IFRS: Fair value measurements at-a-glance. (accessed 29.06.17). URL: http://rsmus.com/pdf/us-gaap-vs-ifrs-fair-value-measurements.pdf.

Ryan, S.G., 2008. Accounting in and for the subprime crisis. The accounting review 83, 1605–1638. URL: https://doi.org/10.2308/accr.2008.83.6.1605.

Smith, A., Wilson, T., 2001. Fitting yield curves with long term constraints. Technical Report. Bacon & Woodrow.

Svensson, L.E., 1994. Estimating and interpreting forward interest rates: Sweden 1992-1994. Technical Report. National Bureau of Economic Research. URL: http://dx.doi.org/10.3386/w4871.

Tukey, J.W., 1977. Exploratory data analysis. Addison-Wesley. URL: http://dx.doi.org/10.1002/bimj.4710230408.

Van der Merwe, C.J., 2019. carelvdmerwe/proxycurve. URL: https://doi.org/10.5281/zenodo.3355465.

APPENDICES

A Additional graphs and results

Various additional graphs and tables of results, referred to in the text, are provided in this appendix.

Table A1.1: *Best performing PIMLR results for the response variables as a function of the $Z^{[\cdot]}$ covariate set.*

$N^{[\cdot]}$	$\hat{Z}_{10}$	$\hat{Z}_{20}$	$\hat{Z}_{30}$	$\hat{B}_{10}$	$\hat{B}_{20}$	$\hat{B}_{30}$	$\hat{L}^{NS}$	$\hat{S}^{NS}$	$\hat{C}^{NS}$	$\hat{S}^{NSL}$	$\hat{C}^{NSL}$
0								+1		+1	
0.1											
0.2				−1	−1	−1	−1		+1		
0.3											+1
0.4											
0.5	−1	−1	−1								−1
0.6											
0.7	+1										
0.8											
0.9		+1	+1	+1	+1	+1	+1				
1	+1	+1	+1	+1	+1	+1	+1	−1	−1	−1	

Table A1.2: *Best performing PIMLR results for the response variables as a function of the $F^{[1M,\cdot]}$ covariate set.*

$F^{[1M,\cdot]}$	$\hat{Z}_{10}$	$\hat{Z}_{20}$	$\hat{Z}_{30}$	$\hat{B}_{10}$	$\hat{B}_{20}$	$\hat{B}_{30}$	$\hat{L}^{NS}$	$\hat{S}^{NS}$	$\hat{C}^{NS}$	$\hat{S}^{NSL}$	$\hat{C}^{NSL}$
0								+1		+1	
0.1									+1		
0.2	+0.2										
0.3											+1
0.4					−1						
0.5		−1	−1	−1		−1	−1				−1
0.6	+0.4	+1	+1	+1	+1	+1	+1				
0.7											
0.8											
0.9	+0.4	+1	+1	+1	+1	+1	+1	−1		−1	
1									−1		

Table A1.3: *Best performing PIMLR results for the response variables as a function of the $F^{[1Y,\cdot]}$ covariate set.*

$F^{[1Y,\cdot]}$	$\hat{Z}_{10}$	$\hat{Z}_{20}$	$\hat{Z}_{30}$	$\hat{B}_{10}$	$\hat{B}_{20}$	$\hat{B}_{30}$	$\hat{L}^{NS}$	$\hat{S}^{NS}$	$\hat{C}^{NS}$	$\hat{S}^{NSL}$	$\hat{C}^{NSL}$
0	+0.25							+1		+1	
0.1									+1		+1
0.2											
0.3											
0.4											−1
0.5											
0.6				−1	−1	−1	−1				
0.7			−1								
0.8		+0.2	+1	+1							
0.9	+0.5	+0.4			+1	+1	+1			−1	
1	+0.25	+0.4	+1	+1	+1	+1	+1	−1	−1		

Table A1.4: *Best performing PIMLR results for the response variables as a function of the $\bar{F}^{[\cdot,\cdot]}$ covariate set.*

$\bar{F}^{[\cdot,\cdot]}$	$\hat{Z}_{10}$	$\hat{Z}_{20}$	$\hat{Z}_{30}$	$\hat{B}_{10}$	$\hat{B}_{20}$	$\hat{B}_{30}$	$\hat{L}^{NS}$	$\hat{S}^{NS}$	$\hat{C}^{NS}$	$\hat{S}^{NSL}$	$\hat{C}^{NSL}$
$\underline{\tau}_{S,1/12}$		-1	-1	-1			-1	$+1$	$+1$	$+1$	
$\underline{\tau}_{M,1/12}$					-1	-1					-1
$\underline{\tau}_{L,1/12}$					$+1$	$+1$		-1	-1	-1	
$\underline{\tau}_{S,6/12}$	$+0.2$	$+1$	$+1$	$+1$			$+1$				
$\underline{\tau}_{M,6/12}$											
$\underline{\tau}_{L,6/12}$	$+0.4$			$+1$			$+1$				
$\underline{\tau}_{S,1}$											
$\underline{\tau}_{M,1}$	$+0.4$										$+1$
$\underline{\tau}_{L,1}$		$+1$	$+1$		$+1$	$+1$					

Table A1.5: *Best performing PIMLR results for the response variables as a function of the $E^{[\cdot]}$ covariate set.*

$E^{[\cdot]}$	$\hat{Z}_{10}$	$\hat{Z}_{20}$	$\hat{Z}_{30}$	$\hat{B}_{10}$	$\hat{B}_{20}$	$\hat{B}_{30}$	$\hat{L}^{NS}$	$\hat{S}^{NS}$	$\hat{C}^{NS}$	$\hat{S}^{NSL}$	$\hat{C}^{NSL}$
D											
L	+0.67	+1	+0.67	+1	+1	+0.67	+1	−1	−1		
M	+0.33	+1		+1						+1	
C		−1		−1				+1	+1	−1	
I											
R			+0.33			+0.33					
N											

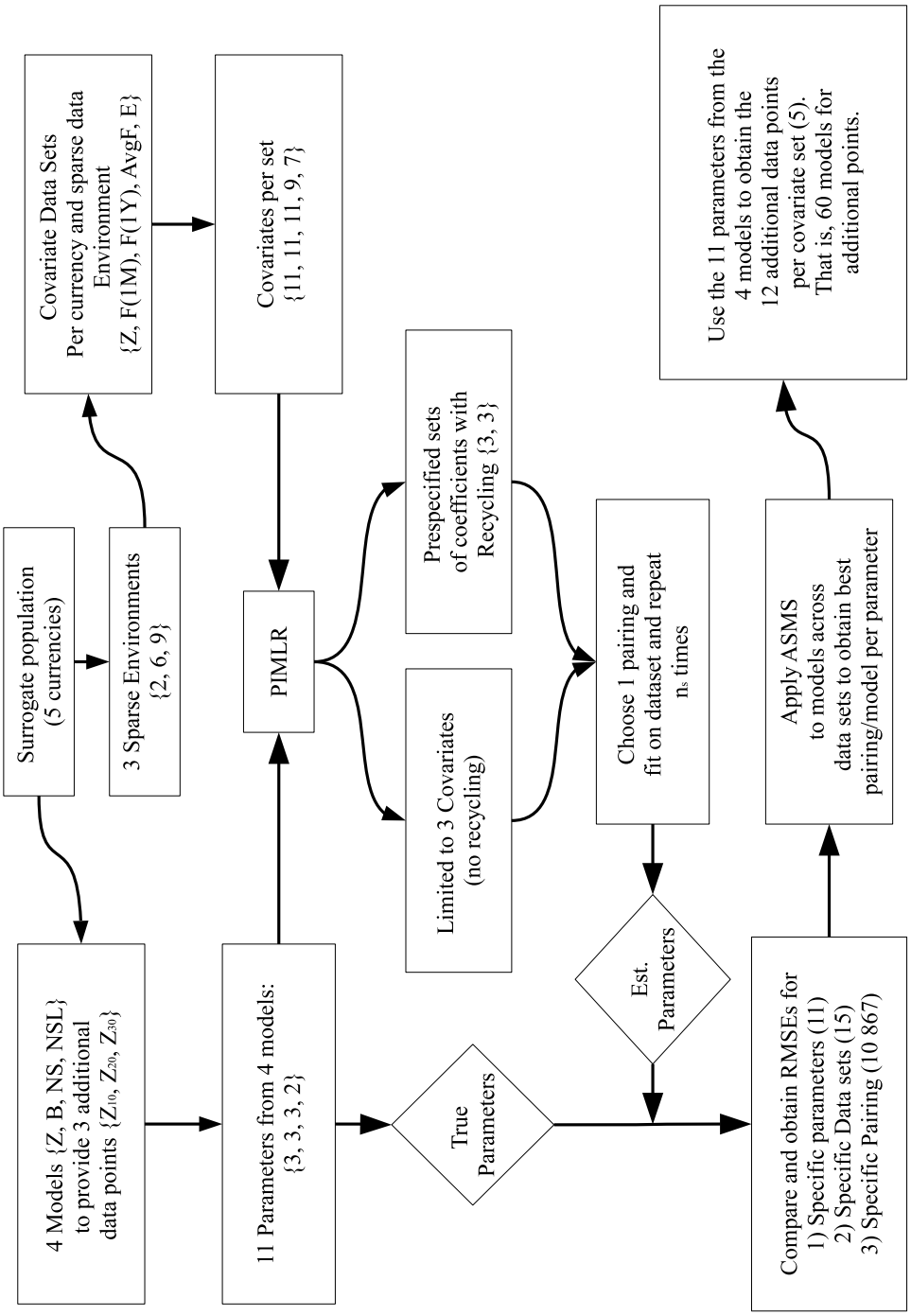


Figure A1.1: *Diagrammatic representation of Phase I.*

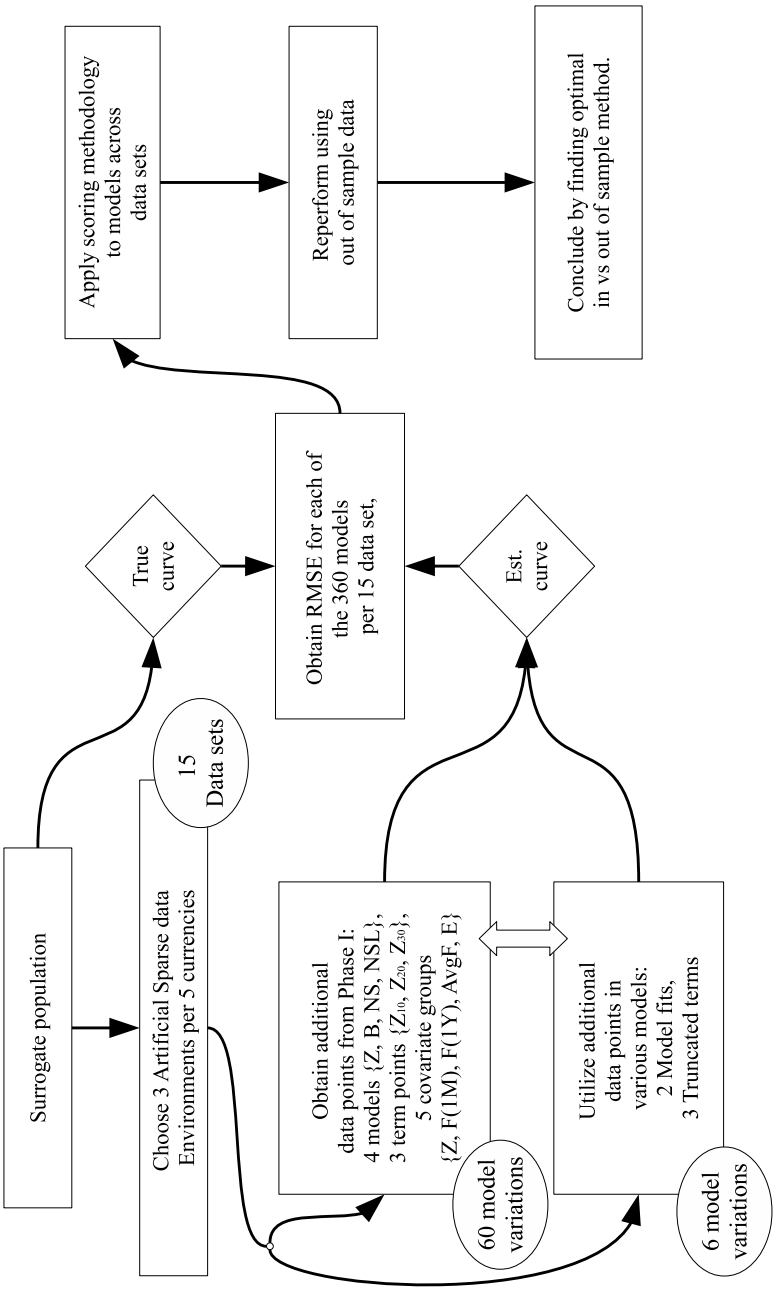


Figure A1.2: *Diagrammatic representation of Phase II.*

CHAPTER 2

TRIPLOT CLASSIFICATION WITH POLYBAGS

This working paper was submitted to an applied statistics journal and two anonymous referee reports were obtained. The suggested changes were incorporated and the paper will be resubmitted together with their addressed comments.

Abstract

Classification techniques do not allow for simple visual interpretation, nor do they usually allow for the limitation of false negative and positive error rates. In this paper, classification techniques are combined with biplots, allowing for simultaneous visual representation and classification of the data, resulting in the so-called triplot. By further incorporating polybags, the ability to limit misclassification type errors is introduced. A simulation study as well as an application is provided, showing that the method provides similar results as compared to existing methods, but with the added benefit of visualisation. A web-based application is also provided, allowing the user to interact with the data sets and methods discussed.

1 Introduction

Visualisation is important for obtaining information from large data sets, but visualisation is complicated for multivariate data. Furthermore, while high accuracy rates for classification techniques are preferred, in some instances, low frequencies of false positives and negatives are more important. A way in which these frequencies can be reduced is by refraining from classifying observations at high risk of misclassification; in other words, shifting the focus from *how*, to instead determining *when* to classify an observation.

The two shortcomings above are particularly relevant in areas where classification is usually treated as a black-box, as well as where the result of obtaining false positives and negatives has a more severe negative impact than the converse positive impact a correct classification would have. A remedy for this is to flag a given observation for further investigation if the classification technique is not powerful enough, thereby reducing the risk of misclassification.

One example of a field of inquiry and practical application where these limitations could be considerable is medicine. As a simple example, one could use data to classify a patient as having a disease or not using some classification technique. The preferred outcome is for the technique to correctly predict the patient's status. However, if a patient is erroneously diagnosed, the resulting adverse effects could be significant for both patient and doctor. Such incorrect classifications can be limited by not classifying the patient, but rather requiring further tests to obtain a more accurate diagnosis.

Additionally, the visualisation of the classification technique allows for further interpretation by the end-user, who would not necessarily possess the underlying mathematical background to understand black-box classification techniques, allowing for better understanding of the resulting classification.

This paper presents a new classification methodology, expanding on the biplot and triplot classification methodologies of Gardner-Lubbe (2016) and Aldrich et al. (2004). Four key properties are incorporated in the new proposed methodology. These are: (i) allowing for various underlying biplot methodologies to be used, (ii) determination of classification regions based on all data points in the training set in contrast to basing it only on class means (as in Aldrich et al. (2004)), (iii) allowing for the limitation of misclassification errors by considering outlying and overlapping observations from different classes, and (iv) the creation of a web-based application for the user to interact with the methodology.

The new methodology uses triplots (a combination of biplots and an underlying classification technique in a two-dimensional graph) which allow for the observations, variables, classes, and classification regions to be observed simultaneously. Through combining these triplots with sample density areas via α-bags, areas where sample points overlap significantly (indicating heightened risk of misclassification) can be identified and treated as 'unclassified'. These areas are termed 'polybags'.

The advantage of this new method of triplot classification with polybags is that it is visually interpretable and limits misclassification errors. For background on biplots and α-bags, see appendix A.

Section 2 discusses current methods that allow for visualisation of classification techniques, including that of Gardner-Lubbe (2016) and Aldrich et al. (2004). Thereafter, the new classification methodology is illustrated and discussed in section 3 using two randomly generated data sets, each with three classes. The first data set will be a 'perfect' data set, where the covariance matrices are the same for the three classes and variances are kept small. The second data set will be much more variable than the first. These two data sets are referred to as the 'small and similarly variable' (SSV) and 'large and differently variable' (LDV) data sets respectively. Both data sets were simulated from multivariate normal distributions for illustration purposes as the CVA biplots have the underlying assumption of normality and would therefore allow for better illustration of the technique. The LDV data set is then analysed in section 4 using the new proposed approach and compared to similar available techniques through a simulation study.

In section 5, a medical research data set is used to further illustrate the performance of the proposed technique. Thereafter, the web-based application that was built supplementary to this paper is discussed. The application allows the user to change the various properties of the proposed technique and the resulting output can then be inspected. Two additional data sets are provided for the user to interact with and the application also allows the users to upload their own training and testing data sets in order to see how the technique would perform. The link to the application can be found at `https://doi.org/10.5281/zenodo.3562013` (Van der Merwe, 2019).

The paper is concluded in section 6.

2 Current methods

Visualisation of multivariate categorical data comprises of how variables, observations, and classifications can be optimally represented in a single graph. Furthermore, classification techniques are not usually designed around limiting misclas-

sification errors. In this section, current techniques for simultaneously visualising and classifying multivariate data are illustrated and discussed with regard to these challenges. None of these methods were designed with limitation of misclassification errors specifically in mind, but a simple way to achieve this is to limit such errors by classifying observations only if the posterior probability is higher than an arbitrarily chosen threshold. This limitation will be applied to the techniques below and included in the subsequent discussion.

In order to illustrate the techniques, two data sets, each of size $n = 150$, were simulated from multivariate normal distributions with each class containing 50 observations. In one data set, all the various responses' covariates have the same underlying covariance matrix, but different means, and the variances for the covariates were also kept small. This data set is referred to as the SSV data set. The other data set was constructed by simulating from normal distributions with the same means and different covariance matrices for all the responses. Here the variances were increased such that discrimination became more difficult. This data set is called the LDV data set.

A more detailed discussion on these data sets along with a discussion on how to interpret their respective biplots can be found in appendix B.

2.1 Correlation multiplots and radar graphs

A traditional scatterplot allows for straightforward visual interpretation of all observations, but is limited to the interaction between two variables. A correlation multiplot, such as the ones in figure 2.1, can be drawn to see all the various interactions, but quickly becomes large as one needs $(p \times (p-1))/2$ to see all the interactions. While this does include all the variables, it remains difficult to consider the full extent of the interactions between variables.

Another simple approach is radar graphs. This graph allows for all the variables to be visualised, but only very few observations. Additionally, interactions between variables are limited to those which are positioned next to each other on the radar graph.

Neither of the above techniques allow for classification. Therefore, a possible approach would be to classify the observations using an external classification

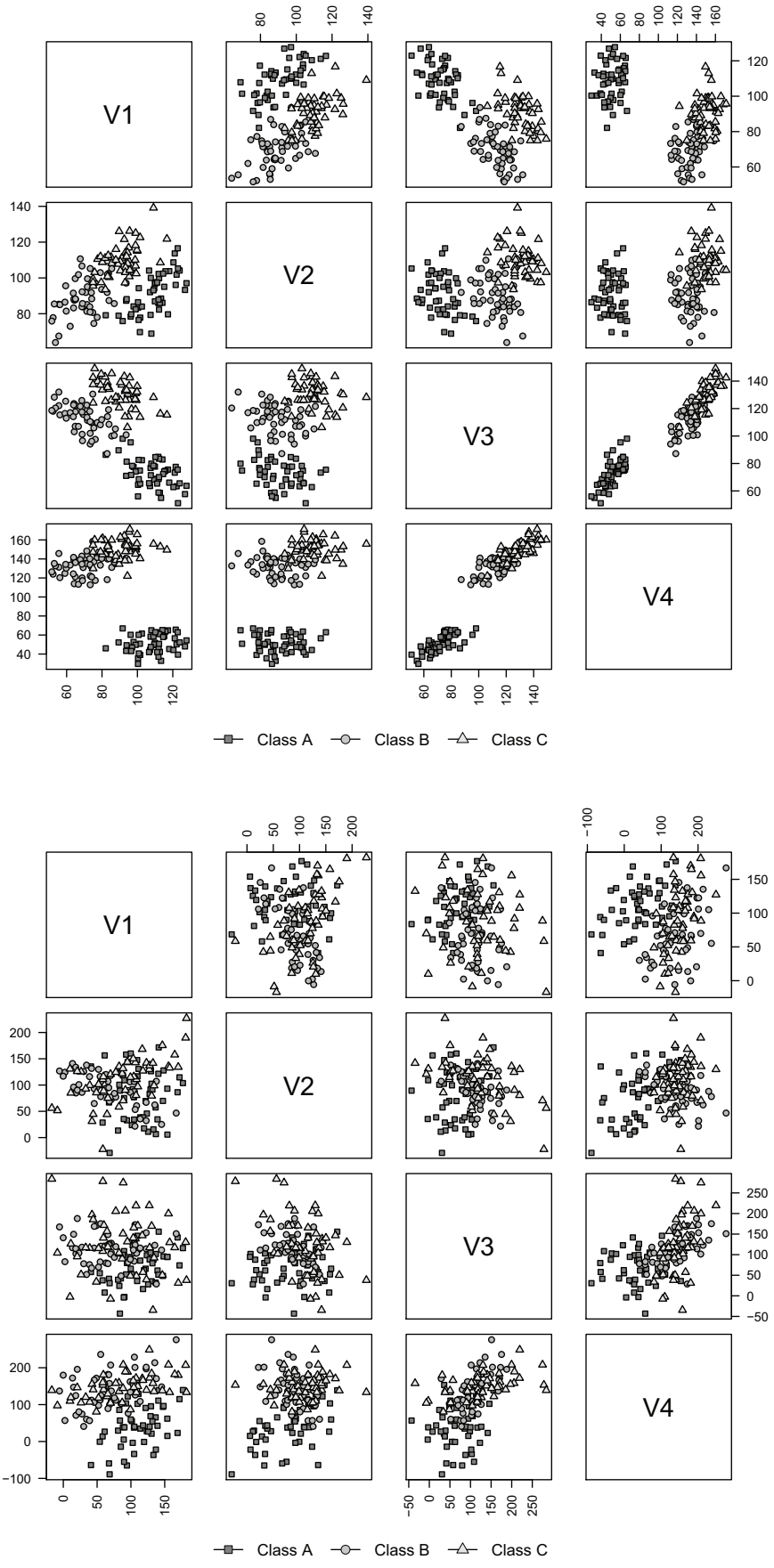


Figure 2.1: *Correlation multiplot of the SSV (top) and LDV (bottom) data sets.*

method, and to indicate the classified value on the graphs subsequently with a specific indicator. Out-of-sample observations would require the external method to be classified and plotted afterwards.

Traditional visualisation and classification techniques for multivariate data have some shortcomings. There are, however, methods available to simultaneously visualise and classify multidimensional data, which are discussed in the next two subsections.

2.2 Biplots with α-bags and classification regions

Initially introduced by Gabriel (1971), the concept of biplots is not new, but has only recently been popularised by e.g. Gower et al. (2011) and Greenacre (2010). The field of biplots has since been extended from simple principal component analysis (PCA) biplots to more complex methods, such as canonical variate analysis (CVA) and analysis of distance (AOD) biplots. Biplots have the benefit of presenting data in two or three dimensions, thus providing the necessary basis for visualising and classifying high dimensional data.

Through combining CVA biplots with classification rules within the biplot plotting space, Aldrich et al. (2004) was able to simultaneously view a full multivariate data set on copper froth containing eight variables and five classes, classifying observations into these classes. They noted however, that some of the classes were indistinguishable. They plotted the means of the classes in the CVA biplot space, and then classified the observations according to the closest class mean in the CVA biplot space.

They then used α-bags (Rousseeuw et al., 1999; Gower et al., 2011) to visualise the sample densities of the various classes to allow for better identification of overlap. The benefit of using α-bags compared to other techniques, such as convex hull peeling, is that the latter does not fully utilise the statistical properties of the data set.

Figure 2.2 shows two CVA biplots for each of the SSV and LDV data sets. The first contains the class means together with the classification region relative to the class means, and the second shows 95%-bags used to determine the extent of the class

overlap. It is clear that for the SSV data set, the classes are well separated, while there is significant overlap between two of the classes in the LDV data set.

There are, however, still some shortcomings to this approach, namely that only CVA biplots were considered; their classification only relied on the class means as opposed to the underlying data for each class; and they included areas where risk of misclassification was high due to overlap.

An alternative approach for visualising classification with biplots would be to use an external classification method trained on the original data, and to plot these observations according to their external classification. However, classifying these out-of-sample points would still require an external classification model, and if points overlap it would be difficult to use the triplot to classify these points.

2.3 Posterior probability log-ratio triplots

The CVA biplot provides a visual representation of the optimal class separation obtained in linear discriminant analysis. A similar concept of multiclass classification visualisation was introduced by Gardner-Lubbe (2016), which consisted of a log-ratio biplot of the posterior probabilities belonging to a class in order to calculate classification regions. This was extended to include information on the underlying variables, ultimately resulting in a triplot.

Posterior probabilities obtained from classification can be considered as compositional data, and log-ratio biplots of compositional data can be constructed easily for two-dimensional visualisation and interpretation (see Aitchison and Greenacre, 2002; Greenacre, 2018). Log-ratio biplots are interpreted differently from the biplots discussed in the previous section. On a log-ratio biplot the observations are considered relative to two variables of the compositional data (or classes, in the case of posterior probability data). Therefore, the outputs of the log-ratio biplot can be used to create classification regions. It is important to note that this biplot on its own would not provide any information of the underlying variables.

Gardner-Lubbe, therefore, uses the log-ratio biplot to create underlying classification regions by considering the axes of the log-ratio biplot, but then discards them, as their information is contained in the classification regions. The remaining data

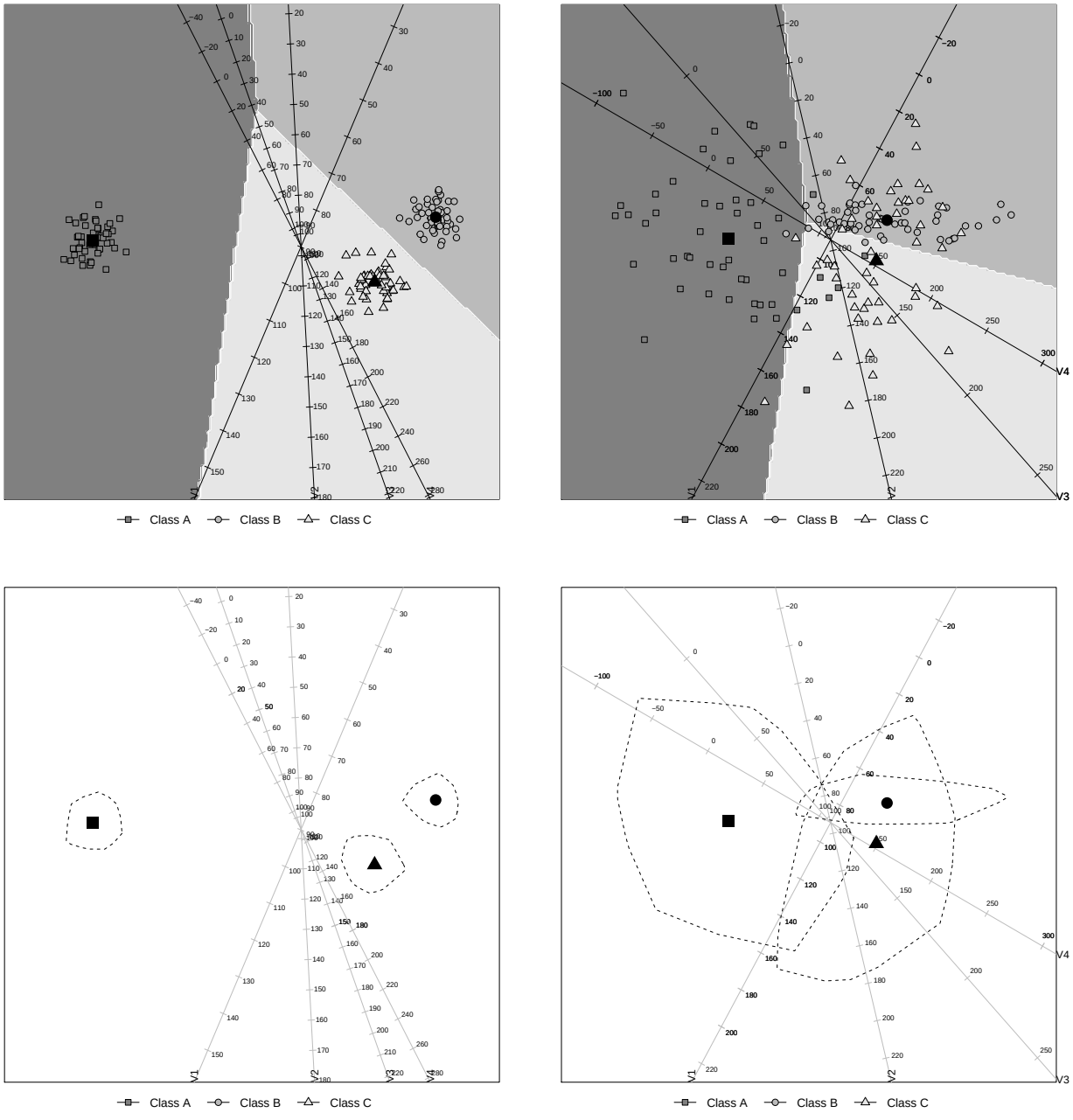


Figure 2.2: *CVA biplots of the SSV (left column) and LDV (right column) data sets. The top row contains the class means along with classification areas relative to the class means, while the bottom biplots contains the class means with 95%-bags.*

are the observations in the log-ratio biplot space along with the classification regions. Information of the underlying variables is then added through regressing the original matrix **X** on the coordinates of the observations in the biplot space, **Z**.

They state, importantly, that the triplot was not designed for optimal classification of samples, but rather to provide visual representation of all three aspects involved in the multiclass classification and their interrelationships. In figure 2.3, the *k*-nearest neighbour (KNN) triplot of Gardner-Lubbe is provided for the SSV and LDV data sets.

In figure 2.3, the SSV data set is visibly well separated. This is due to the posterior probabilities being mostly 100% for certain classes. The problem is that these observations, although different from each other, lie exactly on top of each other in the log-ratio biplot. Other observations that have a 0% posterior probability for a certain class and a split probability for the other two classes would fall between the two regions, which could result in misclassification. Similar observations can be made for the LDV data set, with additional points in the middle of the graph indicating observations for which none of the posterior probabilities are equal to 0% or 100%. It is therefore clear that the log-ratio triplot focuses on the classification values first, thereafter incorporating additional information on the variables.

Additionally, while this approach allows for the classification of out-of-sample data, the posterior probabilities obtained from the underlying classification method would still be required. This is due to the underlying biplot being constructed from these probabilities.

The triplot could also possibly provide an alternative classification to what the underlying classification method renders. This technique would therefore be most useful to investigate and visualise data classified under a specified method, rather than classifying out-of-sample observations.

3 Proposed approach and illustrative results

In this section, the proposed approach is discussed and illustrated on the SSV and LDV data sets. The proposed approach joins three of the above techniques - biplots,

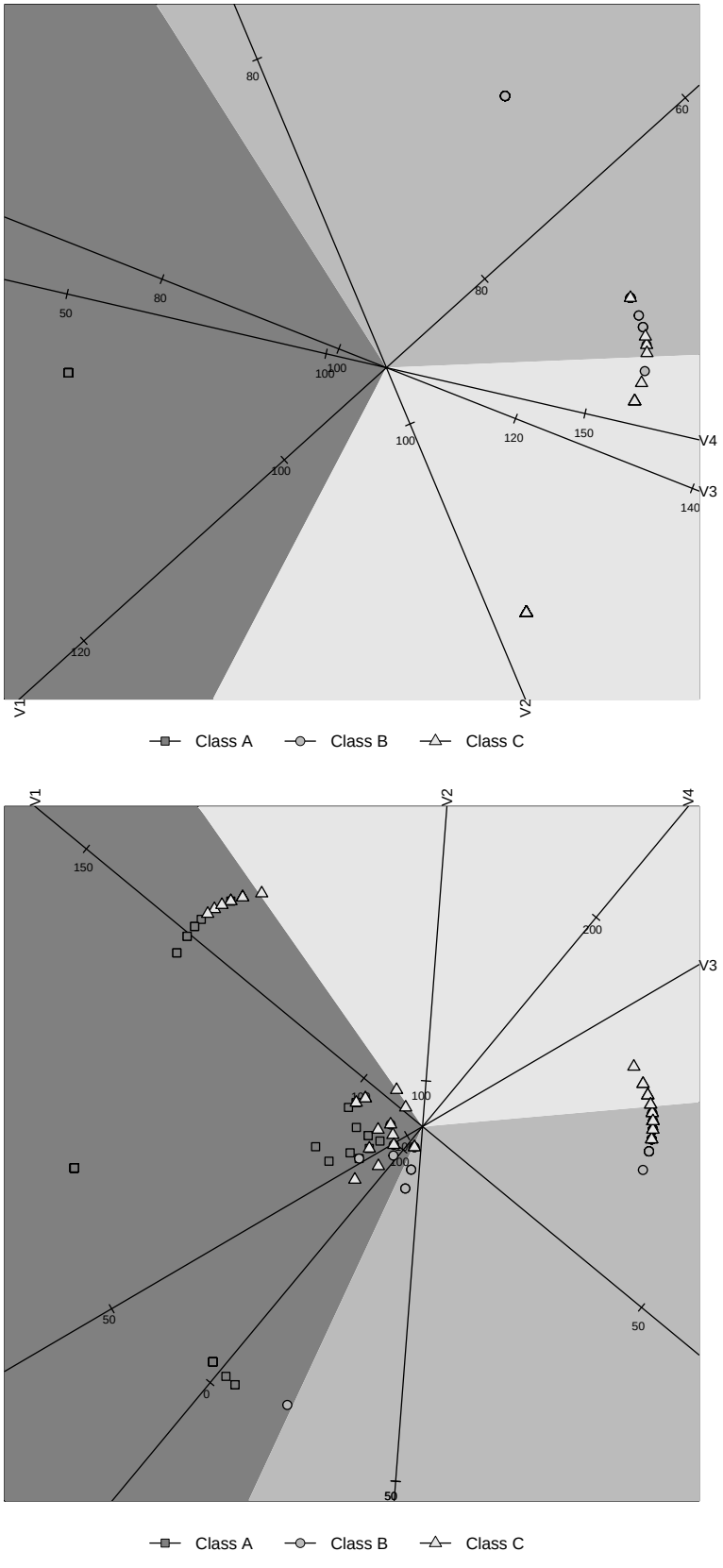


Figure 2.3: *The posterior probability log-ratio triplot with underlying KNN (k = 11) classification of the SSV (top) and LDV (bottom) data sets.*

underlying classification techniques and α-bags - to create a new classification methodology that allows for:

1. visual presentation of the data through a base biplot of the original data;
2. out-of-sample classification that does not depend on an external classification model; and
3. limitation of the classification error due to outliers and overlapping classes.

In the following subsections, a biplot with classification regions is discussed after which the concept of polybags is introduced. The polybags are illustrated by first drawing the outer-polybags, which cater for outliers; and then the inner-polybags, which cater for overlapping data. This is followed by an interpretation of the final triplot with polybags.

3.1 Base biplot with classification areas

In order to illustrate the proposed technique, a simple biplot of the data is drawn to which the various components are added. The proposed approach does not rely on any specific underlying biplot. For illustrative purposes, though, the CVA biplot is shown, and the PCA and AOD biplots are tested for robustness in section 4.1. PCA does not differentiate between classes and AOD is constructed using only the means of the classes - shortcomings mitigated by the CVA biplot. Therefore, the CVA biplot was chosen, as it is expected to render the best results through differentiating between the classes and using more than only the class means.

In order to add the classification region, a similar approach to Aldrich et al. (2004) is followed. Recall that they classified any point in the biplot space according to the closest class mean. This was illustrated on the biplot by drawing a grid in the biplot space and filling the areas of the biplot space with the colour that matches the classification.

An alternative approach is to take all the training data points in the two-dimensional biplot space and train an underlying classification model on these coordinates. The grid points in the biplot space are then classified using this model, and shaded contour areas are created accordingly.

The training models were implemented through the *caret* (Kuhn et al., 2008) package in R. The illustrated classification method for this paper was chosen as KNN. Other available methods that were also tested for robustness in section 4.2 include linear discriminant analysis (LDA), quadratic discriminant analysis (QDA), naive Bayes with Gaussian model (NB with GM), naive Bayes with kernel density estimation (NB with KDE), multinomial, support vector machine with polynomial kernel (SVM with PK), support vector machine with Gaussian kernel (SVM with GK), classification and regression trees (CART), bagging, and random forest (RF).

Figure 2.4 shows a base CVA biplot with underlying KNN ($k = 11$) classification together with 95%-bags. Note here that the resulting triplot looks similar to a combination of the ones illustrated in figure 2.2, however with the LDV data set having visually dissimilar classification regions.

The figure shows that there are areas in the classification region of the SSV data set's triplot that clearly lie outside of the concentration of the data. There is also significant overlap between some of the classes for the LDV data set. These two shortcomings are addressed with the help of inner- and outer-polybags.

3.2 Polybags

The α-bags are used to find areas which can be excluded from the classification, as they lie beyond the classifiable range. The α-bags furthermore help to identify areas where sample points overlap significantly and classification would therefore be imprecise. The expectation is that these areas will not render conclusive classification, and are thus left unclassified pending further investigation. Two types of polybags are used, the first being the compliment of the area containing at least one α-bag, and another where all α-bags overlap, i.e. the union and intersections of the α-bags. These two areas are termed the outer- and inner-polybags.

The sizes of these α-bags are considered tuning parameters to the proposed classification methodology. It should therefore be noted that there is no 'correct' choice of the tuning values. Instead, the user should ensure that the chosen parameter values provide them with a level of misclassification on the test data set that they are comfortable with. One could also investigate the possibility of determining the

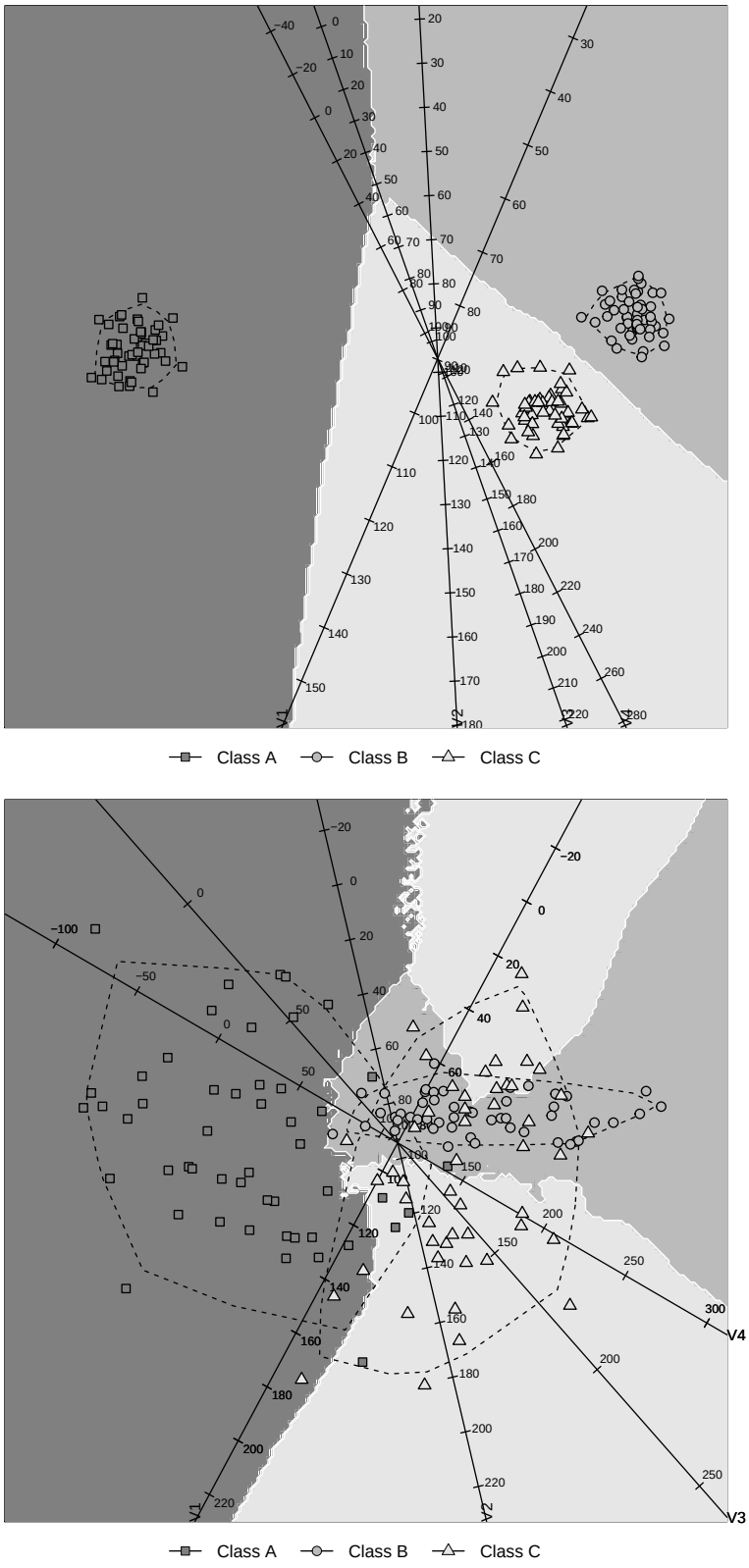


Figure 2.4: *CVA biplots of the SSV (top) and LDV (bottom) data sets with 95%-bags and classification regions drawn based on KNN with k = 11.*

theoretical unclassified rate and consider the relationship between the α-bag sizes and this rate.

Once these polybags are determined, such areas are left unclassified. That is, the intersection of the inner α-bags (inner-polybag) and the complement of the union of the outer α-bags (outer-polybag). The coordinates for the polybags are found through the R package *polyclip* (Johnson and Baddeley, 2017), while determination of whether certain sample points lie within the polybag is facilitated by the package *SDMTools* (VanDerWal et al., 2014).

Outer-polybags

Considering the SSV data set's triplot in figure 2.4, if a 'normal' observation occurs, it is most likely going to lie within the 95%-bag area. If an outlier occurs, it will typically fall outside of the bag, but still relatively close to the bag. This poses a problem as the whole triplot area was classified into one of the three classes. The outer-polybags are introduced to mitigate this. It takes α-bags, inflates them by a predetermined factor, and deems everything outside of the union of these bags unclassified. These outer-polybags, using $1.5 \times$ 95%-bags, are applied and presented in figure 2.5.

Inner-polybags

The second problem only occurs in the LDV data set's triplot: Class B's ($\circ$) observations are almost entirely enveloped by Class C ($\triangle$). This would complicate classification, so ideally these types of areas should remain unclassified. The inner-polybag is therefore constructed by taking the intersection of at least two of the α-bags. This is illustrated as the middle white area in figure 2.6 with 95%-bags. As the SSV data set does not contain overlapping α-bags, only the LDV data set's triplot was redrawn indicating the inner-polybag.

3.3 Interpreting the triplot with polybags

Consider now the interpretations of the two data sets' triplots. For the SSV data set, the same classification as discussed in appendix B.1 holds. Interpretation of

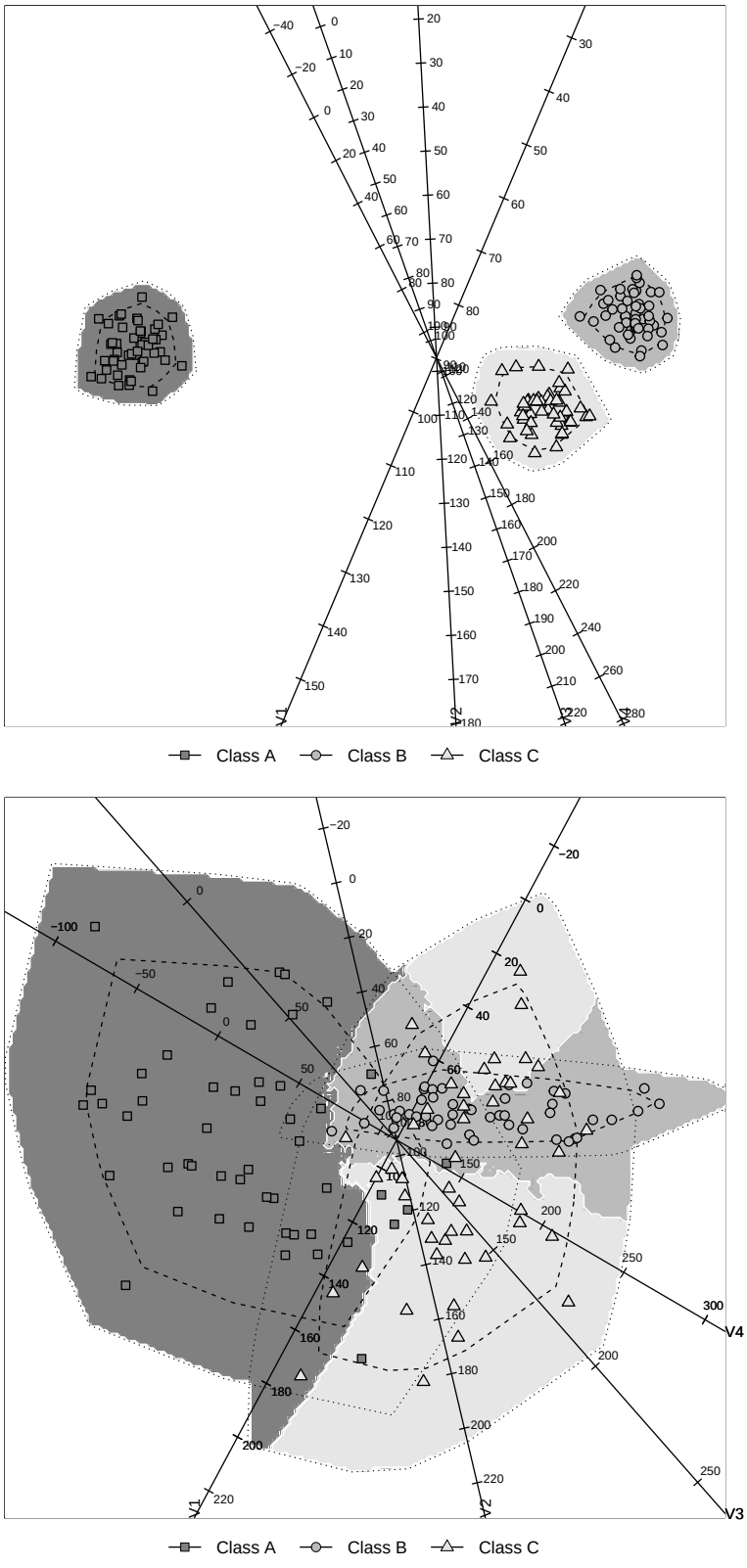


Figure 2.5: *KNN-triplots of the SSV (top) and LDV (bottom) data sets with 95%-bags and each point classified using KNN with k = 11 along with 1.5 × 95% outer-polybags.*

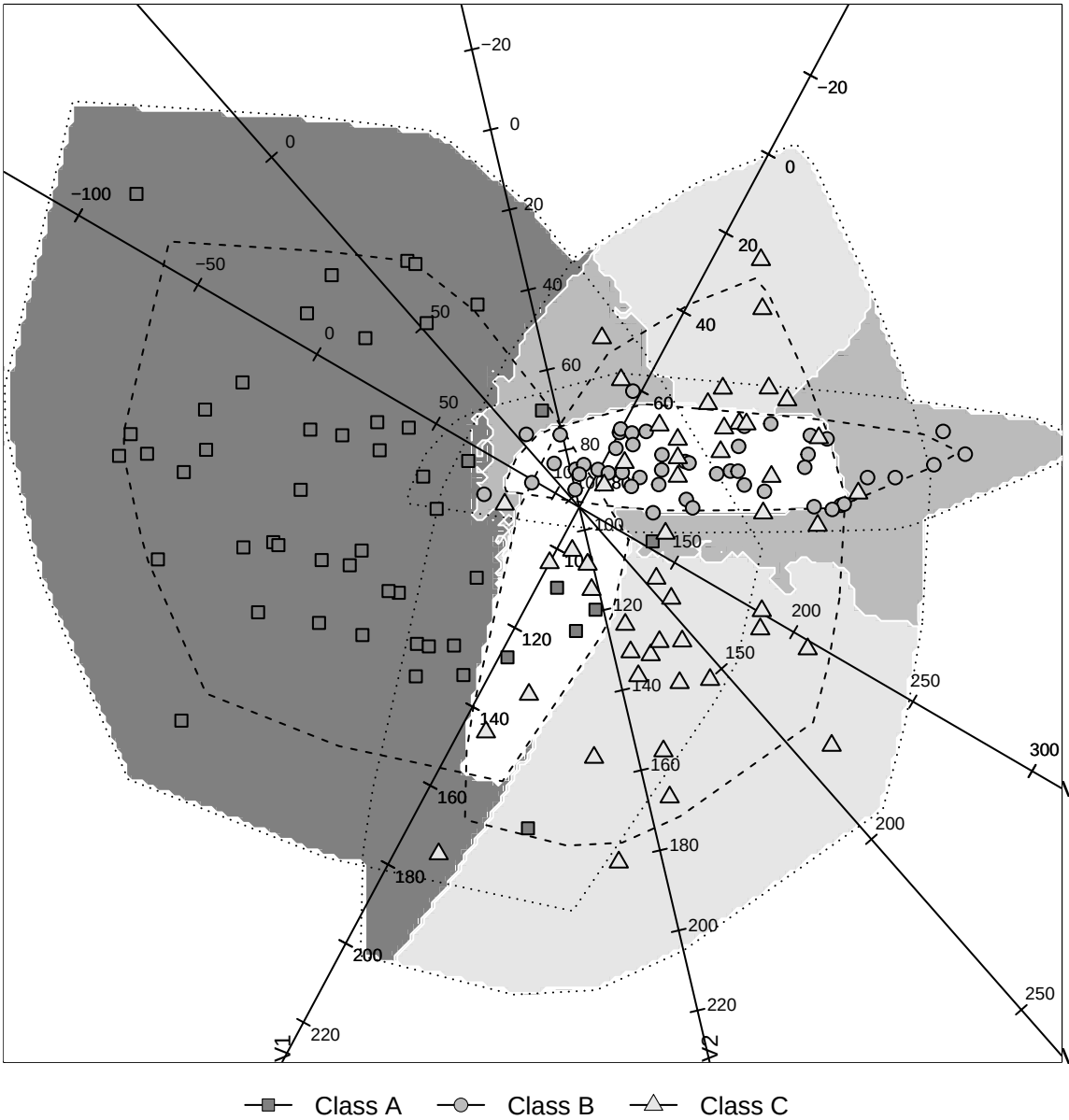


Figure 2.6: *KNN-Triplot of the LDV data set with each point classified using KNN with k = 11 together with 1.5 × 95% outer- and 95% inner-polybags.*

the LDV data set remains arduous, but can be summarised as follows: for values of $-100 < V_4 < 100$ the majority of the points are classified as Class A (□), for values of $130 < V_4 < 250$ and $100 < V_1 < 200$, or $50 < V_4 < 150$ and $0 < V_1 < 50$, Class C (△) is observed, and finally, for values of $150 < V_4 < 250$ and $20 < V_1 < 100$, Class B (∘) is observed. The exact classification can, and should, be obtained by calculating the coordinates of the observation in the biplot plotting space ($\mathbf{Z}^*$). To be specific:

$$
\begin{aligned}
I_{\text{A}} &= \{-100 < V_4 < 100\} \\
I_{\text{B}} &= \{130 < V_4 < 250 \cap 100 < V_1 < 200\} \cup \{50 < V_4 < 150 \cap 0 < V_1 < 50\} \\
I_{\text{C}} &= \{150 < V_4 < 250 \cap 20 < V_1 < 100\}
\end{aligned}
$$

where I_{class} indicates the respective classes that the observation should belong to.

Here the polybags become important: should the new observation fall within one of these white areas, it is as yet unclassified. This provides the benefit of avoiding classification in areas of overlap with no clearly visible differentiation between groups, along with extreme outlier data. It is also possible to train a set of models over various parameters to select an optimal one. In this illustrative application, KNN with $k = 11$ was chosen and therefore any training of the models would add no benefit, as all parameters were chosen at the outset. The choice of k was made as a trade-off between a small and large value. A small value of k would not delineate the classification regions nicely as it would over fit the training data, while a large value of k would result in higher computation time as well as very rigid classification regions. Similar to the sizes of the polybags, the choice of k is also considered a tuning parameter, and the user should be comfortable with the performance of the test data set for the choice of the tuning value.

One could, for instance, train the KNN over a set of values of k to find the value of k for which the out-of-sample test data performs the best. The accuracy of the final classification method is determined through a validation set which is not used in the training of the model. The classification methodology is applied using other classification methods in the simulation study in the following section to check robustness.

4 Robustness checks

Different properties of the new methodology can easily be varied. This includes the underlying biplot, the underlying classification technique, and the size of the inner- and outer-polybags. Additionally, this method could possibly favour certain types of data structures over others. In the previous section, these properties were fixed. Only CVA biplots were used as the underlying base triplots, KNN was used as the underlying classification method, and the size of the inner- and outer-polybags were kept at 95% and 1.5 × 95%, respectively. Additionally, only the SSV and LDV data sets were considered, which in essence only differ in terms of variance size.

In this section, the results of two simulation studies are discussed. The first considers how the misclassification rates vary for different underlying biplots and sizes of the inner- and outer-polybags, applied to the LDV data set. The second investigates how the misclassification rate varies with regards to different underlying classification methodologies and different types of data structures, relative to the posterior probability log-ratio triplot and black-box techniques.

Table 3.1 provides an updated confusion matrix that includes observations that are not classified or predicted using the new method.

Table 2.1: *Confusion matrix, updated to include inconclusive observations.*

TOTAL POPULATION (N)	**True Positive Condition (TPC)**	**True Negative Condition (TNC)**
Predicted Positive Condition (PPC)	True Positive (TP)	False Positive (FP)
Predicted Negative Condition (PNC)	False Negative (FN)	True Negative (TN)
Not Predicted (NA)	NA Positive (NAP)	NA Negative (NAN)

4.1 Effects of using different base biplots and sizes of inner- and outer-polybags

Three types of triplots were discussed in the text, while only the CVA triplot was illustrated. Through sampling 100 validation sets of 25% from the LDV data set, the average effects of varying the size of the inner-polybag as well as the multiple for the outer- polybag over PCA, CVA, and AOD base biplots are considered. The validation samples were kept the same in each instance for consistency.

Figure 2.7 illustrates the results of how the validation data set performed on average over the three different types of triplots with KNN ($k = 11$) as underlying method, varying the size of the inner-polybag from 0% to 95%. The outer-polybag's size varies over 0.5, 1.5, and 3 × 95%.

The lowest bar (indicated with 'Full') illustrates the classification error if all the original data was classified using the traditional KNN ($k = 11$), i.e. circa 27% misclassification. Note that no visualisation (unless only two variables are used) is possible with the traditional classification. Thus it is expected to perform worse, because the reduction in dimensionality results in a loss of information. It is interesting to note, however, that the CVA triplot with outer-polybags with a multiple larger than 1.5 performed better than the full classification.

Additionally, the intersection of the two lines on the graphs can be interpreted as the inner-polybag size at which the triplot method produced a lower misclassification rate than the full model. Accuracy considered, the lower this crossing lies, the more powerful the method becomes.

Of the graphs in figure 2.7, CVA performs the best of the three triplot methods, as it renders the lowest misclassification rate without losing as much accuracy as the other methods. It also provides, given an inner-polybag of 0%, the lowest misclassification rate of all three methods.

The effects of changing the multiple for the outer-polybag can also be seen. A multiple of 0.5 × 95% resulted in a large number of unclassified data points. That is, all the points not classified by an inner-polybag of 0% result from the outer-polybag. As the multiple increases, the effects of the outer-polybag decreases. The incorporation of the outer-polybag is mainly to capture outliers. Too small an outer-

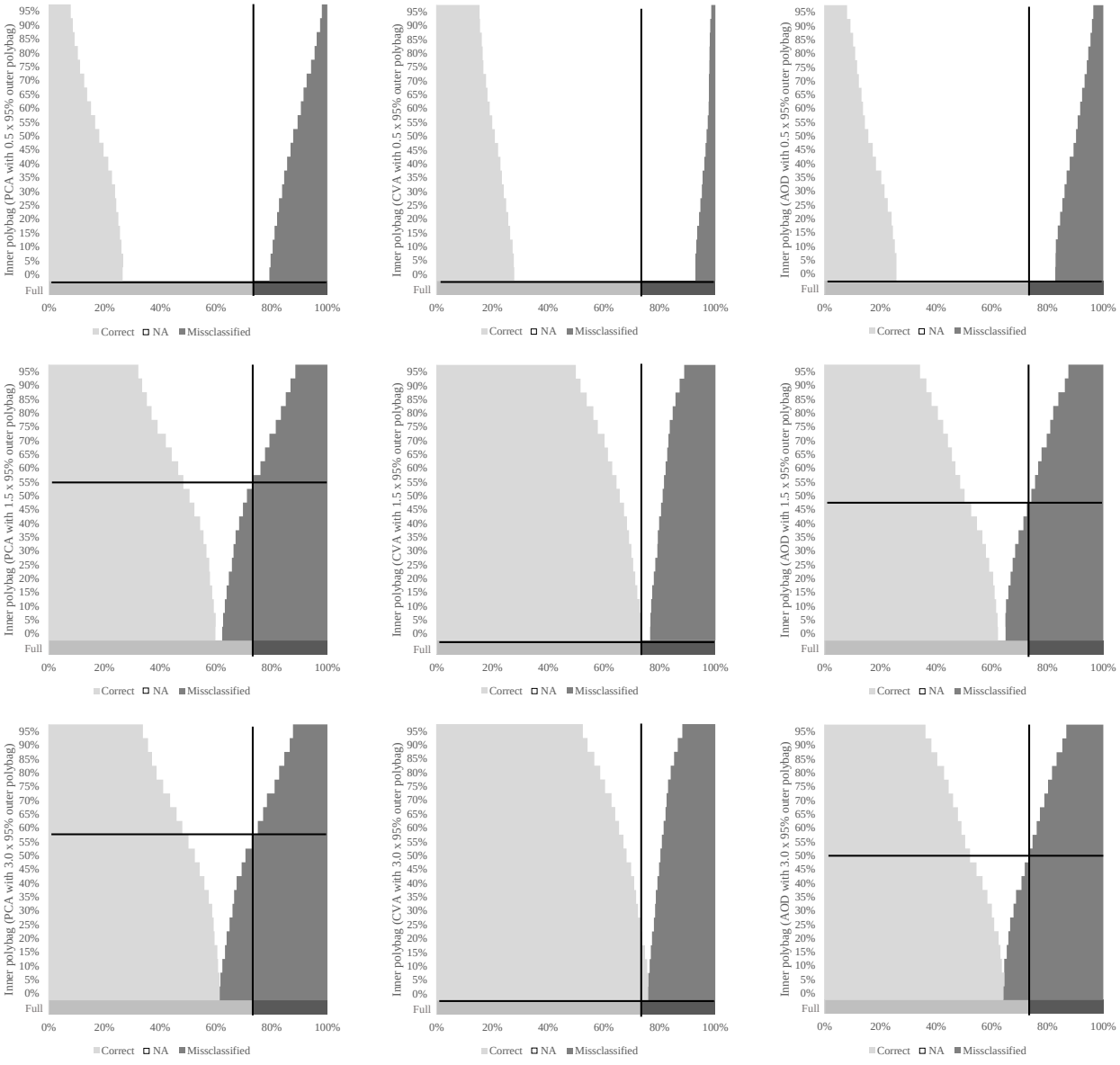


Figure 2.7: *Classification error with varying sizes of the inner-polybag for PCA (left column), CVA (middle column), and AOD (right column) triplots with various outer-polybags equal to 0.5 (top row), 1.5 (middle row), and 3.0 × 95% (bottom row).*

polybag, therefore, is inadequate, as half the observations would be considered outliers; too large and no outliers would be detected.

In the above simulation study, only the LDV data set with underlying KNN classification was considered. In the next subsection the results of a simulation study will be provided. The results show how the overall misclassification rates vary when different underlying classification techniques are used. This will be applied to various types of data structures and the results will be compared with the posterior probability log-ratio triplots and black-box techniques.

4.2 Effects of using different underlying classification techniques and data structures

A second simulation study was done with regards to the underlying classification techniques in order to see the performance of different data structures using triplots with polybags compared to the posterior probability log-ratio triplot and black-box techniques. Throughout this study, a CVA-base biplot was chosen with the inner-polybag kept at 95% and outer-polybag at 1.5 × 95%.

Comparison of various techniques

The posterior probability log-ratio triplot (L) and black-box (B) techniques were adapted to exclude the same number of unclassified observations as with the triplot with polybags (P). This was to ensure comparability between the techniques. The observations with the smallest posterior probability for the classified class were discarded until the same number of observations was excluded as with the triplot with polybags.

The underlying classification methods included all those mentioned in section 3.1. The classification techniques are denoted with M_{m_t,m_c} with $m_t = P, L, B$ indicating the tested techniques, and m_c = KNN, QDA, NB with GM, NB with KDE, multinomial, SVM with PK, SVM with GK, CART, bagging, and RF the respective underlying classification methods.

Combining the various techniques with different underlying classification methods results in 30 different classification methods to be tested.

The different data structures that were used are discussed in the next subsection. Thereafter, a method for comparison is discussed and the results are provided and interpreted.

Simulating different data structures

In order to determine how well triplots with polybags perform over different data structures, a combination of certain attributes was determined and a number of data sets were created from these parameters. It follows a similar approach as the compilation of the SSV and LDV data sets considered above. In each case, 150 observations were sampled from multivariate normal distributions for the three classes. The assumption of multivariate normal distribution might not always hold for real life data and a future extension of the simulation study would be to incorporate other distributions. The benefit, however, of only simulating from one well-known distribution is to see the effect on how different data structures perform in the new classification technique. For the application in the following section it will be seen that the classification method still performs relatively well even though the data is not necessarily normally distributed.

Four main areas were identified to describe the data structures. Once all the properties of the structures were identified, 10 simulated data sets were created with both the parameters and observations randomly generated. The distribution parameters were arbitrarily chosen in such a manner that the simulated data represents the properties of the data structures being simulated.

The first structural property incorporated regards the means of the various classes. Two types of mean structures were identified, including separate (SEPR) and overlapping (OVLP) means. The means for the separate structure were uniformly simulated from three ranges ($[-7.5, -2.5]$, $[-2.5, 2.5]$, and $[2.5, 7.5]$) while the overlapping structure only included class means from a single range, $[-5, 5]$.

The second structural property is with regards to the variance of the classes. In order to ignore the scale of the data, the coefficient of variation (CV) was used. The first type of simulated CVs was small with low variation (SL), while the second type was large and highly variable (VH). The small and low variable CVs were uniformly simulated from $[0.5, 1]$ and the large and highly variable CVs from $[2, 5]$. The

variances were then calculated using the class means that were simulated in the previous structure.

Next the balance of the classes is considered. This structural property considers the proportion of observations in each class. Even and skewed class balance structures were randomly allocated to each class, with the skewed structure allocating 60% to the largest class, 40% to the middle class, and 20% to the smallest class.

The last simulated structural property of the data sets relates to the correlation structure. Three types of correlation structures were considered. These included a low correlation structure (Low) that only contained correlations between -0.5 and 0.5, highly negative correlations (Hneg) with values between -0.5 and -0.9, and highly positive correlations (Hpos), with values between 0.5 and 0.9. Once the correlation structures for each class were simulated, the closest positive definite covariance matrix was calculated using the variances simulated in the previous step.

There are 24 different combinations of data structures, of which 10 data sets were simulated for each. Let these 240 data sets be denoted with D_{d_s,d_n}, with $d_s = 1,\ldots,24$ the various data structure combinations, and $d_n = 1,\ldots,10$ the various simulations from that specific structure.

Simulating misclassification errors

Ten randomly sampled training and test sets from each of the 240 data sets (D_{d_s,d_n}) were used to train and test the 30 classification methods (M_{m_t,m_c}). The confusion matrix from table 3.1 was calculated and certain error metrics were calculated and averaged over the 10 simulations. These are denoted by $x_{d_s,d_n,m_t,m_c,i}$ with x the error metric, $i = 1,\ldots,10$ indicating the sampled data set, and $\bar{x}_{d_s,d_n,m_t,m_c}$ the average of the simulated error metrics. While various misclassification type errors were calculated, only the overall misclassification was considered here. This was calculated as the $\frac{FP+FN}{N}$ as per the notation in table 3.1.

These misclassifications are summarised in table 2.2, starting with the calculation of the mean and variance of the average misclassification rates of the 10 data sets with similar structures. That is, calculate

$$\mu_{d_s,m_t,m_c} = \frac{1}{10}\sum_{d_n=1}^{10} \bar{x}_{d_s,d_n,m_t,m_c}$$

and

$$\sigma^2_{d_s,m_t,m_c} = \frac{1}{10}\sum_{d_n=1}^{10} (\bar{x}_{d_s,d_n,m_t,m_c} - \mu_{d_s,m_t,m_c})^2$$

These two values are then used to calculate the mean squared error (MSE) of the misclassification rate with

$$MSE_{d_s,m_t,m_c} = (\mu_{d_s,m_t,m_c} - \alpha)^2 + \sigma^2_{d_s,m_t,m_c}$$

with $\alpha = 0\%$ the target for the misclassification rate (for accuracy, α would be set at 100%).

Next the MSEs for the proposed technique ($m_t = P$) are compared relative to the minimum of the other two techniques ($m_t = L, B$) and the difference is calculated such that a table can be constructed with the data structures (d_s) and underlying classification methods (m_c). That is, a table that contains the relative MSEs (RelMSE).

$$RelMSE_{d_s,m_c} = MSE_{d_s,P,m_c} - \min\left\{MSE_{d_s,L,m_c}, MSE_{d_s,B,m_c}\right\}.$$

The higher the values for $RelMSE_{d_s,m_c}$, the worse the proposed method performed relative to the better of the other two methods, and vice versa. The benefit of using MSE is that it penalises classification methods that render inconsistent classifications. The results are provided in table 2.2.

The simulation results show that the proposed method with KNN, CART, bagging, or RF as underlying classification methodology performs better than, or relatively similarly to, the other two techniques over all tested data structures apart from low and positively correlated highly variable variables. For multinomial and LDA as underlying, the proposed method performed equally well as the better of the other two methods. It proved difficult for this technique to outperform the benchmarks for

Table 2.2: *The mean squared error (MSE) of the proposed method compared to the minimum MSE of the posterior probability log-ratio triplot and black-box techniques over various data structures and underlying classification methods.*

Means	CV	Class Balance	Correlation	KNN	LDA	QDA	NB with GM	NB with KDE	Multinomial	SVM with PK	SVM with GK	CART	Bagging	RF
SEPR	SL	Even	Low	0.03	0.03	0.03	0.04	0.04	0.01	0.03	0.03	−0.04	−0.01	0.03
			Hpos	0.03	0.04	0.05*	0.03	0.03	0.02	0.04	0.04	0.01	0.03	0.04
			Hneg	0.00	0.00	0.00	0.00	−0.01	0.00	0.00	0.00	−0.07*	−0.03	0.00
		Skewed	Low	0.00	0.01	0.02	0.02	0.03	0.01	0.02	0.01	−0.02	0.00	0.03
			Hpos	0.03	0.03	0.05*	0.03	0.03	0.02	0.04	0.04	0.01	0.03	0.04
			Hneg	0.00	0.00	0.00	0.00	−0.01	0.00	0.00	−0.01	−0.07*	−0.04	0.00
	VH	Even	Low	0.07*	0.02	0.10†	0.10†	0.10†	0.02	0.11†	0.10†	0.04	0.10†	0.12†
			Hpos	0.07*	0.02	0.10†	0.09*	0.09*	0.02	0.09*	0.09*	0.03	0.08*	0.09*
			Hneg	−0.19‡	0.01	0.01	−0.14†	−0.20‡	0.01	0.00	−0.03	−0.25‡	−0.19‡	−0.14†
		Skewed	Low	0.06*	0.02	0.08*	0.08*	0.08*	0.02	0.07*	0.07*	0.01	0.07*	0.08*
			Hpos	0.06*	0.02	0.10*	0.06*	0.07*	0.01	0.07*	0.07*	0.02	0.06*	0.06*
			Hneg	−0.13†	0.00	0.01	−0.16‡	−0.19‡	0.01	0.01	−0.01	−0.23‡	−0.16‡	−0.12†
OVLP	SL	Even	Low	0.00	0.01	0.05	0.05*	0.05*	0.02	0.03	0.02	0.02	0.04	0.05
			Hpos	0.00	0.01	0.03	0.00	−0.01	0.01	0.02	0.02	−0.03	0.00	0.02
			Hneg	−0.03	0.00	0.03	−0.03	−0.03	0.01	0.01	0.01	−0.07*	−0.04	−0.02
		Skewed	Low	0.00	0.01	0.04	0.04	0.05	0.01	0.03	0.03	0.02	0.03	0.04
			Hpos	−0.01	0.01	0.03	−0.01	−0.01	0.01	0.02	0.02	−0.01	0.00	0.02
			Hneg	−0.03	0.01	0.03	0.00	−0.01	0.01	0.02	0.01	−0.05*	−0.03	−0.01
	VH	Even	Low	0.08*	0.01	0.12†	0.14†	0.14†	0.01	0.09*	0.09*	0.09*	0.11†	0.12†
			Hpos	0.06*	0.02	0.14†	0.12†	0.11†	0.02	0.11†	0.11†	0.06*	0.11†	0.11†
			Hneg	−0.11†	0.02	0.12†	−0.02	−0.03	0.05	−0.01	−0.05	−0.10*	−0.07*	−0.07*
		Skewed	Low	0.06*	0.03	0.09*	0.12†	0.12†	0.03	0.10†	0.08*	0.05*	0.10*	0.10*
			Hpos	0.06*	0.02	0.12†	0.08*	0.07*	0.01	0.10†	0.09*	0.04	0.09*	0.10†
			Hneg	−0.09*	0.03	0.10†	−0.02	−0.03	0.05	0.04	−0.04	−0.09*	−0.06*	−0.06*

entry = MSE_P - min(MSE_L,MSE_B)
‡ |entry| ≥ 0.15; † |entry| ≥ 0.10; * |entry| ≥ 0.05

data sets with large and highly variable CVs and for variables that were positively correlated. On the other hand, negatively correlated data sets seem to consistently outperform the other data sets, regardless of other structural properties or underlying classification methodology.

The following section presents a realistic application of the methodology.

5 Application

In this section, the proposed methodology is applied to a well-known machine learning data set relating to vertebral column data obtained from UCI's machine learning repository (Dheeru and Karra Taniskidou, 2017). The web-based application that was created is introduced thereafter.

5.1 Application to medical data set

Application of the classification technique proposed in this paper could greatly benefit the medical sciences, as further testing is much preferred to an erroneous classification, or diagnosis. Visual presentation of the data is an additional advantage.

The selected data set contains values for six biomechanical features used to classify orthopaedic patients into three classes: normal (△), disk hernia (□) or spondylolisthesis (∘), including 310 instances and six attributes - all real numbers. Each patient (or observation) is represented in the data set by six biomechanical attributes derived from the shape and orientation of the pelvis and lumbar spine (in this order): pelvic incidence (V1), pelvic tilt (V2), lumbar lordosis angle (V3), sacralslope (V4), pelvic radius (V5) and grade of spondylolisthesis (V6).

Table 2.3 provides descriptive statistics of the data set. It also includes the p-value for the Shapiro-Wilk (Shapiro and Wilk, 1965) test for normality. The Shapiro-Wilk test tests the null hypothesis that the variable is normally distribution. If the hypothesis is rejected with a low p-value, then the variable can be considered to be non-normally distributed. Additionally, the Mardia (Mardia, 1970) test for multivariate normality showed that at the 5% level neither the Disk Hernia nor Spondylolisthesis

classes are multivariate normally distributed. While this is a slight deviation from the assumption of normality that underlies the CVA biplot, it will be seen in the results that this does not significantly reduce the accurateness of the technique. The *MVN* (Korkmaz et al., 2014) package in R was used to perform the summary and tests.

Figure 2.8 provides an example of the triplot with polybags and underlying KNN classification applied to this data set. The axes were shifted parallel for enhanced visual interpretation.

In this illustration, k was chosen as 21, and the inner- and outer-polybags as 85% and 1.5 × 95%, respectively. These were chosen for illustrative purposes and was found to provide a good fit for the training data. The triplot with polybags using an underlying KNN classification was trained on 232 of the 310 data points (75%). Of the resulting out-of-sample data points, 54 were classified correctly, three incorrectly, and the remaining 21 were inconclusive. Twenty of these data points can be seen in the white area in figure 2.8, while the last point was deemed an outlier, falling outside the plotting area.

This implies that the misclassification rate was only 4%, and that further investigation should be done with regards to the 21 data points.

Additional to the classification, the visualisation proves useful in describing the multivariate data set. The grade of spondylolisthesis (V6) is shown to be a clear indicator of patients suffering spondylolisthesis (∘) (values above 25 are mostly classified as such). It also becomes clear that high values for pelvic incidence (V1), lumbar lordosis angle (V3), and sacral slope (V4) indicate spondylolisthesis (∘).

For low values of the above-mentioned variables, the pelvic tilt (V2) or pelvic radius (V5) determine whether the patient is either normal (△) or suffers from disk hernia (□) - for values of pelvic tilt (V2) below 15 and higher than 120 for pelvic radius (V5), the patient would be classified as normal (△), otherwise with disk hernia (□).

Next, the posterior probability log-ratio triplot of Gardner-Lubbe (2016) was applied and the result is illustrated in figure 2.9. The same number of points was left unclassified by considering their posterior probabilities.

Table 2.3: *Descriptive statistics of the six biomechanical features (V1-V6) split into each of the three diagnostic classes for the medical data set used in the application. The last column shows the p-value for the Shapiro-Wilk test for normality. For low p-values, the variables can be considered to be non-normally distributed.*

Variable	Class	Mean	Std. Dev.	Median	Min	Max	25th perc.	75th perc.	Skewness	Kurtosis	SW p-value
Pelvic incidence	Disk Hernia	47.64	10.70	46.42	26.15	74.43	41.02	53.92	0.37	−0.31	0.44
	Spondylolisthesis	71.51	15.11	72.15	37.90	129.83	60.66	81.08	0.36	1.11	0.00*
	Normal	51.69	12.37	50.13	30.74	89.83	42.82	61.47	0.73	0.34	0.00*
Pelvic tilt	Disk Hernia	17.40	7.02	16.95	3.14	41.56	12.88	22.00	0.64	0.95	0.24
	Spondylolisthesis	20.75	11.51	19.31	−6.55	49.43	13.46	29.23	0.36	−0.34	0.03*
	Normal	12.82	6.78	13.49	−5.85	29.89	8.80	16.79	−0.19	0.06	0.79
Lumbar lordosis angle	Disk Hernia	35.46	9.77	35.17	14.00	62.28	29.04	42.01	0.18	0.07	0.68
	Spondylolisthesis	64.11	16.40	62.56	24.71	125.74	52.00	76.99	0.42	0.31	0.07
	Normal	43.54	12.36	42.64	19.07	90.56	35.00	51.60	0.73	1.02	0.02*
Sacralslope	Disk Hernia	30.24	7.56	30.00	13.37	46.61	25.21	34.97	−0.13	−0.37	0.80
	Spondylolisthesis	50.77	12.32	50.75	19.29	121.43	43.17	56.68	1.16	6.26	0.00*
	Normal	38.86	9.62	37.06	17.39	67.20	32.34	44.61	0.42	0.00	0.26
Pelvic radius	Disk Hernia	116.48	9.36	116.70	84.24	137.54	112.28	122.04	−0.59	1.29	0.14
	Spondylolisthesis	114.52	15.58	114.85	70.08	163.07	104.71	123.34	0.23	0.51	0.32
	Normal	123.89	9.01	123.88	100.50	147.89	118.18	129.04	0.01	−0.01	0.94
Grade of spondylolisthesis	Disk Hernia	2.48	5.53	2.54	−10.68	15.78	−0.91	6.06	−0.21	0.01	0.64
	Spondylolisthesis	51.90	40.11	42.38	1.01	418.54	30.40	61.90	5.36	44.39	0.00*
	Normal	2.19	6.31	1.16	−11.06	31.17	−1.51	4.97	1.64	5.10	0.00*

* p-value < 0.05

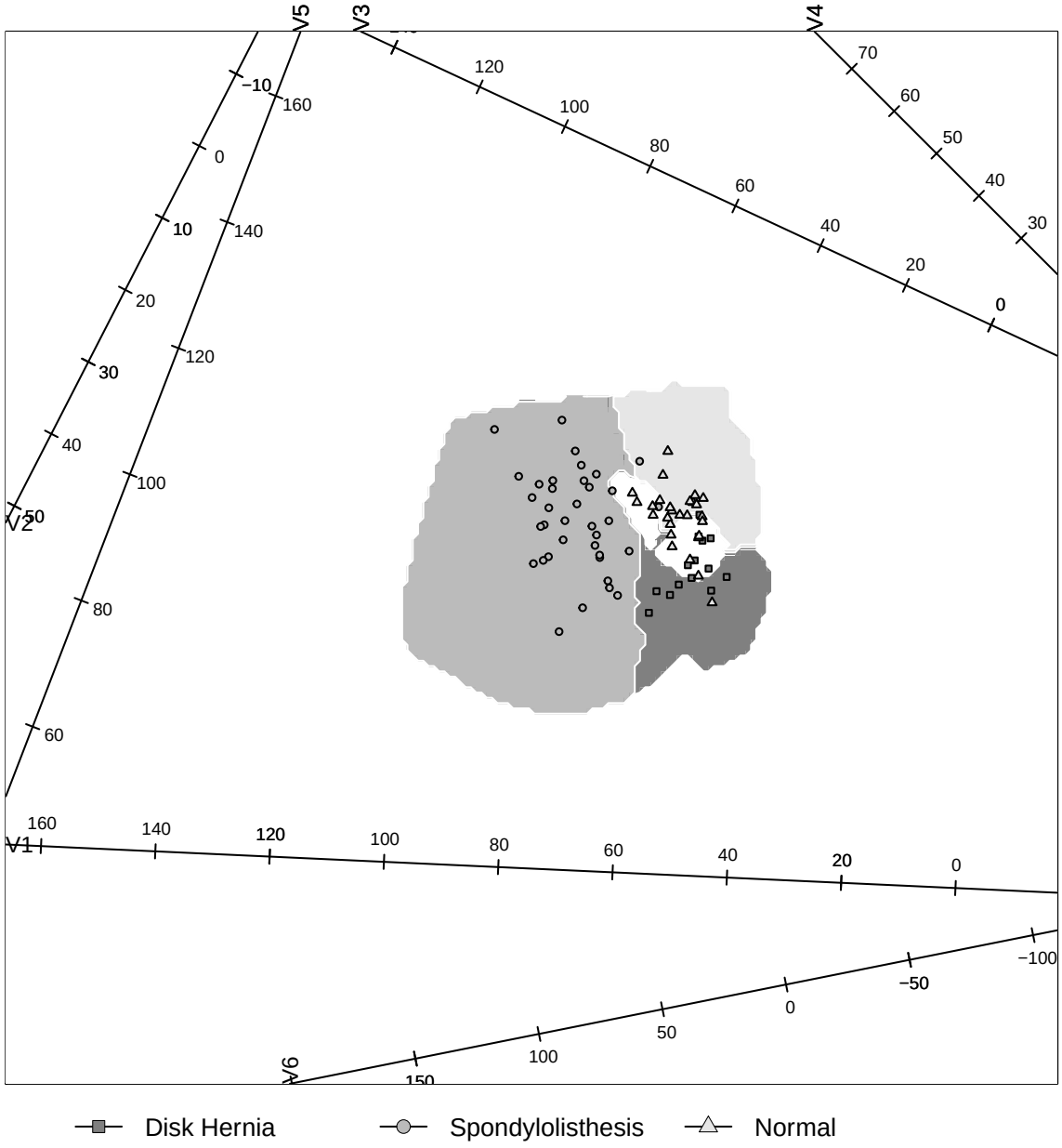


Figure 2.8: *Triplot with polybags and underlying KNN classification of out-of-sample data using polybags with $k = 21$ and 1.5 × 95% outer- and 85% inner-polybags for the vertebral column data.*

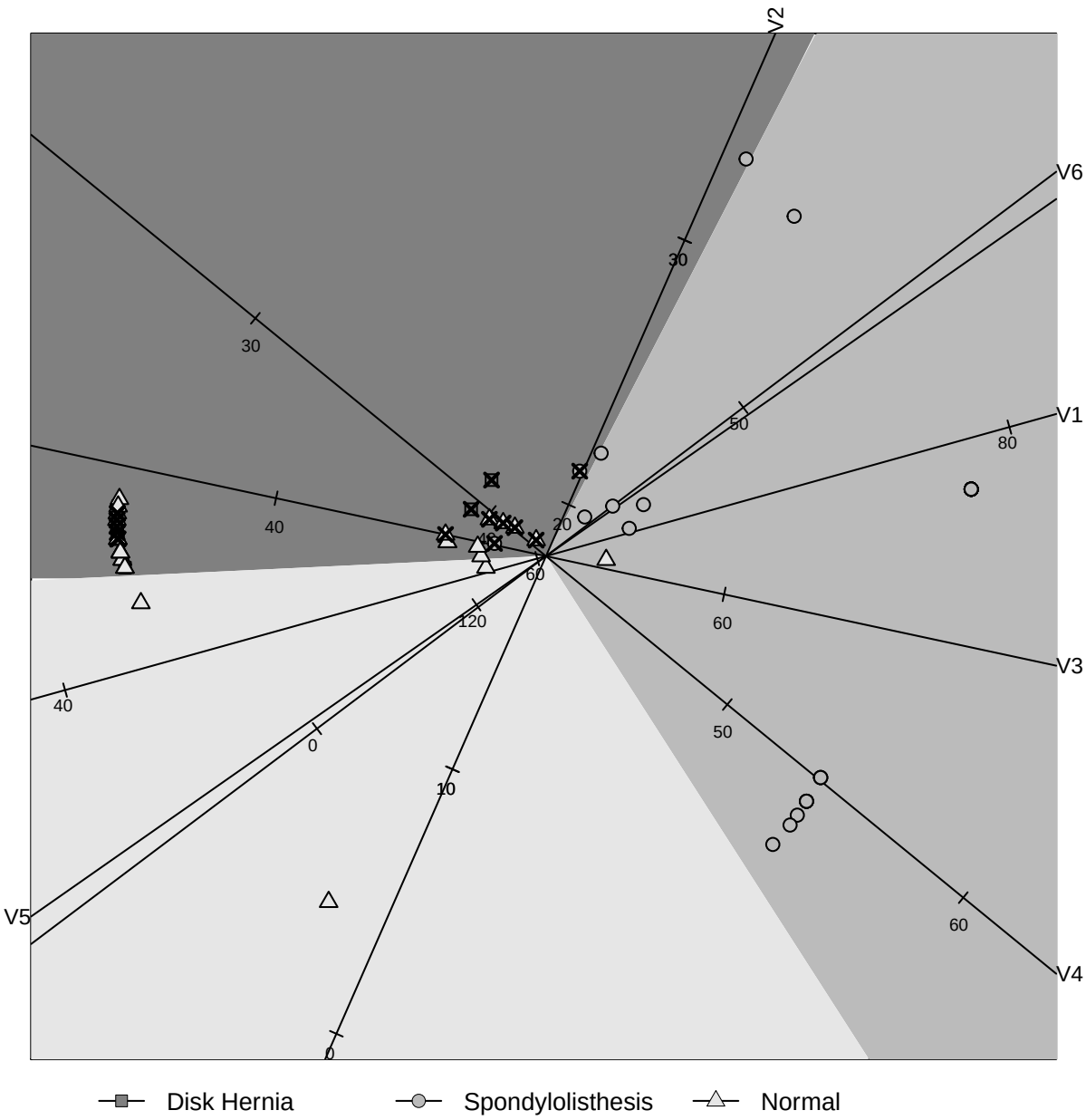


Figure 2.9: *Posterior probability log-ratio triplot with underlying KNN (k = 21) applied to the vertebral column data. Points that were not classified are indicated with a cross.*

Figure 2.9 shows almost no classification for disk hernia class (□) compared to figure 2.8. Another observable benefit of figure 2.8 vis-à-vis figure 2.9 is the underlying CVA biplot, rendering more accurate interpolation of points.

Finally, the data were also tested across all the methods discussed in section 4.2 and the resulting classification metrics are provided in table 2.4.

The proposed technique fares well against the other incorporated techniques, rendering either comparable or reduced misclassification type errors.

The above example illustrates that triplot and polybag methodology allows for intuitive visual classification and limits misclassification. While highly relevant to medical research in particular, the technique is broadly applicable to other areas of research as well.

5.2 Web-based application and replication

The technique was developed through creating and amending various pieces of code in order to produce the final triplot with polybags. As such, instead of leaving the replication exercise for the reader, a web-based application was created as supplementary data which allows the user to interact with all the properties and data sets discussed in this paper. Furthermore, the code underlying the technique is also available.

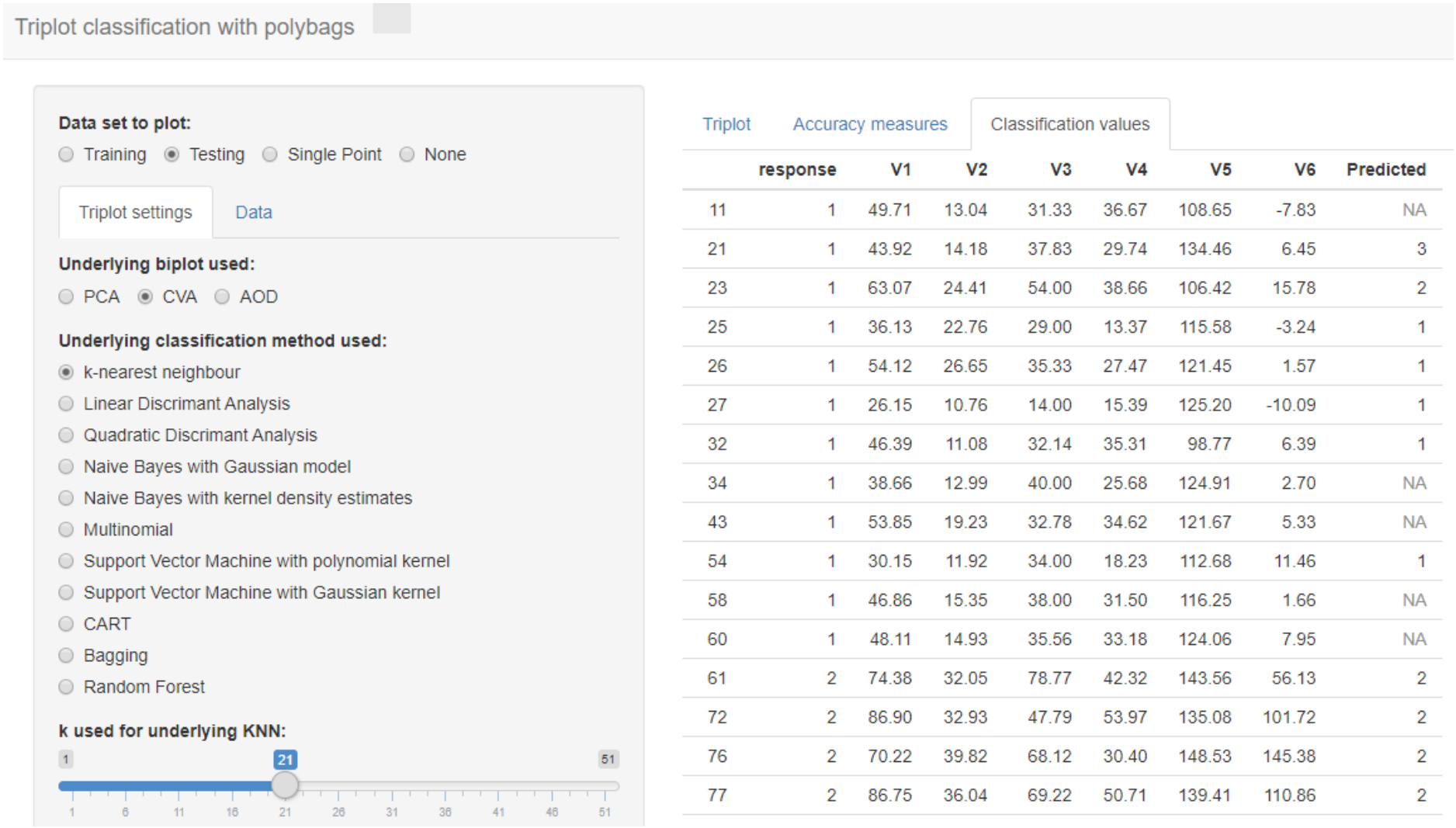
Web-based application

The web-based application allows the user to interact with various data sets as well as the various properties of the technique. The resulting triplot with polybags, together with the classification metrics and out-of-sample classifications, are provided. A screenshot and instructions on how to access it is provided in figure 2.10.

The application provides access to three preloaded data sets. The vertebral column data from section 5.1 is loaded automatically, and there is an additional set of bankruptcy data (Dheeru and Karra Taniskidou, 2017) where three levels (negative, average, and positive) of industrial risk, management risk, financial flexibility,

Table 2.4: *Simulated classification and misclassification type errors for the proposed (P), posterior probability log-ratio triplot (L), and black-box (B) techniques on the vertebral column data. A 75/25 training/validation split was used on the data. The measures were calculated as follows: Total accuracy:* $\frac{TP+TN}{N}$*; Total misclassification:* $\frac{NAP+NAN}{N}$*; Precision:* $\frac{1}{c}\sum_{i=1}^{c}\frac{TP_i}{PPC_i}$*; Negative predictive value =* $\frac{1}{c}\sum_{i=1}^{c}\frac{TN_i}{PNC_i}$*; True positive rate:* $\frac{1}{c}\sum_{i=1}^{c}\frac{TP_i}{TPC_i}$*; False negative rate:* $\frac{1}{c}\sum_{i=1}^{c}\frac{FN_i}{TPC_i}$*; False positive rate:* $\frac{1}{c}\sum_{i=1}^{c}\frac{FP_i}{TNC_i}$*; True negative rate:* $\frac{1}{c}\sum_{i=1}^{c}\frac{TN_i}{TNC_i}$*; with c equal to the number of classes.*

Classification method		**Total accuracy (+)**	**Total misclass. (−)**	**Precision (+)**	**Neg. pred. value (+)**	**True pos. rate (+)**	**False neg. rate (−)**	**False pos. rate (−)**	**True neg. rate (+)**
Average over all	P	0.68	0.09	0.86	0.94	0.62	0.10	0.04	0.71
	L	0.66	0.10	0.81	0.94	0.58	0.11	0.05	0.70
	B	0.69	0.07	0.86	0.96	0.60	0.09	0.04	0.71
KNN	P	0.68	0.08	0.86	0.94	0.62	0.09	0.04	0.71
	L	0.61	0.15	0.75	0.92	0.53	0.16	0.07	0.68
	B	0.69	0.07	0.84	0.95	0.60	0.08	0.03	0.71
LDA	P	0.68	0.08	0.86	0.94	0.62	0.08	0.04	0.71
	L	0.66	0.10	0.82	0.93	0.61	0.10	0.05	0.70
	B	0.64	0.13	0.79	0.93	0.56	0.13	0.06	0.69
QDA	P	0.68	0.08	0.86	0.94	0.62	0.09	0.04	0.71
	L	0.68	0.08	0.84	0.94	0.60	0.09	0.04	0.70
	B	0.69	0.07	0.86	0.96	0.59	0.09	0.04	0.71
NB with GM	P	0.68	0.08	0.86	0.94	0.62	0.09	0.04	0.71
	L	0.68	0.08	0.86	0.95	0.60	0.09	0.04	0.70
	B	0.68	0.08	0.86	0.94	0.62	0.09	0.04	0.71
NB with KDE	P	0.67	0.09	0.85	0.94	0.61	0.10	0.05	0.70
	L	0.64	0.12	0.78	0.92	0.56	0.14	0.06	0.68
	B	0.70	0.06	0.88	0.96	0.60	0.08	0.03	0.71
Multi-nomial	P	0.68	0.08	0.86	0.94	0.62	0.09	0.04	0.71
	L	0.70	0.06	0.89	0.96	0.63	0.06	0.03	0.72
	B	0.69	0.08	0.88	0.96	0.60	0.09	0.04	0.70
SVM with PK	P	0.68	0.08	0.86	0.94	0.62	0.09	0.04	0.71
	L	0.69	0.07	0.87	0.95	0.63	0.07	0.03	0.71
	B	0.67	0.10	0.84	0.94	0.58	0.11	0.05	0.69
SVM with GK	P	0.68	0.08	0.86	0.94	0.62	0.09	0.04	0.71
	L	0.69	0.07	0.86	0.96	0.61	0.09	0.04	0.71
	B	0.71	0.05	0.91	0.97	0.64	0.05	0.02	0.72
CART	P	0.67	0.09	0.85	0.94	0.61	0.10	0.05	0.70
	L	0.56	0.21	0.64	0.90	0.46	0.22	0.09	0.65
	B	0.71	0.05	0.90	0.97	0.64	0.06	0.03	0.72
Bagging	P	0.67	0.09	0.85	0.94	0.61	0.10	0.05	0.70
	L	0.69	0.07	0.84	0.95	0.60	0.08	0.03	0.71
	B	0.69	0.07	0.86	0.96	0.61	0.09	0.04	0.71
Random Forest	P	0.67	0.10	0.84	0.93	0.60	0.11	0.05	0.70
	L	0.64	0.13	0.79	0.93	0.56	0.13	0.06	0.69
	B	0.67	0.09	0.81	0.94	0.56	0.12	0.04	0.70



	response	V1	V2	V3	V4	V5	V6	Predicted
11	1	49.71	13.04	31.33	36.67	108.65	-7.83	NA
21	1	43.92	14.18	37.83	29.74	134.46	6.45	3
23	1	63.07	24.41	54.00	38.66	106.42	15.78	2
25	1	36.13	22.76	29.00	13.37	115.58	-3.24	1
26	1	54.12	26.65	35.33	27.47	121.45	1.57	1
27	1	26.15	10.76	14.00	15.39	125.20	-10.09	1
32	1	46.39	11.08	32.14	35.31	98.77	6.39	1
34	1	38.66	12.99	40.00	25.68	124.91	2.70	NA
43	1	53.85	19.23	32.78	34.62	121.67	5.33	NA
54	1	30.15	11.92	34.00	18.23	112.68	11.46	1
58	1	46.86	15.35	38.00	31.50	116.25	1.66	NA
60	1	48.11	14.93	35.56	33.18	124.06	7.95	NA
61	2	74.38	32.05	78.77	42.32	143.56	56.13	2
72	2	86.90	32.93	47.79	53.97	135.08	101.72	2
76	2	70.22	39.82	68.12	30.40	148.53	145.38	2
77	2	86.75	36.04	69.22	50.71	139.41	110.86	2

Figure 2.10: *A screenshot of the web-application that was created to supplement this paper. The link and code for the application can be found at* `https://doi.org/10.5281/zenodo.3562013` *(Van der Merwe, 2019). Note that this is not a direct link to the application, but rather to the GitHub repository for the code where the link is available at the top of the page.*

credibility, competitiveness, and operating risk are used to predict bankruptcy. The copper froth data set from Aldrich et al. (2004) is also included.

The application allows the user to visualise the training set, a testing set, or a single observation. This can be changed under the 'Data set to plot' radio buttons. Within the 'Data' tab, each of the above data sets' percentage test data can be changed, as well as the single point to be plotted.

The application also allows the user to upload data sets to experiment further with plotting. Note that the web-based application is limited to five classes, and care should be taken that the training and test set files follow the guidelines of the application.

The user can also simulate and analyse all of the 24 various data structures discussed in section 4.2.

Lastly, under the 'Triplot settings' the following can be changed:

- Underlying biplot used,
- Underlying classification method used,
- k used within KNN classification
- The size of the inner-polybag,
- The multiple and size of the outer-polybag,
- Whether the α-bags used in the construction of the polybags should be plotted,
- The eigenvectors used in the construction of the underlying biplot, and
- Size (zoom) and number of points presented on axes. (These two options are mostly relevant if data does not present well on the triplot.)

Underlying R code

The reader can access the underlying code for the Shiny application in the *app.R* file on the GitHub repository that can be found at `https://doi.org/10.5281/zenodo.3562013` (Van der Merwe, 2019). Here one can see the exact implementation of all the underlying source files. For completeness, the purpose of each of these source files is summarised in appendix C.

The code for the log-ratio triplot was obtained from the author.

6 Discussion and conclusion

In this paper a new classification methodology using traditional classification methods combined with biplots and α-bags was introduced. The technique introduces an inconclusive area within a triplot that can be seen as an area where there is insufficient evidence to classify a certain data point, necessitating further investigation. Incorporating this method reduces the misclassification error of the classification technique employed.

The technique and methods were illustrated on two randomly generated data sets, additional robustness checks were done, and a medical application was discussed. Additionally, a web-based application is available supplementary to this paper, allowing the reader to interact with the data and techniques discussed.

The simulation study showed that this technique performs very well compared to others for negatively correlated data, and as good as others for certain underlying classification methods such as KNN, LDA, and multinomial. Future research may focus on combining similar methodologies with triplots to further explore visualisation of classification techniques and related methods, as well as applying the approach to non-linear axes.

REFERENCES

Aitchison, J., Greenacre, M., 2002. Biplots of compositional data. Journal of the Royal Statistical Society: Series C (Applied Statistics) 51, 375–392. URL: https://doi.org/10.1111/1467-9876.00275.

Aldrich, C., Gardner, S., Le Roux, N.J., 2004. Monitoring of metallurgical process plants by using biplots. AIChE Journal 50, 2167–2186. URL: https://doi.org/10.1002/aic.10170.

Bates, D., Maechler, M., 2019. Matrix: Sparse and Dense Matrix Classes and Methods. URL: https://CRAN.R-project.org/package=Matrix. r package version 1.2-17.

Dheeru, D., Karra Taniskidou, E., 2017. UCI machine learning repository. URL: http://archive.ics.uci.edu/ml.

Gabriel, K.R., 1971. The biplot graphic display of matrices with application to principal component analysis. Biometrika 58, 453–467. URL: https://doi.org/10.1093/biomet/58.3.453.

Gardner-Lubbe, S., 2016. A triplot for multiclass classification visualisation. Computational Statistics & Data Analysis 94, 20–32. URL: https://doi.org/10.1016/j.csda.2015.07.014.

Gower, J.C., Gardner-Lubbe, S., Le Roux, N.J., 2011. Understanding biplots. John Wiley & Sons.

Gower, J.C., Krzanowski, W.J., 1999. Analysis of distance for structured multivariate data and extensions to multivariate analysis of variance. Journal of the Royal Statistical Society: Series C (Applied Statistics) 48, 505–519. URL: https://doi.org/10.1111/1467-9876.00168.

Greenacre, M., 2010. Biplots in Practice. Number 2011113 in Books, Fundacion BBVA / BBVA Foundation. URL: https://ideas.repec.org/b/fbb/booklb/2011113.html.

Greenacre, M., 2018. Compositional data analysis in practice. Chapman and Hall/CRC. URL: https://doi.org/10.1201/9780429455537.

Johnson, A., Baddeley, A., 2017. polyclip: Polygon Clipping. URL: https://CRAN.R-project.org/package=polyclip. r package version 1.6-1.

Korkmaz, S., Goksuluk, D., Zararsiz, G., 2014. Mvn: An r package for assessing multivariate normality. The R Journal 6, 151–162. URL: https://journal.r-project.org/archive/2014-2/korkmaz-goksuluk-zararsiz.pdf.

Kuhn, M., et al., 2008. Building predictive models in r using the caret package. Journal of statistical software 28, 1–26. URL: http://dx.doi.org/10.18637/jss.v028.i05.

Le Roux, N.J., Lubbe, S., 2013. UBbipl: Understanding biplots: Data sets and functions. URL: http://www.wiley.com/go/biplots. r package version 3.0.4.

Mardia, K.V., 1970. Measures of multivariate skewness and kurtosis with applications. Biometrika 57, 519–530. URL: https://doi.org/10.1093/biomet/57.3.519.

Rousseeuw, P.J., Ruts, I., Tukey, J.W., 1999. The bagplot: A bivariate boxplot. The American Statistician 53, 382–387. URL: http://doi.org/10.1080/00031305.1999.10474494.

Shapiro, S.S., Wilk, M.B., 1965. An analysis of variance test for normality (complete samples). Biometrika 52, 591–611. URL: https://doi.org/10.2307/2333709.

Van der Merwe, C.J., 2019. carelvdmerwe/triplotsimulation. URL: https://doi.org/10.5281/zenodo.3562013.

VanDerWal, J., Falconi, L., Januchowski, S., Shoo, L., Storlie, C., 2014. SDMTools: Species Distribution Modelling Tools: Tools for processing data associated with species distribution modelling exercises. URL: https://CRAN.R-project.org/package=SDMTools. r package version 1.1-221.

Wolf, H.P., Bielefeld, U., 2014. aplpack: Another Plot PACKage: stem.leaf, bagplot, faces, spin3R, plotsummary, plothulls, and some slider functions. URL: https://CRAN.R-project.org/package=aplpack. r package version 1.3.0.

APPENDICES

A Technical background on biplots and α-bags

One of the key concepts used in the new classification technique is biplots. Only the most important mathematical concepts from biplots will be discussed in this appendix. It is important to note that there are many different types of biplots. The focus here will be on three types: original PCA biplots, CVA biplots, and AOD biplots.

The main concept behind all three of these biplots is the same: reduce the dimensionality of the data in order to visually represent the data in the best manner possible. There are, however, key differences in the various methods. These are discussed in the next subsection.

In addition to the biplot methods, the concept of α-bags (adapted from bagplots) is also used in the new technique. Bagplots can be thought of as analogues to the traditional univariate boxplots for bivariate data.

A.1 Biplots

Biplots combine three statistical and mathematical techniques to visualise high dimensional data.

The first is the concept of scatterplots. Scatterplots allow the user to observe the relationships between variables in two dimensions, not only enabling visual determination of given points on the diagram, but also plotting new values.

The second is dimension reduction of the underlying data set. Now, given that a matrix $\mathbf{X} : n \times p$ with rank r can be written as $\mathbf{AB}$ with $\mathbf{A} : n \times r$ and $\mathbf{B} : r \times p$, one can force the rank to equal two (i.e. $r = 2$), such that $\mathbf{X} \approx \hat{\mathbf{A}}\hat{\mathbf{B}} = \hat{\mathbf{X}}$ with $\hat{\mathbf{A}} : n \times 2$ and $\hat{\mathbf{B}} : 2 \times p$. The reduction of the rank is done in such a way that the resulting approximation meets certain criteria. These two matrices allow for the construction of the biplot.

Lastly, simple geometry allows for the creation of the axes from the matrix $\hat{\mathbf{B}}$. By utilising these, one can graphically predict values from the observations quite easily by drawing perpendicular lines to each axis.

The three main methods for creating biplots of continuous ratio scale data are PCA, CVA, and AOD. PCA does not differentiate between the various classes in the data and the approximation is constructed by taking the singular value decomposition (SVD) of the full data set and subsequently the eigenvectors relating to the two largest eigenvalues to construct the biplot.

AOD and CVA differ from each other in that AOD optimally represents the class means, while CVA optimally separates the class means through maximising the variance between classes and minimising the variance within classes. Mathematically, CVA is equivalent to Fisher linear discriminant analysis, i.e.

$$\max_{\underline{m}} \frac{\underline{m}'\mathbf{S_B}\underline{m}}{\underline{m}'\mathbf{S_W}\underline{m}},$$

while AOD considers only

$$\max_{\underline{m}} \underline{m}'\mathbf{S_B}\underline{m},$$

with $\mathbf{S_B} : p \times p$ and $\mathbf{S_W} : p \times p$ the respective between and within class covariance matrices, c the number of classes, n the number of observations, and p the number of variables.

For PCA biplots, given a standardised matrix $\mathbf{X} : n \times p$ written as $\mathbf{U\Sigma V}$ with $\mathbf{U} : n \times r$, $\mathbf{\Sigma} : r \times r$, and $\mathbf{V} : r \times p$ with r the rank of the matrix $\mathbf{X}$, then the coordinates of the sample points can be approximated by $\mathbf{Z} = \mathbf{XV}'_{[\mathbf{2}]}$, where $\mathbf{Z} : n \times 2$ and $\mathbf{V}'_{[\mathbf{2}]} : p \times 2$. The p rows of $\mathbf{V}'_{[\mathbf{2}]} : p \times 2$ are used to construct the axes that represent the variables of the matrix $\mathbf{X}$. Therefore, should any n^* out-of-sample observations, say $\mathbf{X}^* : n^* \times p$, be plotted on the biplot, the new coordinates, $\mathbf{Z}^*$, can be found as $\mathbf{Z}^* = \mathbf{X}^*\mathbf{V}'_{[\mathbf{2}]}$.

CVA biplots are constructed by transforming the matrix $\mathbf{X} : n \times p$ to the canonical space through $\mathbf{Y} = \mathbf{XL}$ such that $\mathbf{L}'\mathbf{WL} = \mathbf{I}$ with $\mathbf{I},\ \mathbf{L} : p \times p$. If each class's covariance matrix is equal, then this transformation to the canonical space makes each class spherical with a covariance matrix equal to the identity matrix. The matrix $\mathbf{S_W} : p \times p$ therefore needs to be a good estimate of each within class covariance matrix, implying that the underlying class covariance matrices are assumed to be the same. Next, a PCA of the canonical means, i.e. $\bar{\mathbf{Y}} : c \times p$, is performed such that $\bar{\mathbf{Y}} = \bar{\mathbf{X}}\mathbf{L} = \mathbf{U\Sigma V}$. The class means can therefore be approximated with $\bar{\mathbf{Z}} = \bar{\mathbf{Y}}\mathbf{V}'_{[\mathbf{2}]} = \bar{\mathbf{X}}\mathbf{L}\mathbf{V}'_{[\mathbf{2}]} = \bar{\mathbf{X}}\mathbf{M}'_{[\mathbf{2}]}$, with $\mathbf{M}'_{[\mathbf{2}]} : p \times 2$ used for the construction of the variable axis. Subsequently, new out-of-sample observations can be found as $\mathbf{Z}^* = \mathbf{X}^*\mathbf{M}'_{[\mathbf{2}]}$ from the n^* out-of-sample observations $\mathbf{X}^* : n^* \times p$.

There are various ways to construct an AOD biplot, for the application in this paper AOD was applied on a matrix of Euclidean inter-sample distances resulting in the AOD reducing to a PCA biplot of the class means. That is, let $\mathbf{X} : n \times p$ be a matrix with c classes, then a matrix of the means of the classes can be constructed as $\bar{\mathbf{X}} : c \times p$. Similar to the PCA biplot, the matrix $\bar{\mathbf{X}}$ is then decomposed as $\mathbf{U\Sigma V}$ with $\mathbf{U} : c \times r$, $\mathbf{\Sigma} : r \times r$, and $\mathbf{V} : r \times p$ with r the rank of the matrix $\bar{\mathbf{X}}$. The coordinates of the sample points can then be calculated with $\mathbf{Z} = \mathbf{XV}'_{[\mathbf{2}]}$, where $\mathbf{Z} : n \times 2$ and $\mathbf{V}'_{[\mathbf{2}]} : p \times 2$. The matrix $\mathbf{V}'_{[\mathbf{2}]} : p \times 2$ is used to construct the axes that represent the variables of the matrix $\mathbf{X}$. Finally, new sample points can be added in the same way as with PCA biplots. This method does not make assumptions about the covariance structures between groups as CVA does, and brings the additional benefit of optimally representing the class means, while PCA only focuses on the full

data set. Note that, for AOD more complex methods exist such as those where the axes become non-linear. The research can be extended to biplots with non-linear axes where other distance measures are used. See Gower and Krzanowski (1999) for more detail on the analysis of distance for structured multivariate data.

All functions required to construct biplots can be found in the package *UBbipl* (Le Roux and Lubbe, 2013). The code was amended in order to extract certain intermediary calculations for the purpose of this analysis. It should be noted that to overcome that the matrix $\mathbf{S_W} : p \times p$ sometimes has eigenvalues very close to zero (which could result in strange results for its inverse), the function *nearPD* from the *Matrix* (Bates and Maechler, 2019) package was used to find the closest positive definite matrix with eigenvalues that are further from zero.

A.2 Bagplots and α-bags

The concept of α-bags is adapted from bagplots, which were first introduced by Rousseeuw et al. (1999) as the bivariate generalisation of the classic boxplot. A bagplot is constructed in a two-dimensional space and provides similar information of the data set as a boxplot. This includes its location (the depth median), spread (the size of the bag), correlation (the orientation of the bag), skewness (the shape of the bag and the loop), and tails (the points near the boundary of the loop and the outliers). The bag usually contains $n/2$, or 50%, of the data points in the sample, and the loop is constructed by drawing a polygon through the points that lie closest to a fence which is derived by inflating the bag by a factor of three. The R package *aplpack* (Wolf and Bielefeld, 2014) contains the code for generating these bagplots.

One of the key benefits of bagplots compared to other methods of enclosing a configuration of sample points, is that it considers the concentration of the sample points in construction.

Gower et al. (2011) adapted the concept of the bag by changing the percentage of points that lie within the bag itself and subsequently ignoring the fence and loop. This allows for the representation of various levels of concentration (α) of the sample points over different classes in one graphic. These bags are termed α-bags.

B Data utilised

This section provides a more detailed description of the data that was simulated to illustrate the proposed technique. Note that the distribution parameters were randomly chosen for the different data sets. For both data sets the means were kept the same, for the SSV data set the covariance matrices for the classes were randomly chosen and kept the same for each class, while for the LDV data set the variances were changed for each class.

B.1 SSV data set

Three classes were simulated, each from their own multivariate normal distribution. They all have the same underlying variables V_1 to V_4. The different classes are denoted by Class A (□), Class B (∘), and Class C (△). In the distribution below, the superscript of the variables indicates the specific class's underlying variable's distribution. Fifty samples from each class were randomly generated and the means and covariance matrices were arbitrarily chosen. The number of observations were arbitrarily chosen as 50 as a trade-off between having too few data points, producing a volatile illustration of the method, and too many points, hence over-fitting the method.

$$\begin{pmatrix} V_1^A \\ V_2^A \\ V_3^A \\ V_4^A \end{pmatrix} \sim N \left[\begin{pmatrix} 110 \\ 90 \\ 70 \\ 50 \end{pmatrix}, \begin{pmatrix} 100 & 50 & -50 & 20 \\ 50 & 100 & -20 & 0 \\ -50 & -20 & 100 & 70 \\ 20 & 0 & 70 & 100 \end{pmatrix} \right]$$

$$\begin{pmatrix} V_1^B \\ V_2^B \\ V_3^B \\ V_4^B \end{pmatrix} \sim N \left[\begin{pmatrix} 70 \\ 90 \\ 110 \\ 130 \end{pmatrix}, \begin{pmatrix} 100 & 50 & -50 & 20 \\ 50 & 100 & -20 & 0 \\ -50 & -20 & 100 & 70 \\ 20 & 0 & 70 & 100 \end{pmatrix} \right]$$

$$\begin{pmatrix} V_1^C \\ V_2^C \\ V_3^C \\ V_4^C \end{pmatrix} \sim N \left[\begin{pmatrix} 90 \\ 110 \\ 130 \\ 150 \end{pmatrix}, \begin{pmatrix} 100 & 50 & -50 & 20 \\ 50 & 100 & -20 & 0 \\ -50 & -20 & 100 & 70 \\ 20 & 0 & 70 & 100 \end{pmatrix} \right]$$

It is clear from the correlation multiplot in figure 2.1 that the correlation structures of all the responses are the same. For some of the variables, the observations (Class B (∘) and Class C (△)) also tend to overlap.

Comparing the correlation multiplot to the biplot in figure 2.4, the three groups can be differentiated. The groups of data are relatively concentrated and clearly separated from each other, and if any of these data points were not included in the training set, out-of-sample classification would be simple.

Through considering the α-bags, one can interpret the different classes with ease. Consider now drawing lines perpendicular from the α-bags to the various axes. As an example, Class A (□) takes on values of $100 < V_1 < 120$, $80 < V_2 < 100$, $60 < V_3 < 80$, and $30 < V_4 < 70$. These are very much in line with the underlying distribution. Considering the other classes and denoting a particular class by I_{class}, the following holds:

$$I_{\text{A}} = \{100 < V_1 < 120 \cap 80 < V_2 < 100 \cap 60 < V_3 < 80 \cap 30 < V_4 < 70\}$$
$$I_{\text{B}} = \{60 < V_1 < 80 \cap 80 < V_2 < 100 \cap 100 < V_3 < 120 \cap 120 < V_4 < 160\}$$
$$I_{\text{C}} = \{80 < V_1 < 100 \cap 100 < V_2 < 120 \cap 120 < V_3 < 140 \cap 130 < V_4 < 170\}$$

Although visually far superior to a scatterplot, this method of interpretation of the biplot is quite complicated and it would be much easier to utilise the coordinates of the points as inputs into some classification technique.

B.2 LDV data set

A second similar data set is created that has the same means as the SSV data set, but for which the covariance matrices were randomly generated for each of the different classes. Next, the distribution for each of the variables is given. A notation similar to section B.1 is used.

$$\begin{pmatrix} V_1^A \\ V_2^A \\ V_3^A \\ V_4^A \end{pmatrix} \sim N \left[\begin{pmatrix} 110 \\ 90 \\ 70 \\ 50 \end{pmatrix}, \begin{pmatrix} 1225 & -283 & 641 & 1057 \\ -283 & 2025 & -129 & 908 \\ 641 & -129 & 1600 & 303 \\ 1057 & 908 & 303 & 3025 \end{pmatrix} \right]$$

$$\begin{pmatrix} V_1^B \\ V_2^B \\ V_3^B \\ V_4^B \end{pmatrix} \sim N \left[\begin{pmatrix} 70 \\ 90 \\ 110 \\ 130 \end{pmatrix}, \begin{pmatrix} 2025 & -961 & -152 & 1373 \\ -961 & 1225 & -801 & -1036 \\ -152 & -801 & 1600 & 1197 \\ 1373 & -1036 & 1197 & 3025 \end{pmatrix} \right]$$

$$\begin{pmatrix} V_1^C \\ V_2^C \\ V_3^C \\ V_4^C \end{pmatrix} \sim N \left[\begin{pmatrix} 90 \\ 110 \\ 130 \\ 150 \end{pmatrix}, \begin{pmatrix} 2025 & 977 & -391 & 984 \\ 977 & 1600 & -1008 & 253 \\ -391 & -1008 & 3025 & 731 \\ 984 & 253 & 731 & 1225 \end{pmatrix} \right]$$

Figure 2.4 also provides a CVA biplot with 95%-bags of the LDV data. As expected, the ease of graphical interpretation decreases significantly in this case. The correlation multiplot provides very little information apart from Class A (□) being differentiated from Classes B (∘) and C (△) in terms of some variables, and Class B (∘) and Class C (△) overlap significantly.

Considering the biplot with the α-bags, Class A (□) is shown to be clearly distinguished from Class B (∘) and Class C (△), and the majority of Class B (∘) is contained within Class C (△) on the biplot.

C Summary of R code

Below follows a list of all the source files associated with the Shiny application. The files can be found at `https://doi.org/10.5281/zenodo.3562013` (Van der Merwe, 2019).

bipldrawknn.R This is the main function that draws the triplot. It takes the training and test sets, the size of the inner- and outer-polybags, the type of underlying biplot, the underlying classification technique, and the eigenvectors that should be used as inputs. Some visualisation parameters can also be changed.

clipcords.R The function that returns the inner-polybag given the coordinates of the α-bags. The code can take up to 5 classes.

compute.bagplot_C.R Amended the code for calculating the bagplot coordinates from the *aplpack* (Wolf and Bielefeld, 2014) package to extract the α-bags.

confmetrics.R Code for extracting the classification metrics given a confusion matrix that contains observations that were left unclassified.

createdata2.R Code written to create three classes of multivariate normal distributions given a set of means, variances, portions for each class, correlations within classes, and number of observations. It also finds the closest invertible correlation matrix, so the correlations can be randomly simulated.

CVAbipl_C.R and PCAbipl_C.R Amended code from the *UBbipl* (Le Roux and Lubbe, 2013) package in order to extract additional information from the computation of the underlying biplots. The *CVAbipl_C.R* function also includes amended code for finding the closest invertible within class covariance matrix.

Draw.line2.R, Draw.onecmline.R, and Plot.marker.new.R Code from the *UBbipl* (Le Roux and Lubbe, 2013) package, used to redraw the axes on top of the classification area.

drawbipl.bagalpha_C.R Amended code from the *UBbipl* (Le Roux and Lubbe, 2013) package to export the coordinates of the α-bag when drawing the base biplot.

DrawOrthogline.R Code from the *UBbipl* (Le Roux and Lubbe, 2013) package, used to draw orthogonal lines from data points in the triplot perpendicular to the axes.

Eigen.twosided.R and indmat.R Code from the *UBbipl* (Le Roux and Lubbe, 2013) package required for the *CVAbipl_C.R* function to work.

returnconfusion.R Code for returning the confusion matrix of a classification method, given a predetermined percentage of points not to be classified.

unioncords.R The function that returns the outer-polybag given the coordinates of the α-bags. The code can take up to 5 classes.

CHAPTER 3

CLASSIFYING YIELD SPREAD MOVEMENTS THROUGH TRIPLOTS: A SOUTH AFRICAN APPLICATION

This working paper has been submitted to an investments journal. This paper was co-authored with supervisor Tertius de Wet.

Abstract

Significant movements in yield spreads from a sparse data environment are classified using various share, interest rate, financial ratio, and economic type covariates in a visually interpretive manner. This allows for a better understanding of how various factors drive the changes in yield spreads. Additionally, this visualisation technique provides the ability to classify whether an unlisted debt instrument's yield spread had significantly changed or stayed stable during a specific observation period. The analysis was implemented in a web-based application as well.

1 Introduction and background

In lieu of in-house quantitative experts, many corporate entities fail to employ appropriate techniques when calculating the fair value of unlisted debt, thus opting to leave the initial implied spread unchanged. While it is up to management to decide how fair values are to be determined, supporting methodology underlying their decision is still required.

When calculating the fair value of an unlisted debt instrument for reporting purposes, the best approximation is the exit price quoted on a given day. The process of obtaining this actual price from the counterparty is, however, impractical and cumbersome. The yield used for pricing consists of a risk-free portion that can easily be observed, and a spread that is instrument specific. If analysis renders sufficient evidence to show that the spread would have changed from inception of the instrument, the fair value can be investigated further and, for example, an actual updated exit price is obtainable. If, however, there is not sufficient evidence that significant movement had occurred, the spread could be left unchanged and only the risk-free yield curve updated.

Furthermore, management may often be reluctant to incorporate unattributable volatile fair value profit or losses. Additionally, applying some spread model immediately after the inception of an unlisted debt instrument could also create a 'day one' profit or loss, implying that the security was not purchased at market value. These practical challenges emphasise the need for a method providing a simpler explanation why changes in the yields (or valuation) had or had not occurred.

Note also that, even if unlisted debt is held at amortised cost in the financial statements, it is still required - under the International Financial Reporting Standards (IFRS) 7 Article 25, for instance - that the fair value must be disclosed. Companies, however, sometimes simply state that the fair values of instruments are reasonably approximated by their respective amortised cost values. While the fair value could possibly be approximated by the amortised cost in the early or late stages of a debt instrument's term or for certain types of instruments, the assumption will render inaccuracies. Simply stating that the approximation holds could misinform the reader of the financial statements, and careful assessment would be required in auditing to determine whether sufficient evidence is available to support such a claim.

Considering the problem from the perspective of an auditor, there are different effects that incorrect decisions would have. If, for example, the client decided to keep the spread constant, and the auditor erroneously accepts this decision, the balance sheet item would be misstated. IFRS 13 states that financial instruments held at fair value need to be valued at their exit price, implying that if the unlisted debt instrument is sold at year-end, the value (excluding trading cost) should be

very similar to the value that is stated in the balance sheet. That implies that if the asset is sold immediately after year-end, that there could be a considerable profit or loss, only being attributable to the asset not being valued correctly. On the other hand, if the client correctly decides to keep the spread constant and the auditor believes that the spread should have changed, then the client would have to incur indirect costs to obtain an exit-price or updated spread (and in all likelihood have a negative impact on the relationship with the auditor due to the additional work that had to be done).

This research proposes a methodology that can be used to determine whether an increase, decrease, or no change in the spread would have occurred for a specific debt instrument, given observed market conditions. Additionally, the proposed method provides new insights into the determinants of spread changes not discussed in previous research. South African data is used to illustrate the practical application of the methodology, but it would have equal international relevance.

The proposed analysis and methodology differ from previous research in two significant ways. The first is that an indicator for change in spread is used rather than the size of the movement. That is, a test is proposed to establish whether there had been a significant movement in the spread, rather than continually determining new spreads. This allows for the current spread to be kept constant, or, if there is sufficient evidence of a significant move, to choose an appropriate method for determining a new level.

Secondly, the visualisation of the movements and their determinants are incorporated in the analysis through k-nearest neighbour (KNN) triplots with polybags (Van der Merwe, 2019b). This is used as an alternative to other numbered-output, or black-box methods where analyses cannot always be visually interpreted. Triplots allow for the two-dimensional representation and classification of high dimensional data. Polybags then additionally restrict classification in certain areas of the triplot through considering overlapping concentration of the data.

The inner-polybag on a KNN triplot is a non-classifiable area where the overlap of the data is too high for clear differentiation. It is taken as the intersection of the α-bags for any of the classes and is denoted by α_{IN}%.

The outer-polybag places an outer boundary for the classification areas to prevent classification of outliers and is taken as the area outside the overlapping region of multiple inflated α-bags and is denoted by $z \times \alpha_{OUT}\%$, with z the inflation factor.

The triplot therefore contains a classification area where new points can be plotted and classified according to the areas they fall into. The observations' values are read from the graph by drawing perpendicular lines from the observation point to each of the axes - similar to a scatterplot, but with more than two axes.

In this research, the triplot proposed by Van der Merwe (2019b) is expanded to include an additional visually interpretable classification method. The triplots are built on the theory of biplots, for which the reader can refer to Gower et al. (2011) and Greenacre (2010) for a more detailed background.

The response that will be investigated in this research is an indicator of increasing, decreasing, or stable movement of the spread over a certain period, and its relationship with various liquidity, share, interest rate, financial ratio, and economic type covariates. The interpretation of the covariates shows that the incorporation of the stable class for the movements in spreads resulted in some interesting findings: while most variables either have a direct relationship or inverse relationship with the change in spread, some of these also have a split relationship with the stable state. This implies that movements of some factors have a stronger negative impact on spreads than the positive impact they have, or vice versa.

The paper is set out as follows: in the following section an overview of previous literature is provided, followed by a detailed discussion of the data utilised in this study. The variable selection process and results are provided, after which the analysis is given and discussed. Finally, some concluding remarks are given.

2 Literature overview

Determining the factors that drive the change in yield spreads of traded bonds has been the subject of a vast amount of research, with Collin-Dufresn et al. (2001) the first authors to investigate it in detail.

Collin-Dufresn et al. (2001) investigated the determinants of credit spread changes on vanilla industrial bonds using multi-linear regression and principal component

analysis (PCA) applied to the residuals. They found that theoretically explanatory variables offered low explanatory power in their analysis, that the residuals are highly cross-correlated, and that PCA implies that these changes are mostly driven by a single common factor. The authors could not explain the common systematic component, suggesting it is principally driven by local supply/demand shocks that are independent of both credit factors and liquidity proxies. They not only incorporated the default risk, but also the loss given default (in other words, the recovery rate). Since the corporate bond market tends to have relatively high transaction costs and low volume, they also investigated the extent to which the credit spread changes can be explained by proxies for liquidity changes. Their choice of covariates was based on the structural models of default, which also informed the variables used in the regression.

Avramov et al. (2007) expanded on the research of Collin-Dufresn et al. (2001) to include all grades of bonds and more explanatory variables. They found that different sets of explanatory variables have different importance depending on the grade of the bonds and that some of their additional variables yielded significant explanatory power.

Avramov et al. (2007) also pointed out that the difference in studying spread changes rather than spread levels is equivalent to the difference between studying equity prices and equity excess returns - an important insight since, although these two fields draw upon each other, they are two completely different areas of research.

Furthermore, through considering three measures of bond-specific liquidity measures, Chen et al. (2007) found that liquidity is priced into corporate yield spreads and reaffirms that neither the level nor the dynamics of yield spreads can be fully explained by default risk determinants. Bao et al. (2011) incorporated a theoretical measure of illiquidity, namely the amount of price reversals captured by the negative of the autocovariance of price changes and showed that it is both statistically and economically significant with regards to bond prices. They did not investigate the relationship with regards to change in yields.

Most research in changes in yield spreads follows a similar approach: (i) a set of debt instruments is identified, (ii) the spread changes are calculated over a certain period , (iii) a set of covariates is identified, (iv) various multi-linear regression analyses are performed, and (v) the results are interpreted for statistical and economic

significance. While the approach in this paper is very different from the above, the relevant covariates were chosen based on such past research.

The reader is referred to Radier et al. (2016), where the authors provide an extensive overview of literature in this field, as well as a detailed background to the South African bond market, comparing conclusions of previous research in this context.

In all the literature referenced above, the change in spread is presented as the dependant variable that is modelled, mostly over a different group of bonds and over different time periods and time spans. Many of the covariates that were used in the current analysis were also considered in these studies and can be split into five broad categories, namely: liquidity, equity, interest rate, economic, and financial ratio type measures. The literature also contains covariates which would not always be readily observable in sparse data environments, and as such, these were left out of this analysis, therefore allowing for broader application of the methodology.

In the next section the data utilised in this application are discussed in detail.

3 Data utilised

The analysis in this research was performed on data from listed non-state-owned entities' listed vanilla fixed coupon bonds for the period 30 September 2007 to 30 April 2018. Full records prior to this date were not available from the Johannesburg Stock Exchange (JSE). These sample bonds contained no callable, early redemption, nor split maturity features. Only bonds for which all covariate data points were available were included. These various liquidity, share, interest rate, financial ratio, and economic type covariates will be defined in the following subsections.

3.1 The indicator

The most important variable in this analysis is an indicator variable created to denote change in spread over a time period.

It attempts to capture significant movement in the spread. In the application a significant movement in spread was arbitrarily defined as 25 basis points (up or down).

Table 3.1: *The average effect in absolute terms an increase/decrease of 25 basis points in the yield would have on different bonds valued at their respective par yields. From the table it can be seen that bonds with lower coupons and longer maturity would be the most severely impacted by the 25 basis point move.*

	PAR YIELD							
TERM	**2%**	**4%**	**6%**	**8%**	**10%**	**12%**	**14%**	**16%**
1	0.25%	0.24%	0.24%	0.23%	0.23%	0.22%	0.22%	0.22%
5	1.18%	1.11%	1.05%	1.00%	0.95%	0.90%	0.86%	0.82%
10	2.25%	2.03%	1.84%	1.68%	1.54%	1.41%	1.30%	1.21%
15	3.21%	2.78%	2.43%	2.14%	1.90%	1.70%	1.54%	1.39%
20	4.09%	3.40%	2.87%	2.46%	2.13%	1.87%	1.66%	1.48%
25	4.88%	3.91%	3.20%	2.67%	2.27%	1.96%	1.72%	1.52%
30	5.60%	4.33%	3.44%	2.82%	2.36%	2.01%	1.75%	1.54%

Once all bonds to be included in the study were identified, the spread data were obtained for each of them. The spread was calculated as the yield of a bond above the yield of their respective companion government bonds. The trade volume for each trading day was also included. The daily spread was then multiplied by an indicator function which equals one if the bond was traded on a given day and zero if it was not. This new variable - the traded spread - therefore reflects the spread only on days when trading took place and allows for interpolation between traded days. In previous research, the data were only used if a predetermined number of trades took place during the observation period.

Several options regarding treatment of days when no trading took place exist, the first being to ignore them. Even though this would be a convenient approach, given the sparsity of the data, it would leave one with very little data for the research. The second option is to keep the spread constant until the new data point is observed, but this is unlikely to provide a true reflection of the market, as market participants would have anticipated changes in the spread prior to the trade. The other possibility is to apply linear interpolation between the two traded spreads. This approach would increase the number of points available for analysis, but it would imply that the market started moving towards a new spread immediately following a trade, which is also unrealistic. For these reasons, a combination of keeping the spread flat between traded days and linear interpolation towards the next observed point

was employed. The first half of the missing traded spread data points were kept at the previous traded spread, after which the second half was interpolated between that value and the next traded spread. Whilst there is no single correct way to approach this problem, the described approach mitigates the complications of the various approaches to an extent.

Once the missing traded spreads were calculated, they were considered as a time series of value movements. The total observation period of a bond was then divided into time sub-periods that could either be classified as upward (+1), downward (−1), or stable (+0) movement periods. The period was classified as a stable period until the first significant movement of more than 25 basis points from the reference spread (initial spread) was observed. After this observation, the stable period ended and the upward/downward (depending on the direction of the movement) movement period started, together with a new reference spread taken as the first spread in the new period. This period continued until either another significant movement from the new reference spread was observed, or a stable period of spread movements was observed. Once a stable number of trading days were observed (chosen as five in this application), a new stable period was started together with a new reference spread. The process repeated itself until the last observed traded spread was considered.

Each of the bonds were then linked to their listed parent company and 116 one-year rolling window periods taken at month-end, starting 30 September 2007 and ending 30 April 2017, were created. The movements (+1, −1, +0) for all the companies' bonds over those periods were added together. This total movement was then floored and capped at −1 and +1 respectively, thereby indicating whether the aggregate movement for a certain company's bonds was upward, downward, or stable during the one-year observation period. This indicator is the dependent variable investigated in this study.

It is important to note that there are various ways in which the indicator could have been created. The proposed approach was chosen because it allows for continuous observation periods and allows for correction of unaligned movements if some of a company's bonds show a decrease in spread and some an increase. It furthermore incorporates the important stable state when no significant movements were observed.

A graphical illustration of the indicator is presented in figure 3.1 and the results of some robustness checks on the indicator is provided in table 3.2.

There was a total of 256 bonds from 28 different parent companies for which the spread data were available and which were therefore included in the final sample. The majority of the bonds that traded in the period were excluded either for not being vanilla type bonds or having some callable or early redemption feature, resulting in 690 bonds to consider. Once these bonds were excluded, almost another third (296 bonds) were discarded due to not having available spread data. In the final step a filter was added to only include bonds which had a listed non-state-owned entities as their parent companies resulting in the final sample of 256 bonds. These bonds were then analysed according to their movements and grouped according to their parent companies and the overall movement for the parent company for the rolling one-year period were recorded. This resulted in a total of 2369 observations.

In the next subsection, the covariates are discussed.

3.2 Covariates

Data from the parent company for the time period were gathered to provide the relevant information for each of the observation periods. Once all the covariates were sourced, the parent company's reference was removed, and the final data set was used as is. No imputation with regards to missing data was done and the observations were removed if not all the covariates were present. The largest contributors to data that were not available was the financial statement data and underlying share data. This resulted in the removal of 422 observations such that there were 1947 observation in the final data set.

All bond data were obtained from the JSE, with the underlying bond and equity pricing data, as well as financial ratio data from IRESS. Interest rate data were obtained from Reuters, and economic data from Quantec's EasyData platform.

Liquidity measures

Change in liquidity was measured as the change in the percentage of days with no active trades over a one-year period. It is expected that there will be a direct rela-

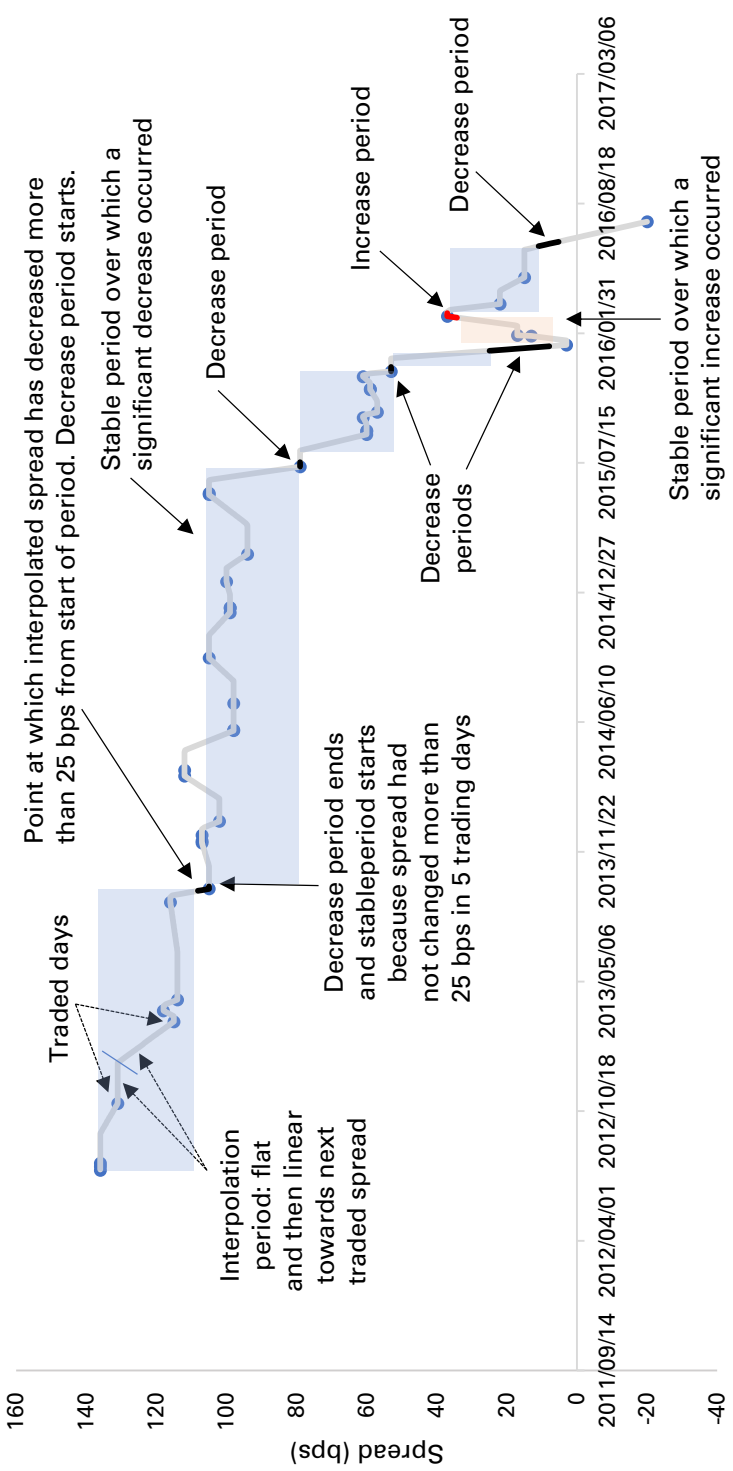


Figure 3.1: *An illustration of the indicator for bond FRX16. The first four arrows on the left indicate how missing spreads were interpolated between traded spreads. It was first kept flat, after which it was linearly interpolated to the next traded spread (the split is indicated by a small line). The next two arrows show how a decreasing period starts and ends after no significant movement is observed. Once no significant movement was observed for five trading days, a stable period started. The remaining arrows on the right point to various increasing (+1), decreasing (−1), and stable (0, shaded) periods for this specific bond.*

Table 3.2: *Results of the robustness checks done on the indicator. Four tests with regards to robustness were performed. The first was to randomly create missing data, the second was to change the number of stable trading days that needs to be observed before a stable period starts, the third was to see the effect of using either a straight line interpolation method and keeping the spread constant from the previous trading date, and lastly the effect of changing the significant basis point (bps) movement value. From the analysis it can be seen that the indicator is relatively robust for missing data, the number of stable days required, and the interpolation method. Furthermore, the indicator is sensitive to the choice of the significant basis point movement.*

Category	**Stress**	**Decr. (-1)**	**Stable (0)**	**Incr. (+1)**	**New NAs**	**Total Δs**	**Δs from +/-1 to 0**	**Δs from 0 to +/-1**	**Sign Δs**	**No Δs**	**New NAs %**	**Δs %**	**No Δs %**
Missing data stress	5%	909	696	338	4	17	7	10	0	1926	0	1	99
	10%	909	692	337	9	21	9	11	1	1917	0	1	98
	20%	901	653	353	40	42	15	26	1	1865	2	2	96
	40%	887	644	349	67	96	47	40	9	1784	3	5	92
	80%	893	615	319	120	297	142	119	36	1530	6	15	79
Stable days stress	1	893	696	358	0	52	21	28	3	1895	0	3	97
	10	894	704	349	0	49	23	22	4	1898	0	3	97
	25	927	660	360	0	92	16	59	17	1855	0	5	95
	50	953	632	362	0	171	45	116	10	1776	0	9	91
	100	1004	580	363	0	274	63	186	25	1673	0	14	86
Interp. method	Straight	926	667	354	0	96	28	64	4	1851	0	5	95
	Stable	893	736	317	1	102	61	28	13	1844	0	5	95
Bps move stress	1	1165	299	483	0	536	12	416	108	1411	0	28	72
	5	1119	388	440	0	436	30	345	61	1511	0	22	78
	10	1064	488	394	1	334	44	258	32	1612	0	17	83
	50	739	958	249	1	363	302	47	14	1583	0	19	81
	100	444	1352	150	1	712	671	22	19	1234	0	37	63

tionship between this measure and the change in spreads. This variable denoted as *d.Illiquidity*.

Year-on-year changes in the difference between the 10-year government bond yield and swap rate (*d.(Y-S)_10*) were chosen as an indication for funding risk. As swaps are not funded but bonds are, an increase in the value would indicate an increase in the spread required to compensate for funding risk. Alternatively, as per Collin-Dufresn et al. (2001), a decrease in this value (thus higher values for the swap rates) could indicate a decrease in the liquidity of the swap market, which could spill over to the bond market and subsequently increase spreads.

Interest rate type measures

Whilst movement in the base interest rate curve would already be incorporated in the companion bond's yield, the change in the government bonds' yield term structure could provide valuable information on the expectation of the economy, and therefore aspects of company growth and recovery rates.

The changes in particular points (levels) on the term structure are denoted by *d.Y_x*, where *x* was chosen as two, five, 10, and the maximum available term in years. The change in the slope of the yield curve was also considered and was calculated as the difference between the 10- and two-year, five- and two-year, maximum available and two-year, and maximum available and 10-year rates, indicated by *d.Slo_(x−y)*. Increases in yield curve slopes are indicative of increases in forward rates - and therefore the level and slopes can be considered together. Avramov et al. (2007) proposed two opposing hypotheses to explain the effect a change in yields could have on credit spreads. The first would be that an increase in the yield curve could provide a higher reinvestment rate for a firm, therefore increasing firm value, which in turn reduces credit spreads. On the other hand, increasing yields imply that the borrowing rates also increase, diminishing the extent to which a firm can take on profitable projects. This could decrease the value of the firm and subsequently increase spreads. The opposite holds for a decrease in yield curves.

The curvature estimation as proposed by Diebold and Li (2006), calculated as twice the two-year rate minus the three-month and 10-year rates, denoted as *d.Cur*(*DL*), was incorporated. A higher curvature parameter indicates a more volatile expected

period in the short term. Therefore, a higher curvature would signal potentially lower recovery rates and therefore higher credit-related spreads.

Additionally, three of the parameters were also estimated based on the sparse data approximation of Van der Merwe et al. (2018). These year-on-year changes in the level, slope, and curvature parameters were calculated as follows: *d.Lev(CvdM)*= $0.9Z+1.0Z-0.2Z$, *d.Slo(CvdM)*= $0.0Z-1.0Z$, and *d.Cur(CvdM)*= $0.2Z-1.0Z$, where xZ represents the $x \times 100^{\text{th}}$ percentile of the available term structure.

Second and third order changes in the 10-year government bond yield were also incorporated to account for non-linear movements. These are indicated as $(d.Y\_10)^x$ with x equal to two and three.

Economic type measures

The South African Reserve Bank (South African Reserve Bank, 2015) publishes three composite business cycle indicators: a leading, coincident, and lagging index. Each of these consist of underlying inputs that indicate which direction the economy is heading (leading), the current state of the economy (coincident), and what realised (lagging). The year-on-year changes of these indices as well as the change in the year-on-year changes of these indices were included to capture any macroeconomic factors and the effects thereof. A positive change in the indices indicates either an expectation or a realisation of an upward turn in the economy. A positive value of these changes should therefore indicate higher expected recovery rates and positive business growth and hence lower credit related spreads (and vice versa). They are denoted by *Leading*, *Coincident*, and *Lagging*.

An increase in the year-on-year percentage movement in the indices would have a similar effect as the level. One should, however, be careful of the state from where the change originates - for example a 1% increase from -10% is better, but still good compared to a 1% increase from a base of 10%. These are denoted by *d.Leading*, *d.Coincident*, and *d.Lagging*.

The components of the various business cycle indicators as per the SARB can be found in the appendix. As these are published with an approximate three month lag, the covariate data coinciding with a certain observation period were also lagged by three months.

Financial ratio type measures

Changes in selected financial ratios, which could indicate the creditworthiness of a firm, were also considered as part of the set of independent variables. These include assets over capital employed (*d.AC*), current ratio (*d.CR*), debt over assets (*d.DA*), debt over equity (*d.DE*), interest cover (*d.IC*), leverage factor (*d.LF*), and long-term loans as percentage of total debt (*d.LTL*).

An increase in assets over capital and the current ratio would indicate a stronger financial position and therefore decrease credit related spreads. An increase in debt over assets and debt over equity would indicate a weaker financial position and therefore increased credit-related spreads.

An increase in interest cover would indicate a stronger position to service current debt and therefore decreased credit related spreads. Furthermore, an increase in the leverage factor would indicate a better utilisation of leverage employed and therefore decreased credit-related spreads.

Finally, a decrease in the long-term loans as percentage of total debt would indicate an increase in short-term debt, putting strain on cash flows to service current debt which would dilute the ability to service current debt and therefore increase credit-related spreads.

Note that in order to incorporate publication lags, a general three month lag was used when considering changes in the financial ratio measures.

Equity type measures

Some measures from the equity market were also included. The change in volatility skew (*d.VolSkew*), as per Collin-Dufresn et al. (2001), was incorporated as a measure for the probability of negative jumps in the market. An increase in the volatility skew would therefore indicate a higher risk of jumps, necessitating additional premiums for investors.

The year-on-year change in the parent company's return (*d.R*), their sector's index (*d.I*), their excess return to their sector's index (*d.(R-I)*), as well as the excess return to market (*d.(R-M)*) were used as an indicator of how well the firm and relevant

sector performed during the period considered. Increases in return and excess returns indicate market sentiment that a company or sector is increasing in financial health, therefore decreasing spreads.

Finally, some volatility measures were considered. Firstly, the year-on-year change in the daily volatility of the parent company (*d.RVol*); secondly, the change in the excess return to index volatility (*d.(R-I)Vol*); and finally, the change in the difference between the firm's volatility and the market's volatility (used as a proxy for idiosyncratic volatility) - *d.(RVol-MVol)*. An increase in any of these measures indicates more risk, and therefore it would increase the spreads required by investors as compensation.

In the next section all these covariates were used together with the dependent variable to perform variable selection to reduce the number of variables in the final analysis. In order to test the accuracy of the analysis, a validation sample of 25% was chosen and the remaining training data were used in the variable selection and subsequent analysis.

4 Variable selection

After all the variables were identified, a subset of the key variables needed to be determined. Numerous variable selection techniques are available, as expounded by Guyon and Elisseeff (2003).

James et al. (2013) note that shrinkage methods for variable selection fit a model containing all p variables and shrink the coefficient estimates towards zero, reducing the variance of the estimates. One such method - the lasso (Tibshirani, 1996) - allows for coefficients to shrink to exactly zero. The adaptive lasso by Zou (2006) further improves on the lasso to include the oracle property, thereby allowing for more consistent selection of variables.

The adaptive lasso estimates are given by:

$$\underline{\hat{\beta}} = \underset{\underline{\beta}}{\operatorname{argmin}} \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{i,j} \right)^2 + \lambda \sum_{j=1}^{p} \hat{w}_j |\beta_j| \right\}$$

with $\hat{w}_j = 1/|\hat{\beta}_j^*|^\gamma$ for $j = 1, 2, ..., p$, $\gamma > 0$, and $\{\hat{\beta}_j^*\}$ a number of non-zero initial estimates of the covariates.

The tuning parameter (λ) was evaluated through calculating the multinomial deviance over a range of values via cross-validation. Each evaluation provides a point estimate of the multinomial deviance together with a one standard deviation range. The final output renders two values: the value of λ providing minimum mean cross-validated error; and the λ, which indicates the most regularised model such that the error is within one standard deviation of the minimum. The latter yields fewer variables and was thus used in this analysis. Once this parameter was obtained, it was used to fit the lasso and the variables included were noted. The package *glmnet* (Friedman et al., 2010) in R was used to implement the variable selection.

Following a similar approach to that of Morozova et al. (2015), the above process was repeated 1000 times (through resampling with replacement of the training set), and a graph of all the variables and their percentage inclusion rates was plotted.

A λ for the initial training data set was obtained through cross validation - this specific λ excluded 18 of the 37 variables. Keeping this λ fixed, 95% confidence intervals for the coefficients were obtained using the bootstrap. The bootstrap was performed by resampling the training data set 1000 times and fitting the adaptive lasso for each sample. If zero was not included in the 95% confidence interval, then the coefficient was considered to be significantly different to zero. Note that each variable has three coefficients because there are three classes per indicator. The number of significant coefficients can therefore either be zero, one, two, or all three of the coefficients. It should be noted that there could be some dependency in some of the observations affecting the results of the bootstrap. The dependency could be addressed through using other bootstrapping techniques such as the block-bootstrap.

The results of the variable selection are provided in figure 3.2. The final variables selection can be seen in the figure as those with a higher inclusion rate than the variable with the lowest inclusion rate that had at least one significant coefficient (i.e. all variables with a higher inclusion rate than *d.Y_max*).

The simulation yielded 19 variables for inclusion - all of which had at least one significant coefficient. While the variable *d.Illiquidity* was included almost all the

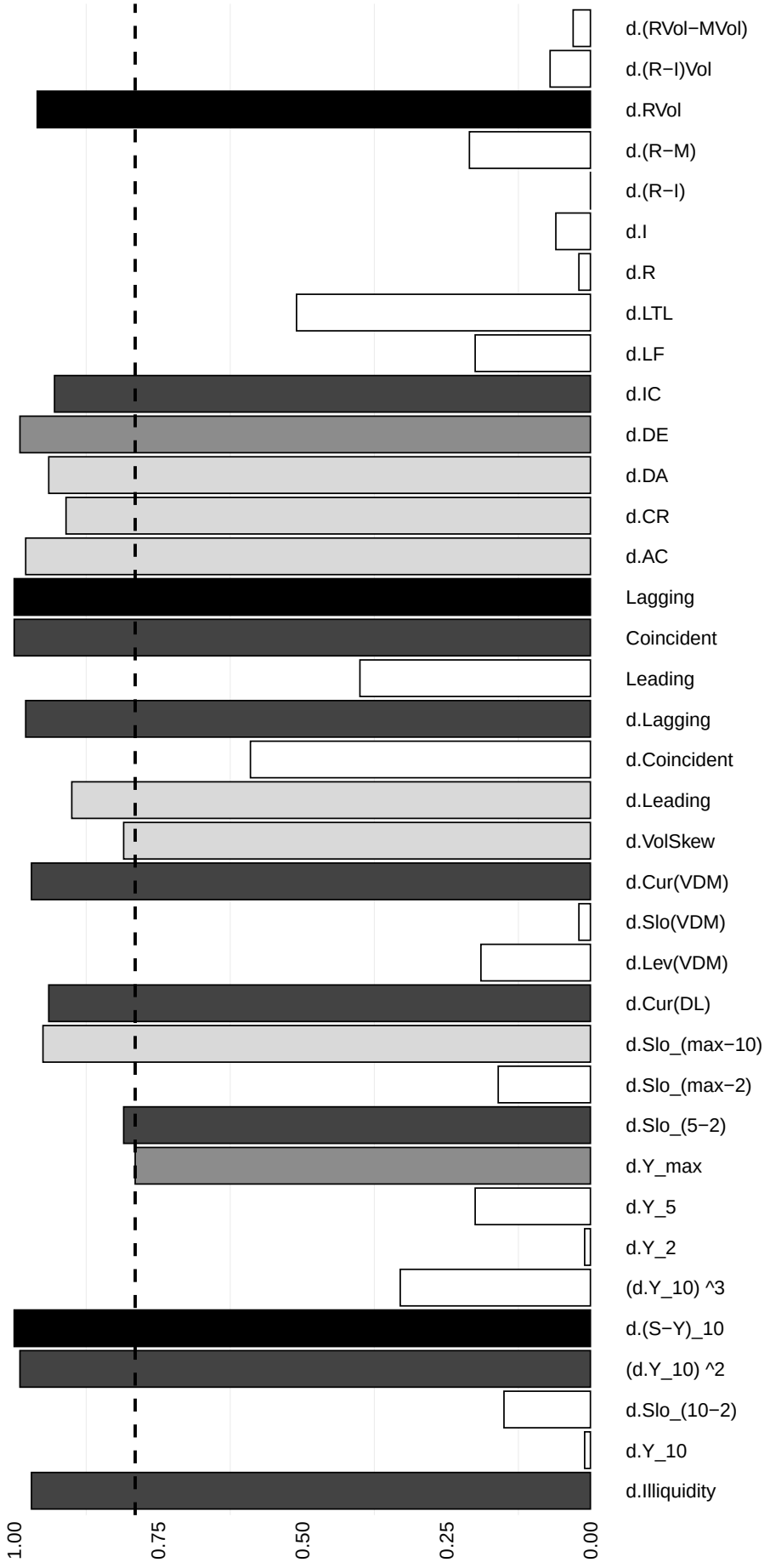


Figure 3.2: *The results of the simulated variable selection process. The bars indicate the fraction of times that a variable was chosen as part of the 1000 resampling iterations. The colour of the bars indicates whether the variable was excluded from the final regression (blank), had none (light grey), one (medium grey), two (dark grey), or three (black) significant coefficients at a 95% confidence interval. The dotted line indicates the fraction of the variable with the lowest inclusion rate that has at least one significant coefficient. All variables with a higher inclusion rate, except for change in illiquidity, were included in the subsequent analysis.*

time, it will not form part of the analysis due to the nature of the variable. That is, it was used to help explain some of the variance, but when used for prediction of unlisted debt instruments it will be inconsequential. Therefore, only 18 variables were left. A summary of these covariates, split according to the movement class, can be found in table 3.3.

The training data's 18 variables are plotted in figure 3.3 in a KNN ($k = 41$) triplot with an inner-polybag of $\alpha_{IN} = 0\%$ and outer-polybag of $z \times \alpha_{OUT} = 2 \times 95\%$. The sizes of the inner- and outer-polybags were chosen for optimal visualisation of the classification area of the triplot. For ease of visualisation the tick marks and labels of the axes are not shown on the triplot.

The tuning parameters k, α_{IN}, and $z \times \alpha_{OUT}\%$, can take on various values. If the choice of k is too small, then the classification region will overfit the training data and out of sample classification will not perform very well. On the other hand, if k is too high, then the classification areas will be too rigid. For the application k was chosen as 41 as it delivered classification regions that were not overly rigid, but still fit the training data relatively well. The size of the inner-polybag was chosen as 0% as a way to make it easier to introduce the adaptation of the triplot in the next section. The outer-polybag was chosen sufficiently large to include all data points.

The triplot shows that the data are concentrated in the centre, but also adequately separated. Furthermore, the axes of only a few variables lie on top of each other, indicating that the variables are mostly not strongly correlated. The problem with the triplot, which will be addressed in the following section, is that it is not visually easily interpretable due to the many axes included. The only method available for classification on this graph is mathematically solving for the position of a new sample point and then determining which coloured area it lies in (see Gardner-Lubbe, 2016).

5 Analysis

While the triplot in figure 3.3 can be used for classification as is, the large number of variables reduces the attractiveness regarding visual interpretability. One could, for example, draw four separate triplots where only certain variables are shown. The

Table 3.3: *Summary statistics of all the covariate data used in the analysis. The Shapiro-Wilk (Shapiro and Wilk, 1965) test for normality was performed on all the marginal distributions and all of the variables were found to be distributed significantly different from the normal distribution. Additionally, the Mardia (Mardia, 1970) test for multivariate normality showed that none of the classes are multivariate normally distributed.*

Variable	Class	Mean	Std. Dev	Median	Min	Max	25th	75th	Skewness	Kurtosis
$(d.Y\_10)^2$	−1	0.84	1.16	0.38	0.00	5.48	0.08	1.19	2.14	4.51
	0	1.08	1.35	0.45	0.00	5.48	0.10	1.54	1.60	1.92
	+1	0.78	1.13	0.33	0.00	5.48	0.08	1.05	2.14	4.48
d.(Y-S)_10	−1	0.07	0.33	0.11	−0.84	0.69	−0.15	0.33	−0.63	−0.20
	0	0.11	0.40	0.19	−0.84	1.36	−0.11	0.36	−0.48	0.16
	+1	0.38	0.32	0.35	−0.84	1.36	0.19	0.54	0.45	1.59
d.Y_max	−1	0.16	0.66	0.10	−1.66	2.15	−0.36	0.56	0.53	0.33
	0	0.14	0.78	0.10	−1.66	2.15	−0.45	0.66	0.27	−0.17
	+1	0.13	0.84	0.14	−1.66	2.15	−0.44	0.60	0.22	−0.15
d.Slo(5-2)	−1	0.06	0.40	0.01	−0.61	2.05	−0.19	0.31	1.37	4.93
	0	0.16	0.51	0.02	−0.64	2.05	−0.13	0.34	1.83	3.85
	+1	0.40	0.75	0.11	−0.64	2.05	−0.10	0.84	0.97	−0.34
d.Slo(max-10)	−1	0.24	0.44	0.30	−1.16	1.00	0.01	0.58	−0.74	0.52
	0	0.07	0.45	0.08	−1.16	1.00	−0.21	0.41	−0.27	0.11
	+1	0.05	0.29	0.07	−1.16	1.00	−0.17	0.21	−0.07	1.46
d.Cur(DL)	−1	−0.01	0.57	0.01	−2.05	1.65	−0.30	0.38	−0.31	0.74
	0	0.11	0.68	0.12	−2.24	2.40	−0.26	0.50	0.01	1.67
	+1	0.11	0.96	0.15	−2.24	2.40	−0.41	0.64	−0.14	0.38
d.Cur(VDM)	−1	−0.27	0.49	−0.24	−1.54	0.75	−0.68	0.10	0.08	−0.52
	0	−0.06	0.51	−0.07	−1.54	0.75	−0.49	0.31	−0.27	−0.67
	+1	−0.05	0.56	0.07	−1.54	0.75	−0.49	0.38	−0.62	−0.27
d.VolSkew	−1	−0.06	1.41	−0.04	−3.54	2.97	−0.96	0.73	0.13	−0.32
	0	−0.22	1.29	−0.26	−3.54	2.97	−1.01	0.39	0.22	0.11
	+1	−0.22	1.20	−0.27	−3.54	2.97	−0.77	0.27	0.30	1.13
d.Leading	−1	1.92	13.52	−0.59	−21.30	37.72	−4.39	6.50	1.00	0.89
	0	0.91	9.77	−0.36	−21.30	37.72	−2.83	2.82	1.53	4.41
	+1	−1.74	9.55	−1.86	−17.18	35.28	−5.82	1.23	1.35	3.47
d.Lagging	−1	1.00	9.79	1.27	−35.77	20.40	−2.43	5.00	−1.04	2.56
	0	−1.07	8.00	−0.51	−35.77	20.40	−3.35	3.39	−1.46	4.74
	+1	−4.62	10.34	−2.43	−35.77	13.77	−6.12	1.14	−1.40	1.46
Coincident	−1	2.75	3.21	2.78	−12.91	6.92	0.79	5.24	−1.87	6.40
	0	0.99	4.13	1.56	−12.91	6.92	0.39	2.95	−1.99	4.16
	+1	−2.13	5.53	0.40	−12.91	6.02	−5.81	1.41	−0.96	−0.61
Lagging	−1	−2.48	5.54	−0.89	−22.31	5.24	−3.14	0.60	−1.87	2.99
	0	−1.12	5.31	−0.59	−22.31	13.46	−2.00	0.62	−1.53	4.66
	+1	−1.18	7.96	−0.60	−22.31	13.46	−3.24	1.12	−0.63	0.80
d.AC	−1	−0.03	0.90	0.00	−4.55	5.34	−0.09	0.12	1.03	21.46
	0	−0.04	0.34	0.00	−4.55	1.29	−0.09	0.05	−3.75	45.28
	+1	−0.04	0.56	0.00	−4.55	1.29	−0.15	0.14	−5.99	46.63
d.CR	−1	−0.05	0.57	−0.01	−4.48	1.99	−0.13	0.08	−5.33	42.17
	0	−0.06	0.44	−0.03	−1.72	1.26	−0.16	0.06	−0.59	4.60
	+1	−0.10	0.97	0.03	−4.37	3.32	−0.05	0.16	−2.44	11.91
d.DA	−1	−0.01	0.07	0.00	−0.49	0.22	−0.03	0.01	−2.70	18.57
	0	0.01	0.08	0.00	−0.49	0.38	−0.02	0.02	0.45	11.86
	+1	0.01	0.06	0.00	−0.10	0.26	−0.02	0.03	1.86	5.81
d.DE	−1	−0.31	0.81	−0.15	−4.27	2.03	−0.55	−0.01	−1.39	5.87
	0	−0.07	1.04	−0.05	−4.92	4.57	−0.27	0.16	−0.30	12.15
	+1	0.01	0.95	0.03	−2.20	3.26	−0.24	0.36	0.43	3.28
d.IC	−1	−0.01	6.24	0.01	−28.28	31.90	−0.18	0.43	0.19	14.69
	0	0.20	9.10	−0.01	−21.03	53.31	−0.89	0.70	3.26	18.02
	+1	−2.51	10.34	−0.21	−47.43	53.31	−1.91	0.02	−1.88	14.58
d.Rvol	−1	−4.81	11.31	−2.63	−70.77	24.73	−10.03	1.52	−1.17	4.10
	0	−0.40	13.38	−0.21	−70.31	66.71	−6.04	7.72	−1.31	7.23
	+1	9.47	22.78	5.51	−56.16	166.62	−1.45	16.76	3.58	21.63

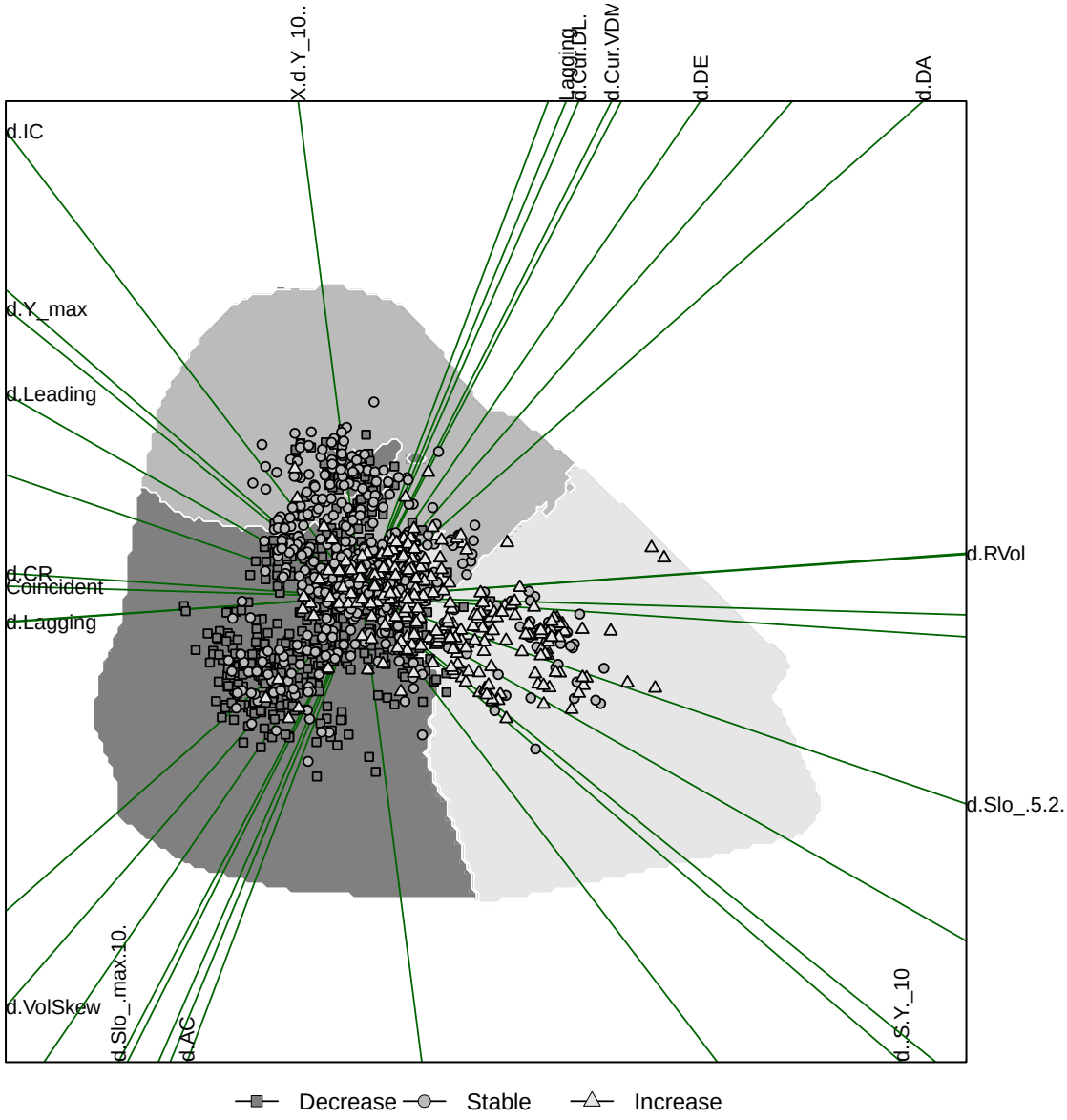


Figure 3.3: *A triplot with underlying k-nearest neighbour classification (k = 41) of the training set data, together with an outer-polybag of 2 × 95% and 0% inner-polybag. The dark grey area indicates a decrease in spread, the medium grey area indicates a period of no significant change in spreads, and the light grey an increase in spreads. By drawing perpendicular lines to the various axes, it becomes clear which variables have discriminatory power with regards to the various classes.*

axes can also easily be shifted parallel and truncated in order to show the centre of the triplot. These separate triplots with parallel shifted and truncated axes are subsequently further improved for visual interpretability, as outlined below.

If one considers any point on any of the axes, a perpendicular line could be drawn at that specific point identifying all areas on the triplot corresponding to that value of the variable. This perpendicular line will cross through the triplot classification region, essentially showing that, for the specific variable's value, there could be various outcomes of the indicator. Additionally, some of the perpendicular lines will not intersect the classification areas of the triplot at all. This is illustrated in figure 3.4 for some values of a specific variable, *d.IC*.

The KNN triplot is therefore adapted (AKNN triplot) to summarise the classification areas of each point on the axes through a multi-shaded bar through orthogonally projecting the outer- and inner-polybags on the variables' axis. That is, to get an idea of the wideness of the classification region for the different values of a variable, the proportion of each class with respect to the intersection of the outer-polybag at a value is plotted on the variables' axis. In addition, the predictivity of each variable is also computed.

This leads to a new method of classification using the triplot. There are four properties in constructing the AKNN that can be varied, and all combinations of these will be tested for accuracy. For the implementation discussed in this paper, k was chosen as 41 for the underlying KNN classification, allowing for sufficiently clear separation of the classes.

The first two properties are the inner- and outer-polybags. For this implementation fairly small and large polybags were chosen. The inner-polybags were chosen as 0% (essentially not including it) and 75% respectively, while the outer-polybags were chosen as $2 \times 95\%$ and $8 \times 95\%$, with the latter being a very large outer polybag taking up almost all the triplot space (essentially not including the effect of the outer-polybag in the triplot). The inner- and outer-polybags change the appearance of the classification region, with the outer-polybag determining the length of each axis.

The third property, illustrated in figure 3.4, is where the classification regions are summarised on each axis through drawing a grid on the classification region per-

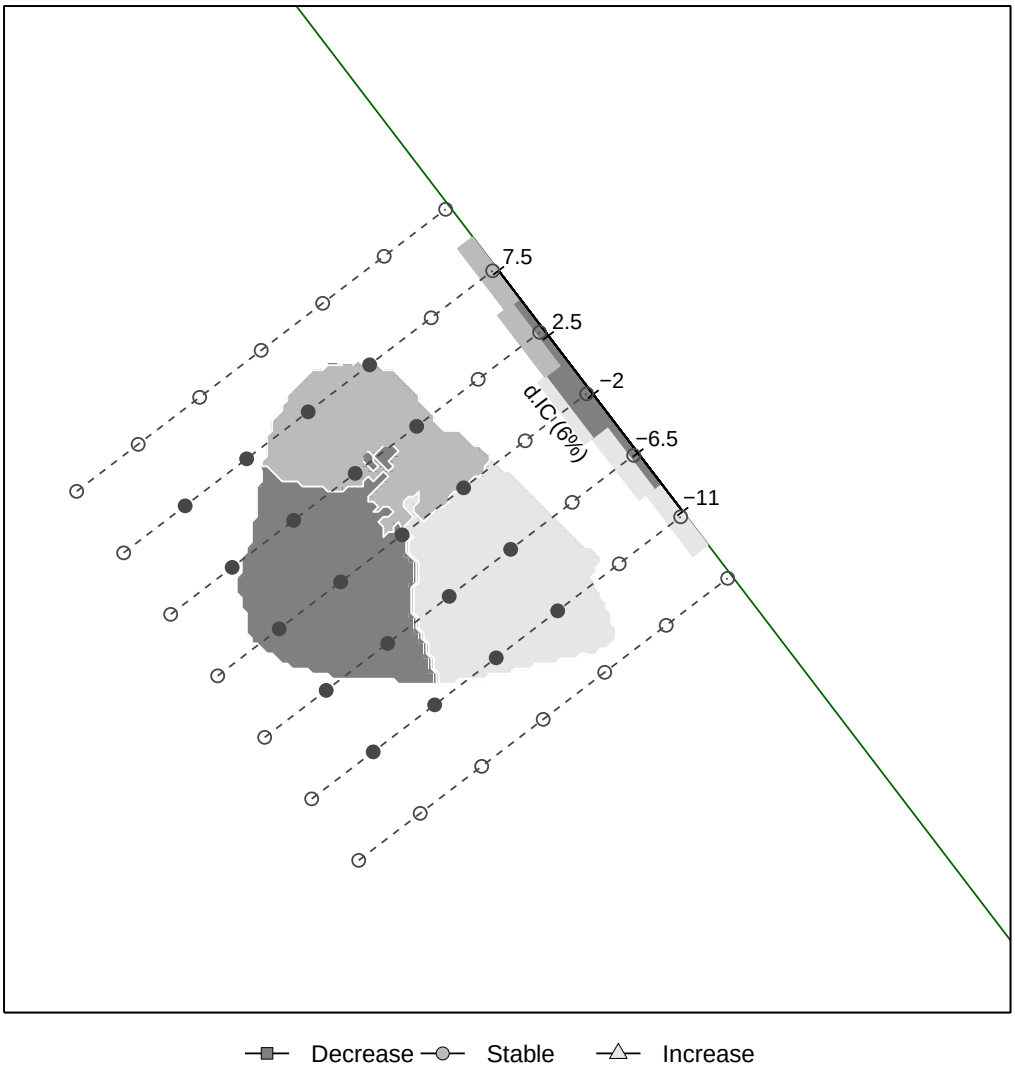


Figure 3.4: *By only considering one (d.IC) of the 18 axes in figure 3.2 and shifting it parallel to the right, the above graph is obtained. Additional to the shifted axis, a 7 × 7 grid is drawn perpendicular to the axis. Each of the points on this grid is subsequently classified according to the classification region. These are then summarised on the axis itself. Here the 'count' type axis is displayed and therefore indicates the number of times each point on the grid is classified in a class per line, and proportions it as a percentage of the grid that crosses the outer-polybag (the black dots). The final implementation uses a much finer grid such that more accurate classification is obtained.*

pendicular to each of the axes. The points on the grid are classified and then summarised on the axis itself. There are two methods to determine these summarised areas: either consider points inside the outer-polybag region exclusively or consider the full grid over the outer-polybag region. The first method therefore considers the number of points per class as proportion of points on that line that lie within the outer-polybag, while the second considers the number of points per class as a portion of the number of points on the grid. The latter method allows for smooth edges of the summarised axes. Note that the inner-polybag would result in a non-classified area in the middle of the axes as part of both summarised areas. The values of the summarised axes are 'scores' that each value obtains for each class at each variable, rendering a value between zero and one for each of the classes. They are multiplied by the predictivity scores of the axes to obtain the final score used in the calculation. The predictivity of the j-th axis is a measure of the accuracy of the two-dimensional approximation associated with the j-th original variable. It is expressed as a proportion of the sum of squares for the prediction versus the sum of squares for the original observed value. Should the prediction be 100% accurate, the proportion will be one, and the larger the difference between the predictions and observed values, the smaller the proportion (Gardner-Lubbe et al., 2008).

This allows for the variables that have better predictivity in the triplot to be allocated higher weights with regards to classification. The two properties are respectively referred to as the ratio- (R-) and count-type (C-type) scoring properties, where the R-type always adds to one, and the C-type only counts the number of points on the grid that touches the classification region.

The last property is the values of the observations that are used in the scoring process. There are two values which can be used, either the actual (A) values from the original data, or the values that are displayed on the triplot itself, namely the predicted (P) values. The predicted values drawn from each axis will intersect precisely at one point on the triplot, while the actual values will not.

Once the observations are scored, their totals for each of the classes, or indicator, are added and the class with the highest score is assigned to that observation.

In the next subsection the results for the classification using the various methods are provided and discussed.

5.1 Comparison of methods

In table 3.4 the various combinations of the AKNN triplot that were tested together with the normal KNN classification (i.e. the black-box model) and the normal KNN triplot with polybags (Van der Merwe, 2019b) are given.

The results reveal that the normal KNN on its own is the best performing model. From the other methods the results are very similar - not only to each other, but to the full KNN classification model as well. Therefore, for this application, at least, the properties can be interchanged easily, and similar results will be obtained.

In the next subsection the AKNN-triplots are illustrated. To further aid in the interpretation and understanding of the adapted triplots, a web-based application was created so that the reader can interact with all the parameters and inputs discussed in this paper (see figure 3.5).

5.2 Interpretation of the graphs

The illustrated triplots contain both the actual and predicted values for one specific observation. The inner-polybag was chosen as 0%, and outer-polybag was chosen as 2 × 95%, with C-type scoring axes used. In figure 3.6, the four different group of covariates' AKNN triplots are provided and the variables will be interpreted in the subsequent sections. Note that, while axes with a low predictivity might not always be interpreted correctly, the influence on the classification will be negligible, as their score incorporates the predictivity.

The share type covariates as shown in figure 3.6a will be used to explain the interpretation of the adapted triplot first. Having only two variables allows for the simultaneous indication of the validation data set and an explanation of the graph's interpretation. For all the covariates that will be discussed, one validation observation is indicated on all the graphs.

The validation sample is plotted in the classification area and the classes can be distinguished from each other by their shapes. The choice of k equal to 41 also allowed for sufficient separation of the classes within the triplot.

Table 3.4: *Classification measures of the various methods compared to the traditional KNN. The scores with a (*) are the ones that perform within 3% of the traditional KNN. The parameter k was chosen as 41 for all methods.*

Method with underlying KNN (k=41)	**Outer Polybag**	**Axes Scoring**	**Inner Polybag**	**Actual / Approx. points**	**Total accuracy (+)**	**Total misclas. (-)**	**Precision (+)**	**False discovery rate (-)**	**False omission rate (-)**	**Neg. predictive value (+)**	**True positive rate (+)**	**False negative rate (-)**	**False positive rate (-)**	**True negative rate (+)**
Black-box	NA	NA	NA	NA	0.60	0.40	0.57	0.43	0.20	0.80	0.56	0.44	0.22	0.78
Triplot	2 × 95%	NA	0%	NA	0.58*	0.42*	0.54*	0.46*	0.22*	0.78*	0.54*	0.46*	0.23*	0.77*
A-Triplot	2 × 95%	Ratio	0%	Actual	0.56	0.44	0.53	0.47	0.23*	0.77*	0.51	0.49	0.24*	0.76*
				Approx.	0.57*	0.43*	0.54*	0.46*	0.23*	0.77*	0.53*	0.47*	0.24*	0.76*
			75%	Actual	0.57*	0.43*	0.54*	0.46*	0.23*	0.77*	0.54*	0.46*	0.23*	0.77*
				Approx.	0.57*	0.43*	0.55*	0.45*	0.23*	0.77*	0.53*	0.47*	0.23*	0.77*
		Count	0%	Actual	0.55	0.45	0.54*	0.46*	0.21*	0.79*	0.49	0.51	0.26	0.74
				Approx.	0.57*	0.43*	0.54*	0.46*	0.23	0.77	0.53*	0.47*	0.24*	0.76*
			75%	Actual	0.55	0.45	0.52	0.48	0.24	0.76	0.51	0.49	0.25*	0.75*
				Approx.	0.56	0.44	0.54*	0.46*	0.24	0.76	0.52	0.48	0.24*	0.76*
	8 × 95%	Ratio	0%	Actual	0.53	0.47	0.51	0.49	0.25	0.75	0.50	0.50	0.25*	0.75*
				Approx.	0.55	0.45	0.56*	0.44*	0.24	0.76	0.53*	0.47*	0.24*	0.76*
			75%	Actual	0.53	0.47	0.53	0.47	0.25	0.75	0.51	0.49	0.24*	0.76*
				Approx.	0.53	0.47	0.55*	0.45*	0.25	0.75	0.51	0.49	0.25*	0.75*
		Count	0%	Actual	0.53	0.47	0.51	0.49	0.25	0.75	0.50	0.50	0.25*	0.75*
				Approx.	0.55	0.45	0.56*	0.44*	0.24	0.76	0.53*	0.47*	0.24*	0.76*
			75%	Actual	0.53	0.47	0.53	0.47	0.25	0.75	0.51	0.49	0.24*	0.76*
				Approx.	0.53	0.47	0.55*	0.45*	0.25	0.75	0.51	0.49	0.25*	0.75*

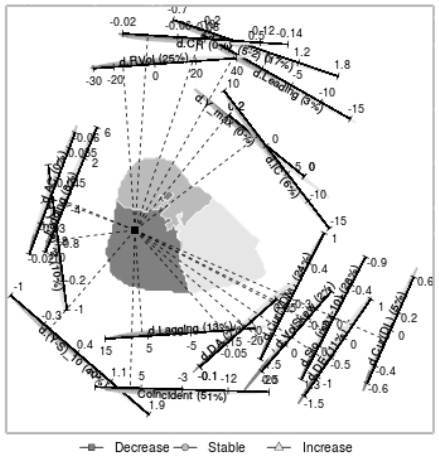
Classifying yield spread movements through triplots: a South African application

Info | Data | Axes Scoring | Base triplot

This Shiny web-based application serves as supplementary data to the paper of Van der Merwe and de Wet on Classifying yield spread movements through triplots: a South African application.

This sidebar panel allows the user to change the values of the specific observation that is analysed under the DATA tab. The type of scoring methods used can be changed in the AXES SCORING tab, and the underlying properties of the base triplot can be changed in the BASE TRIPLOT tab.

Within the MAIN CALCULATION tab in the main panel the user can either view the actual and predicted values of the specific observation together with their respective scores under VARIABLES FINAL SCORE. The VARIABLES RAW SCORE and AXES PREDICTIVENESS tabs provide the values that are used to calculate the final scores. Finally the triplots figures with the various groups of axes can be seen in the INDIVIDUAL TRIPLOTS tab in the main panel.



Main calculation | Individual triplots

Variables final score | Variables raw score | Axes predictiveness

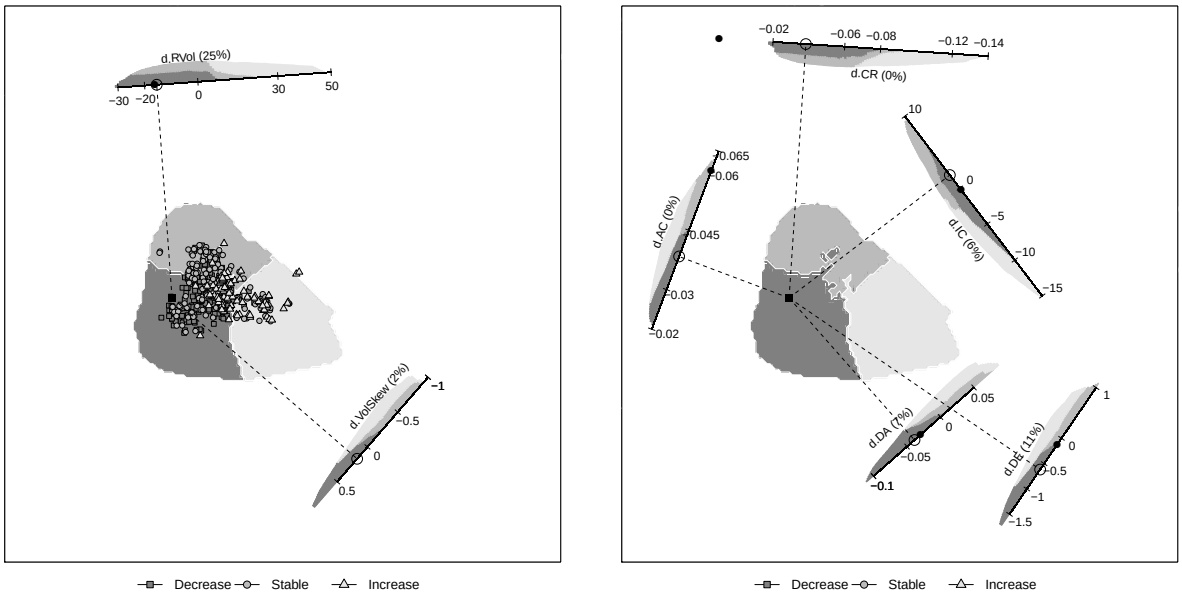
ACTUAL DATA

	Value	Decrease	Stable	Increase
d.VolSkew	2.54	0.00	0.00	0.00
d.RVol	-16.18	0.15	0.08	0.00
d.AC	-0.06	0.00	0.00	0.00
d.CR	0.01	0.00	0.00	0.00
d.DA	-0.03	0.04	0.00	0.01
d.DE	-0.13	0.01	0.03	0.06
d.IC	-0.23	0.03	0.02	0.00
(d.Y_10) ^2	0.17	0.05	0.00	0.05
d.(Y-S)_10	-0.17	0.24	0.13	0.00
d.Y_max	-0.40	NA	NA	NA
d.Slo_(5-2)	0.43	0.07	0.05	0.03
d.Slo_(max-10)	0.02	0.02	0.10	0.17
d.Cur(DL)	0.95	0.00	0.00	0.00
d.Cur(VDM)	0.28	0.01	0.11	0.10
d.Leading	9.35	0.00	0.01	0.00
d.Lagging	16.95	0.00	0.00	0.00
Coincident	6.59	0.30	0.21	0.00
Lagging	-3.20	0.04	0.00	0.04
==TOTAL==	**NA**	**0.95**	**0.74**	**0.46**

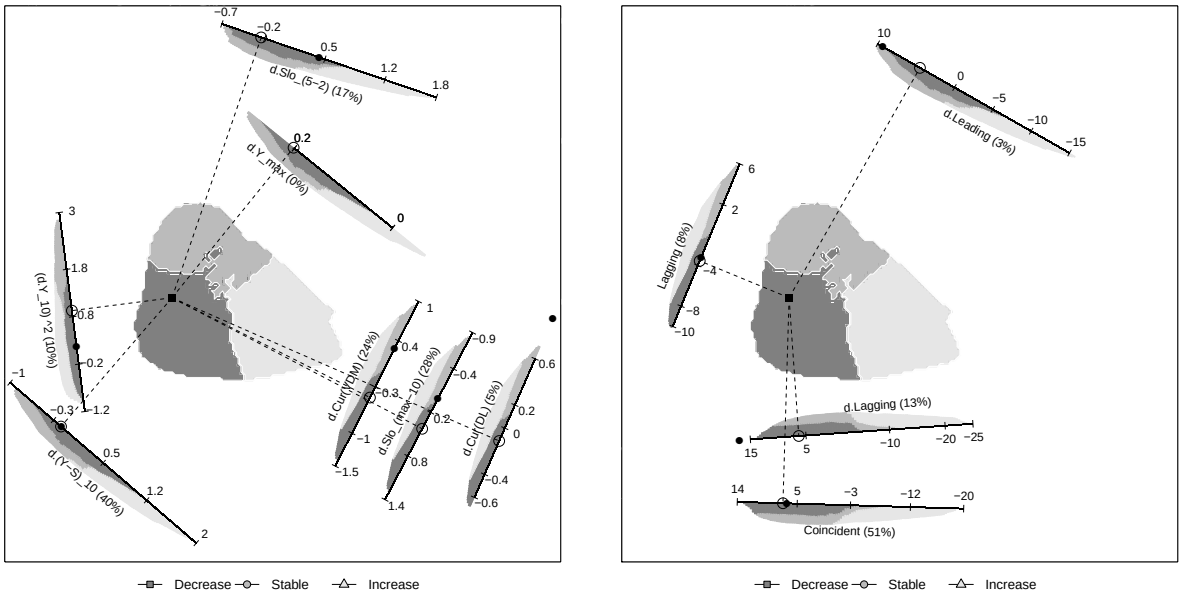
PREDICTED DATA

	Value	Decrease	Stable	Increase
d.VolSkew	0.17	0.01	0.00	0.00
d.RVol	-15.46	0.15	0.08	0.00
d.AC	-0.04	0.00	0.00	0.00
d.CR	-0.04	0.00	0.00	0.00
d.DA	-0.04	0.05	0.00	0.00
d.DE	-0.62	0.05	0.00	0.03
d.IC	1.82	0.04	0.02	0.00
(d.Y_10) ^2	0.93	0.04	0.01	0.04
d.(Y-S)_10	-0.18	0.23	0.14	0.00
d.Y_max	0.20	0.00	0.00	0.00
d.Slo_(5-2)	-0.24	0.09	0.07	0.00
d.Slo_(max-10)	0.42	0.12	0.00	0.12
d.Cur(DL)	-0.11	0.02	0.00	0.02
d.Cur(VDM)	-0.46	0.10	0.00	0.10
d.Leading	4.50	0.02	0.01	0.00
d.Lagging	6.30	0.08	0.04	0.00
Coincident	7.14	0.30	0.19	0.00
Lagging	-3.55	0.04	0.00	0.04
==TOTAL==	**NA**	**1.34**	**0.56**	**0.35**

Figure 3.5: *A screenshot of the Shiny application that was created to illustrate the techniques presented in this paper. The user can change the values of the observation on the left-hand side under 'Data', the methods used in the axes scoring, and settings for the base triplot. The results on the right-hand side includes the scores calculated for the observation (including the raw scores and predictivity used in the calculation) as well as the four groups of adaptive triplots for the group of covariates as shown in this paper. The link and code for the Shiny web-based application can be found at* `https://doi.org/10.5281/zenodo.3565978` *(Van der Merwe, 2019a). Note that this is not a direct link to the application, but rather to the GitHub repository for the code where the link should be available at the top of the page.*



(a) *Share type covariates* ***(b)*** *Financial ratio type covariates*



(c) *Interest rate and liquidity type covariates* ***(d)*** *Economic type covariates*

Figure 3.6: *Adapted KNN (k = 41) triplots with 0% inner- and 2 × 95% outer-polybags, together with C-type scoring axes. Both the actual (black circle) and predicted (empty circle) values for a single observation from the validation sample is shown on the various axes on each graph.*

Next, the two covariates' axes are positioned exterior to the classification area. The first covariate is the change in the underlying company's share price volatility (*d.RVol*), and the second covariate is the change in the volatility skew (*d.VolSkew*). The predictivity of the axes is indicated in brackets next to their names. The change in the volatility skew's axis does not have a high predictivity on the triplot. The change in the underlying share's volatility axis on the other hand has a much higher accuracy. It is very important to note that the number of dimensions of the data is reduced from 18 to two, therefore it is expected that not all variables will be displayed equally well. An example of this would be to observe a three-dimensional plot, with three perpendicular axes, from above - changes in values of the vertical axes will not be observed in this view while changes in the others will be highly visible.

The black square in the classification area indicates a single observation from the validation set. The true value of the observation was a decrease in spread. The actual and predicted values of the covariates of this observation are indicated on the axes. The empty circles are the predicted values. They are the values that are read from the axes if a perpendicular line is drawn from the observation in the triplot - these will almost always (depending on the size of the outer-polybag) fall somewhere on the displayed axes.

The black circles indicate the actual covariate values of the observation. From *d.RVol*, it is clear that the actual value is close to the predicted value, while the actual value for *d.VolSkew* is too far away from the axes to be indicated on the graph (in fact, the value is -16, which is very far outside the range, approximately $[-1, 1]$, of the axis).

Thus, for the change in the underlying share's volatility, the actual and predicted values will have a very similar score allocated to them. In contrast, the actual value for the change in volatility skew will have no score allocated to it, while the predicted value will have a mixture between a decrease and increase score allocated to it. These, however, will be multiplied with a small predictivity value, and therefore will add little towards the final total score.

Next, all four groups of covariates are discussed in terms of what can be observed regarding the interaction of the covariates with the classes.

Share type covariates

Figure 3.6a shows a strong relationship between the increase in the underlying share's volatility (*d.RVol*) and the increase in spreads. However, a decrease in the volatility could signal either a decrease in spread or have no effect.

As already mentioned in the previous section, the change in volatility skew (*d.VolSkew*) does not have a high predictivity in the graph. It shows that an increase in the covariate results in a decrease in spread, and a decrease in the covariate results in either an upward or stable movement of spreads.

From this it can be inferred that an increase in the underlying share's volatility has a stronger negative effect on spreads than the potential positive effect of a decrease.

Financial ratio type covariates

For the financial ratio type covariates shown in figure 3.6b, it can be seen that change in debt over equity (*d.DE*) and debt over assets (*d.DA*) have a similar type of profile. It shows that a decrease in spreads and decrease in variable are related, but the increase of the variable is split between the increase of spreads and stable states.

A decrease in the interest cover ratio (*d.IC*) increases the spreads, while a decrease in this covariate mostly signals a stable state.

The asset over capital (*d.AC*) and current ratio (*d.CR*) has almost no predictivity on the triplot, and therefore will not be discussed.

Given the above, it can be concluded that the weakening of the interest cover ratio of a firm has a stronger negative effect on spreads than an increase would impact positively. Furthermore, the decrease of debt to assets and debt to equity has a stronger positive effect on spreads than the converse. It is further noted that most of these ratios did not have a very high predictivity on the triplot.

Interest rate and Liquidity type covariates

For the interest rate type covariates, shown in figure 3.6c, seven variables were included, one of which was the quadratic movement in the change in the 10-year yield. As this term was used to account for non-linear movements it will not be discussed in more detail. Additionally, the change in the maximum available yield term was also included, but had a predictivity of 0% and will also not be discussed in more detail.

The change in curvatures, *d.Cur(DL)* and *d.Cur(VDM)* have the same profile. It shows that a decrease in the curvature results in either a decrease or increase in the spreads. On the other hand, an increase in the curvature either results in an increase in spreads or the stable state. This indicates that the spread is not as sensitive to small changes in curvature as it would be to larger changes, in either direction.

Two variables are included for the change in slope. The first is the slope between the maximum available point and the 10-year point (*d.Slo(max-10)*), the second is the slope between the five- and two-year points (*d.Slo(5-2)*), which can be interpreted as the long- and medium-term slopes, respectively. An increase in the medium-term slope results in an increase in spreads, while a decrease in slope is shared between a decrease in spread and the stable state. An increase in the long-term slope results in either a decrease or increase in spreads, while a decrease in the long-term slope results in either an increase in spreads or the stable state. These distinct effects on the change in slopes support the initial hypothesis that the change in slope can have various impacts on the change in spreads, but it can also therefore be seen that changes in medium- and long-term slopes have divergent impacts on spread changes.

Finally, the change in the difference between the 10-year yield and swap rates (*d.(Y-S)_10*) is considered. As this difference widens, the spread increases; if it narrows, the spread either stays stable or decreases. This indicates that the widening of the difference between the yield and swap rates have a stronger negative effect on the spreads than the positive effect of a narrowing.

Economic type covariates

Finally, the economic type covariates are discussed. Two of which have the same profile and have relatively high predictivity percentages. The level of the lagging index indicates that a lower level signals either a decrease or increase in spreads, while a higher level signals either an increase or stable spreads.

The profile for the other covariates shows that their decrease results in an increase of spreads, while an increase in the covariates either results in a decrease of spreads or a stable state. This implies that negative changes in economic indicators have a stronger negative effect on spreads than the corresponding positive effects of positive changes in economic indicators.

5.3 Comparison to other literature

Collin-Dufresn et al. (2001, Table X) found that increases in volatility smirk, difference between swap and yield curves, and slopes (10-year minus two-year yields) resulted in increases in yield spreads. If *d.Slo(5-2)* is chosen as a proxy for the slope, and *d(Y-S)_10* as the inverse of the difference between the swap and yield curve, then similar results were obtained for *d.Slo(5-2)*, but different effects for *d.VolSkew* and *d.(Y-S)_10*. This could be due to the low predictivity of *d.VolSkew* on the triplots, and because there are two possible interpretations of *d.(Y-S)_10* as discussed earlier.

The results presented in Avramov et al. (2007, Table 6) show that increases in the long-term slope (30-year minus 10-year yields), and expansionary economic cycles resulted in decreases in spreads. This is similar to the results obtained for *d.Slo(max-10)* and *Coincident*.

Chen et al. (2007, Table VI, Columns 4 and 10) found that decreases in inequity volatility and term slope (10-year minus two-year yields) resulted in increases in yield spreads. No significant coefficients were found for pre-tax interest coverage and long-term debt to assets. The effects of the change in term slope are similar to that found in this research. The difference in *d.RVol* may be due to the authors finding inconsistent statistical evidence regarding equity volatility.

The effects of the change in equity volatility and short-term slope (10-year minus two-year yields) that was reported in Radier et al. (2016, Table 4, Panel A) correspond to the effects reported in this research.

6 Discussion and conclusion

In this paper a visual interpretation and classification methodology was proposed to determine whether an increase, decrease, or no change in spreads have occurred for unlisted debt instruments given observed market conditions. This was done through a new visually interpretable adapted KNN triplot which not only allows for a new classification methodology, but also for interpretation of the sensitivities of the various covariates. The adapted KNN triplot also allowed for new insights into the determinants of spread changes not considered in previous research.

This method can be applied to various other classification problems where visual interpretation is an important aspect and traditional black-box techniques are not sufficient to explain why certain classifications occur.

An interesting finding with regards to the covariates was noted, where it was seen that by incorporating a 'stable' class, some movements in the covariates have stronger negative or positive impact with regards to spread movements than the alternate direction of change.

While this methodology was applied to South African data, it has international application, offering ample future research potential.

Other areas for further research include the automatic parallel shifting of axes (currently these need to be adjusted manually), and the expansion of the research to include more classes.

A web-based application was built supplementary to this paper to provide additional clarity as to how the various properties and inputs affect the analysis on the adapted triplot. The link and code for the *Shiny* web-based application that was built in R can be found at `https://doi.org/10.5281/zenodo.3565978` (Van der Merwe, 2019a). A screenshot is provided in figure 3.5.

REFERENCES

Avramov, D., Jostova, G., Philipov, A., 2007. Understanding changes in corporate credit spreads. Financial Analysts Journal 63, 90–105. URL: http://doi.org/10.2469/faj.v63.n2.4525.

Bao, J., Pan, J., Wang, J., 2011. The illiquidity of corporate bonds. The Journal of Finance 66, 911–946. URL: http://doi.org/10.1111/j.1540-6261.2011.01655.x.

Chen, L., Lesmond, D.A., Wei, J., 2007. Corporate yield spreads and bond liquidity. The Journal of Finance 62, 119–149. URL: http://doi.org/10.1111/j.1540-6261.2007.01203.x.

Collin-Dufresn, P., Goldstein, R.S., Martin, J.S., 2001. The determinants of credit spread changes. The Journal of Finance 56, 2177–2207. URL: http://doi.org/10.1111/0022-1082.00402.

Diebold, F.X., Li, C., 2006. Forecasting the term structure of government bond yields. Journal of econometrics 130, 337–364. URL: https://doi.org/10.1016/j.jeconom.2005.03.005.

Friedman, J., Hastie, T., Tibshirani, R., 2010. Regularization paths for generalized linear models via coordinate descent. Journal of statistical software 33, 1. URL: http://doi.org/10.18637/jss.v033.i01.

Gardner-Lubbe, S., 2016. A triplot for multiclass classification visualisation. Computational Statistics & Data Analysis 94, 20–32. URL: https://doi.org/10.1016/j.csda.2015.07.014.

Gardner-Lubbe, S., Le Roux, N.J., Gowers, J.C., 2008. Measures of fit in principal component and canonical variate analyses. Journal of Applied Statistics 35, 947–965. URL: http://doi.org/10.1080/02664760802185399.

Gower, J.C., Gardner-Lubbe, S., Le Roux, N.J., 2011. Understanding biplots. John Wiley & Sons.

Greenacre, M., 2010. Biplots in Practice. Number 2011113 in Books, Fundacion BBVA / BBVA Foundation. URL: https://ideas.repec.org/b/fbb/booklb/2011113.html.

Guyon, I., Elisseeff, A., 2003. An introduction to variable and feature selection. Journal of machine learning research 3, 1157–1182.

James, G., Witten, D., Hastie, T., Tibshirani, R., 2013. An introduction to statistical learning. volume 112. Springer.

Mardia, K.V., 1970. Measures of multivariate skewness and kurtosis with applications. Biometrika 57, 519–530. URL: https://doi.org/10.1093/biomet/57.3.519.

Morozova, O., Levina, O., Uusküla, A., Heimer, R., 2015. Comparison of subset selection methods in linear regression in the context of health-related quality of life and substance abuse in russia. BMC medical research methodology 15, 71. URL: http://doi.org/10.1186/s12874-015-0066-2.

Radier, G., Majoni, A., Njanike, K., Kwaramba, M., 2016. Determinants of bond yield spread changes in south africa. African Review of Economics and Finance 8, 50–81. URL: https://www.ajol.info/index.php/aref/article/view/162156.

Shapiro, S.S., Wilk, M.B., 1965. An analysis of variance test for normality (complete samples). Biometrika 52, 591–611. URL: https://doi.org/10.2307/2333709.

South African Reserve Bank, 2015. Revisions to the composite leading and coincident business cycle indicators. Quarterly Bulletin No. 276, June 2015. URL: https://www.resbank.co.za/Lists/News%20and%20Publications/Attachments/6776/01Full%20Quarterly%20Bulletin%20%E2%80%93%20June%202015.pdf.

Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B (Methodological) 58, 267–288. URL: http://doi.org/10.1111/j.2517-6161.1996.tb02080.x.

Van der Merwe, C.J., 2019a. carelvdmerwe/triplotsimulation. URL: https://doi.org/10.5281/zenodo.3562013.

Van der Merwe, C.J., 2019b. Triplot classification with polybags. Submitted.

Van der Merwe, C.J., Heyman, D., de Wet, T., 2018. Approximating risk-free curves in sparse data environments. Finance Research Letters 26, 112–118. URL: http://doi.org/10.1016/j.frl.2017.12.016.

Zou, H., 2006. The adaptive lasso and its oracle properties. Journal of the American statistical association 101, 1418–1429. URL: http://doi.org/10.1198/016214506000000735.

APPENDICES

A The South African composite business cycle indicator's components

Below follow the various components of the South African composite business cycle indicator. They were obtained from the South African Reserve Bank on 9 April 2019.

A.1 Components of the Composite Leading Business Cycle Indicator:

- Net balance of manufacturers observing an increase in the average number of hours worked per factory worker: Bureau of Economic Research (half weight)
- Job advertisement space in the *Sunday Times* newspaper: Percentage change over 12 months
- Net balance of manufacturers observing an increase in the volume of orders received: Bureau of Economic Research (half weight)
- Opinion survey of business confidence: Bureau of Economic Research
- Number of residential building plans passed: Flats, townhouses and houses larger than 80m^2
- Number of new passenger vehicles sold: Percentage change over 12 months
- Gross operating surplus as a percentage of gross domestic product
- Interest rate spread: 10-year government bonds minus 91-day Treasury bills
- Real M1 money supply: six-month smoothed growth rate
- Index of commodity prices in US dollar for a basket of South Africa's export commodities
- Composite leading business cycle indicator of South Africa's major trading-partner countries: Percentage change over 12 months

A.2 Components of the Composite Coincident Business Cycle Indicator (Equal weights):

- Gross value added at constant prices, excluding agriculture, forestry and fishing
- Employment in the total formal non-agricultural sector
- Value of retail and new vehicle sales at constant prices
- Industrial production index (comprising the physical volumes of mining, manufacturing and electricity production)
- Utilisation of production capacity in manufacturing

A.3 Components of the Composite Lagging Business Cycle Indicator (Equal weights):

- Value of non-residential buildings completed at constant prices
- Ratio of gross fixed capital formation in machinery and equipment to final consumption expenditure on goods by households
- Ratio of inventories to sales in manufacturing and trade
- Nominal labour cost per unit of production in the manufacturing sector: Percentage change over 12 months
- Predominant prime overdraft rate of banks
- Ratio of consumer instalment sale credit to disposable income of households