CLUSTER ANALYSIS AND FIRM PATTERNS OF EARNINGS PERSISTENCE: A NEW APPROACH

A thesis presented in partial fulfilment of the requirements for the degree of

Doctor of Philosophy

in

Finance

At Massey University, Manawatū, New Zealand

Son Duong Tran

2019

Table of Contents

LIST OF FIGURES

LIST OF TABLES

ABBREVIATIONS

ΔATO	Change in Asset Turnover
ΔPM	Change in Profit Margin
AB_ACC_DEF	Deflated Absolute Value of Accruals
ABS_EARNINGS_DEF	Deflated Absolute Value of Earnings
ACC_DEF	Deflated Accruals
AGE	Firm Age
ATO	Asset Turnover
C_CON	Conditional Conservatism
CA	Cluster Analysis
CAPX_DEF	Deflated Capital Expenditures
CDA	Canonical Discriminant Analysis
CK	Conditional K-means
CR	Current Ratio
DDM	Dividend Discount Model
DIV	Dividend Payout
EM_DN	Downward Earnings Managed
EM_UP	Upward Earnings Managed
FA	Factor Analysis
FE	Forecast error
FLEV	Financial Leverage
HGSC	Heterogeneous-Group Specific Coefficients
iMWK	Intelligent Minkowski Metric Weighted K-Means
INTAN_INT_DEF	Deflated Intangible Investment
MAR	Mean of Absolute Residuals
MBIC	Modified Bayesian Information Criterion
MEAN_AFE	Mean Absolute Forecast Errors
MED_AFE	Median Absolute Forecast Errors
MSR	Mean of Squared Residuals
NBC	Net Borrowing Cost
NO_EM	Without Earnings Managed
OLLEV	Operating Leverage
PCA	Principal Component Analysis
PM	Profit Margin
RIV	Residual Income Valuation
ROWK	Regression-Oriented-Weighted-K_means clustering
SALE_GR	Sales Growth
SIZE	Firm Size
SSD	Sum of Squared Distance
SSR	Sum of Squared Residuals
SYNCLUS	Synthesized Clustering
UN_CON	Unconditional conservatism
VOL_IBC_DEF	Earnings Volatility
WK	Weighted K-Means

ABSTRACT

The development of a method to appropriately address the problem of heterogeneous-group specific coefficients (HGSC) is of paramount importance for any studies where there are concerns of HGSC. Accordingly, the goal of this thesis is to investigate a solution to the prevalent problem of HGSC within the context of the finance discipline. Specifically, this thesis introduces a novel clustering procedure called regression oriented-weighted K-means clustering (ROWK). This new method employs the regression mean absolute residuals (MAR) to inform the cluster analysis identification of optimal feature weights. The performance of ROWK clustering is examined via both simulated and real data. Simulation results show significant improvements from the adoption of ROWK relative to K-means clustering and weighted K-means clustering through three channels. Specifically, through the examination of three case studies, this thesis finds that ROWK places more (less) weight on more (less) relevant features; reduces the influence of multicollinearity by reducing the weights of irrelevant features which are highly correlated with relevant features; and captures relevance not only by its contribution to cluster recognition but also by regression estimation. The thesis further examines the performance of ROWK clustering using real data for earnings persistence models. ROWK outperforms other standard techniques in the sense of correctly identifying the underlying clusters on earnings persistence models. The thesis also documents that analysts' forecasts only partially incorporate the information from cluster patterns in the short run, while ignoring impacts of these patterns on long-term future earnings. As a result, conditioning on such information allows the identification of reliable and economically important patterns in analyst forecast errors.

Keywords: Cluster analysis, K-means, feature weightings, group-specific coefficients, firm patterns, earnings persistence

ACKNOWLEDGEMENTS

First and foremost, I would like to express my sincere acknowledgement to Dr J.G. Chen and Dr Carolyn Wirth for their exceptional guidance throughout this research. I am very impressed with their comprehensible instructions and constructive criticisms. I also gratefully appreciate their unhesitating support, inspiration, and patience. Working with them has been a great opportunity to build up my abilities and enthusiasm for research. A great thanks to J.G. for his help regarding methodology and nice discussions about our lives and families. I am also extremely impressed with Carolyn's excellent skills in writing and time management. I remember how kindly they treated me and the beautiful gifts they gave to my son on his full moon celebration. Again, a genuine thanks from me to them for being great mentors and role models.

I want to express my sincere gratitude to Massey University for providing me with all necessary support during my candidature. I could not have undertaken this research if I had not being awarded a Massey University Doctoral Scholarship. I also thank Dr J.G. Chen, Dr Carolyn Wirth and Mrs Ha Lien Ton for their encouragement that raised my confidence to apply for the scholarship. Likewise, I thankfully acknowledge the kind support and resources given to me by the University of Economics Ho Chi Minh (UEH) and the School of Management, UEH throughout my doctoral program.

Thanks are also due to Ms Fong Mee Chin (Research Officer, School of Economics and Finance-SEF) for her support, advice and pleasant talks. I am very impressed with her expertise in database management and her prompt responses to my queries. I also express my gratitude to Cameron Rhodes and Wood Marks (IT Research Analyst, SEF) for their kind support on IT issues. I also thank Maryke Bublitz (Executive Assistant, SEF) and Sue Edwards (Financial Administrator, SEF) for their assistance regarding my trip to a conference

in Australia. I also express my gratitude to Associate Professor Jeff Wongchoti, Andrea Bennett, Professor Ben Marshall, Associate Professor Candie Chang, Dr Chris Malone and Dr Mui Kuen Yuen for their helpful advice at varying stages of this research.

For their comments and advice, I thank participants at the Massey PhD Symposium (2016, 2017), the 4th Applied Financial Modelling Conference, Melbourne, Deakin University (Feb, 2018), New Zealand Finance Colloquium, Massey University (2018), the 2nd Asia Conference on Business and Economic Studies, Ho Chi Minh City, Vietnam (2019), and several seminars and workshops held by SEF and Massey University.

Special thanks go to Associate Prof Candie Chang, Associate Professor Jeff Wongchoti, Dr M.Humayun Kabir, Associate Professor Thomas Scott and Associate Professor Monique Wan for their useful comments and suggestions in my confirmation and final examination defence. I also thank Andrea Bennett for her excellent comments on my writing. Many thanks are also due to Prof Nhut (Nick) Hoang Nguyen and Prof Nuttawat Visaltanachoti for their constructive comments on application issues for my thesis. I also thank Nader Mahmoudi, the University of Newcastle and Dr Praveen Bhagawan, University of Connecticut, USA for their thorough reviews of my research.

I also sincerely acknowledge the support from my fellow research students in the School. Especially, I would like to thank Lu Wang, Feng Xie, Khairul Zharif Zaharudin, Sulu Manu O'Uiha, Xiping Li, Andrés Camacho, Rui Ma, Ahmad Raza, Dang Thanh Ngo, Xiaoqi Chen, Andryan Setyadharma, Qing Wan, Bilal Hafeez and Suresh Kumar Oad Rajput. I also wish to give a special thanks to Yi Wei for her kind-hearted support.

Finally yet importantly, I want to say a wholehearted thanks to my family for their unconditional love. Thanks to you mum and dad for giving me life, walking along with me,

listening to me, teaching me, sympathizing with me, and giving me your endless love. I am so proud of you. Likewise, I thank my mother-in-law for her dedicated support when my wife had our baby. I also thank my sister for encouraging me during my difficult period. Finally, as ever, I really want to express my tremendous gratitude to my wife (Nhung Pham, Jessica) and my little son (Dang Khoa- Jack). They are my fountain of life and my motivation that has brought me this far in my life.

CHAPTER 1

INTRODUCTION

1.1 INTRODUCTION

Many researchers argue that general econometric models of business and economic processes may not be good representations to depict/forecast actual processes, and that appropriate partitioning of data is necessary to improve model performance (Ou & Penman, 1989; Nissim & Penman, 2001; Jansen, Ramnath, & Yohn, 2012). Put another way, heterogeneous-group specific coefficients (hereafter HGSC) within the econometric models cause instability of the examined relationships between model variables over time and/or across different samples (e.g. different countries or states, different firm groups, in-sample vs. out-of-sample) and poor performance in out-of-sample predictive power.

Indeed, it is not difficult to find evidence against the assumption of parameter homogeneity. It can be found in corporate governance studies where the relationship between a firm's governance and its value depends on whether or not the firm operates in a competitive industry (Giroud & Mueller, 2011). It is also documented in studies of earnings forecasts where earnings persistence coefficients differ on earnings volatility, accruals, firm life cycle or business strategy (e.g. Dichev & Tang, 2009; Dickinson, 2011; Little, Little, & Coffee, 2009; Sloan, 1996). While knowledge of underlying sources breaching the constant-coefficients assumption is well-addressed in existing research, solutions developed in order to mitigate this problem are still restricted to including these sources into predictive regressions, ad-hoc partitioning techniques or in the extreme case, running individual time-series analyses (e.g. Dichev & Tang, 2009; Wang, 2013).

For these reasons, the development of a method to appropriately address the problem of HGSC is of paramount importance for any studies where there are concerns of HGSC.

Such issues are particularly problematic in econometric modelling within the finance discipline (Richardson, Tuna, & Wysocki, 2010). Accordingly, the goal of this thesis is to investigate a solution to the prevalent problem of HGSC within the context of the finance discipline. Identifying the gaps exposed from contemporaneous literature and practices to address the HGSC problem, we proposes a novel solution called Regression-Oriented-Weighted-K_means clustering (hereafter ROWK). It combines K-means clustering and regression analysis. K-means clustering is the most popular method in cluster analysis (hereafter CA), a common technique firstly introduced and developed in the natural sciences from the need to classify data into homogeneous objects. On the one hand, the K-means algorithm iteratively assigns similar observations into clusters. On the other hand, the mean of absolute (squared) residuals (thereafter MAR/MSR) from running regressions is used to guide the process of weight adjustment in clustering and mitigate the problem of multicollinearity. Feature weighting and multicollinearity are two important issues challenging the performance of CA to address the HGSC problem.

The performance of ROWK clustering is examined via both simulated and real data. Simulated data is generated to comprehensively examine different channels through which ROWK improves the performance of CA in dealing with the HSGC problem. The results show significant improvements from the adoption of ROWK relative to K-means clustering and weighted K-means clustering through three channels. Specifically, through the examination of three case studies, this thesis finds that ROWK places more (less) weight on more (less) relevant features; reduces the influence of multicollinearity by reducing the weights of irrelevant features which are highly correlated with relevant features; and captures relevance not only by its contribution to cluster recognition but also by regression estimation.

The thesis further examines the performance of ROWK clustering using real data. Specifically, earnings persistence models are chosen as a potential application of ROWK clustering. Earnings predictability and earnings persistence play a critical role in equity valuation, financial statement analysis, risk management and asset pricing. Several studies document factors that cause persistence and predictability of earnings to vary

across firms, including earnings volatility, accruals, and stages of firm life cycle (e.g. Dichev & Tang, 2009; Dickinson, 2011; Sloan, 1996). These studies provide compelling reasons for the choice of earnings persistence models as a potential candidate for the application of ROWK clustering as introduced in this thesis.

The thesis implements ROWK clustering for the earnings persistence model with 17 examined features. The results show that only five features are relevant for clustering. Examining the earnings persistence patterns across ROWK clusters also reveals differences on coefficients of earnings persistence and intercepts between ROWK clusters. These differences are relatively large in magnitude and suggest that cluster membership is economically important. Moreover, ROWK clustering results in larger differences in earnings persistence between clusters than a single variable cluster partitioning technique. The thesis also documents that analysts' forecasts only partially incorporate the information from cluster patterns in the short run, while ignoring impacts of these patterns on long-term future earnings. As a result, conditioning on such information allows the identification of reliable and economically important patterns in analyst forecast errors.

The remainder of this chapter is organised as follows. Section 1.2 explains the background and research motivation for the study while the specific thesis objective and the related thesis aims are presented in Section 1.3. Section 1.4 highlights the academic contributions of the study. The chapter ends with a description of how the thesis is organised in Section 1.5.

1.2 BACKGROUND AND RESEARCH MOTIVATION

When conducting regression analysis to forecast and/or test hypotheses, researchers usually make some assumptions, for example the Gauss-Markov assumption. If this assumption holds, estimators should be unbiased and efficient. In addition, if proposed relationships are developed through well-established theories, then one expects to observe highly significant coefficients and considerable R-squared values. It is also

expected that reasonable out-of-sample predictive power and stable coefficients will be observed when the regression is run across periods or firm groups.

Yet, low out-of-sample predictive power remains a phenomenon that challenges researchers. A lack of sound theories in constructing empirical tests could be a potential reason, and some evidence supports this view (e.g. Ou & Penman, 1989). Yet even when the research is better-backed by theory, there are still instances where the results lack robustness (e.g. Nissim & Penman, 2001).

A sound explanation for this issue is the dynamic of financial models. For example, even the weak form of the efficient market hypothesis asserts that future stock prices cannot be predicted by movements of past price. The important source underlying this dynamic feature of financial models is the violation of the constant-coefficients assumption. Put another way, "*a general model is not a good representation for all firms (to the extent to which different characteristics generate future earnings in different firms in different ways)"* (Ou & Penman, 1989, p. 299).

Evidence of the violation of the constant-coefficients assumption is found in corporate governance studies where the relationship between a firm's governance and its value depends on whether the firm operates in non-competitive or competitive industries (Giroud & Mueller ,2011). This violation is also documented in studies of earnings forecasts where earnings persistence coefficients differ according to earnings volatility (e.g. Dichev & Tang, 2009), accruals (Sloan, 1996), firm life cycle (e.g. Dickinson, 2011), and business strategy (e.g. Little, et al., 2009). More importantly, the tendency for firms to cluster together (both theoretically and empirically) is of considerable importance in regression analysis. A well-known circumstance of parameter variation in regression models is in time series applications where firms' behaviours exhibit changes between periods. These switching regimes may relate to different states of the business cycle or other more fundamental structural changes (Goldfeld & Quandt, 1973).

As outlined above, the underlying sources of the constant-coefficients assumption violation have been well-identified in existing research. However, approaches to partitioning data that financial researchers employ to address the problem of HGSC have tended to be simplistic, such as the division of a whole sample into different quantiles of certain firm features at which different relationships are expected to be observed. For example, the level of accruals is added into earnings predictive models or firms are assigned into quintiles of earnings volatilities (e.g. Dichev & Tang, 2009; Wang, 2013). These approaches are ad-hoc and cannot be assured to reveal the underlying data patterns. A limit in the number of factors used in the partitioning process is another serious flaw of these approaches.

A common business research practice uses industry classification to partition firms. In particular, model parameters are often estimated using cross-sectional pooled regressions within two-digit SIC codes (Jansen et al., 2012). However, this practice generally leads to imprecise estimates due to the different relations between the dependent variable and its determinants within each industry (Fairfield, Whisenant, & Yohn, 2003). Furthermore, Bernard & Skinner (1996) find that the estimates are even less precise in the case of time-series estimates.

Since the 1970s, there has been increasing attention to the field of econometrics that delves deeper into the HGSC issue (e.g. Goldfeld & Quandt, 1973; Lin & Ng, 2012). Several econometric approaches have been introduced based on sound mathematical foundations that are supported by empirical evidence using either simulated data or real data. Nevertheless, there remain potential drawbacks that need to be addressed. These approaches only focus on the case of two regimes. Yet, when there are more than two regimes, it becomes complicated to construct the likelihood function and to find solutions of optimization. Furthermore, in reality there could be more than one partitioning variable that influences the choice of regimes. A threshold that is estimated as a linear function of several partitioning variables is commonly used to resolve this issue (Lin & Ng, 2012). However, a linear function of several partition variables might not be a good proxy for

the threshold. Two observations with the same score of the function could be very different or far in distance from each other if they are presented as points in space.

To address those concerns, a partitioning technique is needed that is able to (1) incorporate several variables, (2) identity the appropriate numbers of groups to assign; and (3) achieve the highest homogeneities of members within groups. However, existing studies of earnings forecasts pay little attention to the development of such techniques (Richardson et al., 2010).

Searching for other potential solutions, CA emerges as a potential technique to deal with the violation of the constant-coefficients assumption. CA emerged from the need to explore data, and became a common technique for statistical data analysis in areas such as machine learning, pattern recognition, information retrieval and image analysis (Mirkin, 2005). CA attempts to place observations into different clusters such that observations in the same cluster are homogeneous to each other but are different from ones in other groups (Fred & Jain, 2005). However, while a number of studies employ clustering in market segmentation to classify customers into homogeneous groups, little effort has been devoted to using CA within the finance discipline to partition firms (e.g. Epure, Kerstens, & Prior, 2011; Lee, Lee, & Wicks, 2004; Vlckova, Lostakova, Patak, & Tanger, 2014).

Figure 1.1 presents the number of research articles published on clustering since 1917 according to the Web of Science[1]. 'General' indicates all types of published documents, i.e. articles, proceedings papers, review, note and so on. Attention to CA started increasing from the 1970s with around 200 documents in general and 100 published journal articles recorded by the Web of Science. Since then, the number of documents and articles published has increased drastically by 2017 to 8,636 and 5,847, respectively.

[1] To get this number, we use the search tool from Web of Science. All documents having the keyword "cluster*" in their titles are counted. We admit the limitations of this searching method. Nevertheless, we assume that the results from the search are a good overall representation.

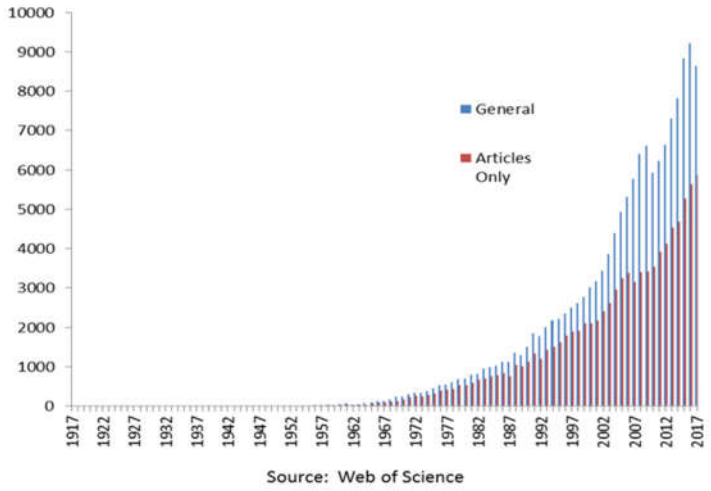


Figure 1.1- Number of Research Publications on CA by Year

Yet, very few studies apply CA to finance. As can be seen from Figure 1.2, there are only 89 articles (i.e. less than 0.1% of total) that have the keyword 'cluster*' in their title that are classified as belonging to the finance domain. However, the need to find better ways to partition firms is typified by the lack of success of earnings prediction studies owing to the violation of the constant β assumption (Nissim & Penman, 2001). Accordingly, this thesis proposes that research in the finance domain could potentially benefit from the use of CA.

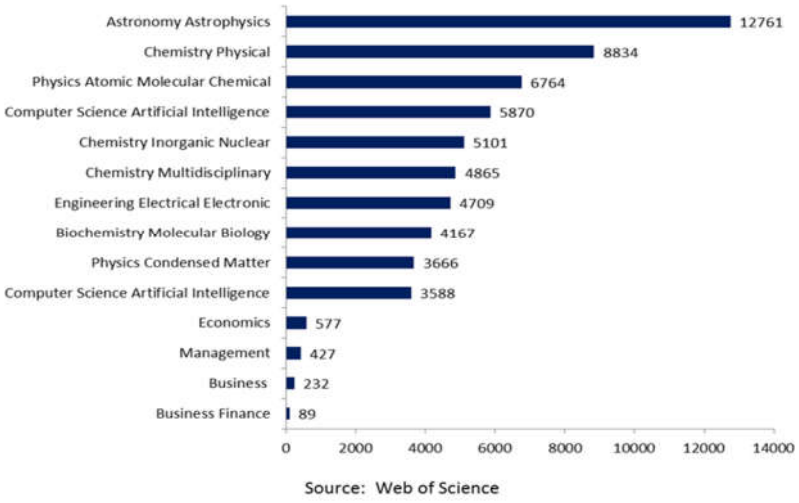


Figure 1.2- Number of Articles on CA by Subject

The extant literature on CA reveals that multicollinearity and feature weighting are among the most important problems challenging the performance of CA to recognize the unknown clusters (Sambandam, 2003; Amorim & Mirkin, 2012). However, the application of CA within the financial discipline focuses only on simple standard clustering techniques, mostly K-means clustering (e.g. Jensen, 1971; Gupta & Huefner, 1972; Epure et al., 2011). As a result, these studies fail to recognise the inherent shortcomings of clustering techniques as discussed above. These deficiencies could reduce or even obscure the performance of CA in addressing the problem of HGSC. This thesis fills this gap by proposing a novel procedure to apply CA in financial research. ***The new procedure attempts to mitigate drawbacks of CA, and consequently helps to reveal underlying clusters of firms where the constant-coefficients assumption is invalid.***

1.3 THESIS OBJECTIVE AND RELATED AIMS

Objective

As mentioned earlier in Section 1.1 and 1.2, ***the thesis objective is to investigate a solution to the prevailing problem of HGSC, and then relate it to an application within the finance domain***. In order to achieve this objective, two thesis aims are established.

Aims

The first research aim of this thesis is to develop a new clustering method that can be applied in financial research to address the problem of HGSC. HGSC remain a real challenge in finance research. Several solutions, ranging from normal partitioning of data to complicated econometric models have been introduced and developed to tackle this problem. This thesis uses CA, particularly K-means clustering, as a basis upon which to develop a new clustering technique to address the problem of HGSC. Note, other potential solutions for the HGSC issue that are not based on CA are considered to be beyond the scope of the thesis. Literature on contemporaneous solutions and the reason

for the choice of CA as a potential solution to resolve the HGSC problem are discussed in Chapter 2.

Furthermore, the thesis focuses only on addressing the HGSC problem. Although there remain several other important issues that finance researchers continue to tackle, such as model misspecification and endogeneity, the development of new solutions to these issues is beyond the scope of the thesis. Rather, in the models developed in the following chapters of the thesis, it is assumed that these issues are able to be satisfied by other means and as such, only the issue of HGSC is addressed.

Due to its simplicity, low computational resources and high popularity among alternative clustering methods, K-means clustering is the basis for the clustering method proposed in this thesis. There exist a variety of other clustering techniques, each of them with different advantages and disadvantages. While it is possible that the thesis innovation of combining CA and regression analysis could be applied using alternate clustering methods other than K-means, such efforts are beyond the scope of this thesis.

Issues concerning K-means clustering and its application across different disciplines are discussed in Chapter 2. This thesis focuses solely on two important issues commonly affecting finance data, i.e. multicollinearity and feature weighting. Furthermore, CA is the only method considered in this thesis to address the problem of HGSC in finance research. Other issues with respect to K-means clustering and its application's purpose are beyond the scope of the thesis.

To achieve the first research aim, the thesis begins Chapter 2 with a thorough review of the theoretical background and literature relating to the HGSC problem. The circumstances in which researchers face the problems of low predictive power and instability of estimates in regression analysis are reviewed. Then, evidence of different behaviours across grouping firms is discussed. Literature on previous econometric approaches proposed to address the problem of HGSC are considered. Potential

drawbacks to these approaches are investigated, suggesting reasons why a new clustering technique is needed to combat the problem of HGSC.

Consequently, a novel clustering technique combining K-means clustering and regression analysis, named ROWK clustering, is proposed in this thesis. To address the first research aim, four hypotheses are developed and subsequently tested. The first three hypotheses investigate those factors that affect the performance of CA with respect to the HGSC issue. From the findings, a new set of guidelines is developed for researchers to consider the feasibility of using CA in general and ROWK clustering in particular to address the HGSC problem. The fourth hypothesis predicts three channels through which ROWK improves the performance of CA with respect to the HSGC problem. The methodology adopted to test the hypotheses relating to the first thesis aim is discussed in Chapter 3, and the empirical results using simulated data sets are presented in Chapter 4 .

The second aim of the thesis is to apply the newly proposed ROWK clustering method to mitigate problems of HGSC in earnings persistence models. Given the theoretical background and empirical evidence of problems of HGSC in earnings persistence models, the thesis demonstrates the application of ROWK clustering to earnings persistence models.

If ROWK clustering can successfully be employed to identify underlying patterns of data, and through that mitigate the problem of HGSC, then it could also have other potential applications in the financial discipline. However, due to time and space limitations, this thesis only investigates the performance of ROWK clustering with respect to the identification of earnings persistence patterns. Although other potential applications of ROWK clustering will be suggested in Chapter 6, the testing of these potential applications is beyond the scope of this thesis.

The thesis application of ROWK clustering to earnings persistence is conducted using US data. There are several reasons for this choice. The availability of sufficient data is essential to conduct ROWK clustering. In case of earnings persistence, there are 17

clustering features that need to be included in the input data. US financial databases permit the availability of sufficient data. More importantly, other studies that address the HGSC problem in earnings persistence are conducted using US data. Hence, the thesis results are more easily compared to US research. The application of ROWK clustering using data from other countries is beyond the scope of this thesis.

To achieve the second research aim, five further hypotheses are developed. Hypotheses H5 to H8 consider the performance of ROWK clustering to identify the heterogeneities of coefficients of earnings persistence models. To test these hypotheses, Chapter 3 explains the research design to apply ROWK clustering to earnings persistence. It begins with steps to conduct ROWK clustering to explore firm patterns on earnings persistence. Adjustments to data processing to deal with the noise of real data are discussed. Models are then presented to estimate firm life cycles, earnings management and accounting conservatism. The final hypothesis tests whether analysts understand the earnings persistence patterns embedded by ROWK clustering by incorporating them into their earnings forecasts. Two approaches are presented to test this hypothesis. The first approach uses portfolio two-way sorting, and the second approach builds a model of analyst forecast errors. The methodology for the second thesis aim is discussed in Chapter 3, and the empirical results relating to the five hypotheses above are presented in Chapter 5.

1.4 CONTRIBUTION OF THE STUDY

This study contributes to both CA and the financial literature in several important ways. This thesis is the first study to systematically apply CA to address the problem of HGSC within financial research. “Systematically” in this context means that the thesis does not simply apply standard techniques of CA to group firms. Instead, the thesis comprehensively examines factors impacting the performance of CA to address the problem of HGSC. Then it discusses shortcomings of CA and proposes a new method to cope with these drawbacks. Finally, it illustrates the utility of this proposed method by examining its performance using both simulated and real data.

This thesis takes a further step of being the first study to mitigate the inherent drawbacks of CA that have not been sufficiently recognized and adjusted for in much of the past research (e.g. Epure et al., 2011, Lee et al., 2004, Li & Li, 2008). In order to achieve this objective, the thesis proposes a novel clustering technique, called Regression-Oriented-Weighted-K_means clustering (ROWK). It combines K-means clustering and regression analysis. Accordingly, the thesis introduces academic researchers to a useful new tool that employs CA to address the problem of HGSC.

This thesis also contributes to research in the finance discipline by introducing a standard procedure to apply CA to solve financial problems. The proposed method has the advantage of being easy to understand and execute using typical statistical analysis programs. Hence the thesis equips researchers with a powerful tool to enhance regression results whenever there are indications of heterogeneous coefficients, which are frequently problematic in financial research.

This thesis is also the first study to investigate (though simulated data) the factors that impact upon the performance of CA on HGSC. This study provides a novel guide for researchers to consider the feasibility of adopting ROWK clustering. It also equips researchers with a powerful tool to empirically explore which features are more important. For example, this thesis determines that, of the examined factors, earnings volatility and accruals are the most relevant to distinguish patterns of earnings persistence.

Additionally, this thesis makes an original contribution with respect to the application of CA (more precisely ROWK clustering) to identify patterns of earnings persistence in business firms. The thesis provides evidence of HGSC on earnings persistence, and shows the usefulness of using information from clusters identified by ROWK clustering to predict analyst forecast errors. The thesis's findings will be of particular interest to both academic researchers and investors who have concerns surrounding earnings forecasting.

Finally, given that ROWK clustering successfully addresses the issue of HGSC in both simulated and real data, the thesis creates promising opportunities for future studies to

apply ROWK clustering in other examined models when there are concerns of HGSC. Some suggestions for areas that could potentially benefit from the application of ROWK clustering will be discussed further in the thesis's conclusion in Chapter 6.

1.5 ORGANISATION OF THE THESIS

The remainder of this thesis is structured as follows. Chapter 2 presents an overview of the theoretical foundations and past empirical research on the solutions to the issue of HGSC and their drawbacks. Research gaps are identified, creating the foundation for the thesis development of a proposed new clustering method, named ROWK clustering. Next, a literature review discusses the HGSC problem in earnings persistence models and past solutions to address it. The literature review reveals an opportunity to apply ROWK clustering to earnings persistence models. Based on this, nine testable hypotheses are developed.

Chapter 3 outlines the data and methodology employed to test the hypotheses relating to the two thesis research aims developed in Chapter 2. It presents the econometric framework for the proposed ROWK clustering technique, which is developed to address the shortcomings of K-means clustering when dealing with HGSC. Then the research design is outlined for testing hypotheses H1 to H4, which relate to the first research aim. Three case studies are presented to shed light on each channel through which ROWK improves the performance of CA when dealing with the HSGC problem. Next, the research design for the application of ROWK clustering is presented to address problems in earnings persistence research. The model for earnings persistence is introduced and some adjustments for data processing are discussed. These adjustments deal with some concerns relating to using real (financial) data to run ROWK clustering. Subsequently, details of the data used in testing are described. Hypotheses relating to the first research aim are tested using simulated data. Then, real data regarding earnings persistence is used to test hypotheses relating the second research aim (hypotheses H5 to H9).

In the first of two empirical chapters, Chapter 4 documents the empirical results for the first research aim. Specifically, it discusses the findings pursuant to hypotheses H1 and

H2 to identify the determinants that improve the performance of K-means clustering for cluster identification and to solve the problem of HGSC. Additionally, evidence is presented of the effect of standardization on the performance of K-means clustering (hypothesis H3). The findings from three case studies are also presented, highlighting the channels through which ROWK improves the performance of CA in dealing with the HSGC issue (hypothesis H4).

In the second results chapter, Chapter 5 presents the outcomes from the investigation of the application of ROWK to address the problem of HGSC in earnings persistence, which is at the heart of the second thesis aim. Specifically, it discusses the findings pursuant to hypotheses H5 to H9 to assess the ability of ROWK clustering to better reveal the earnings persistence patterns that would otherwise remain unclear.

Chapter 6 concludes the thesis. It revisits the research questions and provides a synopsis of the methodology, hypotheses and the results. In addition, contributions and implications of the thesis findings are also stressed. The chapter ends with a discussion of the limitations of the study and suggestions for future research.

CHAPTER 2

LITERATURE REVIEW

2.1 INTRODUCTION

As mentioned in Chapter 1, the first research aim of this thesis is to develop a new method that can be applied in financial research to address the problem of HGSC. To do this, the thesis proposes a novel clustering method that tackles the inherent shortcomings of current solutions to the HGSC problem. For the second research aim, the new clustering method is applied to improve the estimation of firms' earnings persistence. This chapter discusses the theoretical underpinnings and previous empirical evidence related to these research aims. The key purpose of the literature review is to evaluate the existing empirical work, highlight research gaps and derive testable hypotheses.

The literature relating to the first research aim is discussed in Section 2.2, 2.3 and 2.4. Section 2.2 reviews the theoretical background and literature relating to the HGSC problem. It starts with a brief review of circumstances in which researchers face the problem of low predictive power and instability of estimates in regression analysis. It next discusses evidence of different behaviours across grouping firms. Sequentially, reviews of literature on previous econometric approaches proposed to address the problem of HGSC are conducted. Finally, potential drawbacks of these approaches are investigated, suggesting reasons why CA emerges as a potential weapon to combat the problem of HGSC.

Section 2.3 discusses the theoretical issues and existing literature pertaining to CA, a potential tool to address the HGSC problem. It starts with an overview of CA, i.e. its concepts, origin, and different developed clustering techniques. Next, shortcomings of CA and previous solutions are analysed. This section concludes with an overview on applications of CA in business economics research.

Section 2.4 identifies the gaps in the literature on solutions to the HSGC issue. These gaps highlight the need for a new clustering approach. This section presents mechanisms from which the new approach can successfully address the issues of CA, and consequently improve regression estimates. From this analysis, specific hypotheses in relation to the first research aim are developed for empirical testing.

Section 2.5 presents literature relating to the second research aim. The theoretical background and literature relating to earnings persistence is discussed, explaining why this area is considered as a potential application of the newly proposed clustering method. Finally, Section 2.6 provides a summary of the chapter with a list of all hypotheses.

2.2 HETEROGENEOUS-GROUP SPECIFIC COEFFICIENTS IN THE FINANCE AREA

This section briefly summarises evidence of different behaviours across grouping firms and suggests reasons why CA emerges as a potential weapon to combat the heterogeneous coefficients problem.

2.2.1 Low Predictive Power and Instability of Estimates in Regression Analysis

Poor predictability for both in-sample and out-of-sample tests is a phenomenon that researchers may face when conducting robustness checks. A lack of sound theories in constructing empirical tests could be a potential reason, and some evidence supports this view. Earnings predictability, a main component of equity valuation, is an example. Prior to 2000, fundamental analysis or equity valuation, with its aim to identify elements of financial statements that are relevant to assess firm value, attempts without theoretical guidance to select accounting variables or their ratios to predict earnings and consequently stock returns (Richardson et al., 2010). A common facet of these studies is the use of extensive lists of correlated variables in predictive regressions, but without any firm justification for their use, resulting in not only low out-of-sample predictive power,

but also concerns regarding the in-sample estimations (e.g. Lipe, 1986; Ou & Penman, 1989)[2].

However, even when the research is better-backed by theory, there are still instances where the results lack robustness. To correct for a lack of theoretical justifications in previous earnings forecasts research, Nissim & Penman (2001) employ a structural approach to identify value-relevant financial ratios and provide more organised ways to conduct the analytical tasks. They employ in their analytical process the residual income valuation model (RIV) developed by Feltham & Ohlson (1995). By manipulating components in the RIV model (i.e. book value of equity, return on equity and growth in residual earnings), key drivers of each component are uncovered and used as indicators in forecasting residual earnings, and eventually equity value. Not surprisingly, the use of the key drivers identified above results in reasonable R-squared values, and t-statistics from the pooled cross-sectional and time-series analysis. Prediction out of sample, however, demonstrates poor performance similar to that experienced by Ou & Penman (1989), casting doubt on the reliability of the predictive model.

Poor out-of-sample performance has been also documented in research using aggregate predictive models, most notably to estimate aggregate (or market) return and cash flow expectations. The most common approach to estimate is through predictive regressions that are upheld by a well-known theory called the present-value equation, first developed by Campbell & Shiller (1988). It describes the relationship between prices, future cash flows and discount rates. Accordingly, the aggregate price-dividend ratio has been shown as one of the most informative predictive variables of return and cash flow expectations (Kelly & Pruitt, 2013). Similar to the results of Nissim & Penman (2001), Lettau & Van Nieuwerburgh (2008) document reasonable R-squared values for in-sample estimates of predictive models. Typically, about 10% of forecast annual market return can be

[2] Ou & Penman (1989) present one of the first efforts to explore value-relevant attributes of firms using financial statement analysis, but they experience difficulties in their purely empirical results. Particularly, when conducting in-sample estimation of two non-overlapping time periods (i.e. 1965-1972 and 1973-1977), there appears to be little consistency of results in the descriptors across these two periods. This inconsistency in in-sample estimation could be a reason for poor out-of-sample predictive power in their study (Nissim & Penman, 2001).

accounted for by an aggregate value ratio such as price-dividend ratio or book-to-market ratio, but there is still little or even no predictive power in out-of-sample cases.

In summary, prior research has been beset by a host of problems, including instability of the examined relationships between variables over time and/or across different samples (e.g. different countries or states, different firm groups, in-sample vs. out-of-sample) and poor performance in out-of-sample predictive power. The problems are not solely limited to the lack of theoretical foundation, but moreover they challenge the validity of even the more sound theoretically-based models. Examining causes and developing methods to deal with these unsuccessful empirical results should be a matter of urgency. However, there is not much work that explicitly and structurally deals with this.

2.2.2 Evidence of HGSC

The studies of Ou &Penman (1989) and Nissim & Penman (2001) both aim to find value-relevant information from financial statements and come up with a list of accounting variables (or financial ratios) that are used to predict future earnings or residual earnings. Although there are differences in their approach to descriptors' identities, that is, one from purely empirical analysis and the other developed from the RIV model, they both suffer from instability of estimators across different time periods and poor out-of-sample predictive power.

A noticeable point is that both authors from these studies share the same belief that coefficients in their predictive regression are not constant across firms and time. Put another way, "*a general model is not a good representation for all firms (to the extent to which different characteristics generate future earnings in different firms in different ways*)" (Ou & Penman, 1989, p. 299). What their findings suggest is that the relationship between predictors and outcome variables is non-linear, encouraging an urgent call for industry-specific or firm-specific models, careful econometrics and prudent partitioning of the data (Nissim & Penman, 2001). This is a good starting point to take a closer look at the compelling reason underlying these empirical problems. In both of the two studies above, the researchers argue that a general model is not a good

representation for all firms, and appropriate partitioning of data is necessary to improve the results.

Regression models using cross-sectional or panel data often assume homogeneity of coefficients however, it is not difficult to find evidence that refutes this assumption. In a study of corporate governance and firm value, Giroud & Mueller (2011) find a negative relationship only in uncompetitive industries. It indicates that the relationship between corporate governance and firm value is not consistent across different industry groups classified on level of competitiveness. The same can be observed in the relationship between firm investment and profitability. Fu (2010) discovers a U-shaped correlation between investment and operating performance. Particularly, they point out that for firms with positive (negative) abnormal investment after seasoned equity offerings, an increase in investment reduces (improves) firms' operating earnings. Similarly, Hsiao & Tahmiscioglu (1997) document heterogeneous coefficients in a regression describing investment dynamics. Moreover, these parameters' differences cannot be explained by common firm characteristics.

In a similar vein, following the study of Nissim & Penman (2001), subsequent researchers have encountered dissimilar magnitudes in the way key financial ratios predict earnings. Amor-Tapia & Tascón Fernández (2014) detect changes in the signs of the coefficients of key drivers when predicting future profitability, which suggests the presence of non-linearity in the relationships. This is similar to the findings of Nunes, Serrasqueiro & Leitao (2010) who explore nonlinear relationships between the profitability of small- and medium-sized enterprises (thereafter SMEs) in Portugal and various explanatory factors. Specifically, these relations undergo a change across quintiles of the profitability distribution. In follow-up studies, Bauman (2014) observes a differential effect of downward and upward earnings management on one-year-ahead return on net operating assets (thereafter RNOA). He also reveals different magnitudes of profit margin effect on one-year-ahead RNOA, while controlling for earnings management. Earnings persistence

coefficients are also found to be different across quintiles of earnings volatility (Dichev & Tang, 2009) and over the firm life cycle (Dickinson, 2011)[3].

More importantly, there is evidence that firms cluster together, and this is of considerable importance in regression analysis. A well-known circumstance of parameter variation in regression models is in time series applications where firms' behaviours exhibit changes between periods. These switching regimes may relate to different states of the business cycle or other more fundamental structural changes (Goldfeld & Quandt, 1973). In this thesis, the terms '*cluster*', '*group*' and '*regime*' are used interchangeably.

Industry is another example. Firms within (between) an industry typically share common (distinct) characteristics, thus the model parameters are often estimated using cross-sectional pooled regressions within two-digit SIC codes (Jansen et al., 2012). Burnside (1996) observes significant heterogeneity of the production function across industries. Research on earnings management routinely categorises by industry for abnormal accruals estimation (e.g. Cohen & Zarowin, 2008; Hribar & Collins, 2002). The position of a firm relative to the industry average is shown to be more informative rather than using the level itself. For example, Amor-Tapia & Tascón Fernández (2014) document that the relative sign and position of firms' return on equity (thereafter ROE) with respect to the industry contribute additional relevant information about firms' future profitability. Firms are also influenced by actions from other firms in the same industry. Chen, Ho, & Shih (2007) document a negative effect of firms' corporate capital investment announcements on their industry rivals' stock prices. Industry grouping is often chosen as the traditional benchmark that researchers utilise when comparing their proposed partition variables (e.g. Dichev & Tang, 2009).

Firms also exhibit some similarities as they evolve across their life cycles. This process of evolution is determined by both internal factors (i.e. strategy, managerial ability, financial resources) and external factors (i.e. competitive environment and macroeconomics factors). A firm's life cycle represents a set of distinct phases where

[3] Particularly, firms with higher earnings volatility display less earnings persistence than those with lower earnings volatility. For more evidence of heterogeneous group-specific coefficients in panel data, see Lin & Ng (2012) and Hsiao & Tahmiscioglu (1997).

these factors change and then stabilise (Dickinson, 2011). Firms falling into the same phase of life cycle are likely to have the same age and size, hence these are common proxies for life cycle (Bradshaw, Drake, Myers, & Myers, 2012; Desai, Hogan, & Wilkins, 2006). The same pattern is observed for growth in sales and in capital investment, both of which monotonically decrease across the life cycle. Profitability, profit margin, and asset turnover are a function of firm strategy and the competitive environment, and they also show similar (different) patterns within (between) firm life cycle stages (Dickinson, 2011). As a result, these determinants of future profitability (proxied by RNOA) contribute to firm future profit differently across a firm's life cycle.

Firms also cluster based on proximity or through networking structure. A geographical economic factor is an environmental influence crucial to the growth of companies, and is even more important when there are networking activities among them (Kim, Lee, Choe, & Seo, 2014). Networking activities have been found to have a profound impact on firm performance (Ter Wal, 2013). Kim et al. (2014), using Network Structure Analysis claim that a network structure evolves as an endogenous factor of firm growth through close interaction between individual firms.

Briefly, the above findings provide evidence of violations of the coefficients' homogeneity assumption that is usually taken for granted in quantitative research. These could invalidate overall regression results. Hence, instability of coefficients over time and/or across firm groups, low R-squared in in-sample estimation and poor out-of-sample predictive performance are likely to be observed when this violation occurs. Appropriate partitioning of data is suggested as an essential solution for addressing this problem by Ou & Penman (1989) and Nissim & Penman (2001).

2.2.3 Review of Solutions to the HGSC Issue

Approaches to partitioning data that financial researchers employ to address the problem of HGSC have tended to be simplistic, such as by dividing a whole sample into different quantiles of certain firm features at which they expect to observe different relationships. Common partition variables are: firm size and the book to market ratio in the asset pricing model (e.g. Fama & French, 1993); industry competitiveness in corporate governance

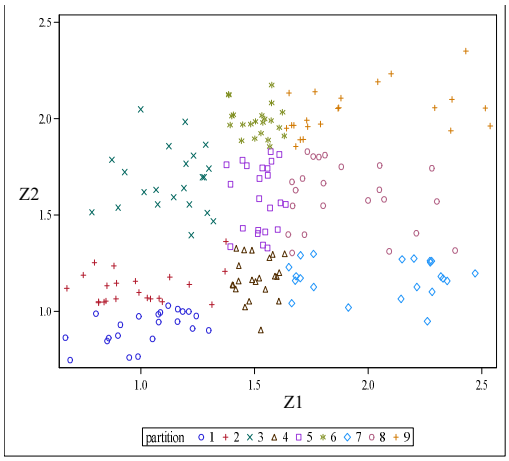
(Giroud & Mueller, 2011); and earnings volatility, firm life cycle and business strategy in equity valuation (Dichev & Tang, 2009; Dickinson, 2011; Little et al., 2009). By dividing data into sub-samples, researchers can gain a deeper understanding of dynamic relationships between concepts, achieve better estimations, increase predictive power, and observe better consistency of estimates across different samples.

However, this approach to partitioning suffers from two shortcomings. First, it is ad-hoc and does not take into account data patterns. Arbitrarily dividing data (firms) into quantiles based on some proposed factors (e.g. size, market to book ratio, earnings volatility, etc.) without considering the nature of the firm data is unlikely to result in an optimal solution. For instance, consider the case in which firms are divided into *3×3* or *5×5* portfolios based on firm size and book to market ratio. If the nature of the data is such that firms are best clustered into *3×2* groups, then employing *3×3* or *5×5* portfolios will not provide an optimal partition. Figure 2.1 illustrates this statement.

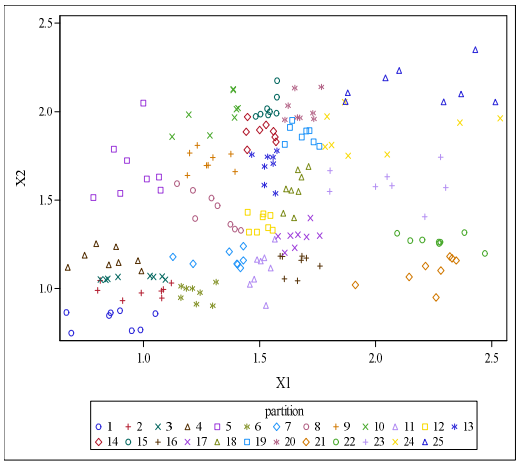
A synthesized dataset is generated which comprises 200 observations divided into six clusters by two partition variables z_1 and z_2. The cluster structure is created to fit the *3×2*-cluster pattern. As can be seen from Figure 2.1a and Figure 2.1b, the *3×3* and *5×5* partitions result in poor grouping. The poor performances are caused from two sources. First, members within a partitioned group include a mix of members from different true clusters. Second, the number of observations in the partitioned groups is less than the true cluster, causing reduced efficiency for subsequent regression estimations.

A limit in the number of factors used in the partitioning process is a second flaw. In the above example, one could argue that a graphical approach could identify the data pattern and guide the appropriate partition. However, when the number of partition variables increases, it becomes impossible to present the data graphically. For example, suppose that we want to split firms by a combination of ten potential features. This is infeasible if firms are allocated to different quintiles corresponding to these ten features. Even if this was possible, it would be a challenge to present the results in tabular form[4].

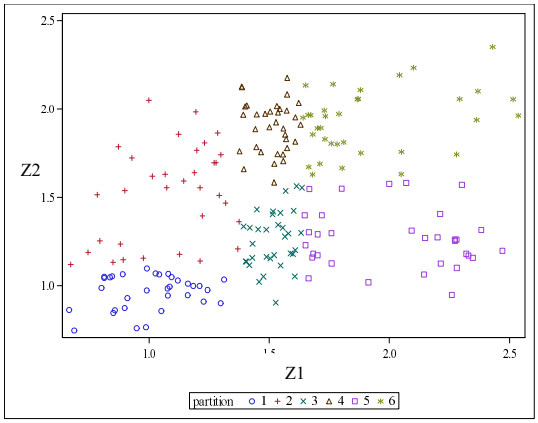
[4] One possible solution to this issue is through the use of an extraction method such as principal component analysis.



a. Result of 3*x*3 Partitioned Sample.



b. Result of 5*x*5 Partitioned Sample



c. Result of 3*x*2 Partitioned Sample

Figure 2.1- Partition Performance on 3x2-Clustered Data

Since the 1970s, there has been increasing attention to the field of econometrics that delves deeper into the HGSC issue. Several econometric approaches have been introduced. These methods differ principally on whether choices between clusters are assumed to be stochastic, i.e. depends on unknown probabilities that an observation belongs to a certain regime, or deterministic in the sense that it depends on unknown thresholds or cutoff values from a list observable variables (Goldfeld & Quandt, 1973).

Assume that there are *N* observations of a dependent variable y that are generated by two distinct regression equations or regimes (i.e. clusters):

(2.1)

$$y_i = x_i'\beta_1 + u_{1i}, i \in \xi_1, \text{ and}$$

(2.2)

$$y_i = x_i'\beta_2 + u_{2i}, i \in \xi_2$$

where *i* indexes observations, x_i', is a row vector including *p* independent variables of the *i-th* observations. ξ_1 and ξ_2 are the groups of observations (i.e. clusters) for which the two different regression equations hold, u_{1i} and u_{2i} are error terms and are assumed to be distributed as $N(0, \sigma_1^2)$ and $N(0, \sigma_2^2)$ respectively, and β_1, β_2 are the vectors of regression coefficients that are assumed to differ between the two equations[5].

When the choice between groups is deterministic (called *D-method*), there is a threshold or a cut-off point, say *i** so that for $z_i \leq i^*$then $i \in \xi_1$ and for $z_i > i^*$then $i \in \xi_2$. z_i, which is called the partition variable, is a variable upon which the value for the selection between regimes is defined. It could be a time index, one of the regressors or an entirely extraneous variable. If z_i is only one variable and there are only two regimes, then Quandt (1958, 1960) proposes to estimate *i** by maximising the likelihood function $L(y|i^*)$:

(2.3)

$$L(y|i^*) = \left(\frac{1}{2\pi}\right)^{N/2} \sigma_1^{-i^*} \sigma_2^{-(n-i^*)} \exp\left\{-\frac{1}{2\sigma_1^2}\sum_{i=1}^{i^*}(y_i - x_i'\beta_1)^2 - \frac{1}{2\sigma_2^2}\sum_{i=i^*+1}^{N}(y_i - x_i'\beta_2)^2\right\}$$

A likelihood ratio test is suggested to test the null hypothesis that no switch took place. It is measured as $\hat{\sigma}_1^{i^*}\hat{\sigma}_2^{(N-i^*)}/\hat{\sigma}^N$, where $\hat{\sigma}_1, \hat{\sigma}_2$ and $\hat{\sigma}$ are the estimated standard deviations of the residuals for Regime 1, Regime 2, and the entire sample, respectively.

If z comprises a list of *V* partition variables, i.e. $z_{i1}, z_{i2}, \ldots, z_{iV}$ then Goldfeld & Quandt (1972) suggest the use of a linear function of several partition variables as a proxy for the threshold. Particularly, this assumes that the selection between Regimes 1 and 2 depends on whether $\sum_{v=1}^{V}\pi_v z_{iv} \leq 0$ or $\sum_{v=1}^{V}\pi_v z_{iv} > 0$, where π_v are unknown coefficients. Letting D_i=0 if $\sum_{v=1}^{V}\pi_v z_{iv} \leq 0$ and D_i=1 otherwise, then the log-likelihood function becomes:

(2.4)

[5] Intercepts are embedded in $\beta_1$and β_2.

$$\log L = -\frac{N}{2} log 2\pi - \frac{1}{2}\sum_{i=1}^{N} log\left(\sigma_1^2(1-D_i)^2 + \sigma_2^2 D_i^2\right) - \frac{1}{2}\sum_{i=1}^{N} \frac{\left(y_i - x_i'(\beta_1(1-D_i) + \beta_2 D_i)\right)^2}{\sigma_1^2(1-D_i)^2 + \sigma_2^2 D_i^2}$$

In the case of a stochastic choice of regimes (called *λ-method*), let λ and *(1-λ)* represent unknown probabilities that an observation belongs to Regime 1 and Regime 2, respectively. Then the appropriate log-likelihood function is measured as:

(2.5)

$$\log L = \sum_{i=1}^{N} \log(\lambda f_{1i} + (1 - \lambda f_{2i}))$$

where f_{1i} and f_{2i} are the probability density function of y_i conditional on x_i' (Quandt, 1972). Again, the natural likelihood ratio could be used to test the null hypothesis that no switch took place.

Following these approaches, Lin & Ng (2012) introduce a similar threshold method called Two-step Pseudo Threshold approach. Unlike Goldfeld & Quandt (1973), this method applies only for panel data ($N \times T$) where regression coefficients can be estimated using individual time-series data. It also differs in the sense that the threshold values are identified without the knowledge of true partition variables. They argue that by the assumption if $i \in \xi_1$ then $\beta_i = \beta_1$ otherwise $\beta_i = \beta_2$ (where $\beta_1 \neq \beta_2$). Therefore, β_i with its threshold $i^* \in [\beta_1, \beta_2]$ could be a potential partition candidate. Based on that reasoning, the two-step Pseudo Threshold approach is developed, comprising two steps.

- In the first step, for each time-series of data for observation *i*, regress y_{it} on x_{it} to obtain $\hat{\beta}_i$. Then order observations based on values of $\hat{\beta}_i$, and let $\hat{\beta}_i$ be the partition variable (i.e. z_i).
- In the second step, the threshold i^* is estimated as a value that minimises the total squared residuals S_{NT}:

(2.6)

$$S_{NT}(\hat{\imath}^*) = \sum_{i|\hat{\beta}_i \leq \hat{\imath}^*} \sum_{t=1}^{T} \left(y_{it} - x_{it}\hat{\beta}_1(\hat{\imath}^*)\right)^2 + \sum_{i|\hat{\beta}_i > \hat{\imath}^*} \sum_{t=1}^{T} \left(y_{it} - x_{it}\hat{\beta}_2(\hat{\imath}^*)\right)^2$$

Recently, researchers have started to recognize the usefulness of cluster-alike algorithms to identify clusters. Lin & Ng (2012) introduce conditional K-means clustering (thereafter CK)[6]. This method employs the algorithm of K-means clustering to assign observations into clusters. However, it differs from K-means clustering in that while the criterion used to assign observations in K-means clustering is to minimise the sum of squared distance (thereafter *SSD*) between observations and clusters' centroids, the creation used in CK relates to the regression model itself (i.e. the sum of the regression squared residuals-thereafter *SSR*).

Particularly, to achieve the minimised *SSR*, the *CK* first randomly assigns *N* observations into *K* groups. Then in the second step, they estimate $\hat{\beta}_{ik}$ which is the pooled estimate of observations in cluster *k* (*k*=1,..., *K*). Observations are then reassigned to a certain cluster, say k' where their $SSR_i^{k'}$ $(= \sum_{t=1}^{T}\left(y_{it} - x_{it}\hat{\beta}_{ik'}\right)^2)$ is minimised for all *k*=1,..., *K*. In a similar vein, Ando & Bai (2016) study panel data models with unobserved group factor structures. This model is slightly different from previous models in the sense that the regression model includes the unobserved group factors and the focus is on how to estimate it[7].

Even though these approaches are proposed based on sound mathematical foundations and are supported by empirical evidence using either simulated data or real data, there are still potential drawbacks that need to be addressed. First, these papers only focus on the case of two regimes. When there are more than two regimes, it becomes complicated to construct the likelihood function and to find solutions of optimization. There are some extensions to these approaches (both *D-method* and *λ-method*) to deal with this issue however the number of regimes has to be defined in advanced.

The second concern relates to optimization problems of the (log) likelihood function. The optimization process becomes intractable if there is no further assumption relating to the

[6] Cluster analysis (and K-means clustering) will be discussed in the next section.

[7] Basically, their examined regression models is as follows:

$$y_{it,k} = x_{it}'\beta_k + f_{k,t}'\gamma_{k,i} + u_{i,t}, i = 1, \dots, N, t = 1, \dots, T, k = 1, \dots, K$$

where $y_{it,k}$ is the value of observation *i* at time *t* which belong to group *k*. $f_{k,t}$ is an vector of unobservable group-specific factors that affect the units only in group k, and $\gamma_{k,i}$ are the factor loadings.

choice of regimes. Particularly, for the case of *D-method*, to make the solution of Equation 2.4 tractable, Goldfeld & Quandt (1972) need a further assumption that D_i follows a continuous function:

(2.7)

$$D(z_i) = \int_{-\infty}^{z_i} \frac{1}{\sqrt{2\pi\sigma}} exp\left\{-\frac{1}{2}\left(\frac{\xi - i^*}{\sigma}\right)^2\right\} d\xi$$

However, this assumption results in some of the estimated $D(z_i)$ not being exactly 0 or 1.

The third concern comes from the fact that in reality there is more than one partition variable that influences the choice of regimes. To tackle this issue, a threshold D_i that is estimated as a linear function of several partition variables as suggested by Goldfeld & Quandt (1972) could be a potential solution. However, this approach has several deficiencies. To be tractable in formulating the log-likelihood function D_i must be assumed to follow the continuous distribution function as in Equation 2.7. More importantly, a linear function of several partition variables is not a good proxy for the threshold. Two observations with the same score of D_i could be very different or in far distance from each other if they are presented as points in space. For example, assume that the estimated linear function is as follows:

(2.8)

$$D(i) = \sum_{v=1}^{V} \pi_v z_{iv} = 0.3z_{i1} + 0.7\, z_{i2} - 0.5z_{i3}$$

Observations i_1 and i_2 have values of z_{iv} as $\{0.1; 0.5; 0.2\}$ and $\{0.1; -0.5; -1.2\}$ respectively. Obviously, these two observations are quite distinguishable since there is a sharp difference in values of z_{i2} and z_{i3} between these two observations. However, both i_1 and i_2 have the same score of D_i (=0.28). As a result, this approach could wrongly assign these two observations into the same cluster.

The final concern is the existence of cluster patterns documented in financial studies as addressed in previous sections. The fact that firms are grouping together based on some

characteristics provides incremental information relative to the regression analysis itself to address the problem of HGSC. An example may clarify the last statement. For Lin & Ng (2012), the target is to assign observations into clusters such that after running a regression within each cluster, the total square of residuals is minimised. To achieve this, they propose an algorithm that repeatedly assigns observation *i* to group k if:

(2.9)

$$SSR_i^k = argmin_k \Sigma_{t=1}^{T} \left(y_{it} - x_{it}{}' \widehat{\beta}_{ik}\right)^2,$$

where $\hat{\beta}_{ik'}$ is the pooled estimate of observations in cluster k.

Now, suppose there are two groups: ξ_1 and ξ_2 with their regression models: $y_i = 0.1 + 0.3x_i + u_i$and $y_i = 0.1 + 0.5x_i + u_i$, accordingly. Let A and B represent two observations that belong to the same group, such as ξ_1. Assume that X_A=1, X_B=1.5, u_A=1, u_B = -1. Even if the estimated $\widehat{\beta_k}, k = 1,2$ are correctly estimated (i.e. 0.3 and 0.5 for ξ_1 and ξ_2, respectively), according to the algorithm, B is correctly identified to ξ_1, but A is not ($SSR_A^{k=1} = 1;\ SSR_A^{k=2} = 0.64;\ SSR_B^{k=1} = 1; SSR_B^{k=1} = 1.69$).

This example illustrates that the performance of grouping methods that merely depend on regression analysis is highly sensitive to interactions between the sign and magnitude of error terms and discrepancies of coefficients across clusters. If observations A and B are at close proximity in space (as is usual) whose dimensions are partitioning features, this meaningful information would be captured by cluster analysis (CA).

In a nutshell, some econometric research has been devoted to address the problem of HGSC, but there are still challenges as discussed above. What is needed is an innovative technique that is able to easily handle multiple clusters and partition variables, is able to account for differences in contribution of partition variables in classification, and more importantly, that not only employs the information from regression analysis but also the rich information from cluster patterns. Hence, a technique that is able to utilise several partition variables and split data (such as firm observations) into meaningful/useful

groups could help researchers to gain new insights into the important features that cluster the data, and consequently improve the performance of statistical tests[8].

CA, in the form of unsupervised classification assigning objects into unlabeled classes, is such a technique. K-means clustering is among the most popular in the CA family. Although CA has been well developed and applied in several fields, there are as yet very few studies in the finance area that employ it as a means to discover firm patterns and enhance the results of empirical tests.

2.3 CLUSTER ANALYSIS

2.3.1 A Tool for Exploring Patterns

CA emerged from the need to explore data, and became a common technique for statistical data analysis in areas such as machine learning, pattern recognition, information retrieval and image analysis[9] (Mirkin, 2005). Attention to CA started to increase from the 1970s with around 200 documents in general and 100 articles published in journals in the Web of Science. Since then, the numbers of documents and articles published has increased drastically by 2017 to 8,636 and 5,847 respectively[10].

CA was first introduced and developed in natural science areas from the need to classify data into homogeneous objects. The five research areas that comprise the most articles on clustering are Astronomy-Astrophysics, Chemistry-Physical, Physics-Atomic Molecular Chemical, Computer Science-Artificial Intelligence and Chemistry-Inorganic Nuclear.

In business, more recently it has been applied to market segmentation studies by managers and analysts (Dolnicar, 2002). Yet, there are very few studies that apply CA to

[8] The purpose of clustering is for either understanding (meaningful clusters) or utility (useful clusters) (Tan, Steinbach, & Kumar, 2005). For the utility purpose, each group (cluster) could be represented by a cluster prototype. Then these prototypes could facilitate the subsequent data analysis or data processing technique such as summarisation or compression. In contrast, for understanding, cluster analysis uncovers meaningful groups whose members share common characteristics. These clusters would help researchers to analyse and describe the true structures underlying the data.

[9] There are various synonyms for methods of classifying similar observations, such as cluster analysis, classification theory, grouping methods, numerical taxonomy and clump theory (Jensen, 1971).

[10] Refer to Figure 1.1 (Chapter 1) that presents the number of published research articles on clustering since 1917 according to the Web of Science.

finance. There are only 89 articles (i.e. less than 0.1% of total) that have the keyword 'cluster*' in their title that are classified as belonging to the finance domain[11]. However, the need to find better ways to partition firms is suggested by the lack of success described above in earnings prediction studies because of the violation of the constant β assumption. Accordingly, this suggests that finance research could potentially benefit from the use of CA.

The aim of CA is to place observations into different clusters such that observations in the same cluster are homogeneous to each other but are different from ones in other groups (Fred & Jain, 2005). There are a number of clustering methods and cluster algorithms that have been developed, and these methods vary depending on the way similarity is defined as well as the assumptions about the 'feature' distributions and the shapes of clusters (Garla, Chakraborty, & Gaeth, 2012)[12]. In this paper, "*a partition variable*" or "*a feature*" or "*a clustering feature*" or "*a clustering dimension*" denotes a characteristic that is used to distinguish clusters. For example, sepal length, sepal width, petal length and petal width are features that is used to classify the well-known Iris dataset of 150 flower specimens (Amorim & Mirkin, 2012)

K-means clustering is the most popular method in the family of centroid approaches. It is credited with simplicity, low computational resources and high popularity among several clustering methods, and accordingly is employed in this thesis as the core technique to explore firm patterns. It was first introduced over four decades ago and is considered the most unsophisticated unsupervised classification algorithm to solve the clustering problem (Sun, Wang, & Fang, 2012). Dolnicar (2002) finds that K-means clustering is the most popular choice of clustering algorithms. Among the segmentation studies that he explored, K-means clustering accounts for 37% (68 out of 184) of all the clustering methods used.

Briefly, the K-means clustering algorithm uses an iterative procedure aimed at minimising the within-cluster dissimilarity as measured by Euclidean distance. The

[11] Refer to Figure 1.2 that presents the number of articles on clustering by subjects.
[12] For details of types of clustering and their corresponding measures' of similarity, see Tan et al.(2005)

procedure can be described as follows: Suppose we have a multivariate input data set Z (i.e. features) that is represented as an $N \times V$ matrix, where N is the number of data points or observations, and V is the number of features, $z_i = (z_{i1}, z_{i2} \dots z_{iV})^T$, $i=1,\dots,N$. The number of clusters is assumed as K. Denote $\xi_1, \xi_2, \dots, \xi_K$ as the corresponding K cluster with centres $c_1, c_2 \dots c_K$ respectively where $c_k = (c_{k1}, c_{k2} \dots c_{kV})^T$, $k=1,\dots,K$. Then, K-mean clustering will attempt to assign N data points into K disjoint clusters such that the sum-of-squares criterion, J, is minimised:

(2.10)

$$J = \sum_{k=1}^{K} \sum_{X_i \in \xi_k} \|z_i - c_k\|^2$$

where $\|.\|$ is the standard Euclidean norm[13] (Qian, 2006; Sun et al., 2012).

It is recognized that the globally minimised J is an *NP*-hard problem, meaning that it is nearly impossible to find any polynomial-time algorithms to solve it (Sun et al., 2012). Therefore, an iterative procedure is used to approximately minimise J. Particularly, at the end of some iteration, such as iteration t, $\xi_k^{(t)}$ and $C_k^{(t)}$ are updated. Then in the next $(t+1)$ iteration, each observation z_i is reassigned to the closest cluster based on its distance to the cluster centroids $C_k^{(t)}$ calculated in the previous iteration t. Subsequently, $\xi_k^{(t+1)}$ and $C_k^{(t+1)}$ are again updated.

(2.11)

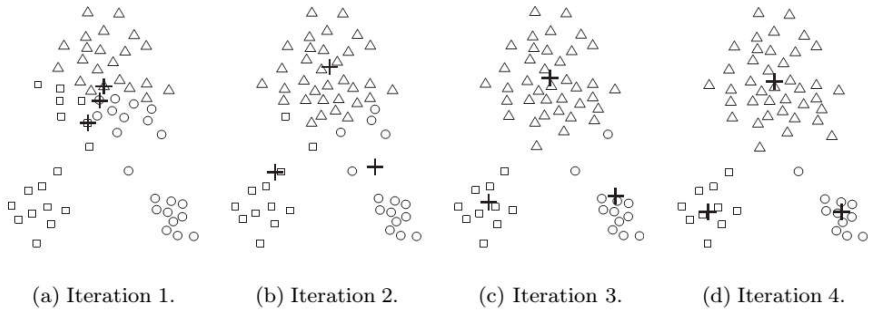
$$C_k^{(t+1)} = \sum_{Z_i \in \xi_k^{(t+1)}} \frac{z_i}{m_k},$$

where m_k is the number of observations in cluster k.

This process is repeated until the change of J when movement to the next iteration converges to the predefined convergence level (convergence criteria). At that point, the lowest J is attained corresponding to the initial K centroids ($C_k^{(0)}$) which are chosen at the beginning of the process. Usually the local minimisation is obtained due to sub-optimal

[13] The standard Euclidean norm is the most used measure of distance in clustering procedures (Gungor & Unler, 2008). It is derived from the Minkowski metric: $d(x,y) = (\sum_{i=1}^{n} |x_i - y_i|^r)^{\frac{1}{r}}$. Replacing r =2, we have the standard Euclidean norm: $d(x,y) = [\sum_{i=1}^{n} (x_i - y_i)^2]^{1/2}$.

initial *K* centroids (Gungor & Unler, 2008). Figure 2.2 illustrates each step of the K-means procedure. The bold crosses denote the cluster centroids.



(a) Iteration 1. (b) Iteration 2. (c) Iteration 3. (d) Iteration 4.

Figure 2.2- Using the K-means algorithm to find three clusters in the sample data.

Source: Tan et al., 2005, p. 498

2.3.2 Reviews of K-means Clustering Shortcomings

For the sake of brevity, this section discusses in-depth those topics relating to the focus of this thesis, i.e. correlated features and feature weighting, which are also the most important issues of K-means clustering. For other problems, see the footnotes below [14].

[14] K-means clustering has several shortcomings. One is that it is sensitive to outliers. The K-means clustering algorithm minimises the sum-of-squares distance (within-cluster-variances), so it tends to amplify outliers, especially in high dimensional data, leading to poor clustering results (SAS Institute Inc., 2009). When outliers are present, researchers are likely to observe clusters that have only a few observations. One remedy to the problem of outliers is the use of absolute distance instead of squared distance. SAS Institute Inc. (2009, p 1993-1672) also propose a solution and SAS codes to address this problem.

K-means clustering exhibits poor results in high-dimensional data. When the number of dimensions within the data is large relative to the number of observations and the cluster difference, then all observations are likely to be assigned into the same cluster. Extracting methods such as Principal Component or Factor Analysis are usually employed to overcome this problem (Sun et al., 2012).

The performance of K-means also depends on the selection of initial centroids. Clustering results vary corresponding to different selections of initial centroids (Gungor & Unler, 2008). This is because the algorithms of K-means clustering execute discrete assignments instead of sets of continuous parameters, resulting in a local (as opposed to global) minimum of the mean square error (Qian, 2006).

One common technique is to randomly choose initial centroids, and run K-means. This process is repeated multiple times, and we select the one with minimum within-cluster variances. However, this method may not be successful if (1) the data set has some pairs of clusters that are farther away from other clusters than clusters within these pairs (2) and the number of cluster is large (Tan et al., 2005). In these situations, it is likely that at least one pair of clusters only has one initial centroid. Because these pairs of clusters are far from other clusters, the K-means algorithm tends to group these clusters as one cluster, leading to a local minimum.

2.3.2.1 Correlated Features

Correlation between variables is common in finance. For example, firms with high operating income volatility tend to display lower financial leverage. In regression analysis, multicollinearity makes it difficult to identify the impact of each independent variable on the dependent variable (Sambandam, 2003). Solutions to deal with highly

---

K++ and bisecting K-means are two common techniques which are developed to address problems of random initial centroids. While the former, proposed first by Arthur, Vassilvitskii, and Siam/Acm (2007), directly improves the quality of the initial selection by spreading out initial cluster centroids, the later mitigates the effect of initialization problems. K++ spreads out initial cluster centroids by subsequently choosing random cluster centroids with a probability proportional to their smallest square distances from the existing cluster centroids. This algorithm has been shown to consistently outperform K-means in both minimum within-cluster variance criterion and running speed (Arthur et al., 2007).

On the contrary, bisecting K-means starts to split the whole data into two clusters by K-means. Then, one of these two clusters is chosen to be further split into two clusters also by K-means. The selected cluster to be split is usually the one with largest-size or highest within variances. This process is repeated and stopped upon reaching K clusters (Tan et al., 2005). Although bisecting K-means does not directly deal with choosing initial centroids, it does mitigate the effect of initialization problems because it conducts many trial bisections and focuses the splitting process in the largest-size clusters or highest within variance clusters.

In spite of improvements to reduce initialization problems, K++ and bisecting K-means still have concerns. Both have to determine the number of clusters (i.e. K) in advance. The so-called "intelligent" K-means (thereafter iK-means) as coined by Mirkin (2005) addresses this issue. This involves a combination of running K-means followed by a process of finding "anomalous clusters". K and initial centroids are determined by the number and centroids of the chosen anomalous clusters, respectively.

K-means also tends to find clusters containing the same number of observations, being compact and roughly hyper-spherical. Pre-defining the number of clusters (K) is another challenge when conducting K-means. Indeed, there is no completely successful method to identify the number of true clusters in any kind of cluster analysis (Bock, 1985).

There is no completely successful method to identify the number of true clusters in any kind of cluster analysis (Bock, 1985). In practice, this could be done by looking at the dataset graphically. This is a good idea if the data only includes two or three features. With higher dimensional data, it is impossible to graphically view the data. However, this can be solved by computing canonical variables from the raw data, and plotting these canonical variables instead of the original data (SAS-Institute-Inc, 2009). However, to compute canonical variables, the researcher must know which clusters are needed to compute the within-cluster covariance matrix. To overcome this problem, Art, Gnanadesikan, and Kettenring (1982) propose an approach to estimate the within-cluster covariance matrix without knowing the clusters. The ACECLUS procedure in the SAS program uses this approach to calculate canonical variables.

Besides looking at data graphically, several criteria have been proposed to find the number of clusters. Among these criteria, the pseudo F statistic, the pseudo t2 statistic and the cubic clustering criteria (CCC) are identified as the best criteria in a simulation study by Cooper and Milligan (1988). To identify the appropriate number of clusters, it is suggested that this number should be a consensus among these statistics. It means that this number has simultaneously (1) the local peak of the CCC, (2) pseudo F statistics and (3) small pseudo t2 statistic preceded by a large pseudo t2 statistic. See Chiang and Mirkin (2010) for the intelligent choice of the number of clusters in K-means clustering.

correlated variables in regression analysis tend to be straightforward, such as dropping one of the collinear variables, transforming the collinear variables or using ridge regression.

However, in CA the multicollinearity problem is different and not easily handled. Sambandam (2003) argues that features that are highly correlated automatically receive higher weights than others. In an extreme case when two features are totally collinear, they represent the same underlying feature and this underlying feature attracts twice the weight than it should. Consequently, the final clustering result tends to over-emphasise this underlying feature. Mitigating the effects of correlation in CA is not easy. Figure 2.3 demonstrates the performance of K-means under highly correlated features.

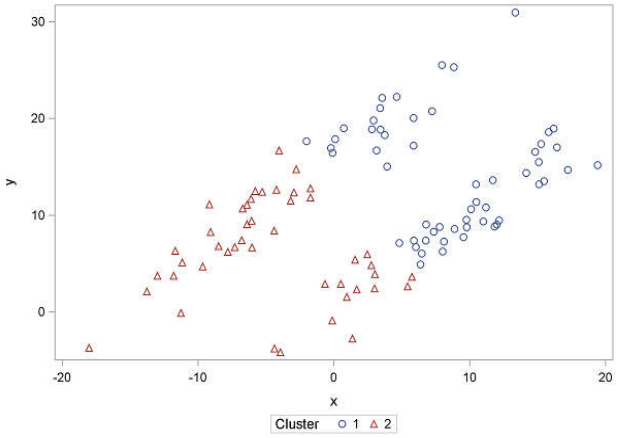


Figure 2.3- K-means with high elongated clusters (Source: SAS-Institute-Inc, 2009, p. 234)

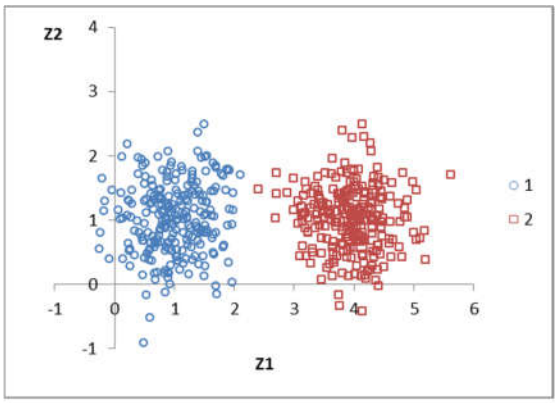
At first glance, one could conduct principal component analysis (thereafter PCA) to extract uncorrelated important components from data, and use these components in subsequent CA. However, there are relatively few first components that are used in clustering, and there is no guarantee that these components contain the target signal that the researcher is seeking to identify using clustering (Zhao & Maclean, 2000, Witten & Tibshirani, 2010). The performance of CA using the extracted components is worse than those using original data if the relevance of these components to identify cluster structures is not in line with the extracting methods. What if the first principle component that accounts for up to (say) 80% of data variation is of less relevance to distinguish the cluster

pattern than the second principle component? Hence, principal components analysis may not help to mitigate the problem of multicollinearity and may indeed exaggerate it.

Within-cluster correlation vis-a-vis *total correlation* is another concern relating to the use of extracting methods in clustering. Total correlation means the correlation coefficients between clustering features are measured using all observations. Figure 2.4a, Figure 2.4b and Figure 2.4c illustrate the differences between these types of correlation. A dataset with 500 observations belonging to two clusters is generated. There are two features, i.e. $z_1$and z_2.

Extracting methods typically deal with total-correlated features, but not with the correlations within clusters. Even when clustering features are correlated as measured by all data but uncorrelated as measured within clusters (Figure 2.4b), there is no need to conduct an extracting process before CA since K-means with the original data could perform well in this case.

The novel idea for the proposed solution in this thesis is that whenever an irrelevant feature is either total- or within-cluster correlated with a relevant feature, including this feature into clustering will negatively affect the following regression estimations, resulting in an unexpectedly large *MAR/MSR*. Therefore, *MAR/MSR* could be used as a criterion to guide the clustering process to identify and reduce the weight of this irrelevant feature.



a, Within- and Total-Uncorrelated Features
$\rho_{z_1z_2}^{Clus1} = 0.127; \rho_{z_1z_2}^{Clus2} = -0.058; \rho_{z_1z_2} = 0.014$

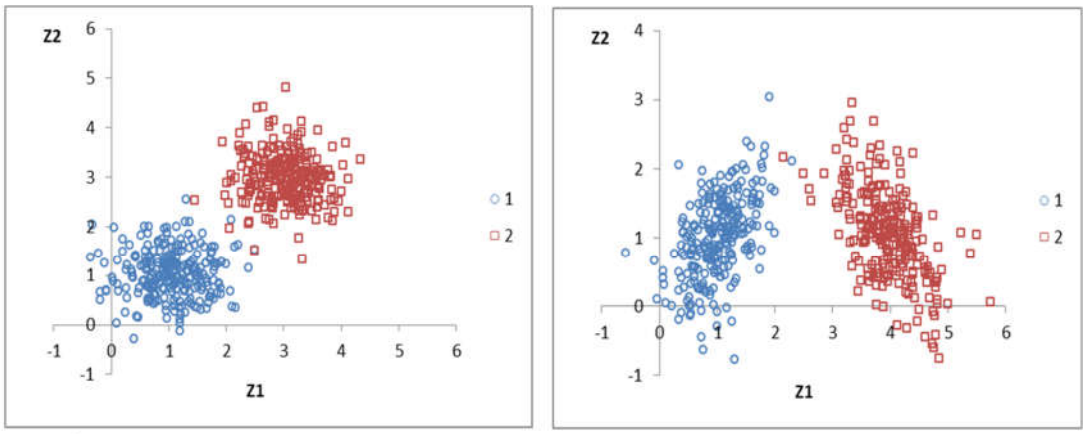


Figure 2.4- Within- vs. Total-Correlated Features

2.3.2.2 Feature Selection and Weighting

The most challenging issues of not only K-means but any CA relate to the feature selection and weighting. The ultimate goal of CA is to discover the true cluster structure. In this regard, choosing relevant features and deciding upon their weights are critical parts to ensure success (Brusco & Cradit, 2001). Indeed, the presentation of irrelevant features that have no contribution to distinguish members between clusters may obscure the cluster structure (Desarbo, Carroll, Clark, & Green, 1984). Figure 2.5 demonstrates this statement and shows how appropriate weighting can improve the performance of K-means. There are two synthesized clusters ($\xi_1$and ξ_2) with 40 members each. There are two features z_1 and z_2 with similar standard deviations ($\sigma_{Z_1} = 0.905$, $\sigma_{Z_2} = 1.09$). z_1 is generated to be an relevant feature to distinguish members from these two clusters, while z_2 is irrelevant[15].

Figure 2.5a exhibits the original data. Observations seem to scatter randomly and it is very difficult to find any clear patterns/clusters from the data. In this case, z_2 obscures the cluster structure, consequently K-means cannot precisely assign members. However, as more weight is placed successively on z_1, observations become distinguishable into

[15] With the knowledge of the true clusters' membership, the relevance of a clustering variable can be tested by ANOVA or discrimination analysis etc.

two clusters as depicted in Figure 2.5b. Consequently, K-means can be used successfully to recognize this pattern. Figure 2.5c shows the true cluster membership. This example demonstrates that the choice of relevant features and the decision concerning their weights are critical components for the successful execution of CA in general and K-means in particular.

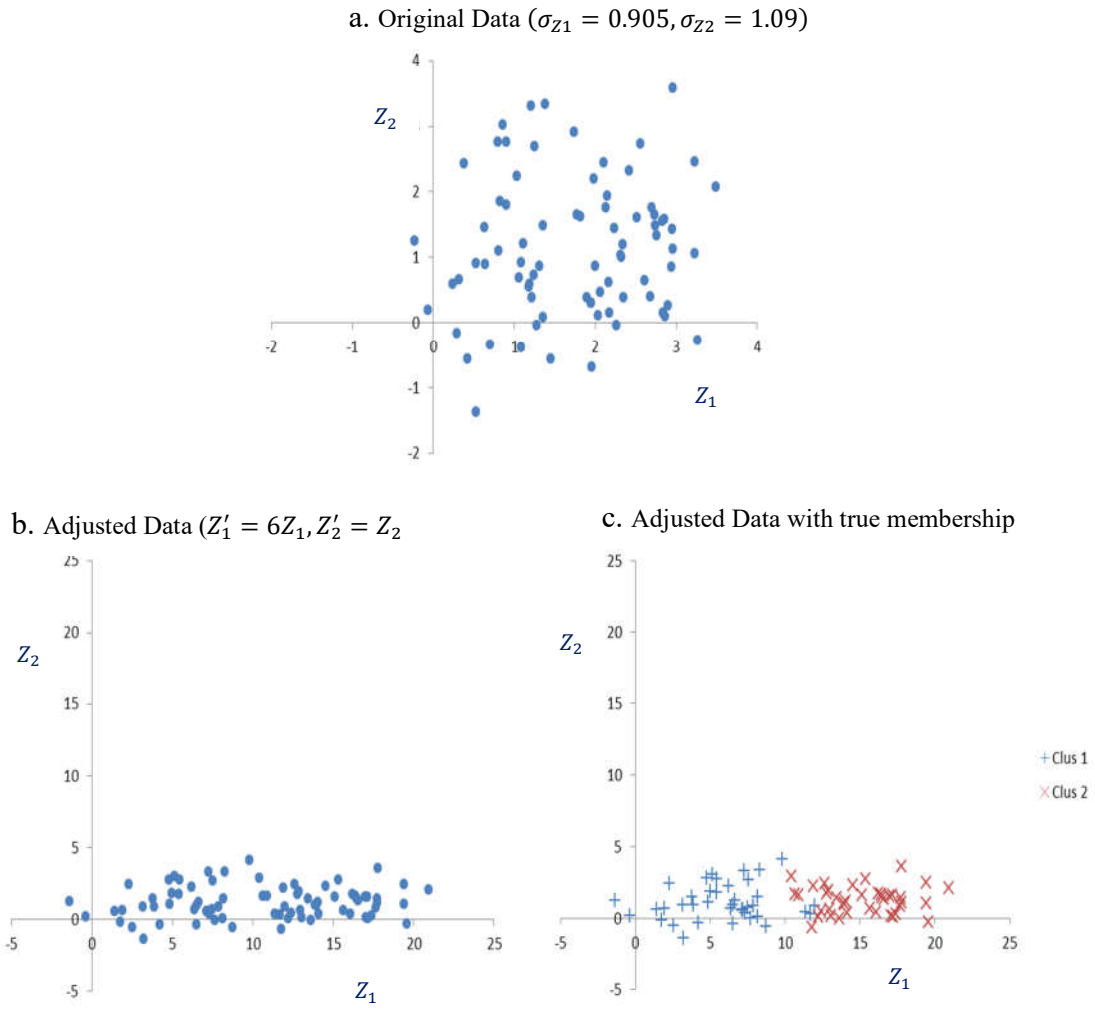


Figure 2.5- Relevant vs. Irrelevant Clustering Features

Enhancement of the selection process can be achieved by a deductive approach that emphasises a strong link between the selection process and theory, leading to a priori expectations on the employed features and the cluster natures (Ketchen & Shook, 1996). There are still, however, existing applications of inductive approaches which do not require any such a priori expectations, resulting in employment of as many features as

possible (Epure et al., 2011). This approach can cause problems of irrelevant features and high dimensionalities.

In feature-weighting clustering methods, features have differential abilities to define cluster patterns. These clustering methods try to find the most appropriate weights in order to eliminate the irrelevant features and consequently strengthen the cluster results (Brusco & Cradit, 2001). A number of feature-weighting clustering methods have been proposed and developed. Among the notable feature-weighting methods applied for K-means analysis is synthesized clustering (thereafter SYNCLUS) introduced by Desarbo et al. (1984). Through an iterative fitting process, optimal weights are estimated using a 'generalized' K-means procedure[16]. However, in a study comparing the performance of various feature-weighting clustering methods, Gnanadesikan, Kettenring, & Tsao (1995) find that the SYNCLUS procedure is less effective than simpler methods such as equal-weight scaling, standardization, and range-scaling. In addition, among six standardization functions, ranking features out-perform the other functions in their ability to deal with high-dimensional data (Tanioka & Yadohisa, 2012). SYNCLUS also requires knowledge

[16] Specifically, with a matrix of features Z, a predefined vector of battery importance weights (W^2) and a predefined number of clusters (K), SYNCLUS tries to find the optimal weights (w) and corresponding clusters' members by minimising the following sum of squares:

$$J = \sum_{b}^{B} \sum_{i<}^{N} \sum_{i'}^{N} w_b^2 \left(\delta_{ii'} - d_{ii'}^{2(b)}\right)^2$$

where:

- N = the number of observations; K = the number of clusters; B = the number of batteries or groups of features;
- N_k= the number of observations in cluster k, so $\sum_{k=1}^{K} N_K = N$;\
- w_b^2 = the predefined weight for the b-th battery. It is normalized so that $\sum_b^B w_b^2 = 1, b = 1, \dots, B$;
- $z_{iv_b}^{(b)}$= the v_b-th feature in the b-th battery describing observation i; $v_b = 1, \dots, V_b$;
- $w_{bv_b}^2$ = the weight for the the v_b-th feature in the b-th battery;
- $d_{ii'}^{2(b)} = \sum_{v_{b=1}}^{V_b} w_{bv_b}^2 \left(z_{iv_b}^{(b)} - z_{i'v_b}^{(b)}\right)^2$= the weighted squared distance between observations i and i' measured for the b-th battery.
- $\delta_{ii'}$ = fitted value of $d_{ii'}^2$, i.e. it is estimated by the following regression:

$$d_{ii'}^2 = \alpha + \beta a_{ii'}^* + \varepsilon_{ii'}$$

where:

$$a_{ii'}^* = \begin{cases} \frac{1}{N_k} & \text{if observations } i \text{ and } i' \text{ are in the same cluster k} \\ 0 & \text{if observations } i \text{ and } i' \text{ are in different cluster k} \end{cases}$$

beforehand of the number of clusters (K) and the vector of battery importance weights (w_b^2).

A recent study on weighted K-means by Amorim & Mirkin (2012) aims to further develop the weighted K-means (thereafter WK) algorithm introduced by Huang, Xu, Ng, & Ye (2008)[17]. They address two important issues of K-means, which are no defense against irrelevant features and no adjustment for the initial location of centroids, by introducing the so-called Intelligent Minkowski metric Weighted K-Means (thereafter iMWK). This is a closed form algorithm analogous to that of Huang et al. (2008) with an adjustment to the distance formulae. Specifically, instead of using the Euclidean metric in the criterion, they utilise the Minkowski metric and sketch out the searching procedure for Minkowski centres as a process of minimisation of a convex function[18]. By simulation, iMWK is shown to outperform both K-means and WK.

However, WK and its developed version-iMWK have three important deficiencies. First, the criterion used to derive optimal weights is totally drawn from the clustering itself. Put another way, its objective is to minimise within cluster distances (measured by Minkowski metric) given that the weights are supposed to be non-negative and sum to

[17] Weighted K-means assigns weights to features by minimising the within-cluster weighted sum-of-squared distance criterion: $\mathrm{W}(\xi, \mathrm{C}, \mathrm{w}) = \sum_{k=1}^{K}\sum_{i\in\xi_k}\sum_{v=1}^{V} w_v^{\beta}(z_{iv} - c_{kv})^2$ constrained to $\mathrm{w_v} \geq 0$ and $\sum_{\mathrm{v}=1}^{\mathrm{V}} \mathrm{w_v} = 1$. The exponent β is a pre-defined parameter. It denotes the intensity of a weight's effect on distance measures (Amorim & Mirkin, 2012). Huang et al. (2008) propose an algorithm which seeks a combination of ξ,C,w to minimise $\mathrm{W}(\xi, \mathrm{C}, \mathrm{w})$ by an iterative process. The algorithm of WK-means is summarised as follows:

- Step 1: Given the initial centroids and weights, assign points to clusters whose centroids are closest to these points. Distance is measured by the weighted formula: $d(z_i, c_k) = \sum_{v=1}^{V} w_v^{\beta}(z_{iv} - c_{kv})^2$
- Step 2: After assigning all points to corresponding clusters, recalculate new cluster centroids. In this case cluster centroids are computed as the means of clusters.
- Step 3: Given clusters and centroids, recompute weights w_v by the following formula:

$$w_v = \frac{1}{\sum_{j=1}^{V}(\frac{D_v}{D_j})^{1/(\beta-1)}}$$

where $D_v = \sum_{k=1}^{K}\sum_{i\in\xi_k}(x_{iv} - c_{kv})^2$. This is the sum of within-cluster variances of feature v. This formula is derived from the first order condition of minimising $\mathrm{W}(\xi, \mathrm{C}, \mathrm{w})$ under the constraints of w_v (i.e. non-negative and unity sum).

[18] A centre or a centroid of a cluster is the point that minimises the sum of squared distances between it and all points within this cluster. For example, if distance is defined as the Minkowski metric with β=1, i.e. the Manhattan distance: $d(i, c_k) = \sum_{v=1}^{V}|z_{iv} - c_{kv}|$, then the median is the centre of cluster. When β=2, i.e. the square Euclidean distance: $d(i, c_k) = \sum_{v=1}^{V}(z_{iv} - c_{kv})^2$, then the gravity mean is the cluster centre. When β>2, finding the cluster centre is more complicated. However, it could be found by using an algorithm as in Amorim and Mirkin (2012, pp. 1065-1066).

unity[19]. The objective is intuitive, but given that it is internally-derived, it fails to define the exponent parameter (β) within the model. The exponent parameter is instead user-defined before running the clustering procedure. Consequently, an optimal β is only identified through the supervised or semi-supervised process (Amorim & Mirkin, 2012). Second, both WK and iMWK only address the problem of noise or irrelevant features, and in their simulated data, each of the clusters is spherical. Problems associated with elongated clusters or correlated dimensions are not considered. Finally, optimal weights as estimated by WK or iMWK do not necessarily coincide with the weights that best improve the regression analysis. As a result, it should be regression analysis that provides the ultimate criteria to guide the cluster analysis and adjust the weights of features, not the internal target of the clustering itself.

Figure 2.6 illustrates the final statement. A synthesized set of data contains three clusters (i.e. ξ_1,ξ_2,ξ_3) with two features z_1 and z_2. While z_1 helps to distinguish ξ_1 from ξ_2, z_2 contributes to identify cluster ξ_3. As the ultimate purpose is to improve regression estimations, it makes sense to emphasise z_2 since the true regression coefficient of cluster ξ_3 (i.e. 1) is distinctly different from the other clusters (i.e. 0.2 and 0.25).

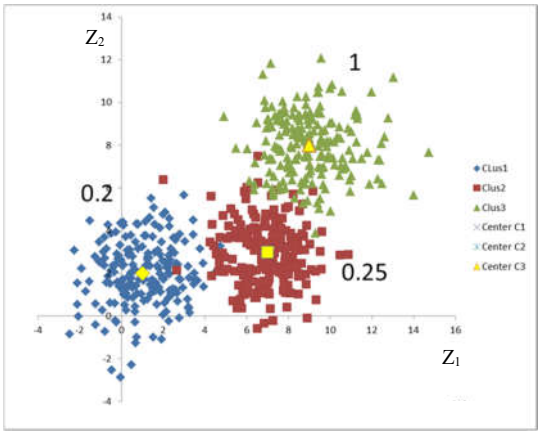


Figure 2.6- Clustering vs. Regression Contribution of Features

[19] Particularly, it minimises J= $\sum_{k=1}^{K}\sum_{v=1}^{V}\sum_{x_i\in\xi_k}\|w_v x_{iv} - w_v c_{kv}\|^{\beta}$ where K is number of clusters; set of V features v, and x_{iv}, c_{kv} are the value of feature v at entity i and centroids $k \in \xi_k$ accordingly. w_v denote feature weights. The exponent β is a pre-defined parameter presenting the rate of effect of the weights on its contribution to the distance.

2.3.3 Cluster Analysis in Business Economics Research

Despite its several inherent drawbacks, the K-means algorithm has been widely employed, and has been found to be successful in many applications (Sun et al., 2012). Market segmentation is a key strategic issue in marketing, and cluster analysis is used to separate customers in a certain market into homogeneous subgroups of customers. Cluster analysis allows suppliers to gain valuable insights into the common characteristics shared by their customers, thereby providing helpful guidance for targeted marketing actions (Garla et al., 2012). Not surprisingly, the large bulk of the research applying cluster analysis in the business economics area is for market segmentation applications. It is also notable that, in a very comprehensive review of cluster analysis in data-driven market segmentation applications, Dolnicar (2002) finds that K-mean clustering is the popular choice among clustering algorithms[20].

One example is from a study by Lee et al. (2004) where the target of segmentation is festival participants. In this study, six motivation dimensions for participants attending the 2002 World Culture Expo are identified. Then these factors are used as inputs in a cluster analysis using the K-means clustering method. This helps to identify four well-defined clusters that are significantly different in three-quarters of their motivation means. By carefully examining the characteristics of each cluster, they argue that they can identify for the event manager more insights into their visitors' motives. Similarly, K-means clustering is employed to segment overseas travelers' motivations (Beh & Bruyere, 2007; Cha, McCleary, & Uysal, 1995), customer behaviours in the metals, metallurgic and metalworking market (Vlckova et al., 2014) and customer behaviours in the real estate market (Park, 2011).

In a similar vein, clustering has also been used in banking research in order to classify customer behaviour with respect to buying financial commodities and services (Lin, Yang, & Ieee, 2006), using free of charge bank services such as internet banking (Maenpaa, 2006) and using the retail banking charges calculator (Soukal & Hedvicakova, 2012). The purpose of these studies is to gain more understanding of customer behaviour

[20] See Dolnicar (2002) for an extensive review of clustering applications in marketing research.

in buying or using bank products by clustering customers into appropriate groups. The insights gained can assist banks to develop better customer-oriented products.

While the above demonstrates research interest in clustering applications in market segmentation research, there is scant published finance research using cluster analysis to group firms or stocks. As stated previously, only 89 articles (accounting for less than 0.1% of all 'clustering' titled articles) in the financial area mention 'cluster*' in their title (see Figure 1.2). Yet, for 60 of these 89 articles, the word 'cluster' mostly refers to price, size or volatility clustering, which has no relationship to CA (e.g. Alexander & Peterson, 2007; Ni, Pearson, & Poteshman, 2005; Ning, Xu, & Wirjanto, 2015). Strikingly, only 29 finance articles mention, apply, or develop cluster analysis.

Jensen (1971) presents one of the first attempts to apply CA in finance. Using a hierarchy clustering method with firms' financial and stock market performance as features, two clusters of companies are identified, i.e. single-firm clusters and multiple-firm clusters. Firms belonging to the former cluster type performed "better-than-average" in the later period (1954-1965). In addition, firms in the multiple-firm clusters exhibited more similar performance than those being clustered randomly. Using the same hierarchy clustering method, Gupta & Huefner (1972) focus on a different objective of how to group industries based on their financial ratios[21].

In another financial application of clustering, Epure et al. (2011) study the changes in productivity and efficiency within the Spanish banking sector. In the first step, key indicators of productivity are identified. These frontier-based productivity drivers are then treated as inputs in a CA to explore dissimilarities among bank groups in terms of productivity performance.

[21] Five industry-level financial ratios are used by Gupta & Huefner (1972): fixed asset turnover, current asset ratio, inventory turnover, average collection period and cash velocity. They document that the clustering results are highly consistent with both the judgmental classifications of economists and with numerous economic characteristics of the industries involved. In addition, the ratios show different contributions on the clustering results. Among clustering ratios, fixed asset turnover yielded the best industry classification results. The findings from this article illustrate that CA can be used to identify surrogate variables, which are good substitutions for sets of characteristics that are extremely difficult to quantify.

As discussed above, extracting methods such as PCA or factor analysis (thereafter FA) are not appropriate tools to deal with the problem of correlated clustering features. However, the recent tendency to employ these methods before running clustering can mainly be explained by researchers' desires to minimise the problem of high numbers of dimensions. For example, Li & Li (2008) perform a combination of FA and CA to investigate the herding behaviour of firms in financing decisions. Starting from the leader-herd index and a list of 13 indicators of corporate financing (measured as the debt to equity ratio), FA is employed to identify six latent factors. Then, the authors use these latent financing decision factors as clustering features to classify firms into three types according to size, growth, and herding behaviour. FA followed by CA is also a main methodology of Di Cimbrini (2015) who tests whether mutual financial support for the poor is just a mask for organising the political activities of the working classes[22].

2.4 GAPS AND MOTIVATIONS FOR A NEW CLUSTERING APPROACH

2.4.1 Gaps and Motivations

Prior research has been beset by a host of problems, including instability of the examined relationships between variables over time and/or across different samples and poor performance in out-of-sample predictive power. Examining the causes of these unsuccessful empirical results and developing methods to deal with them should be a matter of urgency. However, there is a deficit of studies that explicitly and structurally address this.

There is evidence of violations of the coefficients' homogeneity assumption that is usually taken for granted in quantitative research. These could make overall regression results invalid. Appropriate partitioning of data is suggested as an essential solution for addressing this problem (Ou & Penman 1989, Nissim & Penman 2001). The econometric studies reviewed in this chapter attempt to address the problem of HGSC, but challenges still remain. Further work is required to develop a technique that is able to easily handle multiple clusters and features; is able to account for differences on the contribution of

[22] For other examples of articles using a combination of factor analysis and cluster analysis, see Ando & Bai (2016); Mohd-Rahim, Wang, Boussabaine, Abdul-Rahman, & Wood (2014); Nimtrakoon & Tayles (2015).

features in classification; and more importantly, employs not only the information from the regression analysis side but also the rich information from cluster patterns.

CA, particularly K-means, in the form of un-supervised classification assigning objects into unlabeled classes, has the potential to be such a technique. It has been increasingly employed in several disciplines as a powerful tool for partitioning data. A number of studies employ CA in market segmentation to classify customers into homogeneous groups. However, little effort has been devoted within the finance discipline to partitioning firms. Consequently, there appears to be significant untapped potential to use CA to group firms and explore different behaviours of firms across groups. Even in those few studies that undertake firm partitioning, clusters are based only on a single feature (e.g. Gupta & Huefner, 1972), or on a single aspect such as firm productivity (e.g. Epure et al., 2011) and corporate financing (e.g. Li & Li, 2008). It appears that little or no attempt has been made to employ CA to classify firms based on a broader perspective, i.e. firm valuation. Key drivers of firm valuation would provide useful information for cluster analysis to uncover firm patterns (Nissim & Penman, 2001). This key issue of applying CA to firm valuation will be addressed in the sections that follow.

Most importantly, the applications of CA tend to focus mainly on dividing customers or firms into different clusters, and exploring different characteristics among these clusters. However, most studies reviewed above have neglected to recognize the potential usefulness of firm partitioning using CA to address the serious issue of the violation of the constant-β assumption. Yet, as outlined above, CA and particularly, K-means clustering have inherent shortcomings causing incorrect classification of objects. This problem has not been sufficiently recognized and addressed in much of the past research (e.g. Epure et al., 2011, Lee et al., 2004, Li & Li, 2008 etc). This concern motivates the development of the new approach introduced in this thesis.

2.4.2 Regression oriented Weighted K-means

Figure 2.7 establishes a theoretical framework for the development of the novel clustering technique proposed in this thesis. K-means clustering is the most popular method in the family of centroid approaches, being an iterative procedure to minimise the within-cluster

dissimilarity as measured by Euclidean distance. The ultimate goal of K-means clustering is to discover the true cluster structure. In this regard, choosing relevant features and deciding upon their weights are critical parts to ensure success (Brusco & Cradit, 2001).

This thesis proposes an innovative method, namely Regression oriented Weighted K-means (ROWK) that combines K-means clustering and regression analysis. On the one hand, the K-means algorithm iteratively assigns similar observations into clusters. On the other hand, the *MAR/MSR* from running regressions is used to guide the process of weight adjustment in clustering and mitigate the problem of multicollinearity. Accordingly, the ROWK procedure includes four steps. First, the regression model is identified. Second, features are selected. Third, data is pre-processed before being used in the fourth and final step to find the optimal weights. The econometric framework of the ROWK is described in detail in the methodology chapter.

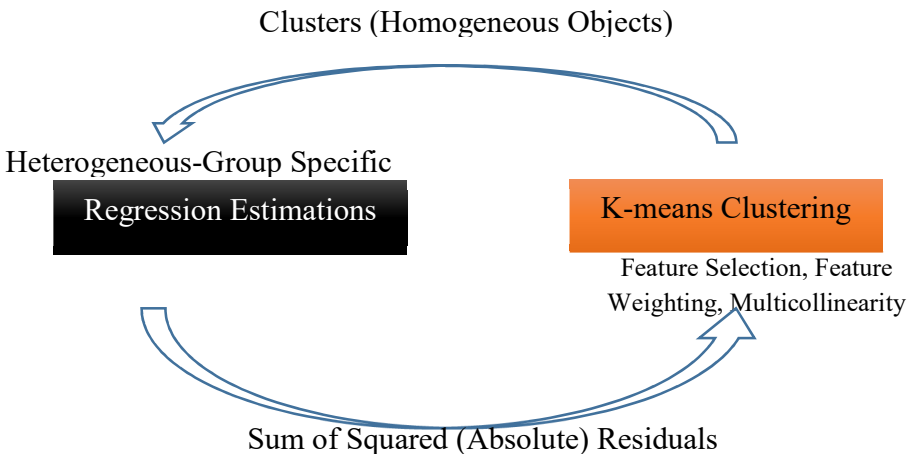


Figure 2.7- Theoretical Framework for the Proposed Clustering

2.4.3 Hypotheses

Clustering algorithms assign observations to the centre of their closest cluster. As a result, if clusters' centres are not far from each other or if observations within clusters are not in close proximity to each other, then it is more difficult to correctly assign entities to their true cluster. Additionally, when clusters' densities (distances between clusters' centres) are low, there is more room to improve the performance of K-means by

increasing the distance between clusters' centres (clusters' densities)[23]. The performance of regression estimations, most particularly the *MAR/MSR*, depends not only on the precision of clustering, but also on the extent of coefficients' differences between clusters. If coefficients are expected to differ significantly between clusters, then running regressions within each cluster can reduce the *MAR/MSR*. Accordingly, the first set of hypotheses to be tested in this thesis involves the identification of those factors that affect the performance of clustering with respect to the precision of cluster recognition and regression estimations. The findings will provide a guide for researchers to consider the feasibility of using CA to solve the problem of HGSC.

H1a: *The higher the clusters' density, the greater the precision for identification of the true clusters when running regressions within clusters. This positive relationship between cluster density and clustering precision is stronger when distances between clusters' centres are lower.*

H1b: *The greater the distance between clusters' centres, the greater the precision for identification of the true clusters when running regressions within clusters. This positive relationship between distances of clusters' centres and clustering precision is stronger when clusters' densities are lower.*

H1c: *The MAR/MSR when running regressions within clusters is lower when (1) the distances between clusters' centres are greater, or (2) the clusters' densities are higher, or (3) the differences of regression coefficients between clusters are larger.*

Correlation between features is a well-known phenomenon, especially in the finance domain. For example, firms with high operating income volatility tend to minimise their financial leverage. Hence these features may be highly (negatively) correlated. While the effect of multicollinearity in regression analysis is well understood, the multicollinearity

[23] The density of a cluster indicates the degree of compactness of the cluster. A cluster with high density indicates that members within the cluster are very close to each other in proximity. In a formal way, a formula to indicate the density at *x, f(x),* is the proportion of entities within a sphere, which is centred at *x* with radius *r*, divided by the volume of the sphere (Wong & Lane, 1983).

problem is different in CA and not easily handled. Features that are highly correlated will automatically get higher weights than others (Sambandam, 2003). In an extreme case when two features are totally collinear, they represent the same underlying feature and this underlying feature attracts twice the weight than it should. Accordingly, the second hypothesis predicts the negative effect of multicollinearity on both CA and regression analysis[24].

H2: *When features are highly correlated, the precision for identification of the true clusters is lower, and the MAR/MSR when running regressions within clusters is higher, as compared to the case of low or uncorrelated features.*

Applications of CA in the business domain typically use K-means with standardized features (e.g. Epure et al., 2011; Li & Li, 2008). With standardization, differences in features' contribution to clustering may not be taken into account. The contribution of a feature to cluster identification is based on two sources, i.e. (1) distances between clusters' centres as measured by this feature and (2) clusters' densities as measured by this feature. In the former case (later case), more relevant features tend to have higher (lower) variances. Recall that standardization tends to reduce the variance of large scale features. As a result, when a feature's weight stems from the *differences of distances between the clusters' centres* as measured by this feature relative to those measured by other features, standardization makes the performance of K-means worse even compared to non-standardization. In contrast, when a feature's weight results from *differences of clusters' densities* measured by this feature relative to those as measured by other features, standardizing clusters' features before running K-means improves the precision for identification of the true clusters because it partly adjusts features' scales in the same direction of the true relative features' weights. Hence, the third hypothesis posits the different effects of standardization on the performance of K-means clustering.

[24] Notice that there are two types of correlation between features: (1) within clusters and (2) within the whole sample. If data are drawn from a distribution where clusters exist, the likelihood of a type (1) correlation is greater. It means that within each cluster, these features are correlated. Although these correlated features are mistakenly over-weighted, they are still relevant features. However, type (2) correlation arises when an irrelevant feature is correlated with a relevant one. Consequently, not only is the relevant feature for clustering over-weighted, but also the irrelevant feature's weight is exaggerated.

H3a*: When a feature's weight results from differences of clusters' densities measured by the feature relative to those measured by other features, standardization improves the performance of clustering as compared to those of unstandardized features.*

H3b*: When a feature's weight results from differences of distances between clusters' centres measured by the feature relative to those measured by other features, standardization decreases the performance of clustering as compared to those of unstandardized features.*

The novel ROWK method proposed in this thesis is developed to tackle these unsolved issues, i.e. the multicollinearity of features and the differences in features' contribution to clustering. To achieve those targets, ROWK employs regression mean absolute (square) residuals as a guide to adjust features' weights and identify clusters when there exist differences between the degrees of contribution of the clustering features. As a result, the weight of a feature reflects its importance not only to cluster identification, but also to improve the regression analysis.

To clarify the last statement, suppose there is a regression model with five group-specific coefficients. Among these five clusters, only regression coefficients in *Cluster 1* significantly differ from those of other clusters. Further, assume that there are five relevant features for clustering, say $z_1, \dots, z_5$. Among them, z_1 and z_2 are more relevant to distinguish the clusters and when running through the WK algorithm, they receive higher weights, say $w_1 = w_2$, and $w_1, w_2 > w_3, w_4, w_5$. Additionally, while z_1 provides relevant information to distinguish between *Cluster 1* and the rest of other four clusters, z_2 helps to distinguish between all clusters, except *Cluster 1*. It turns out that ROWK will place more weight on z_1 than z_2, because its objective is to minimise the *MAR/MSR*. This is rational because the ROWK procedure's ultimate goal is to improve regression results through CA. Accordingly, the fourth hypothesis proposes three channels through which ROWK improves the performance of CA in dealing with the HSGC issue.

H4*: When features have different degrees of contribution to cluster identification and regression estimation, ROWK outperforms generic K-means (both*

standardized and un-standardized) with regard to the precision of cluster recognition and regression estimation. The mechanisms underlying the outperformance of ROWK are through these channels. Specifically, ROWK is hypothesized to:

a. Place more (less) weight on more (less) relevant features.

b. Reduce the influence of the multicollinearity problem by reducing the weights of irrelevant features which are highly correlated with relevant features.

c. Capture relevance not only by its contribution to cluster recognition but also by regression estimation.

2.5 CLUSTERING IN EARNINGS PERSISTENCE - AN APPLICATION OF ROWK CLUSTERING

2.5.1 HGSC on Earnings Persistence

Earnings predictability and earnings persistence play an essential role in equity valuation, financial statement analysis, risk management and asset pricing (Vuolteenaho, 2002). Earnings persistence is a standard proxy to operationalize the concept of earnings quality (Li, 2011). Earnings persistence also affects investors' valuation of stock prices ((Dechow, Ge, & Schrand, 2010). A direct observable channel through which changes in earnings persistence affect firm value is pricing multiples on earnings (Chen, Folsom, Paek, & Sami, 2014). There is empirical evidence on the positive relationship between pricing multiples and earnings persistence (e.g. Ohlson, 1995; Amir, Kama, & Livnat, 2011). Theoretically, the future earnings level and its persistence are key inputs for valuation models, including the widely applied residual income model introduced by Ohlson (1995). In this setting, more persistent earnings lead to an increase in discounted abnormal residual income, and consequently firm value. Empirical research applications include the estimation of implied cost of capital and accounting anomaly studies (Richardson et al., 2010). Given the important role of earnings persistence on equity valuation and asset pricing, it is not surprising that understanding the persistence of earnings and its components has been the focus of several studies (Amir et al., 2011).

Mean-reversion of profitability is well established (Amor-Tapia & Tascón Fernández, 2014; Nissim & Penman, 2001; Wang, 2013). The intuition behind this phenomenon is that competitive forces will gradually erode a firm's abnormal expected risk-adjusted return (Dickinson & Sommers, 2012). However, it is naïve to use a general model of mean-reverting profitability for all firms, given the evidence of heterogeneous group-specific earnings persistence. Earnings volatility is documented as an important variable that negatively correlates with earnings predictability and earnings persistence (Dichev & Tang, 2009). Specifically, assume that firm earnings (E_t) follow a first order autoregressive regression:

(2.12)

$$E_t = \mu + \theta E_{t-1} + \varepsilon_t$$

By taking the variance of this equation and assuming that the variance of earnings is stationary, we have[25]:

(2.13)

$$\frac{d[Var(\varepsilon)]}{dVar(E)} = (1 - \theta^2) - 2\,\theta\, Var(E)\frac{d(\theta)}{d(Var(E)}$$

where *d* denotes the total derivative, *Var(E)* is the proxy for volatility of earnings and *Var(ε)* is inverse proxy for earnings predictability (Dichev &Tang, 2009). From Equation (2.13), the strength of the effect of earnings volatility on earnings predictability depends on earnings persistence itself and the interaction between earnings volatility and earnings persistence. Furthermore, economic or accounting noise in earnings tends to simultaneously increase earnings volatility and reduce earnings persistence (Dichev & Tang, 2009). In addition, high competition between firms also leads to high volatility of operational earnings and lower earnings prescience (Nunes et al., 2010). This negative relation reinforces the base negative relation between earnings volatility and earnings predictability. Indeed, Dichev & Tang (2009) document that earnings of firms with high-

[25] Dichev & Tang (2008) cite prior evidence that over the last 40 years, earnings volatility has roughly doubled. However, for a 1-year horizon, the stationarity assumption is reasonable.

volatility earnings have much lower persistence as compared to firms with low-volatility earnings (0.51 vs. 0.93).

The main point drawn from the above argument is that using the entire sample to estimate the model parameters and make forecasts is not reasonable. In the above example, the true parameters (i.e. earnings persistence coefficients, θ), along with earnings predictability (i.e. $Var\ (\varepsilon)$), are not the same across firms. It is clear that θ and $Var(\varepsilon)$ vary across earnings volatility quintiles as found in Dichev & Tang (2009), leading to invalid regression results when applied to full samples. Consequently, the accuracy of forecasts is poor, not only in in-sample forecasts but also in out-of-sample forecasts.

It is worth emphasising that earnings volatility is not the only factor causing persistence and predictability of earnings to vary across firms. It is well known that the accruals components of earnings are less persistent than those of cash flows and earnings when forecasting earnings (Wang, 2013). This leads to the famous "*accrual anomaly*" originally reported by Sloan (1996) who argues that investors lack an understanding of the low persistence of the accrual component of earnings. The result is inefficient forecasts of future earnings which consequently result in mispricing of stocks. Once again, if we use the general model to forecast earnings as in Equation 2.12, the results would be incorrect because firms with high levels of accruals tend to have less persistent earnings than firms with low levels of accruals.

Firms also exhibit heterogeneity of earnings persistence based on their performance relative to industry average (Amor-Tapia & Tascón Fernández, 2014; Wang, 2013). Notably, earnings persistence of firms varies according to their earnings level relative to the industry average. In particular, there are higher reversion speeds when firm ROEs are lower than industry average. Firms also experience different persistence and convergence patterns of profitability across their life cycles as documented in Dickinson (2011) who employs signals from cash flow patterns to develop a parsimonious indicator of firms' life cycle stages. Using this indicator, she finds that although there is a strong central tendency of profitability, convergence is incomplete because of different patterns of profitability across firms' life cycle stages. Furthermore, there are differences in

persistence of profitability across firm life cycles. Firms in the decline and introduction stages show stronger patterns of mean-reversion than those in mature firms.

Accounting conservatism also affects earnings persistence. Basu (1997) shows one of the first efforts to examine the effect of conditional conservatism on earnings persistence[26]. He finds that conditional conservatism leads to greater timeliness of earnings for bad news over good news. In a subsequent study, Chen et.al. (2014) find that firms with higher conservatism (both conditional and unconditional) have lower earnings persistence.

Earnings management is another factor that negatively effects earnings persistence. Identification of earnings management is important for financial statement users to assess the persistence of earnings. Future profitability is likely to fall (rise) following upward (downward) earnings management since the managed income reverses in the current period (Penman, 2007; Dechow, Richardson, & Tuna, 2003; Jansen et al., 2012).

In summary, there is evidence to support the existence of HGSC in models of earnings persistence. Knowledge of the underlying sources of those heterogeneous coefficients is well addressed in existing research as shown previously. However, solutions developed based on these sources in order to improve forecasting results still represent ad-hoc partitioning techniques or the inclusion of these sources into regression models. For example, to incorporate the effect of earnings volatility on earnings predictability and earnings persistence in order to improve future earnings forecasts, Dichev & Tang (2009) divide firms into quintiles by earnings volatility. They then apply panel data autoregressive regressions of current on past earnings to each quintile. In a similar vein, after controlling for the effect of earnings management, Bauman (2014) addresses the puzzle of persistence of change in profit margin (ΔPM) in forecasting future profitability by partitioning firms based on the direction of ΔPM.

[26] Under current GAAP, conservatism influences the tendency of accountants to practice accounting methods that favour slower recognition of revenues as compared to expenses, and consequently reduce firm net assets (Wolk, Tearney, & Dodd 2001). Condition conservatism is used when certain conditions are satisfied in a specific context (i.e. the arrival of news), while unconditional conservatism is pervasively applied (Chen et.al., 2014).

The intuition behind these methods is that researchers attempt to assign firms into groups according to similar firm characteristics that are considered to impact on firms' earnings persistence. This allows the testing of whether firms in different groups show different patterns in estimated coefficients. Further, more precise coefficients estimates can be obtained in the predictive models because firms in each group are likely to have the same regression coefficients, or at least with little difference. As a result, the models used by both Dichev & Tang (2009) and Bauman (2014) achieve better out-of-sample predictive power than those that aggregate the full sample of firms.

However, there are still concerns with this partitioning method. First, how many groups/clusters/quantiles are needed? Moreover, for a certain number of clusters, how can researchers identify the most homogeneous firms or mathematically the smallest within-cluster variance? These questions cannot be answered simply by assigning firms into quantiles based upon a single variable. Second, it is not only single aspects that contribute to different coefficients in predictive models across firms. Recall that earnings persistence varies across quantiles of earnings volatility and accruals levels. Earnings persistence also differs according to firms' relative earnings positions relative to their industry average and according to stages of firm life cycle. Therefore, to achieve better precision in forecasting future earnings, it is necessary to assign firms into groups such that within groups, firms share similarities with respect to several characteristics such as earnings volatility, earnings management, accruals level, and stage of firm life cycle.

Finally, each partitioning characteristic differs in its ability to identify coefficients' heterogeneity. They also differ in their contribution to revealing the true clusters where regression coefficients are homogeneous. For example, earnings volatility is found to be the most important factor by far to help identify the problem of heterogeneity of earnings persistence (Dichev & Tang, 2009). In this regard, a partitioning method that treats partitioned candidates equally cannot be an optimal choice.

This thesis addresses these issues by employing ROWK clustering as proposed in Section 2.4. There are compelling reasons for using ROWK. First, there is substantial evidence of HGSC in models of earnings persistence. Second, several characteristics explain why models' coefficients differ across groups. Finally, these characteristics exhibit

differences in their ability to identify members of these groups. This thesis introduces the ROWK clustering procedure to address all of these issues.

2.5.2 ROWK's Feature Selection in Earnings Persistence

As discussed in Section 2.3.2.2, the selection process should be based upon underlying theory. Features selected as inputs for clustering must be characteristics that contribute to the distinct behaviours of clusters. In this regard, the selected features for ROWK must relate to characteristics that contribute to distinct firm attributes with respect to persistence of earnings, drivers of future earnings and earnings determinants.

Given the theory and empirical evidence of factors affecting firm earnings persistence, potential features for CA include earnings volatility, firm lifecycle, level of conservatism and earnings management. However, this thesis takes a different approach. Instead, we focus on the use of financial ratios derived from financial statements as a basis for clustering firms. There are several reasons for this. First, firm lifecycle, earnings management and level of conservatism are hard to measure without error (Dickinson, 2011; Chen et al., 2014). Second, several studies have shown the usefulness of financial ratios in predicting future ratios and valuation (e.g. Amor-Tapia & Tascón Fernández, 2014; Bauman, 2014; Fairfield & Yohn, 2001; Lipe, 1986; Ou & Penman, 1989, etc.). In addition, financial ratios are accessible to investors and are easy to calculate with low measurement error. Finally, accounting conservatism, firm lifecycles and earnings management can be identified by using financial statement information. For example, Dickinson (2011) employs cash flow statement information to identify firm life cycles. This approach is superior in identifying differential behaviour in the persistence and convergence patterns of profitability relative to other life-cycle identification methods, such as those of Anthony & Ramesh (1992).

This thesis uses the structural approach of Nissim & Penman (2001) to identify value-relevant financial ratios. These key drivers of firm value are used as a set of relevant features for the application of ROWK. The starting point is from the non-controversial dividend discount model (thereafter DDM) which states that a stock's intrinsic value is measured by the expected future dividends *(E(d))* discounted by a discount rate (*r*) which

reflects security risk. Ou & Penman (1989) state that if the DDM is used as a model for equity evaluation, key drivers of firm value should be determined according to their correlations with *E(d)* and *r*. However, there are concerns with respect to the use of dividends. First, according to Ou & Penman (1989), the ex-ante assessments of dividends by investors cannot be fully observed from the set of realized dividends. Secondly, dividend payouts are arbitrary and driven by tax considerations. These issues, together with the motivation of better utilising accounting information, are put forward as the rationale for the development of the so-called "residual income model" (*RIV*) introduced by Feltham & Ohlson (1995). This model employs a clean surplus relationship to show that equity value can be represented as the sum of the book value of equity (CSE_0) and the present value of residual income or residual earnings:

(2.14)

$$V_0^E = CSE_o + \sum_{t=1}^{\infty}(1 + r_E)^{-t}\,(\overline{CNI_t} - r_E * \overline{CSE_{t-1}})$$

where *CSE* is the book value of equity; *CNI* is the comprehensive income available to common shareholders; and r_E is the required return for common equity. $\overline{CNI_t} - r_E * \overline{CSE_{t-1}}$ is the residual income (RE_t). Bars over variables denote the forecast values. The finite-horizon version of the *RIV* model is presented as:

(2.15)

$$V_0^E = CSE_o + \sum_{t=1}^{T}(1 + r_E)^{-t}\ \overline{RE_t}\ + \frac{\overline{CV_t}}{(1 + r_E)^{-t}}$$

where $\overline{CV_t}$ is the forecast "continuing value" which is the value at time *T* of the residual income beyond *T*. Assuming that the growth rate of residual income is constant at a value of *g* from time *T* onwards, $\overline{CV_t}$ is calculated as:

(2.16)

$$\overline{CV_t} = \frac{\overline{RE_{t+1}}}{r_E - g}$$

Therefore, four key factors of equity value are identified. The first factor comes from the current book value of equity, i.e. CSE_o. The second factor is the expected residual income at the time T horizon. The third factor is the "continuing" value, and the last factor is the required return on common equity. By further decomposing the first three of these factors, the list of key drivers of firm value is as follows[27]: Profit Margin (*PM*); Asset Turnover (*ATO*); Financial Leverage (*FLEV*); Net Borrowing Cost (*NBC*); Operating liabilities leverage (*OLLEV*); Sales Growth (*SALE_GR*). This thesis also adds two more features to proxy for required return on common equity: earnings volatility (*VOL*) and liquidity (*CR*). The addition of these variables avoids the debatable proxies of r_E and instead focuses on accounting variables that affect firm risk as drivers of r_E. The change in asset turnover (ΔATO) and profit margin (ΔPM) are added because ΔATO and ΔPM are shown to provide better information on future profitability and earnings management than their levels (Fairfield & Yohn, 2001; Jansen et al., 2012). Firm size (*SIZE*) and firm age (*AGE*) are also included in the list of features as they are important in identifying firm life cycle and accounting conservatism. Capital expenditure (*CAPX_DEF*) and dividend payout (*DIV*) are further added to the list as indicators of firm investment and dividend decisions. Relative to non-high-tech firms, both economic and accounting (conservatism) factors cause high tech firms to exhibit lower levels of earnings persistence (Kwon & Yin, 2015). Accordingly, the feature list also includes investment in intangible assets (*INTAN_INV_DEF*). Finally, the absolute value of accruals (*AB_ACC_DEF*) and absolute level of earnings (*AB_EARNINGS_DEF*) are included to reflect the importance of conservatism and earnings management. The suffix '*DEF*' denotes that the variable is deflated by average assets. See Table 2-1 for descriptions of the full set of clustering features.

[27] See Nissim & Penman (2001) for details of the decomposition.

Table 2-1: Variable Definitions

#_number denotes the COMPUSTAT data item number.

Variables	Definition and Formula
FA_t	Financial Assets = *cash and short term investments (#1) + investments and advances-other (#32);*
FO_t	Financial Obligations = *debt in current liabilities (#34) + long term debt (#9) + preferred stock (#130) - preferred treasury stock (#227) + preferred dividends in arrears (#242)*
OA_t	Operating Assets = *total asset (AT, #6) – FA_t*
OL_t	Operating Liabilities = *total liabilities (LT, #181) – FO_t*
NFO_t	Net Financial Obligations = FO_t - FA_t
CSE_t	Common Equity = *common equity$_t$ (#60) + preferred treasury stock$_t$ (#227) -preferred dividends in arrears$_t$ (#242);*
NOA_t	Net Operating Assets = *NFO_t + CSE_t*
$\overline{NOA_t}$	Average Net Operating Assets $_t$ = *(NOA_t – NOA_{t-1})/2;*
OI_t	Operating Income = *operating income after depreciation (#178);*
OI_DEF_t	Deflated earnings = OI_t/ $/\overline{AT_t}$
$RNOA_t$	Return on Net Operating Assets = *OI_t /$\overline{NOA_t}$*
$SALE_t$	Net Sales (*#12)*
$SALE\_GR_t$	Sales growth = *($SALE_t$-$SALE_{t-1}$)/$SALE_{t-1}$;*
PM_t,	Profit Margin = *$OI/sales_t$*
ΔPM_t	Change in Profit Margin = *PM_t - PM_{t-1}*
ATO_t	Asset Turnover = *$SALE_t/\overline{NOA_t}$*
ΔATO_t	Change in Asset Turnover = *ATO_t - ATO_{t-1}*
VOL_t	Earnings Volatility = *the standard deviation of the deflated earnings for the most recent 5 years*
CR_t	Current Ratio = *Current Assets$_t$ (#4)/Current Liabilities$_t$ (#5)*
$CAPX\_DEF_t$	Capital Expenditures (#128)/ $\overline{AT_t}$
$INTAN\_INT_t$	Intangible investment = *R&D Expense$_t$ (#46)+ Advertising Expense$_t$ (#45) + Intangibles$_t$ (#33) - Intangibles$_{t-1}$ + Amortization of Intangibles (#65)*
$FLEV_t$	Financial Leverage$_t$ = *NFO_t/NOA_t*
$OLLEV_t$	Operating liabilities leverage$_t$ =*OL_t/OA_t*
NBC_t	Net Borrowing Cost$_t$ = *Interest Income$_t$ (IINT,#62) – Interest Expense$_t$ (XINT, #15*)
$ABS\_ACC\_DEF_t$	Deflated Absolute Level of Accruals = $\left|\right.$[*earnings (IBC,#123)- cash flows from operation (OCF, #308)*]/ $\overline{AT_t}$ $\left|\right.$
ABS_EARNINGS_DEF$_t$	Deflated Absolute Level of Earnings
$SIZE_t$	Natural logarithm of firm size
AGE_t	Firm age since the first time appeared in COMPUSTAT
DIV_t	Dividend payout = Common Dividend (#21)/ Net Income (#172)

2.5.3 Hypotheses

Several factors contribute to differences in earnings persistence. However, some factors are documented to be more relevant to identify differences in earnings persistence than others. Particularly, Dichev & Tang (2009) find that earnings volatility outperforms other examined variables, including level of accruals, cash flow volatility, and earnings level with regard to distinguishing earnings persistence and achieving greater earnings predictability. ROWK clustering is constructed to place more weight upon more relevant features, and to assign less or zero weight to irrelevant ones. Therefore, this thesis hypothesizes that:

H5*: Feature weights identified by ROWK clustering are not equal.*

The ultimate aim of ROWK is to assign firms to clusters such that firms within each cluster are homogenous in term of characteristics that cause differences in earnings persistence. Given the evidence of HGSC in earnings persistence discussed in previous sections, this thesis hypothesizes that:

H6*: Firms exhibit different earnings persistence between ROWK clusters.*

One of the benefits from running ROWK clustering relative to traditional variable partitioning is that ROWK clustering simultaneously utilises information from all cluster features. This benefit is further strengthened by the ROWK mechanism to identify the contribution of relevant features. Consequently, using ROWK clustering in earnings persistence helps to achieve better distinguishable earnings persistence and lower earnings prediction errors. Accordingly, the seventh hypothesis predicts that:

H7*: ROWK clustering results in larger differences in earnings persistence between clusters and lower earnings prediction errors than a single variable cluster partitioning technique.*

Industry classification, conservatism, earnings management, and firm life cycles have been found to affect firm earnings persistence (e.g. Chen et al., 2014; Dickinson, 2011; Gupta & Huefner, 1972; Jansen et al., 2012; McNichols, Rajan, & Reichelstein, 2014). More importantly, accounting information from financial statements has been shown to

be useful to classify industry and firm life cycles, to recognize earnings management and to proxy for the level of conservatism. ROWK clustering aims to find clusters exhibiting differences in earnings persistence based on the information from financial statements. As such, heterogeneities of level of conservatism, earnings management and firm life cycles, and industry classification are likely to be observed across clusters identified by ROWK clustering. Accordingly, the next hypothesis posits that:

H8*: The clusters found in ROWK clustering exhibit heterogeneities with respect to accounting conservatism, earnings management, firm life cycles and industry membership.*

Financial analysts typically proxy for sophisticated users of financial information (Dichev & Tang, 2009). On the one hand, to the extent that financial analysts are able to rationally utilise information from financial statements and are able to understand cluster patterns in earnings persistence, differences in earnings persistence across firms are expected to be embedded in analyst forecasts. On the other hand, a number of studies produce evidence of systematic biases in investors and analysts' forecast errors, which suggests that investors and analysts do not fully impound the implications of existing information (Ball & Bartov, 1996; Dichev & Tang, 2009; Monte-Mor, Galdi, &Costa, 2018; Ulupinar, 2018). Accordingly, if analysts do not fully understand and incorporate the cluster patterns in earnings persistence found by ROWK, then the application of ROWK cluster information would uncover predictable patterns in their forecast errors.

However, another possibility is that analysts understand the implications of clusters on earnings persistence but their forecasts errors are still predictable due to informational advantages, reputation or career incentive concerns (Bartholdy & Feng, 2013; Dichev & Tang, 2009). For example, firms with low earnings persistence and/or exhibiting a downward trend of profitability may suffer from higher information uncertainties and asymmetries. Sell-side analysts may upwardly bias their forecasts for these firms, hoping to gain access to internal data (Dichev & Tang, 2009). Nevertheless, while a finding of predictable forecast biases from the application of ROWK may be evidence of the superiority of this technique in exploring the problem of HGSC on earnings persistence that are not embedded in analysts' models of earnings persistence, this thesis does not

attempt to identify the possible causes underlying analysts' forecast biases. This may be a fruitful area for future studies to investigate. Rather, an aim of this thesis is to identify predictable forecast errors in earnings persistence using the results from ROWK clustering. Accordingly, to the extent that ROWK clustering provides an innovative solution to explore the existence of patterns on earning persistence, the final hypothesis proposes that:

H9*: Information from ROWK's cluster identification predicts analyst forecast errors.*

2.6 CHAPTER SUMMARY

This chapter discussed the theoretical background and the existing literature on the two research aims of this thesis. The resulting hypotheses are summarised in Table 2-2.

Table 2-2 Summary of Research Aims and Hypotheses

RA1: Develop a new clustering approach that addresses the inherent problems of current clustering techniques when dealing with the issue of HGSC.
Factors that affect the performance of cluster analysis to deal with the issue of HGSC
H1a: The higher the clusters' density, the greater the precision for identification of the true clusters when running regressions within clusters. This positive relationship between cluster density and clustering precision is stronger when distances between clusters' centres are lower.
H1b: The greater the distance between clusters' centres, the greater the precision for identification of the true clusters when running regressions within clusters. This positive relationship between distances of clusters' centres and clustering precision is stronger when clusters' densities are lower.
H1c: The MAR/MSR residuals when running regressions within clusters is lower when (1) the distances between clusters' centres are greater, or (2) the clusters' densities are higher, or (3) the differences of regression coefficients between clusters are larger.
H2: When features are highly correlated, the precision for identification of the true clusters is lower, and the MAR/MSR when running regressions within clusters is higher, as compared to the case of low or uncorrelated features.

H3a: When a feature's weight results from differences of clusters' densities measured by the feature relative to those measured by other features, standardization improves the performance of clustering as compared to those of unstandardized features.

H3b: When a feature's weight results from differences of distances between clusters' centres measured by the feature relative to those measured by other features, standardization decreases the performance of clustering as compared to those of unstandardized features.

Channels that contribute to the superior performance of ROWK clustering to address the problem of HGSC

H4: When features have different degrees of contribution to cluster identification and regression estimation, ROWK outperforms generic K-means (both standardized and un-standardized) with regard to the precision of cluster recognition and regression estimation. The mechanisms underlying the outperformance of ROWK are through these channels. Specifically, ROWK is hypothesized to:

a. Place more (less) weight on more (less) relevant features.

b. Reduce the influence of the multicollinearity problem by reducing the weights of irrelevant features which are highly correlated with relevant features.

c. Capture relevance not only by its contribution to cluster recognition but also by regression estimation.

RA2: Apply the proposed innovative clustering method to address the problem of HGSC in the model of earnings persistence.

H5: Firms exhibit different earnings persistence between ROWK clusters.

H6: Feature weights identified by ROWK clustering are not equal.

H7: ROWK clustering results in larger differences in earnings persistence between clusters and lower earnings prediction errors than a single variable cluster partitioning technique.

H8: The clusters found in ROWK clustering exhibit heterogeneities with respect to conservatism, earnings management and firm life cycles and industry membership.

H9: Information from ROWK's cluster identification predicts analyst forecast errors.

In the next chapter (Chapter 3), the data and methodology employed to test the hypotheses relating to the two research aims developed in this chapter will be presented.

CHAPTER 3

DATA AND METHODOLOGY

3.1 INTRODUCTION

This chapter outlines the data and methodology employed to test the hypotheses relating to the two thesis research aims developed in the previous chapter (summarised in Table 2-2 in Section 2.6). It begins with Section 3.2 which is dedicated to the methodology employed in the thesis. Section 3.2.1 presents the econometric framework for the innovative clustering technique, called the Regression Oriented Weighted K-means clustering (ROWK), which is developed to address the shortcomings of K-means clustering when dealing with the heterogeneous group-specific coefficients issue (HGSC). This section is further divided into two sub-sections. Section 3.2.1.1 describes the econometric framework and the procedure for the execution of ROWK. Next, Section 3.2.1.2 discusses the research design for testing Hypotheses H1 to H4, relating to the first research aim. Three case studies are presented to shed light on each channel through which ROWK improves the performance of CA when dealing with the HSGC problem.

Section 3.2.2 presents the research design for the application of ROWK clustering to earnings persistence. The model for earnings persistence is presented and some adjustments for data processing are discussed. These adjustments deal with some concerns relating to using real financial data to run ROWK clustering.

Subsequently, Section 3.3 gives details of the data used in testing. Hypotheses relating to the first research aim are tested using simulated data. Then real data regarding earnings persistence is used to test hypotheses relating the second research aim. Finally, Section 3.4 provides a summary of this chapter.

3.2 METHODOLOGY

This section presents analytical models employed to test the hypotheses relating to the two thesis aims identified in Chapter 1.

3.2.1 ROWK and the Problem of HGSC

3.2.1.1 ROWK- Econometric Framework and the Executing Procedure

This section presents the econometric framework for the proposed innovative clustering method called Regression oriented Weighted K-means (ROWK). Subsequently, a procedure to implement ROWK is introduced. Tests for Hypothesis H1 to H4 using simulated data conclude this section.

3.2.1.1.1 The HGSC model

Let i=1,…,N represent an index of observations. For simplicity, this thesis only considers the case of cross-sectional data. For panel data, nothing changes except that "i" is replaced by "it" where t is an index of time[28]. The response variable of the i-th unit, y_i is expressed as follows:

(3.1)

$$y_i = \alpha_i + x_i'\beta_{(i)} + u_i, i = 1, \dots, N$$

where x_i is a $P\times1$ vector of explanatory variables and u_i is the unit-specific error. $\alpha_{(i)}$ and $\beta_{(i)} = \left(\beta_{(i)1}, \beta_{(i)2}, \dots, \beta_{(i)P}\right)'$ are 1×1 and $P\times1$ vectors of intercepts and slope coefficients for unit i respectively. A group effect is modelled by allowing $\alpha_{\xi_k^0}$ and $\beta_{\xi_k^0} = \left(\beta_{\xi_k^0,1}, \beta_{\xi_k^0,2}, \dots, \beta_{\xi_k^0,P}\right)'$ be 1×1 and $P\times1$ vectors of group-specific intercept and slope coefficients such that $\alpha_{(i)}$ and $\beta_{(i)}$ equal or closely approximate $\alpha_{\xi_k^0}$ and $\beta_{\xi_k^0}$ respectively

[28] Note that the two-step Pseudo Threshold approach introduced by Lin and Ng (2012) is only executed on panel data because it has to run individual time-series regressions. Our framework can apply to both cross-sectional and panel data, so it can be applied in the case of unavailability of individual time-series. Additionally, it allows for the case where firm i can move to different clusters over time.

for all i's in ξ_k^0 [29]. $\xi_1^0, \xi_2^0, \ldots, \xi_K^0$ are the corresponding true K^0 clusters. The components '^{o}' and '$^{\wedge}$' indicate the true and estimated value/membership respectively.

Further, assume that a member of a cluster is represented as a point with V dimensions. Each dimension is a feature that helps to distinguish members between clusters. Specifically, there is a multivariate input data set Z that is represented as an $N \times V$ matrix, where V is the number of cluster features. $z_i = (z_{i1}, z_{i2} \ldots, z_{iV})^T$, $i=1,\ldots,N$. K^0 represents the true number of clusters (which is unknown and fixed). $\xi_1^0, \xi_2^0, \ldots, \xi_K^0$ are the corresponding true K^0 clusters with centres $c_1, c_2, \ldots, c_{K^0}$ respectively where $c_k = (c_{k1}, c_{k2}, \ldots, c_{kV})'$ and $k = 1, \ldots, K^0$. Let N_k^0 be the true number of cross-sectional units within group k so that $N = \sum_{k=1}^{K} N_k^0$.

The following assumptions will be made: (i) $u_i \sim (0, \sigma^2)$ has finite second moments and has cross-sectional and serial independence, i.e. $cov(u_{ij}) = \sigma^2 I$ where I is the identity matrix; (ii) u_i is independent of z_i' for all k=1,…, K^0. Assumptions (i), (ii) imply that the model is correctly specified and can be consistently estimated within each true cluster (Lin & Ng, 2012). Our objective is to estimate $\beta_{\xi_k^0}$ and $\alpha_{\xi_k^0}$ without knowing ξ_k^0. This can be achieved by the proposed ROWK procedure that aims to pool similar observations for estimation.

3.2.1.1.2 The Regression Oriented Weighted K-means (ROWK)

The K-means algorithm is the most popular method in the family of centroid approaches. It is an iterative procedure aiming to minimise the within-cluster dissimilarity as measured by Euclidean distance. The ultimate goal of cluster analysis is to discover the true cluster structure. In this regard, choosing relevant features and deciding upon their weights are critical parts to ensure success (Brusco & Cradit, 2001). This thesis proposes a novel clustering which combines K-means clustering and regression analysis. On the one hand, the K-means algorithm iteratively assigns similar observations into clusters.

[29] We allow for both the intercept and slope coefficients to be group-specific. For panel data with unobserved heterogeneity (α_i), we can transform the original data into demeaned data (i.e. $\overline{y_{it}} = y_{it} - \frac{1}{T}\sum_{t=1}^{N} y_{it}$ and $\overline{z_{it}} = z_{it} - \frac{1}{T}\sum_{t=1}^{N} z_{it}$). Then we have a model with no intercept and only group-specific slope coefficients.

On the other hand, the *MAR/MSR* from running regressions are used to guide the process of weight adjustment in clustering. This innovative procedure is referred to as Regression-Oriented Weighted K-means Clustering. The procedure includes four steps. First, the regression model is identified. Second, features are selected. Then data is pre-processed before being used in the final step to find the optimal weights.

Step 1- Specifying the regression model

The first step in the ROWK procedure is to specify the regression model that exhibits the problem of HGSC. With the focus on the clustering procedure, it is assumed that the problem of HGSC is correctly identified as presented in Section 3.2.1.1.1[30]. It means two things. First, the model is correctly specified and can be consistently estimated if true clusters, which are fixed and unknown, are correctly identified. Second, the cluster memberships are represented by the list of cluster features. The ultimate task, now, is to determine how to assign objects correctly into clusters.

Step 2- Feature selection

After the model of HGSC is presented, the next step of the ROWK procedure is to identify a list of cluster features as the input for the clustering process. As the key input for clustering processes, feature selection is considered an essential element that determines the quality of clustering results. The clustering procedure can be described in Figure 3.1.

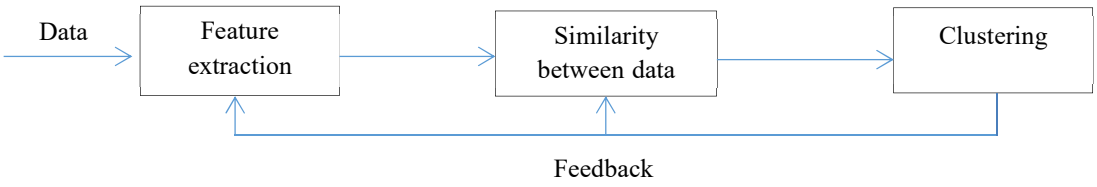


Figure 3.1- Steps of the clustering process.

Source: Gungor & Unler (2008, p. 1116)

[30] It is also assumed that any issue of omitted variables is eliminated through the identification of clusters. In other words, within each cluster the coefficients are homogeneous and error terms are uncorrelated with the independent variables.

The process starts with raw data and ends up with clusters as the result. Feature extraction is comprised of feature selection and weighting (Gungor & Unler, 2008). Feature selection attempts to determine which features are relevant for the true underlying clusters. On the other hand, the weighting process highlights those features that are more important to identify clusters.

For robust clustering, the selection process should strongly connect with underlying theory. Features selected as inputs for clustering have to be characteristics that contribute to the distinct behaviour of clusters. Furthermore, in this thesis it is argued that the problems of inductive approaches can be mitigated, since the ROWK procedure aims to place low or zero weights on irrelevant variables. For the next part, assume that throughout the feature selection process, there are V cluster features[31]. The input variables for the clustering process will be represented by an $N \times V$ matrix, $z_i = (z_{i1}, z_{i2}..., z_{iV})^T$, $i=1,...,N$.

Step 3- Pre-processing data

After specifying the regression models and identifying the list of cluster features, all regression variables and cluster features must be pre-processed before running CA. The purpose of data pre-processing is to mitigate the effect of outliers and to transform the input features for clustering in a fair way. The standard procedure to address the first issue is to winsorize the data, typically at 1% and 99% (e.g. Cooper, Gulen, & Schill, 2008). For the second problem, the reason for standardization before running clustering is simple. Typically, with real data some features display large scales, and consequently the results depend on these features disproportionately more than their true contribution to clustering. The most popular solution is to transform all partitioning features into z-scores with zero means and unit standard deviations. Tanioka & Yadohisa (2012) conduct tests both with real and simulation data using six ways of standardization, and find that a ranking method is the most effective for K-means clustering. However, the performance

[31] For simplicity, this study assumes that after the feature selection process, exactly V features are identified. In real cases, the number of features used in clustering tends to be greater than V. Our assumption is acceptable and less harmful than the case of omitting relevant features because ROWK is built to place low or zero weights on irrelevant features.

of standardization methods seems to be sensitive to the distribution of the features, as Steinley (2004) documents that $z_i/\max(z_j)$ and $z_i/[\max(z_i - \min(z_i)]$ are the most effective when x is drawn from a normal distribution.

However, there are circumstances when it is not necessary to standardize, for example, when variables' scales only differ because of different units of measurement, such as *cm* vs. *m*. As a result, transforming the variables' scales into the same unit is all that is needed. The second situation emerges when the differences in features' scales are reflections of their true nature of weighting. Put another way, a feature may have a large variance because its centroids are far from each other. This feature contributes richer information for clustering compared to others. Consequently, the standardization that aims to lower the scale of this feature leads to poorer performance compared to unstandardized data as proposed in Hypothesis H3.

For these reasons, unless there are large discrepancies between features' scales that do not come from different units of measurement, the unstandardized data (with appropriate winsorizing) is used in clustering. Additionally, as the ROWK procedure adjusts weights according to the features' contributions to identify clusters, it automatically addresses the scale issue.

Step 4- Finding Optimal Weights of Cluster Features

Let w_v $(v = 1, \dots, V)$ denote the corresponding weight of feature z_v. Then, weighted K-means clustering attempts to assign *N* data points into *K* disjointed clusters such that the *SSR* criterion, *J*, is minimised:

(3.2)

$$J = \sum_{k=1}^{K} \sum_{v=1}^{V} \sum_{i \in \xi_{k'}} w_v^{\gamma} \, \|z_{iv} - c_{k'v}\|^{\theta}$$

$$\text{Subject to } w_v \geq 0 \; and \; \sum_{v=1}^{V} w_v = 1$$

where γ is a parameter representing the rate of the effect of the weights on its contribution to the distance. ξ_k' denotes the cluster k resulting from the clustering procedure and does

not necessarily coincide with the true cluster, i.e. ξ_k^0. $\|.\|$ is a norm. Recall that the components 'o' and '' indicate the true and estimated value/membership respectively. If the norm is the Euclidean metric with $\theta = 2$ and $w_v = 1$ for all v=1,…,V, then it is the generic K-means. If the norm is the Euclidean metric with $\theta = 2$ and $\gamma = 1$, then it is the weighted K-means (WK) introduced by Huang et al. (2008). If the norm is the Minkowski θ-metric with $\theta = \gamma$, then it is the Minkowski metric weighted K-means (iMWK) developed by Amorim & Mirkin (2012).

In this thesis, the norm is chosen as the Euclidean metric with θ=2 and γ=1. The main innovation of ROWK is that it does not seek the set of optimal weights $\{w_1^*, \dots, w_V^*\}$ that minimises the function J as in previous research on weighted K-means. Instead, it seeks to identify a set of $w_v^*, v = 1, \dots, V$ such that when applying these weights to K-means clustering, finding clusters and regressing within each cluster, the mean of absolute residuals (*MAR*) is minimised:

(3.3)

$$MAR = \sum_{k=1}^{K} \sum_{i \in \xi_k'} \left| y_i - \hat{\alpha}_{\xi_{k'}} + x_i' \widehat{\beta_{\xi_k'}} \right| / N$$

where $\hat{\alpha}_{\xi_{k'}}$ and $\widehat{\beta_{\xi_k'}}$ are the estimations of $\alpha_{\xi_k^0}$ and $\beta_{\xi_k^0}$, derived from running a regression of model (3.1) within each cluster found by K-means clustering for a certain set of weights, $w_v', v = 1, \dots,$. Applying the weights to K-means implies that features are rescaled based on the square root of corresponding weights. To mitigate the effect of outliers and the brevity of the ROWK algorithm that will be presented later, the thesis chooses the mean of absolute residuals (*MAR*) instead of the mean of squared residuals $\left(MSR = \sum_{i=1}^{N} \left| \left(y_i - \hat{y}_i \right)^2 \right| / N \right)$. In simulation results (unreported), it moderately improves the performance of ROWK. In Chapter 5, presenting an application of ROWK on earnings persistence models, *MSR* is examined since all features are ranked, ruling out the concern of outliers.

The Optimization Problem

Finding solutions of an optimization problem is not an easy task. The optimization process typically involves either regression criteria (i.e. the log-likelihood function as in Equation 2.3 or 2.4 or clustering criterion, i.e. the sum of the weighted Minkowski metric as in Equation 3.2). Typically, these optimization problems do not have analytical solutions, which means that it is impossible to formulate estimated parameters explicitly as a function of data. Instead, these optimization problems require numerical algorithms for the solutions.

Basically, optimization algorithms consist of two important tasks (Amaran, Sahinidis, Bikram Sharda & Bury, 2016). The first task is to generate new guesses of the solution that typically can be derived from either derivatives or numerical derivatives. The second task is to choose the convergence criterion at which the guess is good enough and the algorithm stops. Indeed, the properties of the log-likelihood function to be optimised, together with the algorithm's properties, often ensure that the proposed solution converges to the true solution. In other words, the difference between the proposed solution and the true solution can be achieved as small as the desired value by implementing a sufficiently high number of iterations (Taboga, 2017). However, there are still instances when the convergence value cannot be reached. In fact, this numerical convergence cannot always be guaranteed on a theoretical basis due to the difficulty of understanding the optimization function or the insufficiency to prove numerical convergence, given the chosen algorithm (Taboga, 2017).

For the optimization of regression criteria, this involves a nonlinear optimization problem that is well addressed in the field of econometrics (Quandt, 1972)[32]. Notably, Quandt (1972) documents that the maximisation algorithm of the log-likelihood function in Equation 2.5 fails to reach the convergence criterion in 22 percent of all replications. The failures of the algorithm are caused when either the number of permitted iterations are exceeded or the matrix of second partial derivatives are not negative definite. With respect to the optimization of weighted clustering criteria, numeric algorithms are also employed to solve the (constraint) optimization problem. Specifically, Huang et al.

[32] There is a further assumption on D_i as in Equation 2.7 to make the optimization problems of the (log) likelihood function become computationally tractable.

(2008) propose an algorithm to minimise the sum of weighted squared distance by repeatedly assigning objects to clusters whose centroids are closest to these objects. Then weights are adjusted based on a formula that is basically the solution of w_v given the centroids and clusters use the first order optimality condition for the Lagrange function[33].

In the ROWK procedure proposed in this thesis, finding solutions for an optimization problem is an even more challenging task. While ROWK aims to find the optimal features' weights, the optimization function relates to the regression criterion, (i.e. the *MAR*s as in Equation 3.3. Consequently, the feature weights are not embedded in the optimization function, making it impossible to employ (numerical) derivatives to form new guesses for the parameter value.

To overcome this problem, the thesis pioneers an innovative procedure to find new guesses for the set of cluster feature weights. This basic idea is that the new guess for the set of feature weights must reduce the *MAR* relative to the current set. Specifically, for a starting set of feature weights, say $W^{(0)} = \left\{w_1^{(0)}, w_2^{(0)} \dots, w_V^{(0)}\right\}$, K-means clustering is run using this set of weights (applying the weights into K-means clustering implies that features are rescaled based on the square root of corresponding weights, and J is minimised), the clusters are saved, then the regression is run within each cluster, and the corresponding mean of absolute residual, $MAR^{(o)}$ is calculated. The symbol '$^{(s)}$' indicates the step-sth value during the execution of ROWK procedure. In the next step, each feature weight is adjusted by a certain small percentage change, say Δ:

(3.4)

$$w_v^{(1,u)} = w_v^{(0)}(1+\Delta) \text{ or } w_v^{(1,d)} = w_v^{(0)}/(1+\Delta)$$

Δ is the predefined certain small percentage change in the feature weights. In the thesis's proposed algorithm, the range of Δ is from 10% to up to 1000%. For *V* features, there are a total of *2V* sets of adjusted weights:

[33] See footnote 17 for the formula. For more details, see Amorim and Mirkin (2012, p.1065)

(3.5)

$$W_v^{(1,u)} = \left\{w_1^{(0)}, \dots, w_v^{(1,u)}, \dots, w_V^0\right\} \text{ and } W_v^{(1,d)} = \left\{w_1^{(0)}, \dots, w_v^{(1,d)}, \dots, w_V^{(0)}\right\}, v=1,\dots,V.$$

Applying these $2V$ adjusted set of weights to K-means clustering derives the corresponding $2V$ mean of the absolute residuals, i.e. $\left\{MAR_s^{(1)}\right\}, s = 1, \dots, 2V$. Then the set of adjusted weights corresponding to the lowest MAR among the set of $\left\{MAR_s^{(1)}, MAR^{(o)}\right\}$, is updated as the new starting set of weights. The routine is run iteratively until the convergence criterion is satisfied.

Initial Set of Weights

It is well-documented that a poorly designated initial set of centroids leads to poor performance of K-means clustering (Qian, 2006). The same is true for the ROWK procedure where an inappropriate initial starting set of cluster features' weights causes an undesirable result of local optimization. This negative effect is even stronger because of the inertia of feature weights, especially with the existence of irrelevant features and a high number of features. Particularly, during the execution of the above-proposed procedure, this study finds that when all features are included, the number of iterations to satisfy the convergence criterion is quite small (less than 50 iterations), resulting in fast executions. However, the results are extremely poor in the sense that there is little change in the weights relative to the starting value, and there are still significant weights placed upon irrelevant features.

To address this issue, this thesis proposes two potential solutions. The first solution is to pick a randomly-selected set of weights in the sense that the range of weights could cover as wide a range as possible. This allows us to attain all possible values of sets of weights (which is unlimited) through a limited number of trials. The second solution is the employment of the stepwise (forward) procedure, which is a popular technique of model specification (Hwang & Hu, 2015). The reason for the employment of the stepwise procedure is to mitigate the inertia effect of feature weight adjustments on the optimization function. Specifically, instead of running CA for the whole set of features' weights, the procedure is started at only two features. Subsequently, each feature is added

until all features are employed. The order of features is crucial. In line with standard stepwise regression procedure, the ordering of features is based on the contribution to the reduction of *MAR*. It means that K-means clustering is run with only one feature, and the corresponding *MAR* is calculated. Then the feature with the lowest *MAR* ranks first and vice versa.

Empirical results using simulated data (unreported) show that the random-sets-of-weights solution underperforms the stepwise approach. Therefore, this thesis employs the stepwise approach as the main solution to address the issue of determining the initial set of weights.

Convergence Criterion

The algorithm stops when the new guesses produce only minimal reductions of *MAR*, which means that only negligible improvements of the solution can be achieved by performing new iterations. Given the aforementioned inertia effect of weight adjustment on *MAR*, the convergence value is set as zero without a large negative impact on the time to run the procedure.

Inclusion/Exclusion of Features

In stepwise regression, the inclusion or exclusion of a variable is based on some pre-specified criterion such as F-tests, t-tests or Akaike information criterion (Hwang & Hu, 2015). However, in ROWK, it is not the inclusion or exclusion of a regression variable to consider but rather, a cluster feature. As a result, it is not possible to employ the above criteria to decide which features to include after running ROWK. This study uses two ways to tackle this issue. In the first approach, a feature is only included in clustering if the inclusion of this feature reduces the *MAR* relative to those without this feature. In the second approach, a graph of values of the MAR for a given number of features against this number of features is used. Then, the 'knee point' at which the graph starts to flatten is used to decide which features are relevant and to be included. For brevity, only the first approach is presented for the below ROWK algorithm.

The Proposed Optimization Algorithm

The algorithm for the ROWK procedure is as follows. Let K' and K^0 represent the number of clusters used during the clustering process and the true number of clusters, respectively. K^{max} is the pre-defined maximum number of clusters, and $\xi'_k ; k = 1, \dots, K'$ is the cluster k identified during the clustering process[34]. Optimal weights of clustering features are estimated through the following algorithm:

For each number of clusters K'=1,…,K^{max}, run:

Step 1: Ranking features,

- Step 1.1: For each feature $z_v, v = 1, \dots, V$
  - Run K-means with only one feature- $z_{i,v}$, get $\{\xi'_1, \xi'_2, \dots, \xi'_{K'}\}$
  - Run a regression of Equation 3.1 for each $\xi'_k, k = 1, \dots, K'$, and calculate the mean of the absolute residuals, *MARs* using Equation 3.3.
- Step 1.2: Rank features based on *MAR*. The first ranking is the feature having the lowest *MAR* and vice versa. Without loss of generality, assume $z_1 -$ rank 1^{st}, $z_2 -$ rank 2^{nd}, …, $z_V -$ rank V^{th}.

Step 2: Finding optimal weights. For the ordered list of clustering features $\{z_1, \dots, z_V\}$, let $w^*_{1,2} = 1$. Repeat for *j=2,…,V*.

- Step 2.1: Pick the first j features from the list, i.e. $z_1,\dots, z_j$; set w_1=$w^*_{1,j}$;w_2=$w^*_{2,j}$; …;w_{j-1}=$w^*_{j-1,j}$; $w_{j,j}$= $w^*_{j,j}$=1. Run K-means with this set of $\{w_v\}, v = 1, \dots, V$[35]. Save the clusters $\{\xi'_1, \xi'_2, \dots, \xi'_{K'}\}$, then run the Equation 3.1 regression for each cluster, and calculate the mean of the absolute residual, denoted by MAR^*_j.
- Step 2.2: Run K-means with w_1= $w^*_{1,j}$;w_2=$w^*_{2,j}$; …;w_{j-1}=$w^*_{j-1,j}$; $w_{j,j} = w^*_{j,j}$. Save the clusters $\{\xi'_1, \xi'_2, \dots, \xi'_{K'}\}$ and then run the Equation 3.1 regression for

[34] Typically, 10 are considered as an appropriate maximum number of clusters in finance areas.

[35] Applying the weights to K-means implies that features are rescaled based on the square root of corresponding weights. In addition, to satisfy the constraint that sum of weights equal 1, features are rescaled based on the following formula: $Z_v^{rescaled} = \sqrt{\frac{w_v}{\sum_{v=1}^{V} w_v}}$

each cluster, and calculate the mean of the absolute residual, denoted by MAR_j^0.

- Step 2.3: Denote Δ as the minimum percentage change of weights. Set Δ= 10%. Repeat for η=1,…,j.
  - Fix all weights except w_η, (i.e.$\{w_v\}, v = 1, \dots, j$, and $v \# \eta$). Set $w_\eta = w_{\eta,j}^*(1+\Delta)$ or $w_\eta = w_{\eta,j}^*/(1+\Delta)$. Run K-means with these sets of weights; save sets of clusters$\{\xi_1', \xi_2', \dots, \xi_{K'}'\}$. Run the Equation 3.1 regression for each set of clusters, and calculate corresponding *MAR*s, denoted by$\{MARs_{w_\eta}\}$[36].
- Step 2.4:
  - If $min\{MAR_j^0; MARs_{w_\eta} : \eta = 1, \dots, j\} \neq MAR_j^0$, then let the weights corresponding to the lowest MAR be $\ddot{w}_{1,J}, \dots, \ddot{w}_{J,J}$, and update $w_{1,j}^* = \ddot{w}_{1,J}, \dots, w_{j,j}^* = \ddot{w}_{J,J}$. Then return to Step 2.2.
  - If $min\{MAR_j^0; MARs_{w_\eta} : \eta = 1, \dots, j\} = MAR_j^0$ then repeat Step 2.3, but replace Δ by $\lambda * \Delta$; $\lambda = 2, \dots, 100$.
    - If at $\lambda = \lambda^* \in [2;100]$ that $min\{MAR_j^0; MARs_{w_\eta} : \eta = 1, \dots, j\} \neq MAR_j^0$, stop at $\lambda = \lambda^*$. Let the weights corresponding to the lowest MAR be $\dddot{w}_{1,J}, \dots, \dddot{w}_{J,J}$, and update $w_{1,j}^* = \dddot{w}_{1,J}, \dots, w_{j,j}^* = \dddot{w}_{J,J}$. Then return to Step 2.2.
    - If $min\{MAR_j^0; MARs_{w_\eta} : \eta = 1, \dots, j\} = MAR_j^0$ for all $\lambda = 2, \dots, 100$, then:
      - If $MAR_j^0 < MAR_j^*$, return to Step 2 at *j+1* with the updating weights as $w_{1,\boldsymbol{j+1}}^* = w_{1,j}^*, \dots, w_{j,\boldsymbol{j+1}}^* = w_{j,j}^*$.
      - If $MAR_j^0 \geq MAR_j^*$, return to Step 2 at *j+1* with the updating weights as $w_{1,\boldsymbol{j+1}}^* = w_{1,j}^*, \dots, w_{j-1,\boldsymbol{j+1}}^* = w_{j-1,\boldsymbol{j}}^*; \; w_{j,\boldsymbol{j+1}}^* = 0$.

[36] There are *2*j* sets of weights and *2*j* corresponding sets of clusters. As a result, there are *2*j* $SARs_{w_\eta}$.

At the end of Step 2, a set of optimal weights corresponding to each k ($k = 1, \dots, K^{max}$) is identified $\{w_{1,}^{*}, \dots, w_{V,}^{*}\}$ along with the corresponding set of clusters $\{\xi_1^*, \dots, \xi_K^*\}$.

Number of Clusters

To identify the true number of clusters, i.e. K^0, this study uses two approaches. The first approach is an informal approach that graphs the values of the *MARs* for a given k against k. Then, the chosen k is the 'knee point' at which the graph starts to flatten. The second approach is to use the modified *BIC* criterion as in the work of Lin & Ng (2012):

(3.6)

$$MBIC(K) = \log\left[\frac{\sum_{k=1}^{K}\sum_{i\in\xi_k}\left(y_i - \hat{\alpha}_{\xi_k} + x_i'\widehat{\beta_{\xi_k}}\right)^2}{N}\right] + K(P+1)\frac{\log(N)}{N} + (K-1)\frac{\log(N^2)}{N^2}$$

where $\hat{\alpha}_{\xi_k}$ and $\widehat{\beta_{\xi_k}}$ are coefficient estimates from running the Equation 3.1 regression within cluster k, K is the number of clusters, P is the number of regressors, and N is the number of observations. The number of clusters with the least modified *BIC* is chosen.

Computation time issues

The ROWK procedure repeatedly runs the K-means algorithm with different sets of feature weights. Hence, the computation time of this procedure depends on (1) the time to run each K-means algorithm, and (2) the number of K-means algorithms to run. The time to run each K-mean algorithm is proportional to the product of the number of observations (N) and the dimension of the variable space (V). Adding each feature consecutively, one by one, and running Steps 2.1 to 2.4 can dramatically slow the running time, especially in case of high dimensions. An alternative is to select all features simultaneously and run Steps 2.1 to 2.4. However, while simulation results show that the running time is considerably faster, the results are much worse relative to consecutive approach.

Computation time may be reduced if instead of repeatedly running the whole procedure for each K', first run this algorithm for K^{max}, and get the set of optimal weights, e.g. $\{w_1^{K^{max}}, \dots, w_V^{K^{max}}\}$. Then, for each K', $K' = 2, \dots, K^{max} - 1$, run the ROWK algorithm as presented above, except without the need to run the Step 1 ranking of features. Instead, run Step 2 for all features (as opposed to adding each feature consecutively) using the set of weights $\{w_1^{K^{max}}, \dots, w_V^{K^{max}}\}$ as the starting weight for Step 2.1. Our simulation results show that this adjustment significantly reduces the running time with little difference in performance as compared to the original procedure[37].

3.2.1.2 Research Design for Testing Hypotheses 1 to Hypotheses 4

Simulated data is generated to test the first four hypotheses relating to the first aim of the thesis. From this section onward, whenever the term '***class***' is used, it indicates the ***true cluster***.

3.2.1.2.1 Factors to Address the Problem of HSGC

The first two hypotheses involve the identification of factors that affect the performance of clustering with respect to the precision of cluster recognition and regression estimations. Four factors examined are cluster density, cluster centroid distance, the degree of heterogeneity in regression coefficients and multicollinearity.

Cluster Density

The density of a class indicates the degree of compactness of the cluster. A class with high density indicates that members within the class are very close to each other in proximity. A formula to indicate the density at *x*, *f(x)*, is the proportion of entities within a sphere, which are centred at *x* with radius *r*, divided by the volume of the sphere (Wong & Lane, 1983). To measure the class density, this thesis employs the degree of class compactness, which is measured as follows. Each class member is represented by a point surrounding its class centre:

[37] An alternative option is start from K'=2 instead K^{max}, causing even more reduction of the computation time. However, starting at the lowest *K* first, and then using the optimal weights at this low *K* for the subsequent higher *K* tends to reach the local optimal, consequently resulting in poorer performance.

(3.7)

$$z_{i,v}^{k} = \left(c_{k,v} + \tau_{i.v}^{k}\right); i = 1, \dots, N_k^0; k = 1, \dots, K^0, v = 1, \dots, V$$

where $c_{k,v}$ is the class centre of class k; $\tau_{i.v}^{k} \sim N(0, \sigma_{k,v}^2)$, and $\sigma_{k,v}^2$ signals the degree of density of cluster k with regard to feature v. $\sigma_{k,v}^2$ is further decomposed into two components:

(3.8)

$$\sigma_{k,v}^2 = den_{k,v}^2 * \frac{\theta_k^2}{\sum_{v=1}^{V} den_{k,v}^2}$$

where $den_{k,v}^2$ denotes the density of cluster k with regard to feature v; θ_k^2 is the expectation of the mean squared distances between members of a class k to its centre[38], being an index of class compactness. The relative density of class k of two feature v_1 and v_2 is calculated as $\sigma_{k,v1}^2/\sigma_{k,v2}^2 = den_{k,v1}^2/den_{k,v2}^2$.

Distances between Class Centres

In real data, there are likely to be some classes that are more distinguishable than others. Therefore, in the simulated data, the centre of Class ξ_1^0 is generated to have the largest distances relative to the centres of other classes. Accordingly, the slope coefficient, i.e. $\beta_{\xi_1^0}$ for Class ξ_1^0 has the largest differences relative to those of other classes. The distance between the centre of Class ξ_i^0 and the centre of Class ξ_j^0 is measured as:

(3.9)

$$d_{\xi_i^0, \xi_j^0} = \sum_{v=1}^{V} (c_{iv} - c_{jv})^2$$

To make Class ξ_1^0 the most distinguishable cluster, this study imposes a restriction on d_{Cij}:

(3.10)

[38] See Appendix A1 for the proof.

$$d_{\xi_1^0,\xi_j^0} > \lambda\, d_{\xi_i^0,\xi_j^0} \text{ for all i}\neq 1$$

where λ indicates the extent of differences between Class ξ_1^0 and other clusters.

Precision of Cluster Identification

The precision of assigning entities to their *clusters/classes* is measured by entropy and purity indexes[39]. Let K' and K^0 equal the number of clusters identified through clustering and the number of true clusters (classes), respectively. Also, let p_{kk^o} be the probability that a member of Cluster $\xi_k{}'$ belongs to Class ξ_k^0, and $\overline{p_{kk^o}}$ is the probability that a member of Class ξ_k^0 belongs to Cluster $\xi_k{}'$.

(3.11)

$$p_{kk^0} = \frac{m_{kk^0}}{m_k} \text{ and } \overline{p_{kk^o}} = \frac{m_{kk^o}}{\overline{m_{k^0}}}$$

where $m_k(\overline{m_{k^0}})$ is the number of objects in Cluster $\xi_k{}'$ (Class ξ_k^0) and m_{kk^o} is the number of entities of Cluster $\xi_k{}'$ in Class ξ_k^0. Entropy is the degree to which each cluster consists of objects of a single class. The lower the entropy, the higher the cluster precision. The entropy of Cluster $\xi_k{}'$ is calculated as:

(3.12)

$$entropy_k = -\sum_{k^0=1}^{K^0} p_{kk^0} \log_2 p_{kk^0}$$

The overall entropy is calculated as:

(3.13)

$$entropy_{overal} = \sum_{k=1}^{K'} \frac{m_k}{N} * entropy_k$$

[39] This study only focuses on supervised measures of cluster validity due to the available information of data membership. For a more comprehensive list of measures of cluster validity on supervised and unsupervised clustering, see Tan et al. (2005)

Purity is also a measure of precision of assigning entities to clusters. This study uses three types of purity indexes to explore different aspects of precision. These are defined below. *Purity_ver1* denotes the extent to which a *cluster* contains objects of a *single class*. *Purity_ver1* of Cluster $\xi_k{}'$ and for overall are calculated respectively as:

(3.14)

$$p_{\mathrm{k}}^{ver1} = max_{k^o}(p_{kk^o}) \text{ and } p_{overal}^{ver1} = \sum_{k=1}^{K'} \frac{m_{\mathrm{k}}}{N} * p_k^{ver1}$$

Purity_ver2 denotes the extent to which *a class* contains objects of a *single cluster*. *Purity_ver2* of Class ξ_k^0 and for overall are calculated respectively as:

(3.15)

$$p_{k^0}^{ver2} = max_{\mathrm{k}}(p_{kk^o}) \text{ and } p_{overal}^{ver2} = \sum_{k^0=1}^{K^0} \frac{\overline{m_{k^0}}}{N} * p_{k^0}^{ver2}$$

Purity_ver3 has the same meaning as for version 2 but with a change to the weights of individual cluster purity to calculate overall purity. *Purity_ver3* of Class $\xi_k^0$and for overall are calculated respectively as:

(3.16)

$$p_{k0}^{ver3} = p_{k0}^{ver3} \text{ and } p_{overal}^{ver3} = \sum_{k^0=1}^{K^0} \frac{\widetilde{m_k}}{N} * p_{k0}^{ver3}$$

where $\widetilde{m_{k'}}$ is the number of objects in Cluster $\xi_k{}'$ that corresponds to $max_k(p_{kk^o})$. This version is to correct the *purity_ver2* when the number of objects significantly differs from class to class. *Purity_ver1* is typically used as a measure for supervised clustering validity. However, given that this section focuses more on the precision of assigning members of Class ξ_1^0, versions 2 and 3 are also employed.

Hypothesis Testing

Hypothesis 1a relates to distances between classes. *Class* ξ_1^0 is assumed to be the most noticeable class which is furthest away from the rest of the classes and accordingly, its coefficients are considerably different from those of other classes. As a result, the

performances of clustering and of regression estimations are mainly impacted by whether or not members of *Class* ξ_1^0 are correctly identified. Hence λ (see Equation 3.10) defined as the parameter that represents the extent of differences between *Class* ξ_1^0 and other classes, is used to proxy distances between classes. Hypothesis 1b relates to the densities of classes as measured by θ (see Equation 3.8).

There are 5000 observations (*N*) which belong to five classes ($K^0 = 5$). Each class has 1000 members ($N_k^0 = 1000, k = 1,..,5$); the number of regressors (*P=2*) is two; the intercept ($\alpha_{\xi_k^0}$) and the slope coefficient of X_2 (β_{2,ξ_k^0}) are set to equal 1 and 0.5 for all classes, respectively. Hence, only the coefficient of X_1 varies across clusters. For Hypothesis H1 (a, b and c), the number of clustering features (*V*) is set to four. The *within* covariance matrix of features is set to equal the diagonal matrix $\sum_{v_i v_j} = cov(\tau_{i.v}^k, \tau_{i.v}^k) = D$; $den_{k,v} = 1, \theta_k = \theta$ for all *v*=1,…,*V* and *k*=1,…,K^0. Therefore, $\sigma_{k,v}^2 = \theta^2/4$ and the sum of the squared distance (*SSD*) of each class equals θ^2. This simplifies the analysis and establishes the benchmark to test Hypothesis H2 where features are highly correlated.

For testing Hypothesis H2, all input parameters are the same as in the testing of Hypothesis H1, except for the *within* covariance matrix of features which are generated to differ from the diagonal matrix. Hypothesis H3 examines the impact of standardization on the performance of K-means clustering via different contexts. To test this hypothesis, two scenarios of feature weights are generated. For the first scenario, a feature's weight stems from the differences of distances between the classes' centres as measured by this feature relative to those measured by other features. Since *Class* ξ_1^0 is simulated as the most important and distinguishable class, in the first scenario the weight of a certain feature (w_v) is proxied as the average distance between *Class* ξ_1^0 centres and other classes' centres as measured by this feature relative to those measured by other features:

(3.17)

$$\overline{d_{v,\xi_1^0\xi_j^0}} = \frac{1}{4}\sum_{j=2}^{5}\left(c_{1v} - c_{jv}\right)^2, v = 1, \dots, 4$$

The higher the $\overline{d_{v,\xi_1^0\xi_j^0}}$, the more relevant is the feature for identifying cluster patterns. For the second scenario, a feature's weight results from differences of classes' densities

measured by this feature relative to those as measured by other features. Again, as *Class* ξ_1^0 is simulated as the most important and distinguishable class, the weight of a certain feature (w_v) in the second scenario is proxied as the density index of *Class* ξ_1^0 based on this feature, i.e. $den_{1,v}^2$ as in Equation 3.8.

For testing Hypotheses H1, H2 and H3, 100 simulated data samples are generated for each case. K-means clustering is run for these 100 samples, and the average results are presented. T-tests are used to test for the significance of differences. Each hypothesis is examined by changing the value of appropriate parameters. Table 3-1 summarises these parameters.

Table 3-1: Input Parameters for Hypothesis Testing

Hypothesis	Parameters' Descriptions	Parameters
H1a	λ	Distance between Classes' Centres
H1b	θ	Class Density
H1c	λ θ $\alpha_{\xi_k^0}$ and $\beta_{,\xi_k^0}$	Distance between Classes' Centres Class Density Heterogeneity of Regression Coefficients
H2	$\sum_{v_i v_j} = cov(\tau_{i.v}^k, \tau_{i.v}^k)$	*Within* covariance matrix of features
H3	$\overline{d_{v,\xi_1^0 \xi_j^0}}$ $den_{1,v}^2$	Classes' Densities Measured by a Feature Distance between Classes' Centres Measured by a Feature

3.2.1.2.2 Performance of ROWK

Hypothesis H4 projects three channels through which ROWK improves the performance of cluster analysis with respect to the HSGC problem. Specifically, ROWK is hypothesized to place more (less) weight on more (less) relevant features; reduce the influence of multicollinearity by reducing the weights of irrelevant features which are highly correlated with relevant features and capture relevance not only by its contribution to cluster recognition but also by regression estimation. To test this hypothesis, this study examines three simulated cases. Each case sheds light on each channel through which ROWK improves the performance of cluster analysis with respect to the HSGC problem. Case 1 includes a simple set of simulated data with uncorrelated features. Case 2 employs

the same data as Case 1, but with correlated features. Case 3 analyses a situation where features' weights come from two sources, i.e. contributions to cluster recognition and to regression estimations.

For Case 1, there are 5000 observations belonging to five classes. Each class has 1000 members. There are four features, z_v, v=1,…,4 and a random variable z_5 (~N(0,1)) which is used as an irrelevant feature. For simplicity, only z_3 is more relevant relative to others. For the regression model, there are two independent variables, x_1 and x_2, which are also features of clustering, i.e. $z_1 = x_1$ and $z_2 = x_2$. x_1 and x_2 satisfy assumptions in Section 3.2.1.1.1.

For Case 2, simulated data are generated with the same set of simulated parameters as in Case 1 with an exception that features are within-class correlated. For Case 3, some adjustments to simulated parameters are made as follows. z_3 is generated to be a highly relevant feature to recognize *Class* ξ_1^0's membership, and z_4 is simulated to provide information to distinguish membership of other classes. Clustering features are generated to be uncorrelated. Unlike Case 1 and 2, in Case 3, x_1 and x_2 are not set to be clustering features.

3.2.1.2.3 Canonical Discriminant Analysis

Canonical discriminant analysis (thereafter CDA) is a multivariate technique first introduced by Hotelling (1936) and Fisher (1936). It aims to determine the relationships among a group of independent variables and a categorical variable. One primary purpose of CDA is to reduce dimensional discriminant space in recognizing classes (Zhao & Maclean, 2000). Throughout the thesis, we use canonical discriminant analysis for two purposes. First, results from the canonical (standardized) coefficients can be used to gain more insights on different contributions of cluster feature to distinguish clusters. Second, as a method of reducing discriminant dimensions, CDA is employed in this thesis as a way to facilitate the process of graphing since cluster membership can only be graphed up to three dimensions. Accordingly, the first two or three canonical components are extracted, and are used to demonstrate cluster memberships graphically.

3.2.2 An Earnings Persistence Application of ROWK

This section presents the research design to apply ROWK clustering to earnings persistence. It begins with steps to conduct ROWK clustering to explore firm patterns on earnings persistence. Adjustments to data processing due to the noise of real data are discussed. Models are then presented to estimate firm life cycles, earnings management and accounting conservatism that are the core of Hypothesis H8. Two approaches to test Hypothesis 9 concerning analyst forecasts finalize this section.

3.2.2.1 Executing ROWK Clustering on Earnings Persistence

This section presents steps to conduct ROWK clustering to explore firm patterns with respect to earnings persistence. It includes four steps. The first step introduces the regression model of earnings persistence. The second step presents the list of features that are used in clustering. Pre-processing of clustering data are discussed in the third step. The final step executes ROWK clustering to identify three important aspects, i.e. optimal number of clusters, optimal cluster weights, and cluster membership on earnings persistence.

3.2.2.1.1 Earnings Persistence Model

This thesis follows the earnings persistence model used by Dichev & Tang (2009). Notably, firm-level earnings persistence is measured by:

(3.18)

$$Earnings_{i,t+1} = \mu + \theta\, Earnings_{i,t} + u, i = 1, \dots, N, t = 1, \dots, T$$

This is a first order autoregressive regression (with drift μ and slope θ) of current earnings on 1-year lagged earnings. Given the presence of HGSC, the earnings persistence model now becomes:

(3.19)

$$Earnings_{i,t+1} = \mu_i + Earnings_{i,t}\theta_{(i)} + u_{i,t}, i = 1, \dots, N, t = 1, \dots, T$$

where $\mu_{(i)}$ and $\theta_{(i)}$ are intercepts and slope coefficients for unit i respectively. A group effect is modelled by allowing $\mu_{\xi_k^0}$ and $\theta_{\xi_k^0}$ to be the group-specific intercept and slope coefficients such that $\mu_{(i)}$ and $\theta_{(i)}$ equal or closely approximate $\mu_{\xi_k^0}$ and $\theta_{\xi_k^0}$ respectively for all i's in ξ_k^0. $\xi_1^0, \xi_2^0, \ldots, \xi_K^0$ as the corresponding true K^0 clusters. Hence, the model becomes:

(3.20)

$$Earnings_{i,t+1} = \mu_{\xi_k^0} + Earnings_{i,t}\theta_{\xi_k^0} + u_{i,t}, i = 1, \ldots, N, t = 1, \ldots, T, k = 1, \ldots, K^0$$

The following assumptions will be made: *(i)* $u_i \sim (0, \sigma^2)$ has finite second moments and has cross-sectional and serial independence, i.e. $cov(u_{ij}) = \sigma^2 I$ where I is the identity matrix; *(ii)* u_i is independent of $Earnings_t$ for all k=1,…, K^0. Assumptions *(i)* and *(ii)* imply that the model is correctly specified and can be consistently estimated within each true cluster (Lin & Ng, 2012). Our objective is to estimate $\theta_{\xi_k^0}$ (and $\mu_{\xi_k^0}$) without knowing ξ_k^0. This can be achieved by applying the proposed ROWK procedure to the earnings persistence model.

Recent literature on earnings predictability tend to use return on net operating assets (RNOA) as the measure of firm earnings (e.g. Amor-Tapia & Tascón Fernández, 2014; Bauman, 2014; Fairfield & Yohn, 2001). Operating earnings is consistent with prior research on earnings persistence and earnings forecasts, which emphasise the role of operating activities in creating value (Fairfield & Yohn, 2001)[40]. This thesis follows this convention. However, to illustrate the contribution of the proposed ROWK clustering technique, for consistency we next adopt the same earnings measure as Dichev & Tang (2009) who report earnings volatility as the sole dominant factor to distinguish earnings persistence. Accordingly, income before extraordinary items (***IBC, Compustat #123***) is now used as a proxy for earnings, and earnings volatility is measured as the standard

[40] Nissim & Penman (2001) give a reasonable explanation of the importance of future RNOA forecasts. Particularly, given that all net financial obligations are measured as the market value, then the value of common equity depends only on its current book value and the present value of operating residual income: $V_o^E = CSE_o + \sum_{t=1}^{\infty} \frac{ReOI_t}{(1-r_w)^t} = CSE_o + \sum_{t=1}^{\infty} \frac{(RNOA_t - r_w) * NOA_{t-1}}{(1-r_w)^t}$; where r_w is the cost of operations.

deviation of deflated *IBC* for the most recent 5 years. In robustness testing, deflated *IBC* is replaced by RNOA to test the sensitivity of the results to a different measure of earnings.

3.2.2.1.2 Feature Selection

The selection process should be strongly grounded by underlying theory. Features selected as inputs for clustering have to be characteristics that contribute to the distinct behaviours of clusters. In this regard, the selected features for ROWK must relate to characteristics that contribute to distinct firm behaviours relating to persistence of earnings. As discussed in Section 2.5.2, the list of cluster features used in ROWK clustering is described in Table 3-2. See Section 2.5.2 for details of the theoretical reasoning and formulae for the variables.

Table 3-2: List of Cluster Features on Earnings Persistence

No.	Cluster Features	Definition
1	PM_t,	Profit Margin
2	ATO_t	Asset Turnover
3	ΔPM_t	Change in Profit Margin
4	ΔATO_t	Change in Asset Turnover
5	$VOL\_IBC\_DEF_t$	Earnings Volatility
6	CR_t	Current Ratio
7	$CAPX\_DEF_t$	Deflated Capital Expenditures
8	$INTAN\_INT\_DEF_t$	Deflated Intangible Investment
9	$FLEV_t$	Financial Leverage
10	$OLLEV_t$	Operating Leverage
11	$SALE\_GR_t$	Sales Growth
12	NBC_t	Net Borrowing Cost
13	DIV_t	Dividend Payout
14	AB_ACC_DEF	Deflated Absolute Value of Accruals
15	$SIZE_t$	Firm Size
16	AGE_t	Firm Age
17	ABS_EARNINGS_DEF	Deflated Absolute Value of Earnings

3.2.2.1.3 Data Processing for Cluster Features and Regression Variables

The purpose of data pre-processing is to mitigate the effect of outliers and to transform the input features for clustering in a fair way. The standard procedure to address the first issue is to winsorize the data, typically at 1% and 99% (e.g. Cooper et al., 2008). Accordingly, variables in the model of earnings persistence are winsorized at 1% and

99%. However, clustering is more sensitive to the presence of noise, which is prevalent in real financial data. Furthermore, standardization using z-scores fails to tackle the noise problem. Tanioka & Yadohisa (2012) conduct tests both with real and simulated data on six different ways of standardization and observe that a ranking method is the most effective way to deal with noisy data. Accordingly, we follow this procedure by transforming all cluster features annually into *percentile-ranked features*.

3.2.2.1.4 ROWK Clustering Execution

After winsorizing the earnings persistence model variables and transforming cluster features into percentile-ranks, the optimal feature's weights ($w_v \geq 0 \; and \; \sum_{v=1}^{V} w_v = 1$) are identified by minimising the mean of the absolute residuals, *MAR*:

(3.21)

$$MAR = \sum_{i=1}^{N} \left| Earnings_{i,t+1} - \widehat{Earning}_{i,t+1} \right| / N$$

where $\widehat{Earning}_{i,t+1}$ is the estimate of $Earnings_{i,t+1}$, i.e :

(3.22)

$$\widehat{Earning}_{i,t+1} = \hat{\mu}_{\xi_k} + Earning_{i,t} \widehat{\theta_{\xi_k}}, \text{ k=1,…,K}$$

where $\hat{\mu}_{\xi_k}$ and $\widehat{\theta_{\xi_k}}$ are the estimates of μ_{ξ_k} and θ_{ξ_k} from running the regression Equation 3.18 for each cluster. Cluster memberships are found by assigning *N* data points into *K* disjointed clusters such that the sum-of-squares criterion, *J*, is minimised:

(3.23)

$$J = \sum_{k=1}^{K} \sum_{v=1}^{17} \sum_{i \in \xi_{k'}} \| w_v z_{iv} - w_v c_{kv} \|^2$$

Minimising *J* is equivalent to running K-means clustering with the weighted cluster features, i.e. $w_v z_{iv}$. The steps to minimise *MAR* are described in detail in Section 3.2.1.1.2 (Step 4- Finding Optimal Weights of Cluster Features). The end of this step results in the set of optimal weights, providing more insights on the different contributions of features

to distinguish clusters with respect to earnings persistence. It also provides the number of clusters identified and the corresponding cluster membership.

3.2.2.2 The Performance of ROWK Clustering

This section presents tests of the performance of ROWK clustering to better understand the patterns with respect to earnings persistence, and consequently, earnings predictability.

3.2.2.2.1 Heterogeneities in Optimal Feature Weights

Some factors are more relevant to identify the differences in earnings persistence than others. Dichev & Tang (2009) find that earnings volatility outperforms other examined variables, including level of accruals, cash flow volatility, and earnings level with regard to distinguishing earnings persistence and achieving higher earnings predictability. Accordingly, Hypothesis H5 predicts that feature weights identified by ROWK clustering are not equal. It is not possible to derive the empirical distribution for the estimated optimal cluster features unless we can simulate the data sample. Jackknifing is a solution to estimate the variance of feature weights (Efron & Stein, 1981). However, given the time needed to conduct ROWK clustering, it is infeasible. Hence, we cannot directly test H5. Instead, we compute *MAR* in case of equal weights (*MAR_EQW*). Then we compare the *MAR* of optimal weights (*MAR_ROWK*) to those of equal weights. If H5 is supported, then a significant difference between *MAR_EQW* and *MAR_ROWK* is expected.

3.2.2.2.2 Earnings Persistence across Clusters

After running ROWK clustering for the earnings persistence model, *K* estimated slope coefficients of earnings persistence are observed in accordance with K clusters, i.e. $\widehat{\theta_{\xi_k}}$, k=1,…,K. To test hypothesis H6 which posits that firms exhibit different earnings persistence between ROWK clusters, this thesis follows Dichev & Tang (2009). Specifically, Equation 3.24 below is a regression model that combines observations from the two clusters that have the highest (HIGH) and lowest (LOW) slope coefficients, with dummy intercept and slope variables receiving a value equal to 1 for the observations in the lowest earnings persistence cluster. A t-test is used to test for differences in earnings

persistence across clusters. Thus, the main coefficient of interest, γ, is expected to be significantly negative.

(3.24)

$$Earnings_{i,t+1} = \mu + D_{it,LOW} + \theta Earnings_{i,t} + \gamma Earnings_{i,t} D_{it,LOW} + u_{i,t},$$

where $D_{it,LOW}$ is a dummy variable = 1 if at time t, firm i belongs to the *LOW* cluster, and 0 if it belongs to the *HIGH* cluster.

3.2.2.2.3 Earnings Predictability across Clusters

To complement hypothesis H6, we conduct further tests relating to the performance of ROWK clustering with respect to earnings predictability. For in-sample data, earnings predictability is proxied by the model goodness of fit, i.e. adjusted R^2, being a "relative" measure of predictability (Dichev & Tang, 2009). This is in contrast to the *SSR*, representing "absolute" predictability, which is more suitable to proxy for earnings predictability when using out-of-sample data. Testing adjusted R^2 across different clusters is problematic since R^2 are compared across different regression data samples. Moreover, the dependent variable (i.e. earnings) exhibits large discrepancies in its variation across clusters, invaliding the Vuong test (Dichev & Tang, 2009).

For these reasons, this thesis follows the bootstrapping approach employed in Dichev & Tang (2009). Specifically, the null hypothesis is that cluster membership is unrelated to earnings predictability, and the test statistic is the difference in adjusted R^2 between HIGH and LOW clusters. To simulate the empirical distribution of the differences under the null, the full sample is randomly split into pseudo K clusters. Among these K clusters, two random clusters are selected as HIGH and LOW clusters. Then, the earnings persistence regression is run within the pseudo HIGH and LOW clusters to observe any difference in R^2 between these two clusters. Repeating this procedure 1000 times yields a 1000-observation empirical distribution of R^2 differences under the null. The formal statistical test compares the actual observed difference in adjusted R^2 against the simulated distribution of differences.

3.2.2.2.4 Benchmark

The testing of Hypothesis H7 compares the performance of ROWK clustering with respect to earnings persistence and earnings predictability to that of other benchmark techniques. The first benchmark model results from running the earnings persistence model for the full sample. This means that there is no concern of HGSC in the earnings persistence model. The second benchmark model is standard K-means using all listed features. The third benchmark model is again K-means clustering but with only listed features that have non-zero ROWK optimal weights. The fourth benchmark model involves normal partitions using only one cluster feature. For the sake of brevity, only the first two most important features are chosen. The firth benchmark model is weighted K-means as developed by Huang et al. (2008).

To test for earnings predictability, the thesis also includes an earnings prediction model that incorporates many earnings predictors covered by recent studies of earnings predictability. Specifically, the model of one-year-ahead earnings is as follows[41]:

(3.25)

$$\begin{aligned}\Delta\text{Earnings}_{i,t+1} = {} & \alpha + \beta_1 \text{Earnings}_{i,t} + \beta_2 \Delta\text{Earnings}_{i,t} + \beta_3 \Delta\text{COA}_{i,t} + \beta_4 \Delta\text{NCOA}_{i,t} \\ & + \beta_5 \Delta\text{ATO}_{i,t} + \beta_6 \Delta\text{PM}_{i,t} + \beta_7 EM\_UP_{i,t} + \beta_8 \Delta\text{PM}_{i,t} * EM\_UP_{i,t} \\ & + \beta_9 EM\_DN_{i,t} + \beta_{10} \Delta\text{PM}_{i,t} * EM\_DN_{i,t} + \varepsilon_{i,t+1}\end{aligned}$$

where:

$\text{EM\_UP}_{i,t} = 1$ if $\Delta\,\text{PM}_{i,t} > 0$ and $\Delta\,\text{ATO}_{i,t} < 0$ and $EM\_DN_{i,t} \neq 1$, and 0 otherwise

$\text{EM\_DN}_{i,t} = 1$ if $\Delta\,\text{PM}_{i,t} < 0$ and $\Delta\,\text{ATO}_{i,t} > 0$ and $EM\_UP_{i,t} \neq 1$, and 0 otherwise

See Appendix A2 for details of this model. Table 3-3 summarises the benchmark models for comparison with the proposed ROWK clustering.

[41] Current literature on earnings predictability focuses on predicting one-year-ahead RNOA. However, to be aligned with the earnings persistence model, we replace RNOA by IBC_def. The results are remain unchanged when RNOA is used in Equation 3.25.

Table 3-3: Comparison Benchmarks for ROWK Clustering

No.	Abbreviation	Definition
		Patterns of Earnings Persistence (and Intercept)
1	ALL	Run the earnings persistence model without clusters (i.e. assume homogeneous coefficients across groups/clusters)
2	K_ALL	Run the earnings persistence model for each cluster found by K-means clustering using all features
3	K_NONZERO	Run the earnings persistence model for each cluster found by K-means clustering using only features that have non-zero ROWK weights
4	PART	Run the earnings persistence model for each partitioned sample based on a single feature
5	W_K	Run the earnings persistence model for each cluster found by weighted K-means developed by Huang et al. (2008)
		Earnings Predictability
1,2,3,4,5		As defined in previous rows of the table
6	E_MOD	Run the earnings prediction model (Equation 3.25)

3.2.2.3 Testing for Heterogeneity of Industry, Firm Life Cycles, Earnings Management, and Conservatism across Clusters from ROWK Clustering

This section presents the research design to test hypothesis H8 that conjectures the presence of heterogeneity for industry membership, firm life cycles, earnings management and conservatism across ROWK clusters.

3.2.2.3.1 Construction of Industry Classification, Firm Life Cycles, Earnings Management and Conservatism

Industry Classification

Fama-French 12-industry classifications are used to divide firms into different industries (Fama, & French, 1997). See Appendix Table A3 for details of the Fama-French 12-industry classification.

Firm Life Cycles

The thesis follows Dickinson (2011), who identifies firm life cycles using information from cash flow statements. Using the sign of three net cash flow activities (operating, investing, and financing), firms are assigned into one of the five stages, i.e. introduction, growth, mature, shake-out and decline. Life cycles as proxied by cash flow patterns are

predicted to identify differential profitability persistence between clusters. See Appendix Table A4 for the life cycle classification using cash flow patterns.

Earnings Management

Recent studies use accruals models to identify the presence of earnings management (e.g. Dechow et al., 2003; Kothari, Leone, & Wasley, 2005). An abnormal earnings management component is identified as the difference between total accruals and a measure of the normal non-discretionary component. However, there is controversy over how accruals should behave in the absence of discretion (McNichols, 2000) and serious inference problems have emerged over the use of the accruals models (Fields, Lys, & Vincent, 2001).

Acknowledging the usefulness of financial ratios to identify the existence of earnings management, Jansen et al. (2012) propose a method to identify earnings management using the sign of ΔPM and ΔATO. They argue that due to the articulation of the balance sheet and income statement, opposite changes in ATO and PM could be a signal of earnings management. Specifically, a firm is identified as upwardly managing earnings ($EM\_UP_{i,t}$) if a) $\Delta PM_{i,t} > 0$ and $\Delta ATO_{i,t} < 0$ and b) no downwardly managed earnings are observed in the previous year. In a similar vein, a firm is identified as downwardly managing earnings ($EM\_DN_{i,t}$) if a) $\Delta PM_{i,t} < 0$ and $\Delta ATO_{i,t} > 0$ and b) they have not upwardly managed earnings in the previous year. This simple diagnostic not only demonstrates the usefulness of financial ratios, but also generates incremental information over abnormal accruals on future profitability. The extension of $EM\_DN_{i,t} \neq 1$ ($EM\_UP_{i,t} \neq 1$) in the case of upward (downward) earnings management ensures against the possibility of earnings reversal.

Conservatism

There are two types of conservatism. One is conditional conservatism (C_CON) and the other is unconditional conservatism (U_CON) (Chen et al., 2014). Conditional conservatism emerges when certain circumstances happen. In this context, the circumstance is the arrival of news where bad news is recognized in a more timely manner

than good news (Basu, 1997). This thesis follows Khan & Watts (2009) to estimate conditional conservatism. The approach is as follows:

(3.26)

$$X_{it} = \alpha_0 + \alpha_1 DR_{it} + R_t(\beta_0 + \beta_1 Size_{it} + \beta_2 MTB_{it} + \beta_3 Lev_{it}) + DR_{it} * R_t(\gamma_0 + \gamma_1 Size_{it} + \gamma_2 MB_{it} + \gamma_3 Lev_{it}) + u_{it}$$

where X_{it} denotes earnings before extraordinary items per share for year *t*, deflated by the stock price per share at the beginning of year *t*; R_t denotes the annual stock return from nine months before the fiscal year end *t* to three months after the fiscal year end *t*; DR_{it} is a dummy variable equal to 1 if R_t is negative, and 0 otherwise; $Size_{it}$ is the natural logarithm of total assets at the end of year *t*; MTB_{it} is the ratio of market value of equity to book value of equity at the end of year *t*; and Lev_{it}denotes the ratio of total liabilities to total assets at the end of year *t*. Each year, Equation 3.26 is estimated for all firms to obtain estimates of γ_0, γ_1, γ_2 and γ_3, namely $\widehat{\gamma_0}$, $\widehat{\gamma_1}$, $\widehat{\gamma_2}$, $\widehat{\gamma_3}$. Conditional conservatism for firm i at time *t* is measured as follows:

(3.27)

$$C\_CON_{it} = \widehat{\gamma_0} + \widehat{\gamma_1} Size_{it} + \widehat{\gamma_2} MB_{it} + \widehat{\gamma_3} Lev_{it}$$

Firms with high *C_CON* have higher conditional accounting conservatism. Unconditional conservatism, in contrast, is not dependent on news and is pervasively applied. Following Givoly & Hayn (2000) and Chen et al., (2014), unconditional conservatism is measured as negative cumulative non-operating accruals, which are defined as the difference between total accruals before depreciation and operating accruals.

(3.28)

Non-operating accruals = Total accruals before depreciation - Operating accruals

where:

- *Total accruals before depreciation = (Net Income +Depreciation) – Cash flows from operations;*

- *Operating accruals = ΔAccounts Receivable + ΔInventories + ΔPrepaid Expenses - ΔAccounts Payable - ΔTaxes Payable;*

Non-operating accruals is then deflated by total assets, and is cumulated over the last five years (including the current year). Finally, we multiply this result by negative one to make the sign of the unconditional conservatism proxy (*U_CON*) in the same direction as for unconditional conservatism.

3.2.2.3.2 Tests for Heterogeneity

Pursuant to the classification of industry and measurement of earnings management, firm life cycles and conservatism, Hypothesis H8 is tested as follows. Two-way frequency tables are created in turn for cluster membership by each of industry, firm life cycles, earnings management and conservatism. A chi-squared test for two-way classification is used to test whether cluster membership is independent of industry classification, life cycles, earnings management or conservatism. Specifically, the test statistic χ^2 test of independence in a contingent table is as follows (Bishop, 1969).

(3.29)

$$\chi^2 = \sum_{i \in rows} \sum_{j \in column} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

where O_{ij} is the number of observations in row i and column j; and E_{ij} is the expected number of observations in row i and column j.

3.2.2.4 Analyst Forecasts and Earnings Persistence Patterns

This section describes the research design to test Hypothesis H9 that posits that information from ROWK cluster identification predicts analyst forecast errors. Two approaches are presented to test this hypothesis. The first approach uses portfolio sorting, and the second approach builds a model of analyst forecast errors.

3.2.2.4.1 Portfolio Sorting

Each year, ROWK clustering is run for a rolling last 5-year window, and firms are assigned into clusters *HIGH* and *LOW*. Realized earnings at time *t* of firms in clusters *HIGH* and *LOW* are calculated. The earnings distribution of firms in clusters *HIGH* and *LOW* are not similar, so two way sorting is conducted to make the earnings patterns of *HIGH* and *LOW* more comparable. Specifically, firms are sorted annually into twenty portfolios by current earnings, and ROWK clustering is executed within each portfolio. Accordingly, firms are designated as belonging to the high earnings and high earnings persistence cluster (*HE_HP*) when they fall in the intersection of *HIGH* cluster and earnings level portfolio 17-20. In a similar vein, firms are designated as members of the high earnings and low earnings persistence cluster (*HE_LP*) if they fall in the intersection of *LOW* cluster and earnings level portfolio 17-20. The same construction is for firms that have low earnings-high earnings persistence (*LE_HP*) and low earnings-low earnings persistence (*LE_LP*).

By this construction, it is expected that the realized earnings at time t of firms in *HE_HP* vs. *HE_LP* (and *LE_HP* vs. *LE_LP*) will exhibit similar distributions. Consequently, any observed differences in future earnings distributions between *HE_HP* and *HE_LP* (*LE_HP* vs. *LE_LP*) are more likely to have resulted from differences in earnings persistence (and intercepts), and not from differences in earnings levels. Then, earnings forecasts from analysts are studied to determine whether analysts understand the different patterns of future earnings caused by HGSC in the earnings persistence model.

3.2.2.4.2 The Model of Analyst Forecast Errors

The second approach requires the model of analyst forecast errors to control for positive auto-correlations of analyst forecast errors (Abarbanell & Bernard, 1992, Dichev & Tang (2009). Thus, it is necessary to control for earnings surprises at time *t* to predict future analyst forecast errors. Accordingly, the following regression is run to test whether analysts fully incorporate the cluster earnings persistence information identified by ROWK clustering:

(3.30)

$$FE_{t+1} = \alpha_0 + \alpha_1 HIGH\_P + \alpha_2 FE_t + \alpha_3 HIGH\_P * FE_t + u_t$$

where *HIGH_P* is a dummy variable equal to 1 if a firm is in the cluster having the highest earnings persistence, and 0 if a firm is in the cluster having the lowest earnings persistence. Analyst forecast error for year *t* (FE_t) is measured as the difference between the actual earnings for year *t* and the *last* median analyst forecast for year *t prior* to the announcement of earnings for year *t*. Analyst forecast errors for year *t+1*, (FE_{t+1}) are defined as the differences between the actual earnings for year t+1 and the *first* median analyst forecast of year *t+1* immediately following the announcement of earnings for year *t*. Analyst forecast errors for year *t+2, ..., t+5* are constructed in the same manner as for *t+1*. The coefficients of interest are α_1 and α_3.

3.3 DATA

3.3.1 ROWK and the Problem of HGSC

As discussed in the methodology section, simulated data are generated for tests of Hypothesis 1 to Hypothesis 4. The dependent variable (y_i) and independent variables (x_i) are generated based on the model in Equation 3.1:

$$y_i = \alpha_{\xi_k^0} + x_i' \beta_{\xi_k^0} + u_i, i = 1, \dots, N$$

where x_i is a $P\times1$ vector of explanatory variables and u_i is the unit-specific error; $\alpha_{\xi_k^0}$ and $\beta_{\xi_k^0}$ are 1×1 and $P\times1$ vectors of class-specific intercepts and slope coefficients for unit *i* respectively; $\xi_k^0, k = 1, \dots, K^0$ represent class (true cluster) *k*'s membership. Each class member is represented as a point in a *V*-dimensional Euclidean space where each dimension is a feature, denoted by $z_{i,v}$, *v*=1,…, *V*. In real data, there are likely to be some classes that are more distinguishable than others. Therefore, in the simulated data, the centre of *Class* ξ_1^0 is generated to have the largest distances relative to the centres of other classes. Accordingly, the slope coefficient for *Class* ξ_1^0, i.e. $\beta_{\xi_1^0}$ has biggest differences relative to those of other classes. Members of classes are generated as follows:

Step 1: In a V-dimensional Euclidean space, randomly generate K^0 points, representing K^0 cluster's centres; $c_k = (c_{k1}, c_{k2}, \dots, c_{kV})'$ and $k = 1, \dots, K^0$. Denote d_{Cij} as the distance between the centre of *Class* ξ_i^0 and the centre of *Class* ξ_j^0 measured as in Equation 3.9:

$$d_{\xi_i^0, \xi_j^0} = \sum_{v=1}^{V} (c_{iv} - c_{jv})^2$$

To ensure that the centre of *Class* ξ_1^0 has the largest distance relative to other classes' centres, this study imposes a restriction on $d_{\xi_i^0, \xi_j^0}$ as described in Equation 3.8:

$$d_{\xi_1^0, \xi_j^0} > \lambda \, d_{\xi_i^0, \xi_j^0} \text{ for all } i \neq 1$$

where λ indicates the extent of differences between *Class* ξ_1^0 and other classes.

Step 2: Create members of classes. Each point surrounding its class centre represents each class member as described in Equation 3.7:

$$z_{i,v}^k = \left(c_{k,v} + \tau_{i.v}^k\right); i = 1, \dots, N_k^0; k = 1, \dots, K^0, v = 1, \dots, V$$

where $\tau_{i.v}^k \sim N(0, \sigma_{k,v}^2)$, $\sigma_{k,v}^2$ signals the degree of density of cluster k with regard to feature v. Note that the weight of a feature depends on two components: the distances between classes' centres measured by the feature (λ), and densities of classes measured by the feature($\sigma_{k,v}^2$), which are further decomposed into two components as in Equation 3.8:

$$\sigma_{k,v}^2 = den_{k,v}^2 * \frac{\theta_k^2}{\sum_{v=1}^{V} den_{k,v}^2}$$

Cluster density is set to be the same across clusters, i.e. $den_{k,v} = den_v$ and $\theta_k = \theta$ for all k=1,…,K^0. To test the proposed hypotheses, different sets of simulated parameters are used, and are described immediately before the results in the next section.

3.3.2 ROWK and Earnings Persistence

Analyst forecast data is obtained from I/B/E/S (Thomson Reuters Database) covering the period 1980-2011. Monthly price data is obtained from CRSP. Annual data from financial statements is from the COMPUSTAT annual industrial and research files between 1988

to 2011. CUSIP is used to merge the content from these databases. The sample starts from 1988 in order to replicate the work of Dichev & Tang (2009) who focus on the use of cash flow statements (the data from which became more widely available from 1988) to measure earnings, volatility of earnings, operating cash flows and accruals. Replicating the work of Dichev & Tang (2009), we conduct the main tests on the ***1988-2004*** period. The full sample is then randomly split into two sub-samples for separate in-sample and out-of-sample tests. The results from the 2005-2011 period are presented for robustness testing.

The sample is constrained to firms that have data available with respect to earnings (*IBC_def, Compustat #123*) and all cluster features (see Table 3-2). Following Dichev & Tang (2009), to reduce the impact of noise, the sample is limited to economically substantial firms having a minimum of $100 million in assets. The sample is further restricted to 12/31 fiscal year-end firms to simplify the tests and the interpretation of the results.

To avoid the influence of extreme observations in the earnings persistence model, the top and bottom 1% of earnings is truncated. For cluster features, it is more complicated. Figure 3.2 displays the distributions of cluster features winsorized at the 1% top and bottom percentile. There are two concerns. The first is that there is a 10% reduction in the sample size if *DIV* (dividend payout) and *NBC* (net borrowing cost) are measured as in Table 2-1[42]. To overcome this problem, the new measures of *DIV* and *NBC* are estimated as follows:

(3.31)

$$\text{DIV\_RANK} = \text{rank (DVC)} - \text{rank (NI)}, \text{ and}$$

(3.32)

$$\text{NBC\_RANK} = \text{rank (XINT-IINT)} - \text{rank (NFO)}$$

[42] The sample size reduction arises from the measurement of these features. The formulae of DIV and NBC require the denominators be positive. However, several firms have negative net income or negative net financial obligations.

where *rank(x)* or x^q is the 100qth percentile, *DVC* is common dividends, *NI* denotes net income, *XINT* represents interest expense, and *IINT* denotes interest income.

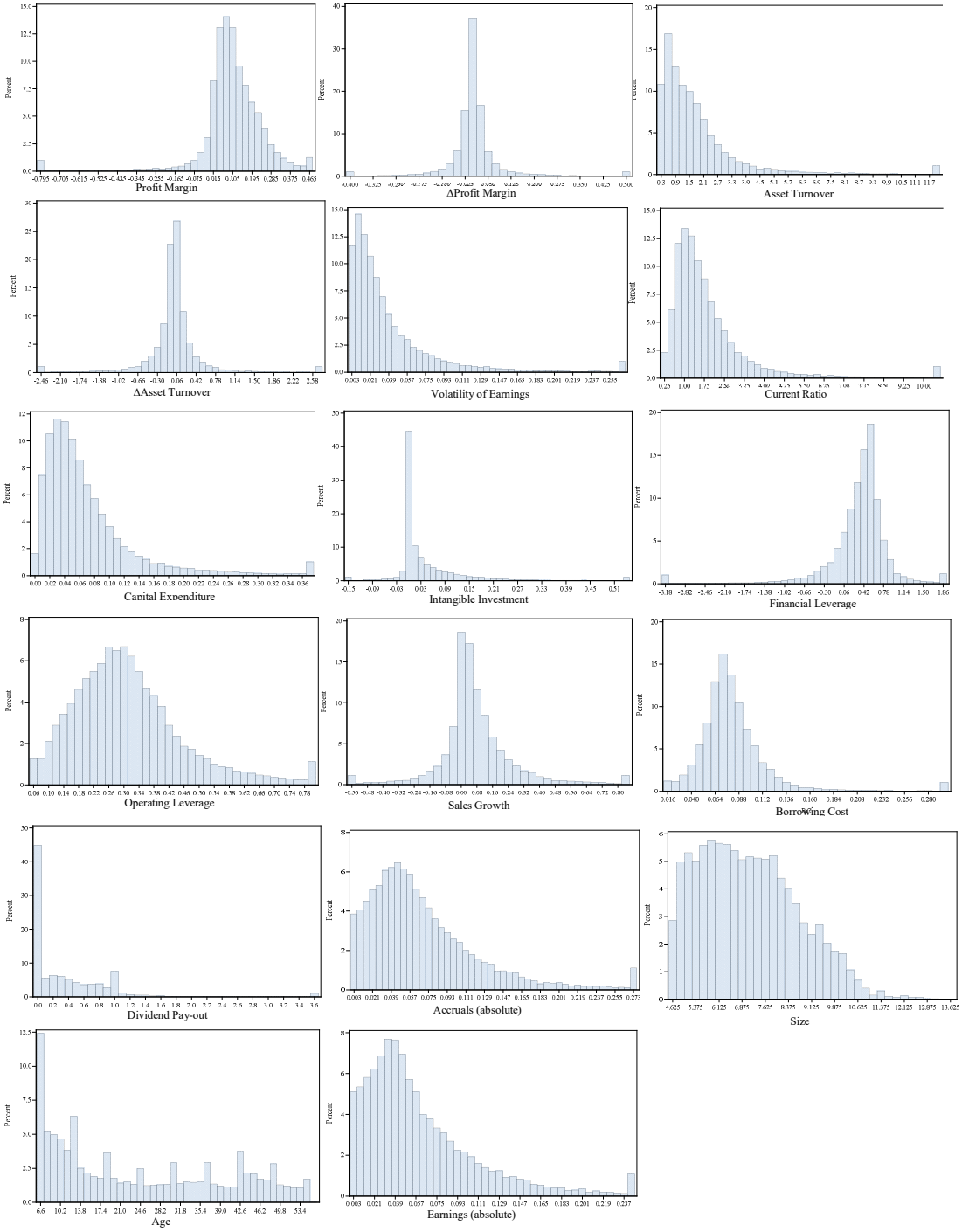


Figure 3.2- Distribution of Cluster Features Winsorized at 1% Top and Bottom

The second concern relates to the noise of cluster features' distribution. As mentioned in Section 3.2.2.1.3 (*Data Processing for Cluster Features and Regression Variables*), truncating or winsorizing all features is not ideal. Truncating all features at the top and bottom 1% shrinks the sample dramatically[43]. Winsorizing cluster features is also problematic. Distributions of real financial data (especially financial ratios) are very noisy and are dissimilar in the tails. As a result, applying a single winsorizing threshold (normally 1% top and bottom) for all features is not feasible, and several features' distributions remain noisy (long tails) after winsorizing. Moreover, winsorization causes extreme observations to be assigned the same value. Hence, these observations will have zero distances, distorting the results from cluster analysis.

To avoid this problem, this thesis follows Tanioka & Yadohisa (2012). Testing six ways of standardization of both real and simulated data, they observe that a ranking method is the most effective way to deal with noise data. Accordingly, all cluster features are transformed into ranked features. There are two versions of ranking. The first version ranks features by percentiles as in Tanioka & Yadohisa (2012). The second version is a normal-distribution ranking where cluster features are first ranked and then transformed into normal distributions. This aligns well with the original data for our features which tend to display relatively normal distributions, after the skewness is removed[44]. For brevity, only the second version is presented here. The results remain unchanged if the first version is used. These are available upon request from the author.

3.4 CHAPTER SUMMARY

This chapter describes empirical specifications and the data employed to test hypotheses developed in Chapter 2 to address the two research aims identified in Chapter 1.

Section 3.2.1 presents the econometric framework and the procedure to execute the newly proposed ROWK clustering technique. The research design uses simulated data to test

[43] If we truncate V variables with a% top and bottom in sequence, the original sample, S_0 will reduce to $S_R = S_0(1-2a)^V$. In our case with V=17 and a=1, $S_R = 0.793\ S_0$, sample size is reduced by over 20%.
[44] The thesis uses the SAS syntax PROC RANK with option NORMAL=BLOM to transform features into the second version format.

the hypotheses relating to the first research aim. Specifically, the ROWK technique is developed to address the problem of HSGC (Hypotheses H1 to H3) and improve the performance of clustering (Hypothesis H4). Next, Section 3.2.2 presents the econometric framework to test hypotheses relating to the second research aim. The steps to apply ROWK clustering in earnings persistence models are presented. Then it discusses tests of the performance of ROWK clustering to identify earnings persistence patterns (Hypothesis H5), identify different contributions of features (Hypothesis H6), and more accurately predict future earnings relative to other benchmarks (Hypothesis H7). Measures of earnings management, firm life cycles and conservatism are introduced to support the testing of hypothesis H8. Finally, two approaches are presented to test hypothesis H9 to assess whether analysts are able to recognize earnings persistence patterns identified by ROWK clustering. Section 3.3 explains the data sources, sample composition and coverage.

The next two chapters present and discuss the empirical results of hypotheses testing by applying the data to the models described in this chapter. Specifically, Chapter 4 describes the simulation results with respect to three channels through which the new ROWK clustering method is posited to demonstrate superiority relative to current clustering techniques in addressing the problem of HGSC (Research Aim 1). Chapter 5 then presents the results from the application of ROWK clustering to explore the patterns of earnings persistence (Research Aim 2).

CHAPTER 4

ROWK AND THE PROBLEM OF HGSC-SIMULATION RESULTS

4.1 INTRODUCTION

This chapter presents the empirical results for the first research aim. Specifically, it discusses the findings pursuant to hypotheses H1 and H2 to identify the determinants that improve the performance of K-means clustering for cluster identification and to solve the problem of HGSC. Additionally, evidence is presented of the effect of standardization on the performance of K-means clustering (hypothesis H3). The findings from three case studies are also presented, highlighting the channels through which ROWK improves the performance of cluster analysis in dealing with the HSGC issue (hypothesis H4).

First, Section 4.2 involves the identification of factors that affect the performance of clustering with respect to the precision of cluster recognition and regression estimations. Four examined factors are class density, class centroid distance, the degree of heterogeneity in regression coefficients and multicollinearity. Then, the empirical results regarding hypothesis H3 which asserts the impact of standardization on the performance of K-means clustering via different contexts are presented in Section 4.3.

Section 4.4 presents the findings for hypothesis H4, which posits three channels through which ROWK improves the performance of cluster analysis with respect to the HSGC issue. Section 4.5 displays the results of several robustness tests. Finally, Section 4.6 presents a summary of this chapter and confirms the findings with respect to the first research aim.

4.2 DETERMINANTS OF K-MEANS CLUSTERING PERFORMANCE

The results of the simulation data analysis include three parts. The first part tests hypotheses H1 and H2 to identify the determinants of K-means clustering performance. Hypothesis H1a relates to distances between classes. *Class* ξ_1^0 is assumed to be the most noticeable class, which is furthest away from the other classes and accordingly, its coefficients are considerably different from the others. As a result, the performances of clustering and of regression estimation (i.e. *MAR*) are mainly impacted by whether or not the members of *Class* ξ_1^0 are correctly identified. Hence λ, the parameter that represents the extent of differences between *Class* ξ_1^0 and other classes is used to proxy distances between classes. Hypothesis H1b relates to the densities of classes as measured by θ. Measures of cluster validation are entropy and purity indexes, as modelled in Equations 3.12 to 3.16. The mean of the absolute residuals (*MAR*) is employed as an indicator of regression performance[45].

4.2.1 Class Density, Class Centroid Distance and Heterogeneity of Regression Coefficients

For hypotheses H1a, H1b and H1c, K-means clustering is run across different sets of simulation parameters. There are 5,000 observations (N) which belong to 5 classes (K^0). Each class has 1,000 members (i.e. $N_{\xi_k}^0 = 1000, k = 1, \dots, 5$). There are two independent variables $x_p, p = 1,2$. Regression error terms (u_i) follow an $N(0,1)$ distribution[46]. For all classes, the intercept ($\alpha_{\xi_k^0}$) and slope coefficient of x_2 (i.e. β_{2,ξ_k^0}) are set to equal 1 and 0.5 respectively. The number of clustering features (V) is set to 4. The within covariance matrix of features is set to equal the diagonal matrix, $\sum_{v_i v_j} = cov(\tau_{i.v}^k, \tau_{i.v}^k) = D$.

Because these hypotheses focus on clustering in general, rather than setting relative weights, feature weights are set to be equal, i.e. $den_{k,v} = 1, \theta_k = \theta$ for all v=1,…,V and k=1,…,K^0. Hence $\sigma_{k,v}^2 = \theta^2/4$ and the sum of the squared within-class distances equals θ^2. This simplifies the analysis and establishes the benchmark to test hypothesis

[45] For conciseness, only the mean of absolute residuals (*MAR*) is presented when testing hypotheses H1 to H4. The findings (untabulated) remain unchanged when the mean of squared residuals (*MSR*) is used.
[46] For simplicity, it is assumed that variations of error terms are homogeneous across classes. Given ROWK aims to identify class membership, this assumption can be relaxed without any material effect on the final result.

H2 when features are highly correlated. For each set of parameters, 100 simulated data samples are generated. Then K-means with unstandardized features is run using these 100 samples. The averages are then reported. Consequently, paired t-tests are employed to test for differences in means.

Class Density

Suppose the number of classes is correctly identified ($K' = K^0 = 5$). Table 4-1 presents an analysis of the effect of class distance (λ) and class density (θ) on performance of clustering and regression estimation. For brevity, only purity versions 2 and 3 are presented. However, results are unchanged if entropy and purity version 1 are included. Consistent with hypothesis H1a, the purities (both for *Class* ξ_1^0 and overall) monotonically reduce when class densities (as proxied by θ) decrease[47]. Particularly, Column 9 shows that at λ=2, p_1^{ver3} significantly diminishes from 100% to only 68% when θ increases from 1 to 4. The reduction in p_1^{ver3} is significant at the 1% level. Additionally, when λ reduces to 1.5 and 1, the reductions in p_1^{ver3} are even significantly larger, i.e. from 99.6% to 58.4% and from 99.1% to 51.5%, respectively. This evidence supports the proposition in hypothesis H1a that the positive relationship between cluster density and clustering precision is stronger when distances between classes' centres are lower.

[47] Note that θ is an inverse measure of density. Note also that as expected, the mean squared distances between members and their class centres are approximately θ^2, as observed in column 6.

Table 4-1: Performance of K-means Clustering For Different Class Centroid Distances and Densities Cluster Validation

Regression Model:
$y_i = \alpha_{\xi_k^0} + x_i'\beta_{\xi_k^0} + u_i, i = 1, \dots, 5000; k = 1, \dots, 5$; where $\alpha_{\xi_k^0}$ and $\beta_{\xi_k^0}$ be 1x1 and Px1 vectors of group-specific intercept and slope coefficients; $u_i \sim N(0,1)$ and $cov(u_i, u_j) = I$(identity matrix); $cov\left(x_{i,\xi_k^0}, u_{j,\xi_k^0}\right) = 0$; $\alpha_{\xi_k^0}$=1 and β_{2,ξ_k^0}=0.5 for all k=1,…,5. Each class has 1000 observations, $N_k^0 = 1000$.
Class Membership:
There are 4 features, i.e. $Z_v, v = 1, \dots, 4$; $z_{i,v}^k = \left(c_{k,v} + \tau_{i.v}^k\right)$ where $c_{k,v}$ is the centre of class k measured by feature Z_v; $\tau_{i.v}^k \sim N(0, \theta^2/4)$ where θ^2 is the expectation of the mean squared distance between members to its corresponding centre; $\sum_{v_i v_j} = cov\left(\tau_{i.v}^k,\ \tau_{i.v}^k\right) = D$; $d_{\xi_1^0,\xi_j^0} > \lambda\, d_{\xi_i^0,\xi_j^0}$ for all $i \neq 1$ where $d_{\xi_i^0,\xi_j^0}$ denotes the distance between Class ξ_i^0 and Class ξ_j^0 and λ indicates the extent of differences between Class ξ_1^0 and other classes. $Z_1 = X_1$ and $Z_2 = X_2$.
Assume that number of classes is correctly identified (K'=5). For each set of {λ,θ), K-means clustering with unstandardized features is run one hundred times, then the average of cluster purity across these one hundred runs is calculated and reported.

Parameters		Average Distances Between Classes Centres				Class Purity (%)			
λ (1)	θ (2)	Class ξ_1^0 (3)	Class ξ_{2-5}^0 (4)	All classes (5)	Mean squared distance (6)	Ver2_Classξ_1^0 (7)	Ver2_all_class (8)	Ver3_Classξ_1^0 (9)	Ver3_all_class (10)
1.0	1.0	3.020	1.992	2.404	0.985	98.8	85.7	99.1	85.6
	2.0				3.941	82.3	61.4	81.7	61.4
	3.0				8.868	63.7	43.3	61.6	43.0
	4.0				15.765	54.7	38.7	51.5	36.8
1.5	1.0	3.839	1.356	2.349	0.985	50.2	74.7	99.6	68.3
	2.0				3.941	85.6	56.2	95.4	55.1
	3.0				8.868	77.9	45.4	72.8	43.8
	4.0				15.765	66.5	39.7	58.4	37.7
2.0	1.0	4.119	1.310	2.433	0.985	100.0	76.9	100.0	76.6
	2.0				3.941	92.9	56.8	97.1	57.0
	3.0				8.868	78.0	45.8	84.8	45.8
	4.0				15.765	68.4	38.7	68.0	37.3

Class Centroid Distance

Regarding distances between class centres (λ), holding the level of classes' densities θ fixed, p_1^{ver3} is significantly lower when the distances between the centre of *Class* ξ_1^0 and the centres of other classes are higher (as λ moves from 1 to 2). Specifically, at θ=1, weak evidence is presented of an increase from 99.1% to 100% in p_1^{ver3} (p<0.1). However, when class densities diminish at θ=3(4), p_1^{ver3} rises significantly (p<0.01) from 61.6%

(51.5%) to 84.8% (68.0%). The findings are consistent with hypothesis H1b that the higher the distances between class centres, the higher the precision of class membership identification. Furthermore, this positive relationship between distances between class centres and clustering precision is stronger when class densities are lower.

The findings remain unchanged when p_1^{ver3} is replaced by $p_{overall}^{ver3}$ and $p_{overall}^{ver2}$. However, there is an exception for the case of p_1^{ver2} at *{λ,θ} = {1.5, 1}*. In this case, instead of increasing, p_1^{ver2} drops drastically from 98.8% to only 50.2% when λ increases from 1 to 1.5. Consequently, the relationships between density, centre distance and purity contradict hypotheses H1a and H1b. To clarify this, two points are considered. First, the contradictory result at *{λ,θ} = {1.5, 1}* only occurs if p_1^{ver2} is used. The results for other purity measures are as expected. Second, p_1^{ver2} indicates the extent to which *Class* ξ_1^0 contains objects from a single cluster. Therefore, the low value of p_1^{ver2} at *{λ,θ} = {1.5, 1}* may be a signal that members of *Class* ξ_1^0 are *appropriately* divided into two or more clusters. The intuition is that although members of *Class* ξ_1^0 are assigned to more than one cluster, these clusters comprise members mostly from *Class* ξ_1^0 with few or no members from other classes. As a result, regression estimations using these cluster memberships consistently achieve good results. Panel A of Table 4-2 supports this argument. As can be seen from this table, 1,000 members of *Class* ξ_1^0 are assigned almost equally (502 and 498) into two clusters, i.e. ξ_2 and ξ_3. More importantly, members of *Class* ξ_1^0 make up nearly 100% of members of *Clusters* ξ_2 and ξ_3. As a result, the true slope coefficients of *Class* ξ_1^0 can be consistently estimated when running regressions within each of *Clusters* ξ_2 and ξ_3[48].

[48] The results (unreported) support this statement.

Table 4-2: Frequency of Class Membership by Cluster

Frequency of class membership by clusters at *(λ,θ)=(1.5;1)*

Panel A: Number of clusters (K') = 5

Class/cluster	1	2	3	4	5	Sum
1	0	**502**	**498**	0	0	1000
2	942	0	0	28	30	1000
3	91	0	0	683	226	1000
4	12	2	0	755	231	1000
5	31	0	0	117	852	1000
Sum	1076	504	498	1583	1339	5000

Panel B: Number of cluster (K') = 6

Class/cluster	1	2	3	4	5	6	Sum
1	0	0	511	0	0	489	1000
2	895	1	0	82	22	0	1000
3	54	224	0	597	125	0	1000
4	6	600	0	278	116	0	1000
5	23	91	0	128	758	0	1000
Sum	978	916	511	1085	1021	489	5000

Overall, the results presented in Panel A of Table 4-2 are consistent with hypotheses H1a and H1b. The contradictory value of p_1^{ver2} at *{λ,θ} = {1.5, 1}* does not rule out these hypotheses, but rather it provides an important signal to a situation where a certain class (i.e. *Class* ξ_1^0) is sufficiently far from other classes to allow K-means clustering to divide this class into smaller pure clusters. Regression estimates are still consistent if it is run within these clusters. More importantly, these findings shed light on those cases where the optimal number of clusters in K-means should not coincide with the true number of classes. For example, as can be seen from Panel B of Table 4-2, with respect to the case where *{λ,θ} = {1.5, 1}*, the precision of K-means clustering improves when the number of clusters used in clustering is set to six. Particularly, members of *Class* ξ_1^0 are correctly assigned to *Clusters* ξ_3 and ξ_6. Additionally, all members of these two clusters also belong to *Class* ξ_1^0. The improvements displayed here also derive from the fact that *Clusters* ξ_1, ξ_2, $\xi_4$and ξ_5 correctly identify the majority of members of *Classes* ξ_2^0, ξ_4^0, ξ_3^0 and ξ_5^0, respectively.

Heterogeneity of Regression Coefficients

Table 4-3 provides the results relating to the regression estimations. Consistent with hypothesis H1c, *MAR* significantly reduces when λ increases while holding θ fixed, θ reduces while holding λ fixed, or when differences in regression coefficients between classes are larger. Specifically, when θ is fixed at 4 and λ increases from 1 to 2, *MAR* significantly declines from 1.072 to 1.049 and from 1.447 to 1.340 when $\left\{\beta_{1,\xi_1}, \beta_{1,\xi_{2-5}}\right\}$= {1,0.5} and {1,0} respectively. In relation to class density, at λ=1, an increase of θ from 1 to 4 leads to a significant increase of *MAR* from 0.962 to 1.072 and from 0.964 to 1.447 when $\left\{\beta_{1,\xi_1}, \beta_{1,\xi_{2-5}}\right\}$ moves from {1,0.5} to {1,0}.

The degree of coefficients' heterogeneity is an important determinant of how K-means clustering improves the regression estimations. When there are no HGSC (i.e. = $\left\{\beta_{1,\xi_1}, \beta_{1,\xi_{2-5}}\right\}$ = {1,1}, for all combinations of (λ, θ), *MAR* remains almost unchanged at around 0.962. This is nearly equal to the results of regressions without clustering (hovering at 0.965) and to those of the ideal case (0.961)[49]. However, when β_{1,ξ_1} differs from those of $\beta_{1,\xi_{2-5}}$, regression estimations following clustering results have significantly lower *MARs* compared to those not exploiting clustering results. In addition, the reductions of *MARs* are larger when differences between β_{1,ξ_1} and $\beta_{1,\xi_{2-5}}$ increase. For example, at (λ, θ) = (2, 4) and $\left\{\beta_{1,\xi_1}, \beta_{1,\xi_{2-5}}\right\}$ = {1,0.5}, the reduction in *MAR* is 0.054 (=1.103-1.049) which is strongly significantly lower than 0.228 (=1.568-1.340) as in the case of $\left\{\beta_{1,\xi_1}, \beta_{1,\xi_{2-5}}\right\}$ = {1,0}. Hence, there is strong evidence to support hypothesis H1c which states that the MAR when running regressions within clusters is lower when the distances between clusters' centres are greater, or the clusters' densities are higher, or the differences of regression coefficients between clusters are larger.

[49] The ideal case is where knowledge of classification is already known.

Table 4-3: Performance of K-means Clustering For Different Class Centroid Distances and Densities

Regression Model:

$y_i = \alpha_{\xi_k^0} + x_i'\beta_{\xi_k^0} + u_i, i = 1,\dots,5000; k = 1,\dots,5$; where $\alpha_{\xi_k^0}$ and $\beta_{\xi_k^0}$ be 1*x*1 and *Px*1 vectors of group-specific intercept and slope coefficients; $u_i \sim N(0,1)$ and $cov(u_i, u_j) = I$(identity matrix); $cov\left(x_{i,\xi_k^0}, u_{j,\xi_k^0}\right) = 0$; $\alpha_{\xi_k^0}$=1 and β_{2,ξ_k^0}=0.5 for all k=1,…,5. Each class has 1000 observations, $N_k^0 = 1000$.

Class Membership:

There are 4 features, i.e. $Z_v, v = 1,\dots,4$; $z_{i,v}^k = \left(c_{k,v} + \tau_{i.v}^k\right)$ where $c_{k,v}$ is the centre of class *k* measured by feature Z_v; $\tau_{i.v}^k \sim N(0, \theta^2/4)$ where θ^2 the is expectation of mean squared distance between members to its corresponding centre; $\sum_{v_i v_j} = cov(\tau_{i.v}^k, \tau_{i.v}^k) = D$; $d_{\xi_1^0,\xi_j^0} > \lambda\, d_{\xi_i^0,\xi_j^0}$ for all $i \neq 1$ where $d_{\xi_i^0,\xi_j^0}$ denotes the distance between Class ξ_i^0 and Class ξ_j^0 and λ indicates the extent of differences between *Class* ξ_1^0 and other classes. $z_1 = x_1$ and $z_2 = x_2$.

MAR is mean absolute residuals; ideal *MAR* is *MAR* if members are 100% correctly assigned; *CLUSTER (ALL)* are *MAR*s of running a regression of Equation 3.1within each cluster (for the all data sample). Assume that number of classes is correctly identified (*K'*=5). For each set of {*λ,θ*), K-means clustering with unstandardized features is run one hundred times, then the average of cluster purity and *MAR* across these one hundred runs are calculated.

Parameters		Mean Distances Between Classes Centres				Mean Absolute Residuals $(\beta_{1,\xi_1^0}, \beta_{1,\xi_{2-5}^0})$							
					Mean	(1,1)		(1,0.5)		(1,0)	(1,1)		
λ	θ	Class ξ_1^0	Class ξ_{2-5}^0	All classes	squared distance	Cluster	All sam	Cluster	All sam	Cluster	All sam	Ideal *MAR*	
(1)	(2)	(3)	(4)	(5)	(6)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	
1.0	1.0	3.020	1.992	2.404	0.985	0.961	0.965	0.962	0.976	0.964	1.019	0.961	
	2.0				3.941	0.962	0.965	0.973	1.001	1.020	1.131	0.961	
	3.0				8.868	0.962	0.965	1.013	1.044	1.196	1.317	0.961	
	4.0				15.765	0.963	0.965	1.072	1.106	1.447	1.580	0.961	
1.5	1.0	3.839	1.356	2.349	0.985	0.962	0.965	0.962	0.990	0.963	1.089	0.961	
	2.0				3.941	0.963	0.965	0.972	1.025	1.008	1.242	0.961	
	3.0				8.868	0.963	0.965	1.011	1.073	1.177	1.446	0.961	
	4.0				15.765	0.964	0.965	1.072	1.137	1.447	1.716	0.961	
2.0	1.0	4.119	1.310	2.433	0.985	0.961	0.964	0.961	0.971	0.961	1.006	0.961	
	2.0				3.941	0.961	0.965	0.967	0.996	0.984	1.117	0.961	
	3.0				8.868	0.961	0.965	0.997	1.040	1.116	1.304	0.961	
	4.0				15.765	0.964	0.965	1.049	1.103	1.340	1.568	0.961	

4.2.2 Multicollinearity

Table 4-4 presents the performance of K-means clustering when features that are used in clustering are highly correlated. All parameters are kept the same, with an exception that the correlation matrix of features is set to be not equal to the identity matrix as in Table 4-1 and Table 4-3. For brevity, this section only reports the simulation case where correlations within class are generated. The results for full sample correlations between features remain the same, and therefore are not reported. Since the correlation matrices are assumed to be identical across different classes, only the correlation matrix is reported for *Class* ξ_1^0.

Panel A of Table 4-4 displays correlation matrices of features for both within Class ξ_1^0 and the full sample. Within *Class* ξ_1^0, there are highly significant correlations between z_1 and z_2 ($\rho_{Z1Z2}^{\xi_1^0}$=0.705), z_3 and z_4 ($\rho_{Z3Z4}^{\xi_1^0}$=0.731), whereas there are no significant correlations between other pairs of features. However, when the full sample is used to calculate correlations, the correlation patterns differ considerably. All pairs of features are significantly correlated at the 1% level. This evidence suggests the need to identify whether the correlations of features stem from co-movements of the features within classes or across the full sample. Specifically, $\rho_{Z1Z2}^{\xi_k^0}$ and $\rho_{Z3Z4}^{\xi_k^0}, k = 1, \dots, 4$ are identified as potential multicollinearity concerns to control in clustering, while other pairs of correlations are not. This provides an important explanation of why the use of FA cannot resolve the problem of multicollinearity in K-means clustering when features' correlations originate within clusters and not in the full sample. Notably, FA is used to deal with only full-sample correlations. Moreover, even when features' correlations stem from co-movement in the full sample, the use of extracted principal components in subsequent CA may not ensure good performance. This can be explained by the relatively few first components that are used in CA, whereupon there is no guarantee that these components contain the target signal that the researcher is seeking to identify using CA (Witten & Tibshirani, 2010, p. 713).

Panel B of Table 4-4 displays classes' purities and *MARs* when λ is fixed as 1, and θ*=1,…,4*. The findings, which are unreported for the sake of brevity, are similar when

λ=1.5 and 2. Panel C of Table 4-4 compares classes' purities and *MARs* between the cases of uncorrelated features (as presented in Table 4-1and Table 4-3) and correlated features. Consistent with Hypothesis H2, for most cases, classes' purities (*MARs*) are significantly lower (higher) when features are highly correlated as compared to those of uncorrelated features.

In addition, the effects of multicollinearity on class purity do not follow a monotonic pattern. As can be seen in Column 3, Panel C, the difference of $p_{\substack{\xi_0 \\ \xi_1}}^{ver2}$ is modest at 0.7% when θ=1 (high density), jumps drastically to 16.3% when θ=2 and declines when θ increases to 3 and 4 (low density). The interpretation is that when overlapping of class membership is extremely low or high, the effect of multicollinearity on clustering precision is trivial. However, the effect is stronger when the blurriness of class pattern is low, but not so low that there are more chances that the loss due to multicollinearity significantly impacts the precision of classification. Overall, there is evidence to support hypothesis H2 that multicollinearity negatively impacts K-means clustering performance, and consequently hampers regression estimations.

Table 4-4: Effects of Multicollinearity on Performance of K-means Clustering

Regression Model:

$y_i = \alpha_{\xi_k^0} + x_i'\beta_{\xi_k^0} + u_i, i = 1,\dots,5000; k = 1,\dots,5$; where $\alpha_{\xi_k^0}$ and $\beta_{\xi_k^0}$ be 1x1 and Px1 vectors of group-specific intercept and slope coefficients; $u_i \sim N(0,1)$ and $cov(u_i, u_j) = I$(identity matrix); $cov\left(x_{i,\xi_k^0}, u_{j,\xi_k^0}\right) = 0$; $\alpha_{\xi_k^0}$=1 and β_{2,ξ_k^0}=0.5 for all k=1,…,5. Each class has 1000 observations, $N_k^0 = 1000$.

Class Membership:

There are 4 features, i.e. $Z_v, v = 1,\dots,4$; $z_{i,v}^k = \left(c_{k,v} + \tau_{i.v}^k\right)$ where $c_{k,v}$ is the centre of class k measured by feature Z_v; $\tau_{i.v}^k \sim N(0, \theta^2/4)$ where θ^2 is expectation of the mean squared distances between members to its corresponding centre; $\sum_{v_i v_j} = cov(\tau_{i.v}^k,\ \tau_{i.v}^k) \neq D$ and is described in Panel A of the table; $d_{\xi_1^0,\xi_j^0} > \lambda\, d_{\xi_i^0,\xi_j^0}$ for all $i \neq 1$ where $d_{\xi_i^0,\xi_j^0}$ denotes the distance between *Class* ξ_i^0 and *Class* ξ_j^0 and λ indicates the extent of differences between *Class* ξ_1^0 and other classes; $z_1 = x_1$ and $z_2 = x_2$.

MAR is mean absolute residuals; *CLUSTER (ALL)* are *MARs* from running regressions of Equation 3.1 within each clusters (for the all data sample). Assume that number of classes is correctly identified (K'=5). For each set of {λ,θ), K-means clustering with unstandardized features is run one hundred times, then the average of cluster purity and *MAR* across these one hundred runs are calculated. Paired t-tests are used to test for significant differences in means. P-values and t-statistics are presented in brackets in Panels A and C, respectively. *, **, *** denote significance at 10%, 5% and 1%, respectively.

Panel A : Pearson Correlation Coefficients; Prob > |r| under H0: Rho=0

Class ξ_1^0 (N=1000)					All sample (N=5000)				
	X1	X2	X3	X4		X1	X2	X3	X4
X1	1	0.7046	0.0026	-0.0225	X1	1	0.6772	0.1052	0.3233
		(<.0001)	(0.9334)	(0.4770)			(<.0001)	(<.0001)	(<.0001)
X2		1	-0.0405	-0.0404	X2		1	-0.4164	0.2501
			(0.2005)	(0.2010)				(<.0001)	(<.0001)
X3			1	0.7314	X3			1	0.3841
				(<.0001)					(<.0001)
X4				1	X4				1

Panel B : Purities and MSEs When Clustering Features Are Correlated

Parameters		Class Purity (%)				Mean Absolute Residuals $(\beta_{1,\xi_1^0}, \beta_{1,\xi_{2-5}^0})$		
		Ver2_ Classξ_1^0	Ver2_ all class	Ver3_ Classξ_1^0	Ver3_ all class	(1,1)	(1,0.5)	(1,0)
λ (1)	θ (2)	(3)	(4)	(5)	(6)	Cluster (7)	Cluster (9)	Cluster (11)
1.0	1.0	98.1	81.5	97.2	81.5	0.963	0.963	0.966
	2.0	66	46.9	84.2	51.4	0.963	0.977	1.038
	3.0	50.6	37.2	54.5	37.9	0.963	1.018	1.216
	4.0	39.8	31.8	43.5	32.5	0.962	1.080	1.476

Panel C: Differences of Purities (%) and MSRs Between Cases of Uncorrelated and Correlated Clustering Features

1.0	1.0	**0.7*** *(1.43)*	**4.2***** *(4.51)*	**1.9***** *(2.67)*	**4.1***** *(3.97)*	-0.002 *(-0.49)*	-0.001 *(-0.67)*	-0.002 *(-1.13)*
	2.0	**16.3***** *(9.72)*	**14.5***** *(8.19)*	**-2.5*** *(-1.34)*	**10***** *(6.54)*	-0.001 *(-0.52)*	**-0.004*** *(-1.33)*	**-0.018***** *(-13.52)*
	3.0	**13.1***** *(9.55)*	**6.1***** *(4.55)*	**7.1***** *(4.73)*	**5.1***** *(5.31)*	-0.001 *(-0.63)*	**-0.005**** *(-2.26)*	**-0.02***** *(-15.28)*
	4.0	**14.9***** *(7.80)*	**6.9***** *(3.45)*	**8***** *(5.43)*	**4.3***** *(3.81)*	0.001 *(0.43)*	**-0.008***** *(-4.46)*	**-0.029***** *(-21.23)*

4.3 EFFECT OF STANDARDIZATION

K-means clustering with standardized features is a conventional way to apply clustering in the business domain (e.g. Li & Li, 2008). However, the ranking method is documented to be the most effective method of data pre-processing for K-means clustering (Tanioka & Yadohisa, 2012). Hypothesis H3 further argues that when a feature's weight results from differences of distanc

es between class centres measured by the feature relative to those measured by other features, standardization makes the performance of clustering worse, even compared to that of non-standardization[50].

Hypothesis H3a posits that when a feature weight results from differences of class densities measured by the feature relative to those measured by other features, the standardization method (thereafter *STD_K*) improves the performance of CA as compared to those of unstandardized features (thereafter *UNSTD_K*), since standardization partly adjusts the features' scales in the same direction as the true relative features' weights. Panel A of Table 4-5 displays the results of K-means clustering performance when feature z_1 is generated to have the highest class densities among clustering features ($den_{k,1} = 0.5$ and $den_{k,v} = 1, v = 2, \ldots, 4$; λ=1 and θ=2). As expected, z_1 (0.978) has the lowest standard deviation among the four features (1.266 for z_2, 1.557 for z_3 and 1.209 for z_4). Given that the purpose of standardization is to reduce (increase) variances of large (small) scale features, it increases the scale of z_1 in clustering, relative to others. As a result, the precision of clustering, as proxied by p_1^{ver2}, $p_{overall}^{ver2}$, and $p_{overall}^{ver3}$, increases significantly at 4%, 2.18% and 3.71% respectively, when features are standardized, which is consistent with hypothesis H3a.

To test hypothesis H3b, feature z_1 is set to have more weight than other features via the following inequality:

(4.1)

[50] In this chapter, without further reference, standardization means transforming original features into corresponding z-scores.

$$d^1_{\xi_1^0,\xi_j^0} > \lambda_1\, d^v_{\xi_i^0,\xi_j^0} \text{ for all } i \neq 1 \text{ and } v \neq 1$$

where $d^v_{\xi_i^0,\xi_j^0}$ denotes the distance between *Class* ξ_i^0 and *Class* ξ_j^0 as measured by feature z_v.

Panel B of Table 4-5 reports the results when λ_1=2, λ=1 and θ=2. The first important finding is that, as expected, z_1 has the largest standard deviation (i.e. 1.696 vs. 1.126, 1.135 and 1.036). Second, opposite to the results of Panel A, for all four measures of class purity, the performance of *STD_K* is significantly inferior as compared to those of *UNSTD_K*. The differences are -6.20%, -2.58%, -1.54% and -2.40% when the purity measures are p_1^{ver2}, $p_{overall}^{ver2}$, p_1^{ver3} and $p_{overall}^{ver3}$, respectively. Consequently, standardization of features leads to significantly larger *MARs* than those of unstandardized features. This negative effect is even worse when regression coefficients differ considerably across classes. For example, the increase in *MAR* is 0.011 when $\left\{\beta_{1,\xi_1}, \beta_{1,\xi_{2-5}}\right\}$ = {1,0.5}, then its jumps considerably to 0.048 when $\left\{\beta_{1,\xi_1}, \beta_{1,\xi_{2-5}}\right\}$={1,0}. Overall, the aforementioned evidence supports hypothesis H3b which proposes that when a feature weight results from differences of distances between class centres measured by the feature relative to those measured by other features, standardization decreases the performance of clustering as compared to those of an unstandardized feature.

In summary, the findings from Table 4-5 are consistent with hypotheses H3a and H3b. This underscores the necessity for K-means clustering users to scrutinize the underlying causes of differences in feature scales. Standardization may improve the precision of K-means clustering when a feature weight results from differences of features' class densities, but it also makes clustering performance worse when a feature weight results from differences of distances between class centres measured by the feature. In addition, even when standardizing of features enhances K-means clustering as in the first case, it does not fully adjust the scale of features to their truly relative weights. In contrast, this is appropriately addressed by the innovative ROWK method proposed in this thesis. The findings presented in the next section support this statement.

Table 4-5: Effects of Feature Standardization on Performance of K-means Clustering

Regression Model:
$y_i = \alpha_{\xi_k^0} + x_i'\beta_{\xi_k^0} + u_i, i = 1,\dots,5000; k = 1,\dots,5$; where $\alpha_{\xi_k^0}$ and $\beta_{\xi_k^0}$ be 1x1 and Px1 vectors of group-specific intercept and slope coefficients; $u_i \sim N(0,1)$ and $cov(u_i, u_j) = I$(identity matrix); $cov\left(x_{i,\xi_k^0}, u_{j,\xi_k^0}\right) = 0$; $\alpha_{\xi_k^0}$=1 and β_{2,ξ_k^0}=0.5 for all k=1,…,5. Each class has 1000 observations, $N_k^0 = 1000$.
Class Membership:
There are 4 features, i.e. $Z_v, v = 1,\dots,4$; $z_{i,v}^k = \left(c_{k,v} + \tau_{i.v}^k\right)$ where $c_{k,v}$ is the centre of class k measured by feature Z_v; $z_{i,v}^k = \left(c_{k,v} + \tau_{i.v}^k\right)$ where $\sigma_{k,v}^2 = den_{k,v}^2 * \frac{\theta_k^2}{\sum_{v=1}^{V} den_{k,v}^2}$; $\sum_{v_i v_j} = cov\left(\tau_{i.v}^k,\ \tau_{i.v}^k\right) = D$; $Z_1 = X_1$ and $Z_2 = X_2$. In Panel A, $den_{k,1} = 0.5$ and $den_{k,v} = 1, v = 2,\dots,4$ and $k = 1,\dots,5$. In Panel B, $den_{k,v} = 1, v = 1,\dots,4$; $d_{\xi_1^0,\xi_j^0} > \lambda\, d_{\xi_i^0,\xi_j^0}$ and $d^1{}_{\xi_1^0,\xi_j^0} > \lambda_1\, d^v{}_{\xi_i^0,\xi_j^0}$ for all i≠1 and v≠1 where $d_{\xi_i^0,\xi_j^0}$ ($d^v_{\xi_i^0,\xi_j^0}$) denotes the distance between Class ξ_i^0 and Class ξ_j^0 as measured by all features (as measured by feature Z_v). λ (λ_1) indicates the extent of differences between Class ξ_1^0 and other classes as measured by all features (as measured by feature Z_v).
MAR is mean absolute residuals. Assume that number of classes is correctly identified (K'=5). For each set of {λ,θ), K-means clustering with unstandardized features is run one hundred times, then the average of cluster purity and *MAR* across these one hundred runs are calculated. '*S'* stands for K-means with standardized features and '*U'* stands for K-means with unstandardized features. Paired t-tests are used to test for mean differences. T-statistics are presented in brackets.
*, **, *** denote significance at 10%, 5% and 1%, respectively.

Panel A

Parameters: λ =1; θ=2; $den_{k,1} = 0.5$; $den_{k,v} = 1, v = 2,\dots,4$

Data	Class Purity (%)				MAR ($\beta_{1,\xi_1^0}, \beta_{1,\xi_{2-5}^0}$)			Standard Deviation			
	Ver2_ Class ξ_1^0	Ver2_ all class	Ver3_ Class ξ_1^0	Ver3_ all class	(1,1)	(1,0.5)	(1,0)	Z_1	Z_2	Z_3	Z_4
Std	82	63.3	80.9	63.6	0.962	0.967	0.986				
Unstd	78	61.1	81.2	59.9	0.962	0.967	0.989	0.98	1.27	1.56	1.21
Diff(S-U)	**4***** *(11.34)*	**2.18***** *(4.38)*	-0.208 *(-0.57)*	**3.708***** *(5.89)*	0.000 *(0.00)*	0.000 *(0.00)*	-0.003 *(-0.72)*				

Table 4-5 (cont.) Panel B

Parameters: λ_1=2; λ =1; θ=2 ; $den_{k,v} = 1, v = 1,\dots,4$

Data	Class Purity (%)				Mean Absolute Residuals ($\beta_{1,\xi_1^0}, \beta_{1,\xi_{2-5}^0}$)			Standard Deviation			
	Ver2_ Class ξ_1^0	Ver2_ all class	Ver3_ Class ξ_1^0	Ver3_ all class	(1,1)	(1,0.5)	(1,0)	Z_1	Z_2	Z_3	Z_4
Std	80	54.4	89.5	53.8	0.961	1.004	1.140				
Unstd	86.2	56.9	91	56.2	0.96	0.993	1.092	1.67	1.13	1.14	1.04
Diff(S-U)	-6.2*** *(-15.79)*	-2.58*** *(-7.91)*	-1.54*** *(-5.35)*	-2.40*** *(-7.90)*	-0.005 *(-0.23)*	0.011** *(1.87)*	0.048*** *(11.15)*				

4.4 ROWK PERFORMANCE WITH RESPECT TO THE HGSC PROBLEM

The innovation of the ROWK procedure results from using the regression *MAR* as a guide to adjust feature weights. As illustrated in a previous example (see Section 2.2.3), the performance of grouping methods which merely depend on regression analysis is highly sensitive to interactions between the sign and magnitude of error terms and discrepancies of coefficients across classes. Instead, the ROWK procedure addresses the issue of HGSC by taking into account the rich information from CA, particularly K-means clustering. Furthermore, by using the *MAR* from regression estimations, the ROWK procedure addresses the problem of multicollinearity that is not well-resolved by current weighted K-means clustering procedures (e.g. Amorim & Mirkin, 2012). More importantly, the optimal weights that minimise a regression's *MAR* reflect its importance not only to class identification, but also to improve the regression analysis.

To test these statements, this study uses three case studies. Case 1 includes a simple set of simulated data with uncorrelated features. Case 2 employs the same data as Case 1, but with correlated features. Case 3 analyses a situation where feature weights come from two sources, i.e. contributions to cluster recognition and contributions to regression estimations.

4.4.1 Case Study 1-The First Channel

Panel A of Table 4-6 displays simulated parameters for Case 1. There are 5000 observations which belong to five classes. Each class has 1,000 members. There are four features, z_i, i=1,…,4 and a random variable z_5 (~N(0,1)) which is used as an irrelevant clustering feature.

For simplicity and without loss of generality, only z_3 is generated to have more weights relative to others. To generate more weights for z_3, the study employs two approaches. The first is from class density as measured by a feature, i.e. $den_{k,3} = 0.5$; $den_{k,v} = 1, v \neq 3$. The second stems from distances between class centres as measured by a feature, i.e. $\lambda = \lambda_3 = 0.9$. For the regression model, there are two independent variables x_1 and x_2 which are also features of clustering, i.e. $z_1 = x_1$ and $z_2 = x_2$. x_1 and x_2 satisfy regression assumptions in section 3.2.1.1.1. Regarding the regression model, *Class* ξ_1^0 is generated

to be the most distinguishable class, consequently has its corresponding regression coefficients significantly differentiated from those of other classes. For simplicity, the heterogeneity of regression coefficients only applies for x_2. The intercept and slopes of x_1 are set to be homogeneous. Particularly, $\alpha_{\xi_k^0}$=1 and β_{2,ξ_k^0}=0.5 for all k=1,…,5; $\beta_{1,\xi_1^0} = -1$; $\beta_{1,\xi_2^0} = 0.5$; $\beta_{1,\xi_3^0} = 0.2$; $\beta_{1,\xi_4^0} = 0.1$; $\beta_{1,\xi_5^0} = 0$.

Panel B of Table 4-6 presents descriptive statistics of clustering features for Case 1. Standard deviations of the four features are similar ranging from 1.494 for z_2 to 1.743 for z_3. Panel C of Figure 4.6 exhibits full sample (lower triangle) and within *Class* ξ_1^0 (upper triangle) correlation matrices of clustering features. While all pairs of features, as expected, display no significant correlation within *Class* ξ_1^0, there are significant correlations (five out of six pairs of correlations) between features for the full sample.

Table 4-6: Descriptive Statistics of Clustering Features (Case Study 1)

Regression Model:

$y_i = \alpha_{\xi_k^0} + x_i'\beta_{\xi_k^0} + u_i, i = 1, \ldots, 5000; k = 1, \ldots, 5$; where $\alpha_{\xi_k^0}$ and $\beta_{\xi_k^0}$ be 1*x*1 and *Px*1 vectors of group-specific intercept and slope coefficients; $u_i \sim N(0,1)$ and $cov(u_i, u_j) = I$(identity matrix); $cov\left(x_{i,\xi_k^0}, u_{j,\xi_k^0}\right) = 0$; $\alpha_{\xi_k^0}$=1 and β_{2,ξ_k^0}=0.5 for all k=1,…,5. Each class has 1000 observations, $N_k^0 = 1000$

Class Membership:

There are 4 features, i.e.$Z_v, v = 1, \ldots, 4$ and a random variable $Z_5 \sim N(0,1)$; $z_{i,v}^k = \left(c_{k,v} + \tau_{i.v}^k\right)$ where $c_{k,v}$ is the centre of class *k* measured by feature Z_v; $\tau_{i.v}^k \sim N(0, \theta^2/4)$ where θ^2 is the expectation of the mean squared distances between members to its corresponding centre; $\sum_{v_i v_j} = cov(\tau_{i.v}^k, \tau_{i.v}^k) = D$; $d_{\xi_1^0,\xi_j^0} > \lambda\, d_{\xi_i^0,\xi_j^0}$ and $d^3{}_{\xi_1^0,\xi_j^0} > \lambda_3\, d^v{}_{\xi_i^0,\xi_j^0}$ for all $i \neq 1$ and $v \neq 1$ where $d_{\xi_i^0,\xi_j^0}$ ($d^v_{\xi_i^0,\xi_j^0}$) denotes the distance between Class ξ_i^0 and Class ξ_j^0 as measured by all features (as measured by feature Z_v). λ (λ_1) indicates the extent of differences between Class ξ_1^0 and other classes as measured by all features (as measured by feature Z_v).

Panel A: Simulated parameters

N=5000; $N_k^0 = 1000$, k=1,…,5 ; P=2 and V=5;
$den_{k,3} = 0.5$; $den_{k,v} = 1, v \neq 3$;
$\lambda = \lambda_3 = 0.9$; θ=2.5;
$\alpha_{\xi_k^0}$=1 and β_{2,ξ_k^0}=0.5 for all k=1,…,5;
$\beta_{1,\xi_1^0} = -1$; $\beta_{1,\xi_2^0} = 0.5$; $\beta_{1,\xi_3^0} = 0.2$; $\beta_{1,\xi_4^0} = 0.1$; $\beta_{1,\xi_5^0} = 0$

Panel B: Summary statistics of clustering features

	Mean	Std	Skewness	Percentiles		
				1%	50%	99%
Z1	0.2317	1.5167	-0.0416	3.7233	0.2366	-3.3355
Z2	0.4127	1.4939	-0.0715	3.8130	0.4445	-3.1268
Z3	-0.1668	1.7429	0.4557	3.6322	-0.5002	-3.1200
Z4	0.2100	1.5855	0.0622	4.0103	0.1836	-3.3534
Z5	0.0044	1.0160	-0.0255	2.3451	0.0202	-2.2985

Table 4-6 (cont.) Panel C: Correlation Matrix

The upper triangle displays correlations within Class $\boldsymbol{\xi_1^0}$ (N=1000), and the lower triangle displays correlations for all data sample (N=5000). Correlations at the 1% significance level are bolded.

	Z1	Z2	Z3	Z4	Z5
Z1	1.0000	-0.0632	-0.0327	0.0381	0.0522
Z2	-0.0070	1.0000	0.01993	-0.0095	0.0123
Z3	**-0.0766**	**0.1703**	1.0000	-0.0316	0.0083
Z4	**0.0380**	**0.0893**	**0.3372**	1.0000	0.0183
Z5	0.0097	0.0042	0.0008	-0.0003	1.0000

Figure 4.1 demonstrates the distribution of clustering features. Using tests for multimodality (not reported here), there is evidence of multimodality only for z_3. Figure 4.2a and Figure *4.2*b graphically present simulated data based on the first two canonical variables. The CDA is performed on the list of features with the knowledge of true class membership to derive canonical variables. PROC SCANDIC in the SAS program is used to run CDA. Figure 4.2 shows observations by the first two canonical variables derived from CDA with true membership. As can be seen from Figure 4.2a, there is much overlap of members across classes. It seems that there are only three clusters, the most distinguishable on the far right, one in the middle and a bundle of observations on the left. This is confirmed by Figure 4.2b, which is similar to Figure 4.2a accompanied with the class membership. The one on the right is *Class* ξ_1^0, which is by far the most distinguishable class. It is also not difficult to identify members of *Class* ξ_3^0, which is in the middle. However, there is much overlap between *Class* ξ_2^0, *Class* ξ_4^0 and *Class* ξ_5^0, challenging the performance of clustering in general and ROWK in particular as evidenced in Section 4.2. Also, the first canonical variable is good enough to represent the class membership, and z_3 contributes the most to the first canonical variable, which is as expected. See Appendix Table B1 for detailed results of the CDA.

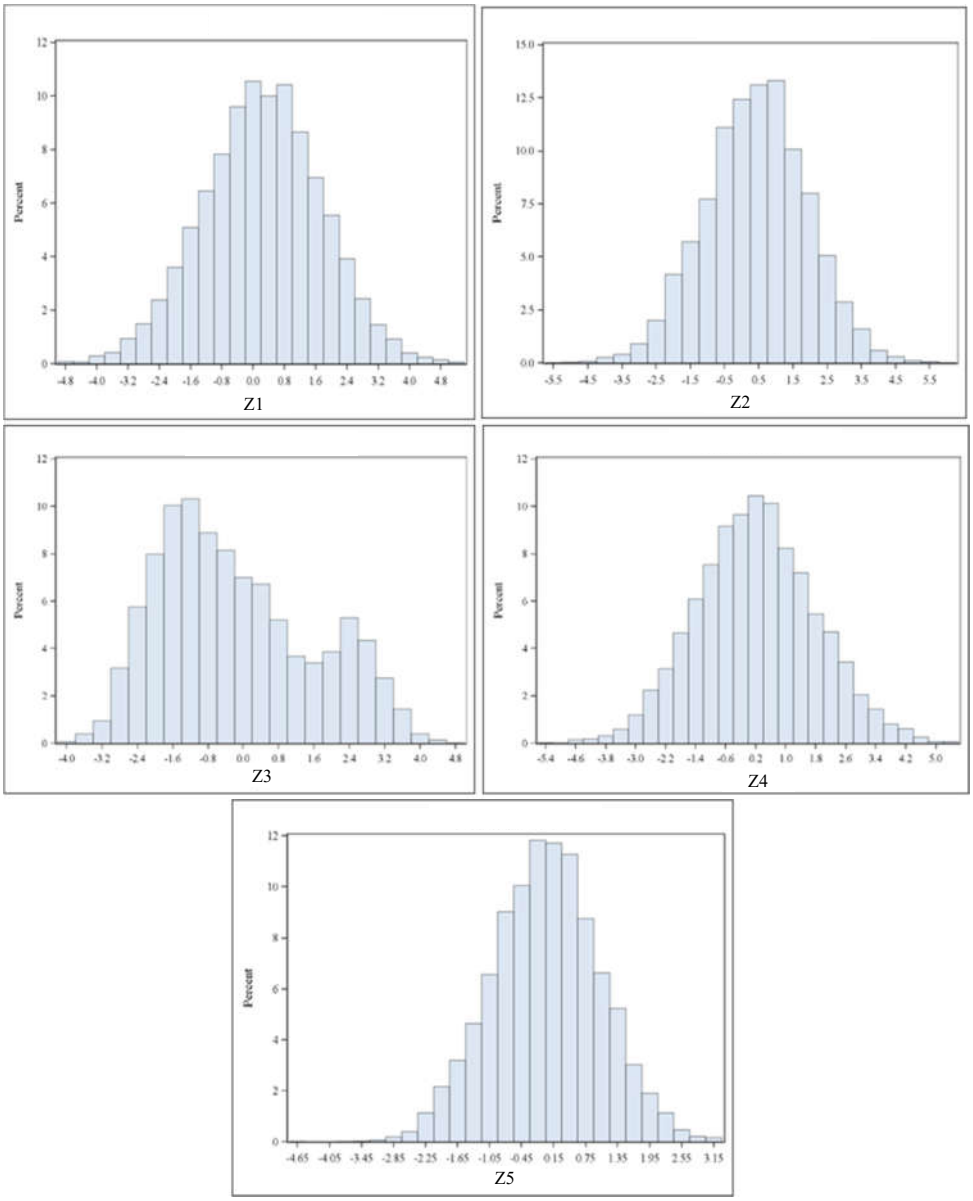
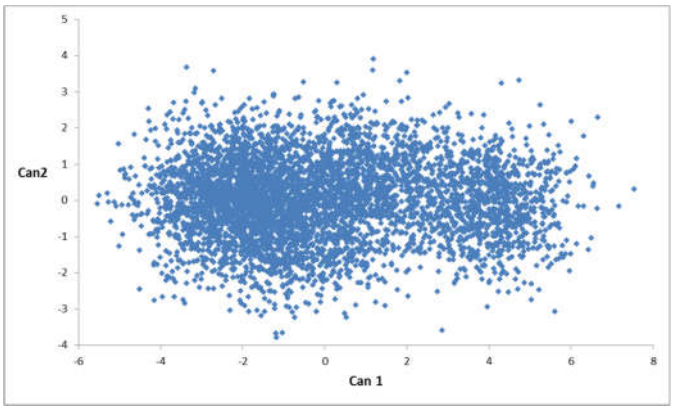
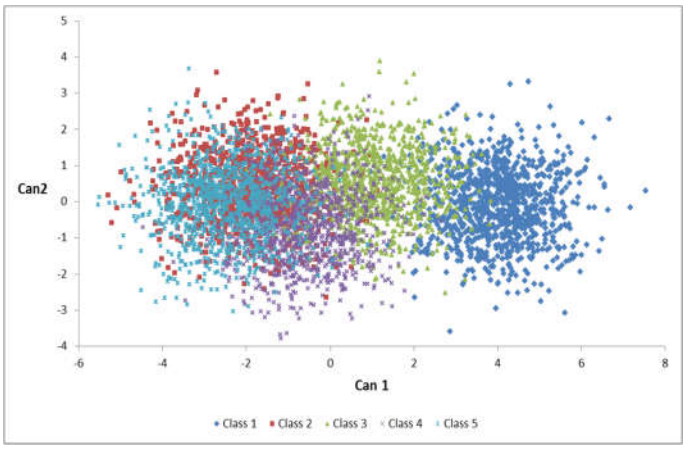


Figure 4.1- Distributions of Clustering Features - Case Study 1



a. Observations by the First Two Canonical Variables Derived from CDA with True Membership – Case 1



b. Class Membership by the First Two Canonical Variables Derived from CDA with True Membership- Case 1

Figure 4.2- Observations by CDA with True Membership- Case 1

The first step of the ROWK procedure relates to ranking features based on regression *MARs* calculated by running K-means only for each feature. The results of feature ranking at K^{max}=10 are presented in Figure 4.3[51]. As expected, z_3 has the lowest *MAR* (0.837), and accordingly ranks first. Noticeably, z_2, which is generated as a clustering feature, has a lower ranking (5th) than z_5, a random feature (ranked 4th). This is possibly due to the fact that with the exception of z_3, other features are less relevant to identify the class membership. Either class densities measured by these features ($den_{k,v}, v \neq 3$) are low or

[51] Assume that the possible maximum number of clusters is 10.

distances between class centres measured by these features (θ=2.5) are high, leading to poor class identification if only using these features individually.

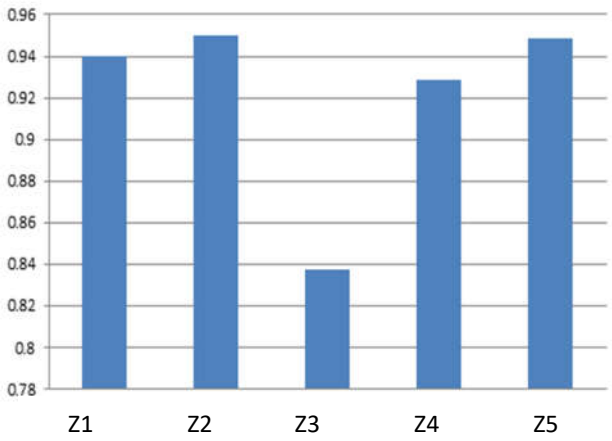


Figure 4.3- MARs for Each Feature at K'=10 (Case 1)

As stated previously in Section 3.2.1.1.2 (Step 4- Finding Optimal Weights of Cluster Features), the ROWK procedure determines a new guess of the set of feature weights as the one from the list of adjusted weights to minimise *MAR*. Furthermore, to overcome the local optimum issue caused by poor initial starting weights, this study employs a stepwise (forward) procedure, which is a popular technique of model specification in regression analysis (Hwang & Hu, 2015). Panel A of Table 4-7 presents the stepwise results of ROWK at K^{max}=10. The time to run ROWK at K^{max}=10 is 15 minutes using the standard SAS program and CPU (Intel® Core™ i5-4570 CPU @ 3.20GHZ). As expected, at K^{max} =10 the optimal set of weights is found as $\{w_1, w_2, w_3, w_4, w_5\}$ = {0.268, 0.047, 0.499, 0.186, 0}.

Panel B of Table 4-7 exhibits the optimal weights using different numbers of clusters. Consistent with hypothesis H4a, for all examined numbers of clusters (K'=2,…,10), while z_3 consistently receives the highest weights among five cluster features, z_5 has no contribution to clustering results. Using the graph in Figure 4.4, the number of clusters is found to be three. As can be seen from the graph, *MAR* drops dramatically for k=2 or k=3, continues to reduce but at a slower pace at k=4 and k=5 and remains unchanged for k>5. Using modified *BIC*s (not reported), the number of clusters are found to be five, which is

chosen. The results (unreported) remain unchanged if the optimal number of clusters is chosen to be three.

Accordingly, the final set of optimal weights is found to be $\{w_1, w_2, w_3, w_4, w_5\}$ = {0.132, 0.145, 0.551, 0.172, 0}. Hence, using the ROWK procedure, the most important feature (i.e. z_3) receives the highest weight. In contrast, the irrelevant (random) feature z_5 receives no weight after running ROWK. This evidence supports hypothesis H4a which posits that the mechanisms underlying the outperformance of ROWK are by placing more (less) weight on more (less) relevant features.

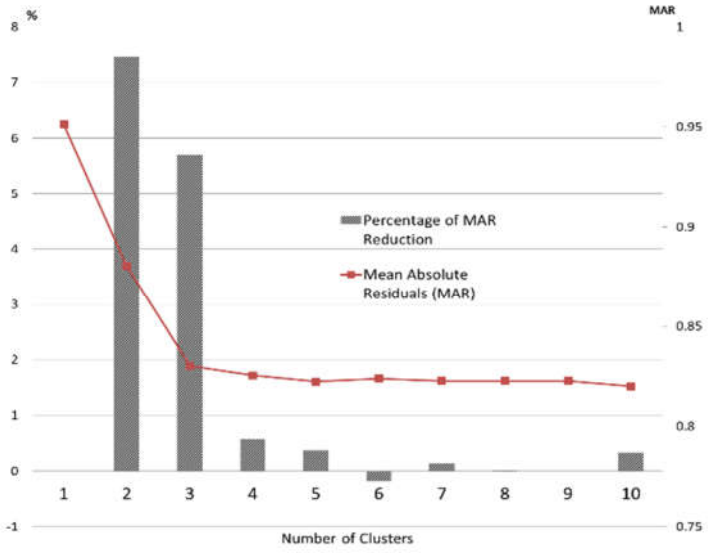


Figure 4.4- MARs at Different Number of Clusters-Case 1

Table 4-7: Optimal Weights by ROWK Clustering-Case1

Panel A: ROWK Results at K^{max}=10 for Each Number of Features (j=1,…,5)						
No. of features (*j*)	*MAR*	*Z1*	*Z2*	*Z3*	*Z4*	*Z5*
1	0.8375	0	0	1	0	0
2	0.8323	0	0	0.7642	0.2358	0
3	0.8238	0.2535	0	0.5704	0.1761	0
4	0.8208	0.2533	0	0.5699	0.1759	0.0010
5	**0.8201**	**0.2683**	**0.0466**	**0.4989**	**0.1863**	**0**
Panel B: ROWK Results at Each Number of Clusters K'						
Number of clusters (K')	*MAR*	*Z1*	*Z2*	*Z3*	*Z4*	*Z5*
1	0.9513	x	x	x	x	x
2	0.8803	0.4700	0	0.2902	0.2398	0
3	0.8302	0.2092	0.0309	0.6101	0.1498	0
4	0.8254	0.1811	0.2126	0.4541	0.1522	0
5	**0.8223**	**0.1320**	**0.1448**	**0.5508**	**0.1724**	**0**
6	0.8239	0.1669	0.0573	0.6243	0.1515	0
7	0.8228	0.176	0.0059	0.6583	0.1598	0
8	0.8227	0.3253	0.0025	0.5409	0.1313	0
9	0.8228	0.3897	0.003	0.4501	0.1572	0
10	0.8201	0.2677	0.0465	0.5000	0.1859	0

Table 4-8 compares the performance of ROWK relative to other standard methods. "*R_sq"* denotes the index of model fitness (=1-*MSR/VarY*) where *VarY* denotes the total variance of dependent variables; and "*vs. IDEAL"* denotes the relative performance of a procedure's *MAR* relative to that of the ideal case, computed as (MAR_{all} - MAR_i)/(MAR_{all}-MAR_{ideal})[52]. When using optimal weights, the *MAR* of ROWK clustering (ROWK) is significantly lower at 1% level (0.8224) compared to methods using all sample regression (*ALL*, 0.9513), feature standardized K-means (*STD_K*, 0.8455) and feature

[52] The ideal case is where 100% members are correctly assigned.

unstandardized K-means (*UNSTD_K*, 0.8399)[53]. With respect to class purity_version3, 71.3% of members are correctly assigned. Results remain unchanged for purity version 1 and 3 (see Appendix Table B2). In contrast, only 57.48% of members are correctly assigned when using conventional *UNSTD_K*. Performance of *STD_K* is even poorer with only 49.56% members precisely assigned[54]. This is consistent with the fact that the standard deviation of z_3 is slightly larger than those of other features (i.e. the different weights of features arise mainly from differences in distances between class centres).

Table 4-8: Performance of ROWK (Case 1)

"DIF" denotes differences between ROWK clustering and its benchmarks Paired t-tests are used to test for mean differences. T-statistics are presented in brackets. *IDEAL* denotes the case of 100% correctly-assigned members; *ALL* denotes the case of running regressions without clustering; *ROWK, STD_K* and *UNSTD_K* denote the cases of running regressions within each cluster found in ROWK clustering, standardized K-means clustering and unstandardized K-means clustering respectively.

$p_{overall}^{ver3}$ denotes the purity index version 3 based on five classes. *Var Y* denotes total variance of dependent variables. *R_sq* denotes the index of model fitness (*=1-MSR/Var Y*); *vs.IDEAL* denotes the relative performance a procedure's *MAR* relative to that of the ideal case, computed as 100*(MAR_{all} - MAR_i)/(MAR_{all}-MAR_{ideal}).

*, **, *** denote significance at 10%, 5% and 1%, respectively.

	ROWK	IDEAL	ALL	STD_K	UNSTD_K
W1	**0.132**	_	_	_	_
W2	**0.1448**	_	_	_	_
W3	**0.5508**	_	_	_	_
W4	**0.1724**	_	_	_	_
W5	**0**	_	_	_	_
MAR	**0.8223**	0.7837	0.9513	0.8455	0.8399
DIF_MAR	_	-0.0385*** *(4.54)*	0.1289*** *(7.73)*	0.0232*** *(3.36)*	0.0176*** *(2.52)*
MSR	**1.0740**	0.962	1.4941	1.1600	1.1350
DIF_MSR	_	0.1120*** *(5.71)*	0.4201*** *(11.54)*	0.0860*** *(3.36)*	0.061*** *(2.71)*
$p_{overall}^{ver3}$	**0.7130**	1.0000	0.2000	0.4956	0.5748
R_SQ (%)	**0.5035**	0.5553	0.3091	0.4637	0.4751
vs. IDEAL	**0.7696**	1.0000	0.0000	0.6313	0.6667
N	**5000**	5000	5000	5000	5000

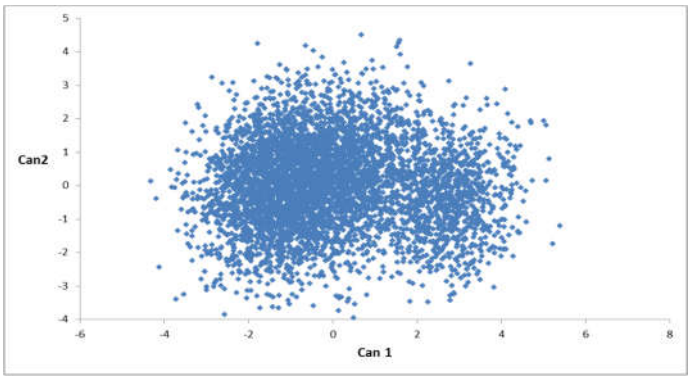
Figure 4.5a depicts observations by the first two canonical variables derived from CDA with cluster membership identified via standardized K-means clustering. It is very hard to discern cluster patterns from this figure, except for a group of observations on the left of the image. Figure 4.5b displays *class* membership for Figure 4.5a. As expected, the

[53] This study uses paired t-tests to test the significance of differences in mean absolute residuals (and squared residuals) between compared methods.

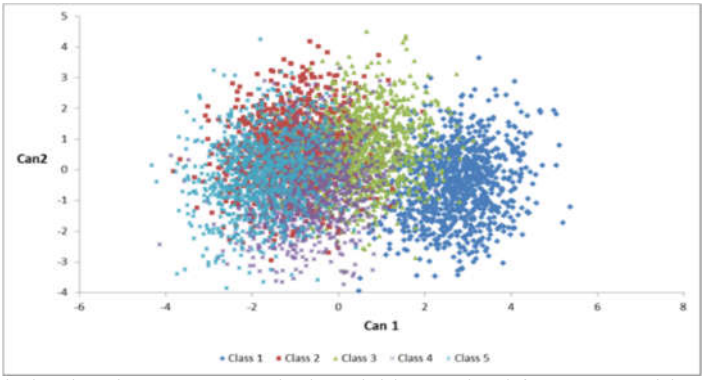
[54] It is infeasible to compute statistics for differences in purity indexes, *R_sqr* or *vs.IDEAL* given only one sample.

aforementioned group of observations belong mainly to *Class* ξ_1^0. However, there remains a large overlap between *Class* ξ_1^0 and *Class* ξ_3^0, and more so between *Class* ξ_3^0 and other remaining classes.

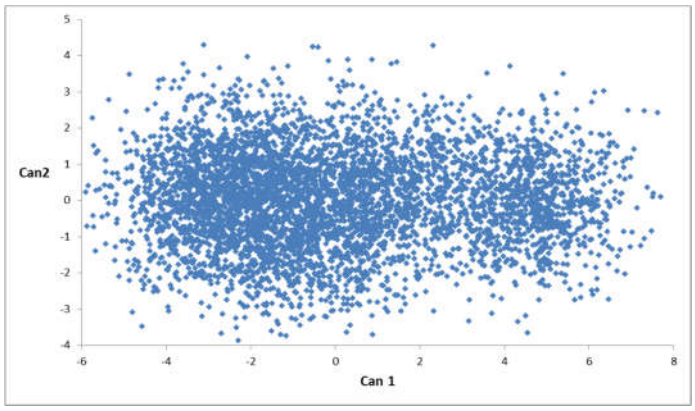
As can be seen from Figure 4.5c and Figure *4.5*d, the performance of ROWK is much better. In both these figures, it is easier to identify members of *Class* ξ_1^0 and *Class* ξ_3^0. There is little overlap between *Class* ξ_1^0 and *Class* ξ_3^0. Furthermore, members of *Class* ξ_3^0 are considerably distinguishable relative to *Class* ξ_2^0, *Class* ξ_4^0 and *Class* ξ_5^0. These observations contribute to significantly higher overall class purities (71.3%) for ROWK clustering as compared to *STD_K* (only 50.8%). Appendix Table B2 gives more details of class purities for ROWK clustering vis-a-vis standardized K-means clustering.



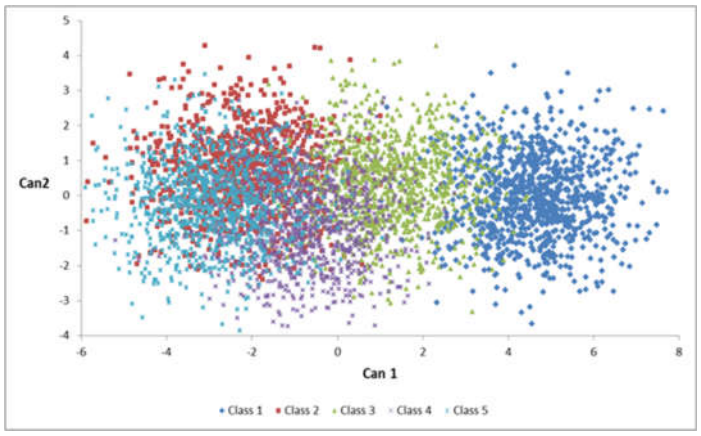
a. Observations by the First Two Canonical Variables Derived from CDA with Cluster Membership Identified via Standardized K-means Clustering-Case 1



b. Class Membership by the First Two Canonical Variables Derived from CDA with Cluster Membership Identified via Standardized K-means Clustering-Case 1



c. Observations by the First Two Canonical Variables Derived from CDA with Cluster Membership Identified via ROWK Clustering-Case 1



d. Class Membership by the First Two Canonical Variables Derived from CDA with Cluster Membership Identified via ROWK Clustering-Case 1

Figure 4.5- Observations by CDA with Cluster Membership Identified by Different Techniques - Case 1

In summary, the findings from the first case study strongly support hypothesis H4a which states that ROWK clustering places more (less) weight on more (less) relevant features when features have different degrees of contribution to cluster identification and regression estimation. Consequently, ROWK outperforms generic K-means (both standardized and un-standardized) with regard to the precision of class recognition and regression estimation.

4.4.2 Case Study 2-The Second Channel

Correlation between features is a well-known phenomenon, especially in the finance domain. Features that are highly correlated will automatically get higher weights than others (Sambandam, 2003). In an extreme case when two features are totally collinear,

they represent the same underlying feature and this underlying feature attracts twice the weight than it should. As documented in Section 4.2.2, multicollinearity negatively impacts K-means clustering performance, and consequently regression estimations.

Weighted K-means clustering techniques introduced by Huang et al. (2008) and further developed by Amorim & Mirkin (2012) were created to tackle the weighting issue of clustering. However, given the optimized criteria merely stem from clustering itself, the problem of multicollinearity is not addressed by these weighted clustering methods. In contrast, a novelty of the ROWK clustering proposed in this study is to employ an external criterion from regression analysis (i.e. *MAR*) to guide the clustering process to identify and reduce the weight of irrelevant correlated features. This leads to the second channel for the superior performance of ROWK as postulated in hypothesis H4b.

To test this hypothesis, this study uses Case Study 2. Simulated data are generated with the same set of simulated parameters as in Case 1 with an exception that features are *within-class* correlated. Recall that in Case 1, z_4 has the second-highest weight. Therefore, to make testing more conservative, in Case 2, z_4 is chosen to be strongly positively correlated with z_3, the most relevant feature. Panel A of Table 4-9 documents the simulated parameters, which are basically identical to those of Case 1. Panel B presents summary statistics for the clustering features. Again, the standard deviations of the four features are not very different, ranging from 1.49 for z_2 to 1.74 for z_3. The difference between Case 2 and Case 1 is apparent in Panel C. The within-*Class* ξ_1^0 correlation between z_3 and z_4 is significantly positive (ρ=0.537). Other within-class correlations are insignificantly different from zero. Importantly, even though only z_3 and z_4 are generated to be correlated only within-class, the lower triangle in Panel C displays several significant correlated features if measured using the full sample (5 out of the total of 10 pairs), challenging the feasibility of factor analysis in mitigating the issue of multicollinearity[55].

[55] The problem of using factor analysis to address multicollinearity is discussed in Section 4.5.5

Table 4-9: Descriptive Statistics of Clustering Features (Case 2)

The regression model and class members are generated to be identical to those of Case 1, with an exception that the within-class correlation matrix is not the identity matrix.

Panel A: Simulated parameters

N=5000; $N_k^0 = 1000$, k=1,…,5 ; P=2 and V=5;
$den_{k,3} = 0.5$; $den_{k,v} = 1, v \neq 3$;
$\lambda = \lambda_3 = 0.9$; θ=2.5;
$\alpha_{\xi_k^0}$=1 and β_{2,ξ_k^0}=0.5 for all k=1,…,5;
$\beta_{1,\xi_1^0} = -1$; $\beta_{1,\xi_2^0} = 0.5$; $\beta_{1,\xi_3^0} = 0.2$; $\beta_{1,\xi_4^0} = 0.1$; $\beta_{1,\xi_5^0} = 0$.

Panel B: Summary Statistics of Clustering Features

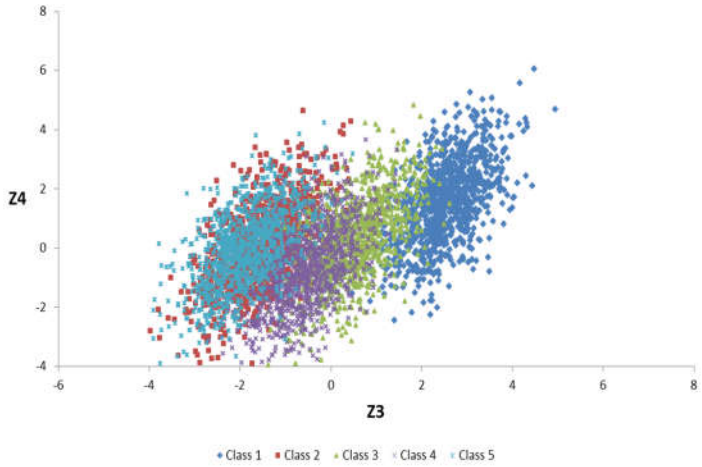
	Mean	Std	Skewness	Percentiles		
				1%	50%	99%
Z1	0.2317	1.5167	-0.0416	-3.3355	0.2366	3.7233
Z2	0.4102	1.4918	-0.0739	-3.1327	0.4489	3.7729
Z3	-0.1668	1.7429	0.4557	-3.1200	-0.5002	3.6322
Z4	0.2114	1.5827	0.0274	-3.4728	0.1941	3.9693
Z5	0.0044	1.0160	-0.0255	-2.2985	0.0202	2.3451

Panel C: Correlation Matrix

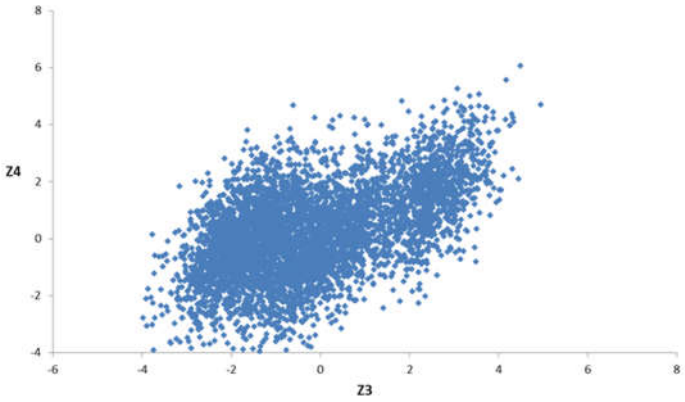
The upper triangle displays correlations within Class $\boldsymbol{\xi_1^0}$ (N=1000), and the lower triangle displays correlations for all data sample (N=5000). Correlations at the 1% significance level are bolded.

	Z1	Z2	Z3	Z4	Z5
Z1		0.0397	-0.0327	0.0137	0.0522
Z2	**0.0794**		0.0166	0.0047	0.0177
Z3	**-0.0766**	**0.1700**		**0.5366**	0.0083
Z4	0.0285	**0.0743**	**0.5464**		0.0201
Z5	0.0097	0.0060	0.0008	-0.0082	

Figure 4.6 helps to clarify this statement graphically. Figure Figure *4.6*a (Figure *4.6*b) displays all data by z_3 and z_4 with (without) class membership. As can be seen from Figure 4.6a, within each class, these two features are highly correlated, as expected. However, the positive correlation between z_3 and z_4 in the absence of class membership is also observed in Figure 4.6b. This correlation is of the same magnitude as the within-class correlation (i.e. 0.5366 vs. 0.5464).



a) Observations by Cluster Features Z_3 and Z_4 with Class Membership - Case 2



b) Observations by Cluster Features Z_3 and Z_4 without Class Membership - Case 2

Figure 4.6- Observations by Cluster Features Z_3 and Z_4

Given the high correlation between z_4 and the most important feature, z_3, it is expected that running K-means clustering using Z_4 alone will result in a lower *MAR* than that of Case 1. Figure 4.7 supports this statement (i.e. 0.929 vs. 0.905 for Case 1 and Case 2, respectively). This raises an issue concerning the use of the *MAR* results to rank the cluster features for the ranking step of ROWK clustering. It could be that some irrelevant features receive high ranks due to high correlations with highly relevant features. However, it is expected that during the process of running ROWK procedure, these irrelevant features will be eliminated.

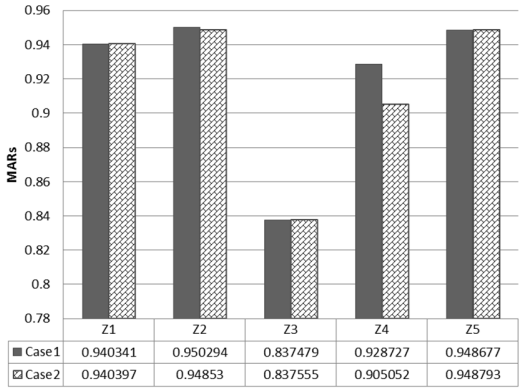
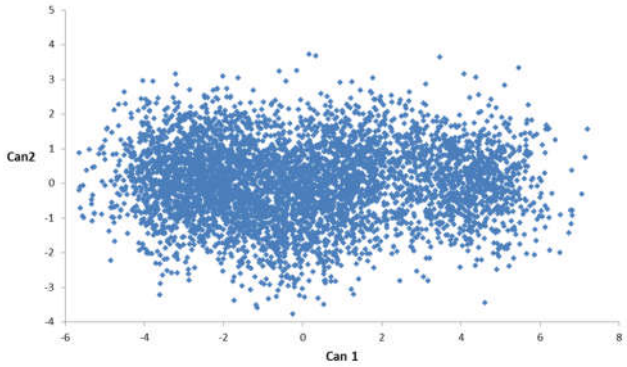
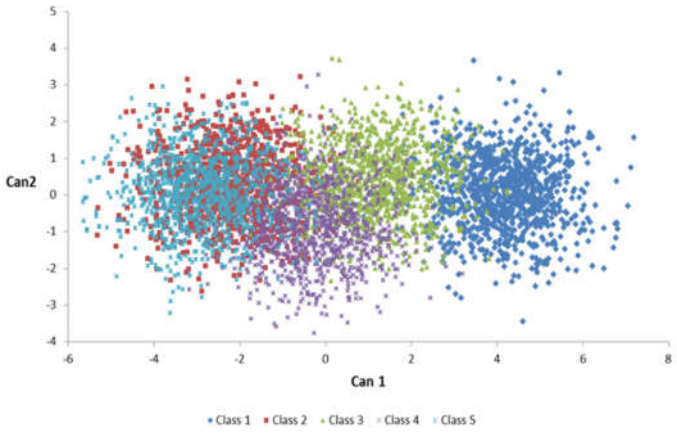


Figure 4.7- MARs for each Feature at *K'*=10 (Case 2 vs. Case 1)

Figure 4.8 graphically presents simulated data based on the first two canonical variables. Basically, given the same parameter input as in Case 2 with only an adjustment of within-class correlation matrix, it is unsurprising that there is no considerable difference between Figure 4.8 and Figure 4.2. See Appendix Table B3 for detailed results of CDA.



a) Observations by the First Two Canonical Variables Derived from CDA with True Membership-Case 2



b) Class Membership by the First Two Canonical Variables Derived from CDA with True Membership-Case 2

Figure 4.8- Observations by CDA with True Membership-Case 2

Panel A of Table 4-10 presents the ROWK results at K^{max} =10 for each number of features. The running time for ROWK at K^{max} =10 is approximately the same as in Case 1. As expected, at K^{max} =10 the optimal set of weights is found to be $\{w_1, w_2, w_3, w_4, w_5\}$ = {0.096, 0.004, 0.711, 0.189, 0}. Panel B of Table 4-10 exhibits the optimal weights given different numbers of clusters. Similar to the Case 1 findings, for all examined numbers of clusters (K'=2,…,10), z_3 consistently receives the highest weight among five cluster features, while z_5 has no contribution to clustering results. Using the graph in Figure 4.9, the optimal number of clusters is again found to be three. The result from Figure 4.9 is reasonable due to the fact that the data are generated to have only two distinguishable classes (see Figure 4.8b, i.e. *Class* ξ_1^0, *Class* ξ_3^0). Consequently, only three distinguishable clusters are identified based on the knee point in Figure 4.9. However, although there is still considerable overlap between members of *Class* ξ_4^0 and other classes, to some extent, a proportion of *Class* ξ_4^0 remains separated from those of the other classes. Hence, modified *BIC*s (unreported) captures this evidence and concludes with four clusters. Based on modified *BICs*, this study selects the optimal number of clusters as four. Similar results are also observed for three clusters.

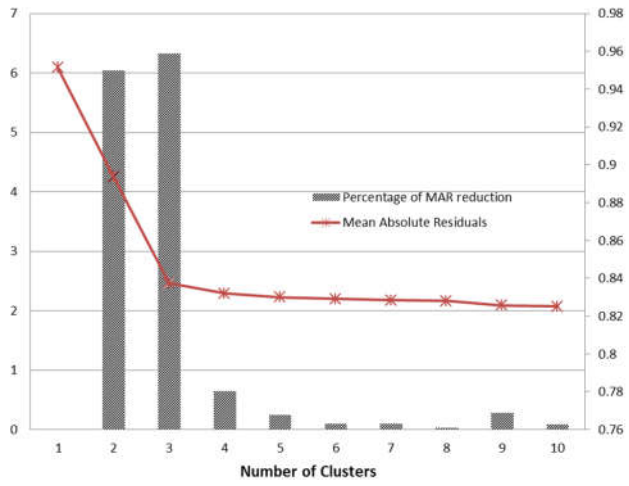


Figure 4.9- Results of ROWK Clustering at Different Numbers of Clusters

Table 4-10: Optimal Weights by ROWK Clustering-Case2

Panel A: ROWK Results at K^{max}=10 for each Number of Features (j=1,…,5)						
No. of features (j)	MAR	Z1	Z2	Z3	Z4	Z5
1	0.8376	0	0	1	0	0
2	0.8355	0	0	0.8592	0.1408	0
3	0.8251	0.1147	0	0.6996	0.1857	0
4	**0.8244**	**0.0963**	**0.0037**	**0.7111**	**0.1888**	**0**
5	0.8252	0.1138	0.0005	0.6941	0.1843	0.0074
Panel B: ROWK Results at each Number of Clusters K'						
Number of clusters (K')	MAR	Z1	Z2	Z3	Z4	Z5
1	0.9513					
2	0.8912	0.4284	0.1579	0.2644	0.1493	0
3	0.8369	0.2312	0.0608	0.6623	0.0457	0
4	**0.8313**	**0.1806**	**0.0684**	**0.7447**	**0.0064**	**0**
5	0.8283	0.233	0.0373	0.7218	0.0079	0
6	0.8259	0.1624	0.1458	0.6898	0.002	0
7	0.8283	0.3182	0	0.6779	0.0039	0
8	0.8281	0.1117	0	0.6982	0.1901	0
9	0.8258	0.1011	0	0.7191	0.1798	0
10	0.8244	0.0963	0.0037	0.7111	0.1888	0

Recall in the Case 1 set of optimal weights, ROWK assigns the second highest weight (0.172) to z_4. In Case 2, however, z_4 is strongly correlated with z_3, the most relevant feature. This could cause K-means clustering to incorrectly amplify the weight of z_4, leading to poor performance of cluster identification and regression estimation (Sambandam, 2003). However, as stated in hypothesis H4b, a novel aspect of ROWK clustering as proposed in this study is to employ an external criterion from regression analysis (i.e. *MAR*), guiding the clustering process to identify and reduce the weight of irrelevant correlated features. Consistent with hypothesis H4b, Panel B of Table 4-10 highlights the final set of optimal weights for K'= 4 as $\{w_1, w_2, w_3, w_4, w_5\}$ = {0.181, 0.068, 0.745, 0.006, 0}, indicating that ROWK clustering does mitigate the effect of multicollinearity by lessening the weight of z_4 from 0.172 in the absence of multicollinearity to only 0.006 in the presence of multicollinearity.

Panel A of Table 4-11 presents the performance of ROWK clustering relative to other methods. Using optimal weights found by ROWK clustering, *MAR* is significantly lower (0.8319) than those of *ALL* (0.9513), *STD_K* (0.8615) and *UNSTD_K* (0.8637).

Regarding cluster validation, 60% of observations are correctly assigned. The corresponding values for *STD_K* and *UNSTD_K* are lower at 43.7% and 47.9%, respectively. Similar findings could be observed for the mean square of residuals and the pseudo R-squared. This evidence lends further support for hypothesis H4b. By mitigating the problem of multicollinearity, ROWK clustering improves the precision of cluster identification, and consequently improves the results of regression estimations. See Appendix Table B4 for the purity indices for each cluster/class and the frequency of class membership by cluster.

Note that relative to the case of uncorrelated features (i.e. Case 1), the performance of K-means (both standardized and unstandardized), is significantly worse when features are correlated. To make comparisons between two cases in a fair way, the numbers of clusters need to be identical across the two cases. Hence, the results of Case 2 with five clusters are compared with those of Case 1. These results are unreported and are available upon request. Specifically, at *0.86*, the *MAR* of *STD_K* with correlated features is considerably higher than that with uncorrelated features (*0.845*). These findings are consistent with the argument that K-means clustering does not address the problem of multicollinearity, and incorrectly amplifies the weights of irrelevant but correlated features. In contrast, the performance of ROWK clustering is unimpaired with or without multicollinearity. For example, the difference in *MAR* between Case 1 and Case 2 is only 0.006 (=0.8283-0.8223), and statistically is no different from zero.

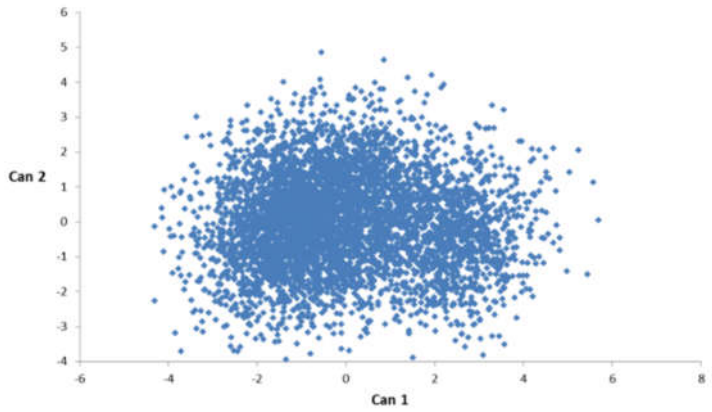
Table 4-11: Performance of ROWK (Case 2)

"*DIF*" denotes differences between ROWK clustering and its benchmarks. Paired t-tests are used to test for mean differences. T-statistics are presented in brackets. *IDEAL* denotes the case of 100% correctly-assigned members; *ALL* denotes the case of running regressions without clustering; *ROWK, STD_K* and *UNSTD_K* denote the cases of running regression within each cluster found in ROWK clustering, standardized K-means clustering and unstandardized K-means clustering respectively. $p_{overall}^{ver3}$ denotes the purity index version 3. *Var Y* denotes total variance of dependent variables. *R_sq* denotes the index of model fitness (*=1-MSR/Var Y*); *vs. IDEAL* denotes a procedure's *MAR* relative to that of ideal case, computed as $100*(MAR_{all} - MAR_i)/(MAR_{all}-MAR_{ideal})$..

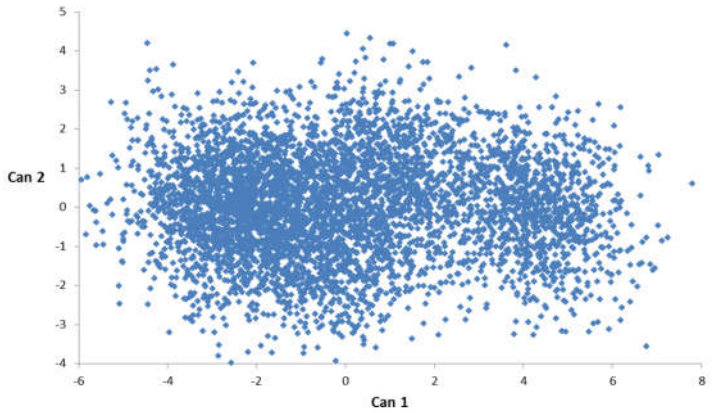
*, **, *** denote significance at 10%, 5% and 1%, respectively.

	ROWK	IDEAL	ALL	STD_K	UNSTD_K
W1	**0.1806**	_	_	_	_
W2	**0.0684**	_	_	_	_
W3	**0.7447**	_	_	_	_
W4	**0.0064**	_	_	_	_
W5	**0**	_	_	_	_
MAR	**0.8313**	0.7837	0.9513	0.8615	0.8637
DIF_MAR	_	-0.0482***	0.1194***	0.0296***	0.0318***
		(-3.24)	*(6.73)*	*(2.78)*	*(2.99)*
MSR	**1.1113**	0.9621	1.4952	1.2082	1.2141
DIF_MSR	_	-0.149***	0.384***	0.097***	0.103***
		(-6.66)	*(8.15)*	*(4.32)*	*(3.18)*
$p_{overall}^{ver3}$	**0.623**	1	0.2	0.471	0.515
R_SQ	**0.4844**	0.5536	0.3063	0.4393	0.4365
vs. IDEAL	**0.7124**	1	0	0.5358	0.5227
N	**5000**	5000	5000	5000	5000

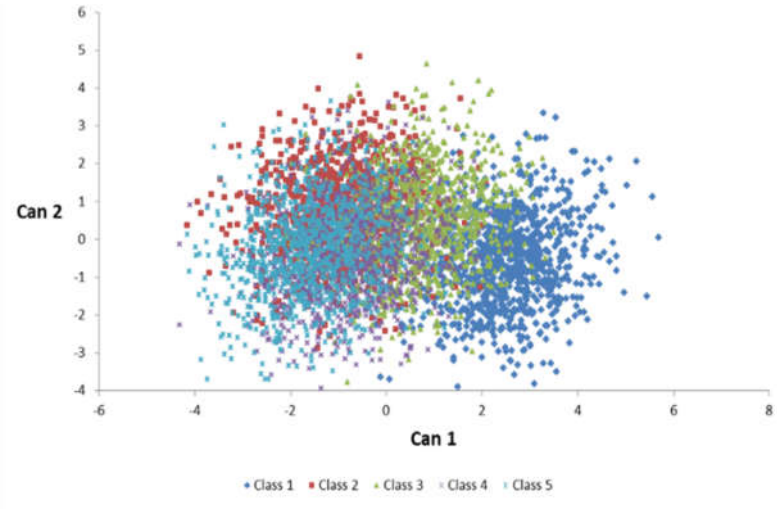
Figure 4.10 graphically presents observations based on the first two canonical variables derived from CDA using cluster membership from ROWK clustering and standardized K-means clustering. With the negative impact of multicollinearity as documented in Panel A of Table 4-11, a blurring of cluster patterns identified using standardized K-means clustering is expected. Figure 4.10a, which exhibits observations by the first two canonical variables derived from CDA with cluster membership identified via standardized K-means clustering, supports this statement. All observations appear to belong to a single group. Figure 4.10c displays class membership for Figure 4.10a. Unsurprisingly, there is considerable overlap between members of classes, even for Class ξ_1^0.



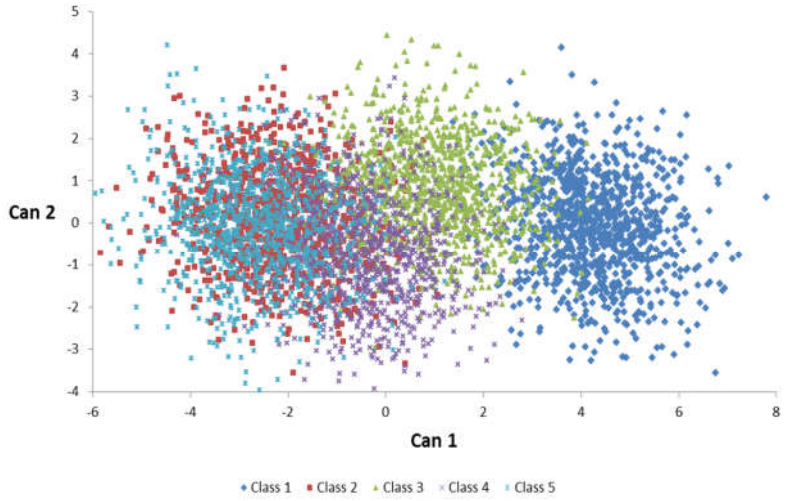
a. Observations by the First Two Canonical Variables Derived from CDA with Cluster Membership Identified via Standardized K-means Clustering – Case 2



b. Observations by the First Two Canonical Variables Derived from CDA with Cluster Membership Identified via ROWK Clustering – Case 2



c. Class Membership by the First Two Canonical Variables Derived from CDA with Cluster Membership Identified via Standardized K-means Clustering – Case 2



d. Class Membership by the First Two Canonical Variables Derived from CDA with Cluster Membership Identified via ROWK Clustering – Case 2

Figure 4.10- Observations by the First Two Canonical Variables Derived from CDA with Cluster Membership Identified via Different Techniques-Case 2

As depicted in Figure 4.10b and Figure *4.10*d, the performance of ROWK is much better. In both these figures, it is easier to distinguish members of *Class* ξ_1^0 and *Class* ξ_3^0. Figure 4.10d reveals that there is little overlap between members of these two classes. More convincingly, comparing Figure 4.8b and Figure 4.10d, the cluster patterns that are

displayed by the first two canonical variables derived from a CDA using ROWK's cluster membership are as good as those using true cluster memberships.

In summary, the findings from the second case study support hypothesis H4b. The second channel through which ROWK clustering outperforms generic K-means (both standardized and un-standardized) with regard to the precision of cluster recognition and regression estimation is by reducing the influence of multicollinearity. This is achieved by reducing the weights of irrelevant features which are highly correlated with relevant features. This suggests that ROWK clustering represents an important advancement given the failure of contemporary clustering techniques to deal with the problem of multicollinearity (e.g. Huang et al., 2008; Amorim & Mirkin, 2012).

4.4.3 Case Study 3-The Third Channel

The main innovation of ROWK is the connection between the problem of feature weighting in weighted clustering and the problem of HGSC in regression analysis. Hence, it mitigates the issue of HGSC by not only providing the information from CA, but also adjusting the feature weighting process to be aligned with regression estimation, reflecting ROWK's ultimate objectives. This mechanism is distinguished from other traditional weighted clustering techniques that only focus on the clustering side (e.g. Desarbo et al., 1984; Huang et al., 2008; Amorim & Mirkin, 2012). Accordingly, Hypothesis H4c proposes that when features have different degrees of contribution to class identification and regression estimation, ROWK outperforms generic K-means (both standardized and un-standardized) and weighted K-means clustering (WK) in terms of precision of class recognition and regression estimation. It does this by better capturing feature relevance through its contribution to class recognition and regression estimation.

To test this statement, Case Study 3 is introduced. Similar to Case 1 and 2, there are 5,000 observations which belong to *five classes*. Each class has 1,000 members. There are four features, z_i, i=1,…,4 and a random variable z_5 (~$N(0,1)$) which is used as an irrelevant

clustering feature. Features are generated to be uncorrelated. Unlike Case 1 and 2, in Case 3, x_1 and x_2 are not set to be clustering features[56].

For Case 3, adjustments of simulated parameters are made as follows. Instead of generating z_3 alone to be the most relevant feature, in this case both z_3 and z_4 have the same degree of contribution to the identification of class patterns. However, the relevance of these features differs in the sense that z_4 is for only class recognition, but z_3 is for both class recognition and regression estimation. Specifically, z_3 is generated to be a highly relevant feature to recognize members of *Class* ξ_1^0 , *Class* ξ_2^0 and *Class* ξ_4^0 (z_1 and z_2 have similar contribution to z_3 but to a lesser degree), while z_4 is simulated to provide significant information to identify membership of *Class* ξ_3^0 and *Class* ξ_5^0. The important difference between these two groups of classes is their relevance to regression estimations. Specifically, while regression coefficients are homogeneous (heterogeneous) within (between) *Class* ξ_1^0 , *Class* ξ_2^0 and *Class* ξ_4^0, members within *Class* ξ_3^0 or *Class* ξ_5^0 do not display homogeneous regression coefficients. Consequently, only the identification of *Class* ξ_1^0 , *Class* ξ_2^0 and *Class* ξ_4^0 is important for solving the problem of HGSC.

Panel A of Table 4-12 presents the details of the input parameters for Case Study 3. The HSGC problem only exists for *Class* ξ_1^0 , *Class* ξ_2^0 and *Class* ξ_4^0 with their corresponding coefficients of x_2 as *-1, 0* and *0.3* respectively. The members within *Class* ξ_3^0 and *Class* ξ_5^0 do not exhibit homogeneous regression coefficients. Instead, its coefficients are *-1, 0* or *0.3* if the true class it belongs in (determined only by z_1, z_2 and z_3) is *Class* ξ_1^0 or *Class* ξ_2^0 or *Class* ξ_4^0, respectively. As a result, identification of *Class* ξ_3^0 and *Class* ξ_5^0 does not help to alleviate the problem of HGSC. From this point of view, an appropriate clustering procedure needs to place low or zero weight to z_4, which only helps to identify membership of *Class* ξ_3^0 and *Class* ξ_5^0. This is confirmed in later findings.

[56] In Case Study 3, there are two reasons why x_1 and x_2 are not set to be clustering features. First, Cases 1 and 2 are built to compare with each other, consequently the simulated setting needs to be comparable. However, Case 3 is built to examine the third channel of ROWK outperformance with no requirement for connection with the previous cases. As a result, setting x_1 and x_2 to not be clustering features does not impact the cohesion of the thesis. Second, in real data, it is more likely that cluster features are separate from regression variables.

Panel B of Table 4-12 presents summary statistics of clustering features. Standard deviations for the four features are similar ranging from *0.9* for Z_3 to *1.41* for Z_4. Note that in this case, Z_3 has the lowest variance. Therefore, superior performance is expected for *STD_K* over *UNSTD_K*. Panel C of Table 4-12 exhibits the full-sample (lower triangle) and within *Class* ξ_1^0 (upper triangle) correlation matrices of clustering features. As expected, all pairs of features display no significant correlation within *Class* ξ_1^0, while there are significant correlations (three out of six pairs of correlations) between features for the full sample. Panel D displays distances between class centres.

Table 4-12: Descriptive Statistics of Clustering Features (Case Study 3)

Regression Model:

$y_i = \alpha_{\xi_k^0} + x_i'\beta_{\xi_k^0} + u_i, i = 1,\dots,5000; k = 1,\dots,5$; where $\alpha_{\xi_k^0}$ and $\beta_{\xi_k^0}$ be 1x1 and Px1 vectors of group-specific intercept and slope coefficients; $u_i \sim N(0,1)$ and $cov(u_i, u_j) = I$(identity matrix); $cov\left(x_{i,\xi_k^0}, u_{j,\xi_k^0}\right) = 0$; $\alpha_{\xi_k^0}$=1 and β_{2,ξ_k^0}=0.5 for all k=1,…,5. Each class has 1000 observations, $N_k^0 = 1000$

Class Membership:

There are 4 features, i.e.$Z_v, v = 1,\dots,4$ and a random variable $Z_5 \sim N(0,1)$; $z_{i,v}^k = \left(c_{k,v} + \tau_{i.v}^k\right)$ where $c_{k,v}$ is the centre of class k measured by feature Z_v; $\tau_{i.v}^k \sim N(0, \theta^2/4)$ where θ^2 is expectation of the mean squared distances between members to its corresponding centre; $\sum_{v_i v_j} = cov\left(\tau_{i.v}^k,\ \tau_{i.v}^k\right) = D$; $d_{\xi_1^0,\xi_j^0} > \lambda\, d_{\xi_i^0,\xi_j^0}$ and $d^3{}_{\xi_1^0,\xi_j^0} > \lambda_3\, d^v{}_{\xi_i^0,\xi_j^0}$ for all $i \neq 1$ and $v \neq 1$ where $d_{\xi_i^0,\xi_j^0}$ ($d^v_{\xi_i^0,\xi_j^0}$) denotes the distance between *Class* ξ_i^0 and *Class* ξ_j^0 as measured by all features (as measured by feature Z_v). λ (λ_1) indicates the extent of differences between *Class* ξ_1^0 and other classes as measured by all features (as measured by feature Z_v). $z_1 = x_1$ and $z_2 = x_2$

Panel A: Simulated parameters

N=5000; $N_k^0 = 1000$, k=1,…,5 ; P=2 and V=5;
$den_{k,v} = 0.8, v \neq 3$; $den_{k,3} = 0.5$
$\lambda = \lambda_3 = 0.7$; θ=2.2;
$\alpha_{\xi_k^0}$=1 and β_{2,ξ_k^0}=0.5 for all k=1,…,5;
$\beta_{1,\xi_1^0} = -1$; $\beta_{1,\xi_2^0} = 0$; $\beta_{1,\xi_4^0} = 0.3$

Panel B: Summary Statistics of Clustering Features

	Mean	Std	Skewness	Percentiles		
				1%	50%	99%
Z1	1.7411	1.3595	0.0615	-1.2966	1.7392	4.9768
Z2	1.9262	1.2723	-0.0514	-1.1668	1.9176	4.8519
Z3	2.0544	0.9093	0.1554	0.0754	2.0219	4.2642
Z4	2.2022	1.4139	0.0150	-1.0335	2.2169	5.5527
Z5	0.0035	1.0009	0.0428	-2.3626	-0.0049	2.4068

Table 4-12 (cont.) Panel C: Correlation Matrix

The upper triangle displays correlations within Class $\boldsymbol{\xi_1^0}$ (N=1000), and the lower triangle displays correlations for the full data sample (N=5000). Correlations at the 1% significance level are bolded.

	Z1	*Z2*	*Z3*	*Z4*	*Z5*
Z1	1.0000	-0.0385	-0.0020	-0.0117	0.0184
Z2	**-0.1851**	1.0000	0.0355	-0.0345	0.0260
Z3	**0.2271**	**-0.1697**	1.0000	-0.0059	-0.0119
Z4	-0.0106	-0.0033	0.0140	1.0000	-0.0088
Z5	0.0104	0.0003	-0.0015	0.0041	1.0000

Panel D: Distances Between Class Centres

	Class ξ_1^0	Class ξ_2^0	Class ξ_3^0	Class ξ_4^0	Class ξ_5^0
Class $\boldsymbol{\xi_1^0}$	0.0000				
Class $\boldsymbol{\xi_2^0}$	1.6031	0.0000			
Class $\boldsymbol{\xi_3^0}$	2.4920	1.7205	0.0000		
Class $\boldsymbol{\xi_4^0}$	2.1119	0.8307	2.0322	0.0000	
Class $\boldsymbol{\xi_5^0}$	2.0785	1.3602	2.6325	1.8709	0.0000

Figure 4.11 clarifies these statements graphically. Figure 4.11a displays all data by the first two canonical variables derived from CDA with all features. The class pattern is extremely unclear. However, Figure 4.11b reveals that when CDA is conducted for only *Class* ξ_1^0 , Class ξ_2^0 and *Class* ξ_4^0 using the first three features (z_1, z_2 and z_3), although class membership still overlaps considerably, it is now much easier to discern [57]. Furthermore, Figure 4.11c shows the class patterns for *Class* ξ_3^0 and *Class* ξ_5^0 more clearly when displayed by z_4. See Appendix B5 for detailed results of CDA relating to Figure 4.11a.

[57] There are two reasons for generating indistinct class patterns based on the first three features. First, class patterns in real data, when they exist, tend to be blurry (Amorim & Mirkin, 2012). Second, the results are more conservative in the sense that Z_4 will receive more weight in the standard weighted clustering procedures.

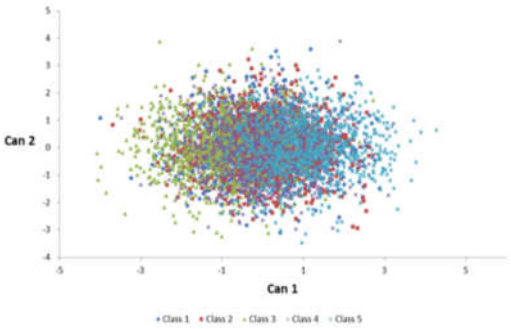


Figure 4.11a.

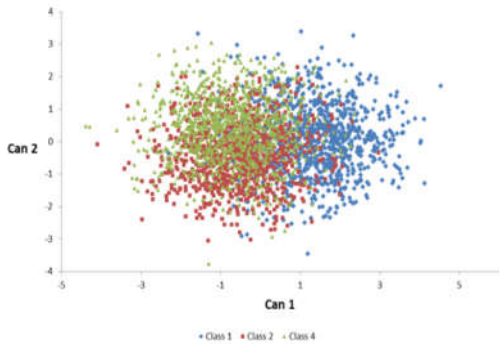


Figure 4.11b.

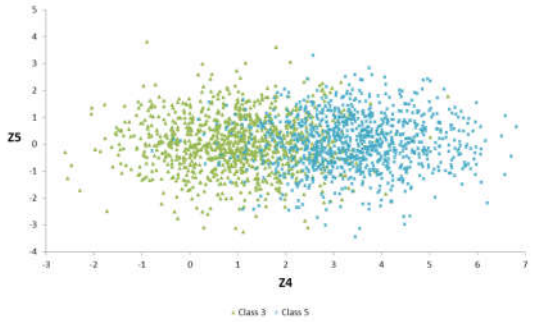


Figure 4.11c.

Figure 4.11- Observations by CDA- Case 3. Figure 4.11a presents class membership by the first two canonical variables derived from CDA with five-class membership-Case 3. Figure 4.11b presents CDA with Class ξ_1^0, Class ξ_2^0 and Class ξ_4^0 by z1, z2 and z3-Case 3. Figure 4.11c presents members of Class ξ_3^0 and Class ξ_5^0 displayed by z4 and z5-Case 3

Running K-means clustering using z_4 alone is only able to identity members of *Class* ξ_3^0 and *Class* ξ_5^0, which are irrelevant for regression estimation. Accordingly, it is expected that the *MAR* of z_4 will be as high as that of z_5, the random feature. Figure 4.12 supports this statement. The *MAR* of Z_4 is 1.1725, which is even slightly higher than that of z_5 (1.1714). Consequently, z_4 and z_5 rank last. As expected, z_3 ranks first, while the second and third are z_1 and z_2, respectively.

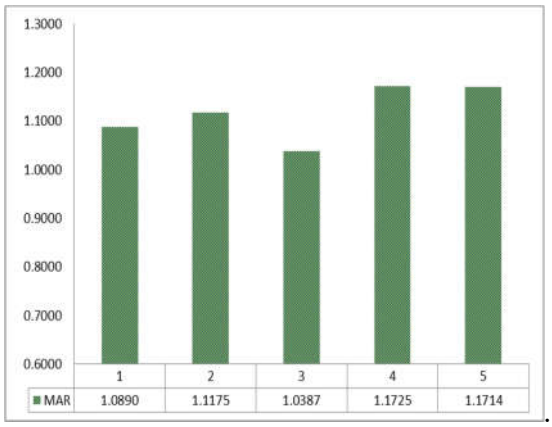


Figure 4.12- MAR for Each Feature at *K'*=10 (Case 3)

Panel A of Table 4-13 presents the stepwise results of ROWK at K^{max} =10 for each number of features. Figure 4.13 graphs the results from Panel A. Based on the knee point in Figure 4.13, only z_1, z_2 and z_3 are chosen. As a result, at K^{max} =10 the optimal set of weights is found as $\{w_1, w_2, w_3, w_4, w_5\}$ = {0.219, 0.116, 0.666, 0, 0}. Panel B of Table 4-13 exhibits the optimal weights at different numbers of clusters. For all examined number of clusters (K' = 2,…,10), z_3 consistently receives the highest weights among five cluster features, while z_5 does not contribute to the clustering results. The results of modified *BICs* (not reported) and of the knee point in Figure 4.14 coincide at three clusters, which is consistent with the simulated data.

Table 4-13: Optimal Weights by ROWK Clustering-Case 3

Panel A: ROWK Results at K^{max}=10 for Each Number of Features (j=1,…,5)						
No. of features (j)	MAR	Z1	Z2	Z3	Z4	Z5
1	1.0387	0	0	1	0	0
2	0.9927	0.1076	0	0.8924	0	0
3	**0.9845**	**0.2186**	**0.1156**	**0.6658**	**0**	**0**
4	0.9828	0.1769	0.1132	0.6521	0.0577	0
5	0.9822	0.1143	0.354	0.4213	0.0373	0.0731
Panel B: ROWK Results at Each Number of Cluster *K'*						
Number of clusters (K')	MAR	Z1	Z2	Z3	Z4	Z5
1	1.1768	_	_	_	_	_
2	1.0081	0.2307	0.1787	0.5906	0	0
3	**0.9864**	**0.2656**	**0.2322**	**0.5022**	**0**	**0**
4	0.9908	0.2589	0.2337	0.5074	0	0
5	0.9878	0.1847	0.1304	0.6849	0	0
6	0.9804	0.1412	0.1436	0.7152	0	0
7	0.9843	0.1119	0.101	0.7872	0	0
8	0.9882	0.1401	0.0989	0.761	0	0
9	0.9819	0.0601	0.0513	0.8886	0	0
10	0.9846	0.0429	0.0443	0.9128	0	0

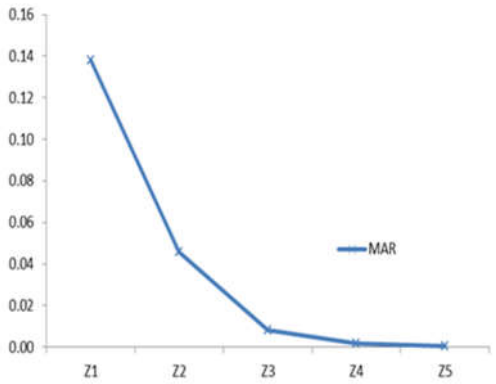


Figure 4.13- MARs across Features at K'=10 (Case 3)

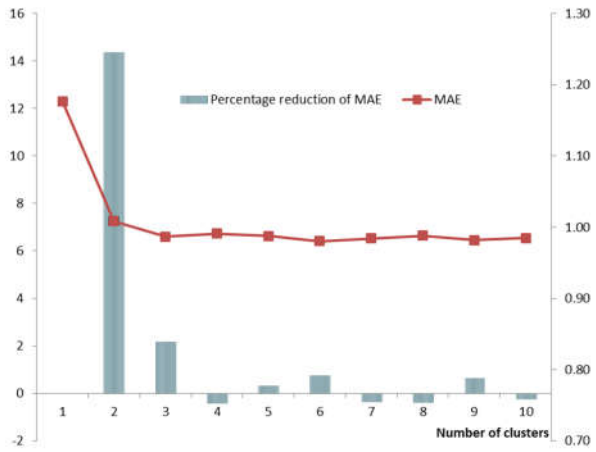


Figure 4.14- MARs for Each Feature at K=10 (Case 3)

Accordingly, Panel B of Table 4-13 highlights the final set of optimal weights for K'=3 as $\{w_1, w_2, w_3, w_4, w_5\}$ = {0.266, 0.232, 0.502, 0, 0}. While ROWK assigns the highest weight to the most relevant synthesized feature, z_3, it places zero weight on z_5, the random feature. More importantly, ROWK clustering correctly assign zero weight to z_4, a feature that is only relevant to class identification, but not to regression estimation. In other words, the optimal weights that minimise a regression's *MAR* reflect its importance not only to class identification, but also to improving the regression analysis, which is consistent with hypothesis H4c. As a robustness check, 100 simulated samples are created with the same set of parameters and the *MARs* are calculated for two sets of weights for z_3 and z_4: one with the above set of optimal weights {0.266, 0.232, 0.502, 0, 0} and the other with equal weights {0.266, 0.232, 0.502, 0.502, 0}. The results (unreported) show that the average *MARs* for the optimal set of weights for z_3 and z_4 is significantly lower than the set with equal weights.

Table 4-14 presents the results of the performance of ROWK clustering relative to other methods. In this case study, the weighted K-means (*WK*) introduced by Huang et al., 2008 is also included as a benchmark. Given the aim of minimising the sum of weighted squared distances, it is expected that *WK* tends to assign a substantial weight to z_4. The first finding from Table 4-14 is the superior performance of *STD_K* relative to those of *UNSTD_K*. This gives further evidence in support of hypothesis H3a, since the standard deviation of z_3 is the lowest of all five features (see Panel B, Table 4-12). As a result, the

standardization process coincidently aligns with the true weighting, leading to the superior results of *STD_K* over *UNSTD_K*. The second finding relates to the optimal weights found by *WK*. As expected, while correctly identifying the noise feature Z_5, *WK* incorrectly assigns considerable weight to z_4 (0.332). Consequently, the performance of *WK* is not much different from *UNSTD_K* and significantly worse than *STD_K*.

In contrast, ROWK clustering with its innovation to recognize a weight to a feature based on its contribution to both class recognition and regression estimation is predicted to dominate generic K-means clustering, irrespective of standardization. The results presented in Table 4-14 support this prediction. Using optimal weights found by ROWK clustering, *MAR* is significantly lower (0.9864) than those from *ALL* (1.1716), *STD_K* (1.0141), *UNSTD_K* (1.0786) and *WK* (1.0645). Regarding cluster validation, if *purity_ver3* indexes are computed using the *five-class* patterns, only 31% of observations are precisely assigned by ROWK, which is not much different from those of *STD_K* and *UNSTD_K*. This is rational since ROWK clustering does not aim to discover the *real* class patterns (i.e. five classes). Its ultimate goal is to explore class patterns that reduce the problem of HGSC, here *Class* ξ_1, *Class* ξ_2 and *Class* ξ_4. Consistent with our expectations, if only these class members are considered, 62.47% of observations are correctly assigned by ROWK, which is higher than the classifications using *STD_K* (54.27%), *UNSTD_K* (48.91%) and *WK* (48.15%). Similar findings are observed for the *MSRs* and the *pseudo R-square*. This evidence gives further support for hypothesis H4c. By assigning a weight to a feature based on its contribution to both class recognition and regression estimation, ROWK clustering improves the results of regression estimates. See Appendix Table B6 for the purity indices for each cluster/class and the frequency of class membership by clusters.

Table 4-14: Performance of ROWK (Case 3)

"*DIF*" denotes differences of mean absolute residuals (*MAR*) between ROWK clustering and its benchmarks Paired t-tests is used to test for mean differences. T-statistics are presented in brackets. IDEAL denotes the case of 100% correctly-assigned members; *ALL* denotes the case of running regressions without clustering; *ROWK*, *STD_K*, *UNSTD_K and WK* denote the cases of running regressions within each cluster found in ROWK clustering, standardized K-means clustering, unstandardized K-means and weighted K-means clustering respectively.

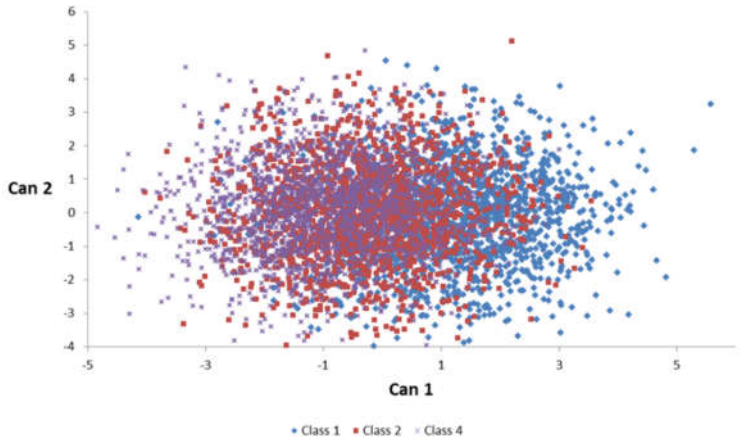
$p^{ver3}_{overall\_5class}$ indicates the purity index version 3 based on five classes. $p^{ver3}_{overall\_3class}$ denotes the purity index version based on three classes (*Class* ξ_1^0 *, Class* ξ_2^0 *and Class*$\xi\ _4^0$). *Var Y* denotes total variance of dependent variables. *R_sq* denotes the index of model fitness *(=1-MSR/Var Y*); *vs.IDEAL* denotes how good of a procedure's *MAR* relative to that of ideal case, computed as 100*(*MAR_{all} - MAR_i)/(MAR_{all}-MAR_{ideal}).

*, **, *** denote significance at 10%, 5% and 1%, respectively.

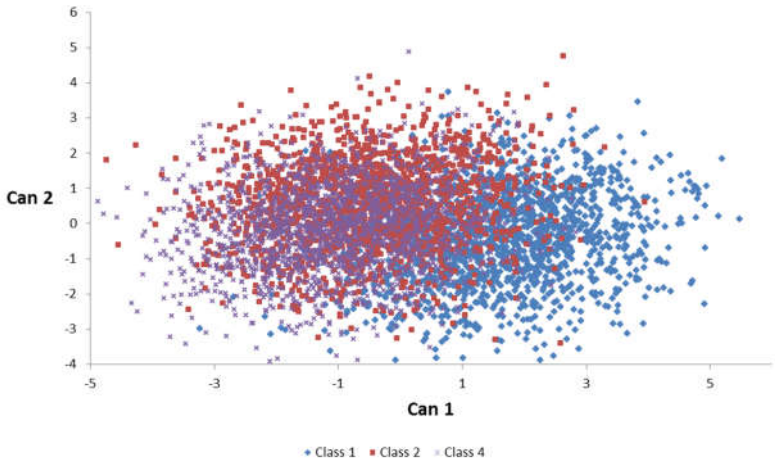
	ROWK	IDEAL	ALL	STD_K	UNSTD_K	WK
W1	**0.2656**	_	_	_	_	0.1295
W2	**0.2322**	_	_	_	_	0.0207
W3	**0.5022**	_	_	_	_	0.5181
W4	**0**	_	_	_	_	0.3316
W5	**0**	_	_	_	_	0
MAR	**0.9864**	0.7879	1.1768	1.0141	1.0786	1.0645
DIF_MAR	_	-0.1985***	0.1904***	0.0277**	0.0922***	0.0781***
		(-5.87)	*(6.11)*	*(1.89)*	*(4.37)*	*(3.25)*
MSR	**1.6763**	0.9579	2.2415	1.7590	1.9439	1.9264
DIF_MSR	_	-0.7184***	0.5652***	0.0827**	0.2676***	0.2501***
		(-6.76)	*(5.79)*	*(1.78)*	*(2.75)*	*(2.44)*
$p^{ver3}_{overall\_5class}$	**0.3100**	100.00	20.00	0.3141	0.3277	0.3403
$p^{ver3}_{overall\_3class}$	**62.47**	100.00	33.33	54.27	48.91	0.5137
RSQ	**0.5488**	0.7422	0.3967	0.5265	0.4768	0.4815
vs. IDEAL	**48.96**	100.00	0.00	41.84	25.25	28.88
N	**5000**	5000	5000	5000	5000	5000

Figure 4.15 graphically presents three-class membership (*Class* ξ_1*, Class* ξ_2 and *Class* ξ_4) based on the first two canonical variables derived from CDA using cluster membership from *ROWK* and *STD_K*. As before, the simulated data are created to mirror real data where class patterns, if any, are relatively indistinct. Figure 4.15a presents the results from *STD_K*. No clear patterns are discernable, and extreme overlap is evident between members of these classes. In contrast, a clearer picture is observed for the results of *ROWK* in Figure 4.15b. More importantly, Class ξ_1's members that exhibit the most distinguishable coefficients from those of other classes are reasonably distinct when expressed by the first two canonical variables using cluster identification from ROWK clustering.

In summary, the findings from Case Study 3 support hypothesis H4c that proposes a third channel to explain the superiority of ROWK clustering relative to K-means clustering when dealing with the HGSC problem. Specifically, the ROWK clustering mechanism by which a weight is assigned to a feature based on its contribution to both class recognition and regression estimation leads to improved identification of the class patterns that address the HGSC problem.



a. Three-Class Membership by the First Two Canonical Variables Derived from CDA with Clusters from Standardized K-means Clustering-Case 3



b. Three-Class Membership by the First Two Canonical Variables Derived from CDA with Clusters from ROWK Clustering-Case 3

Figure 4.15- Three-Class Membership by CDA with Clusters from Different Techniques- Case 3

4.5 ROBUSTNESS TESTS

4.5.1 Out-Of-Sample Results

Pursuant to the strong evidence found in three aforementioned case studies, this thesis undertakes out-of-sample robustness tests. The reasons to conduct out-of-sample tests are straightforward. First, only one simulation data set is generated for each case study above. This means that the results discussed so far are results from the training data set. Furthermore, low out-of-sample predictive power in out-of-sample tests is a problem that researchers may find that casts doubt on the reliability of the predictive model (e.g. Lipe, 1986; Ou & Penman, 1989, Nissim & Penman (2001). Second, the ability to generate multiple simulation data sets allows inferences to be computed with more predictive power. Moreover, some estimators cannot make inferences using a single data set; for example class purity, pseudo R-squared etc. This issue can be solved by creating several random data sets with the same parameters. Accordingly, for each examined case above, 100 simulated data sets are generated with the same set of parameters. Then, K-means clustering is run using the *optimal weights found in the corresponding case*. The averages are computed and reported. Paired t-tests are used to test the significance of mean differences between ROWK clustering and other benchmarks[58].

Table 4-15 presents the out-of-sample performance of ROWK relative to other benchmarks for Case study 1. For all indicators, i.e. class purity, *MAR, MSR*, *Rsquare* and *vs.IDEAL*, ROWK clustering achieves significantly better results relative to other methods. Strikingly, the performance of ROWK is even more compelling for out-of-sample data sets. This evidence confirms the validity of ROWK performance to identify appropriate feature weights. This is in contrast to data mining, which is a common issue challenging data exploration techniques (Epure et al., 2011).

[58] At first glimpse, this is not the correct procedure to conduct out-of-sample tests. Rather, it should be as follows. First, 100 out-of-sample data sets are simulated. Then for each data set, run ROWK clustering and attain the corresponding set of optimal weights and *MARs*. However, at this stage, the timing issue of running the ROWK procedure (it takes over two hours for one run) makes this infeasible. Instead, the study examines how the set of optimal weights found in the training sample performs using simulated out-of-sample data sets. It is analogous to the cross-validation techniques employed in regression analysis (i.e. training samples vs. testing samples).

Table 4-15: Out-Of-Sample Performance of ROWK (Case 1)

100 simulated data samples are generated with the same parameters as Case 1. Reported results are the averages of 100 simulations. "*DIF*" denotes differences between ROWK clustering and its benchmarks. Paired t-tests are used to test for mean differences. T-statistics are presented in brackets. *IDEAL* denotes the case of 100% correctly-assigned members; *ALL* denotes the case of running regressions without clustering; *ROWK, STD_K* and *UNSTD_K* denote the cases of running regressions within each cluster found in ROWK clustering, standardized K-means clustering and unstandardized K-means clustering respectively. $p_{overall}^{ver3}$ denotes the purity index version 3. Var Y denotes total variance of dependent variables. *R_sq* denotes the index of model fitness (*=1-MSR/Var Y*); *vs.IDEAL* denotes the relative performance a procedure's *MAR* relative to that of the ideal case, computed as 100*(*MAR_{all} - MAR_i)/(MAR_{all}-MAR_{ideal}*).
*, **, *** denote significance at 10%, 5% and 1%, respectively.

	ROWK	IDEAL	ALL	STD_K	UNSTD_K
W1	**0.132**	_	_	_	_
W2	**0.1448**	_	_	_	_
W3	**0.5508**	_	_	_	_
W4	**0.1724**	_	_	_	_
W5	**0**	_	_	_	_
MAR	**0.8183**	0.7842	0.9416	0.849	0.8389
DIF_MAR	_	-0.034***	0.1233***	0.0307***	0.0206**
		(3.65)	*(6.43)*	*(3.35)*	*(2.18)*
MSR	**1.068**	0.9627	1.4536	1.1755	1.1389
DIF_MSR	_	-0.1053***	0.3856***	0.1074***	0.0709
		(-3.54)	*(15.44)*	*(6.48)*	*(1.89)***
$p_{overall}^{ver3}$	**0.7196**	1	0.2	0.5182	0.621
*DIF_*$p_{overall}^{ver3}$	_	0.2804***	-0.5196***	-0.2014***	-0.0986***
		(6.55)	*(-12.51)*	*(-5.65)*	*(3.11)*
RSQ (%)	**0.4944**	0.5442	0.3119	0.4435	0.4608
DIF_ RSQ	_	0.0498***	-0.1825***	-0.0509***	-0.0336**
		(3.35)	*(-11.23)*	*(-3.49)*	*(-2.23)*
vs. IDEAL (%)	**0.7837**	1	0	0.5884	0.6528
DIF_ vs. IDEAL	_	0.2163***	-0.7837***	-0.1953***	-0.1308**
		(4.33)	*(-12.31)*	*(-3.68)*	*(-2.21)*
N	**5000**	5000	5000	5000	5000

Table 4-16 presents the out-of-sample performance of ROWK relative to other benchmarks for Case study 2. Briefly, the findings in Case study 2 using an in-sample data set remain unchanged for out-of-sample data sets.

Table 4-16: Out-Of-Sample Performance of ROWK (Case 2)

100 simulated data samples are generated with the same parameters as Case 2. Reported results are the averages of 100 simulations. "DIF" denotes differences between ROWK clustering and its benchmarks. Paired t-tests are used to test for mean differences. T-statistics are presented in brackets. *IDEAL* denotes the case of 100% correctly-assigned members; *ALL* denotes the case of running regressions without clustering; *ROWK, STD_K* and *UNSTD_K* denote the cases of running regressions within each cluster found in ROWK clustering, standardized K-means clustering and unstandardized K-means clustering respectively. $p_{overall}^{ver3}$ denotes the purity index version 3. *Var Y* denotes total variance of dependent variables. *R_sq* denotes the index of model fitness *(=1-MSR/Var Y)*; *vs.IDEAL* denotes the relative performance a procedure's *MAR* relative to that of the ideal case, computed as 100*(*MAR_{all} - MAR_i)/(MAR_{alle}-MAR_{ideal}).
*, **, *** denote significance at 10%, 5% and 1%, respectively.

	ROWK	IDEAL	ALL	STD_K	UNSTD_K
W1	**0.1806**	_	_	_	_
W2	**0.0684**	_	_	_	_
W3	**0.7447**	_	_	_	_
W4	**0.0064**	_	_	_	_
W5	**0**	_	_	_	_
MAR	**0.8365**	0.7851	0.9649	0.8717	0.8682
DIF_MAR	_	-0.0514***	0.1284***	0.0352***	0.0316***
		(3.13)	*(8.90)*	*(2.81)*	*(2.75)*
MSR	**1.126**	0.9631	1.5458	1.2493	1.2386
DIF_MSR	_	-0.1629***	0.4197***	0.1233***	0.1126***
		(-3.88)	*(7.89)*	*(3.23)*	*(2.51)*
$p_{overall}^{ver3}$	**0.638**	1	0.2	0.456	0.532
DIF_$p_{overall}^{ver3}$	_	0.362***	-0.438***	-0.182***	-0.106**
		(5.35)	*(-6.11*	*(-3.65)*	*(2.01)*
RSQ (%)	**0.4845**	0.5591	0.2923	0.428	0.433
DIF_ RSQ	_	0.0746***	-0.1922***	-0.0565***	-0.0515***
		(5.26)	*(-10.99)*	*(-3.19)*	*(-2.83)*
vs. IDEAL (%)	**0.7139**	1	0	0.5184	0.5379
DIF_ vs. IDEAL	_	0.2861***	-0.7139***	-0.1955***	-0.176***
		(6.23)	*(-15.53)*	*(-2.97)*	*(-2.77)*
N	**5000**	5000	5000	5000	5000

Table 4-17 presents the out-of-sample performance of ROWK relative to other benchmarks for Case study 3. Generally, the in-sample superior performance of ROWK clustering in Case study 3 persists for out-of-sample data sets. However, the degree of in-sample improvement from *ROWK* relative to *STD_K* is less in the out-of-sample case. An explanation for this is differences in feature variances. The standard deviation of z_3 is by far the lowest among other features. In contrast, z_4 has the highest standard deviation. Therefore, standardizing the features is coincidently analogous to placing more weight to z_3 and lower weight to z_4. To clarify this statement quantitatively, the effect of

standardization could be transformed into the corresponding effect of assigning weights to features. Recall that ROWK clustering finds the set of optimal weights ($w_{v,\ v=1,\ldots,V}$) such that when this set of weights is applied to K-means clustering and derives the cluster results, the *MAR* from running the regression estimation within each cluster is minimised. Furthermore, this study follows the weighted distance formula from Huang et al. (2008):

$$J = \sum_{k=1}^{K} \sum_{v=1}^{V} \sum_{i \in \xi_k} w_v \left\| z_{iv} - c_{kv} \right\|^2$$

$$\text{Subject to } w_v \geq 0 \; and \; \sum_{v=1}^{V} w_v = 1$$

Hence applying the set of weights ($w_{v,\ v=1,\ldots,V}$) into K-means clustering is equivalent to rescaling the original features by the corresponding factors of $w_v^{1/2}$. On the other hand, the effect of standardization is basically the rescaling of features by the factors of its reversed standard deviation ($Z_{v,stand} = (Z_v - \overline{Z_v})/std(Z_v)$). As a result, the effect of standardization is transformed into the equivalent clustering weights as $\left\{\frac{[std(Z_v)^{-1}]}{\sum_1^V [std(Z_v)^{-1}]}, v = 1, \ldots, 5\right\}$. The average standard deviations of features for 100 simulated data sets are $\{1.38, 1.29, 0.84, 1.68, 1.05\}$ (untabulated). Consequently, the effect of standardization is equivalent to assigning weights of $\{0.17, 0.18, 0.28, 0.14, 0.22\}$. Accordingly, except for the weight of z_5, the relative weights of other features in *STD_K* are similar to those of ROWK's optimal weights, i.e. {0.266, 0.232, 0.502, 0, 0}. This fact explains why the out-of-sample performance of *STD_K* is superior to that of *UNSTD_K* and only modestly inferior to those of ROWK clustering.

In summary, the out-of-sample results for the three case studies remain unchanged. This evidence supports hypotheses H4a, b and c of the channels that lead to the superior performance of ROWK clustering as compared to either standardized on unstandardized K-means clustering.

Table 4-17: Out-Of-Sample Performance of ROWK (Case 3)

100 simulated data samples are generated with the same parameters as Case 3. Reported results are the averages of 100 simulations. "DIF" denotes differences between ROWK clustering and its benchmarks. Paired t-tests are used to test for mean differences. T-statistics are presented in brackets. IDEAL denotes the case of 100% correctly-assigned members; ALL denotes the case of running regressions without clustering; *ROWK, STD_K* and *UNSTD_K* denote the cases of running regressions within each cluster found in ROWK clustering, standardized K-means clustering and unstandardized K-means clustering respectively. $p^{ver3}_{overall\_3class}$ denotes the purity index version based on three classes (*Class* ξ^0_1 *, Class* ξ^0_2 *and Class*ξ^0_4*). Var Y* denotes total variance of dependent variables. *R_sq* denotes the index of model fitness *(=1-MSR/Var Y*); *vs.IDEAL* denotes the relative performance a procedure's *MAR* relative to that of the ideal case, computed as *100*(*MAR_{all} *-* MAR_i*)/(*MAR_{all}*-*MAR_{ideal}*)*.

*, **, *** denote significance at 10%, 5% and 1%, respectively.

	ROWK	IDEAL	ALL	STD_K	UNSTD_K
W1	**0.2656**	_	_	0.172	1
W2	**0.2322**	_	_	0.181	1
W3	**0.5022**	_	_	0.284	1
W4	0	_	_	0.140	1
W5	0	_	_	0.221	1
MAR	**1.0047**	0.7925	1.1730	1.0160	1.0792
DIF_MAR		-0.2122***	0.1683***	0.0114***	0.0746***
		(-25.18)	*(14.70)*	*(3.55)*	*(4.28)*
MSR	**1.7276**	0.9852	2.2392	1.7584	1.9500
DIF_MSR		-0.7425***	0.5116***	0.0308***	0.2224***
		(-21.05)	*(10.63)*	*(2.63)*	*(5.14)*
$p^{ver3}_{overall\_3class}$	**0.6103**	1.0000	0.3333	0.5454	0.5001
DIF_$p^{ver3}_{overall_3class}$		0.3897***	-0.2770***	-0.0649***	-0.1102***
		(36.78)	*(-26.14)*	*(-4.86)*	*(-5.77)*
RSQ (%)	**0.5021**	0.7160	0.3548	0.4932	0.4381
DIF_ RSQ		0.2140***	-0.1473***	-0.0089***	-0.0640***
		(22.83)	*(-11.92)*	*(-2.56)*	*(-5.31)*
vs. IDEAL (%)	**0.4420**	1.0000	0.0000	0.4122	0.2462
DIF_ vs. IDEAL		0.5580***	-0.4420***	-0.0299***	-0.1958***
		(24.71)	*(-19.58)*	*(-3.51)*	*(-4.41)*
N	**5000**	5000	5000	5000	5000

4.5.2 Different Class Sizes

K-means clustering tends to find clusters containing the same number of observations (Tan et al., 2005). For all three examined case studies, classes are generated to have equal numbers of members. To the extent that these findings could be sensitive to the relative size of classes, this thesis next examines the performance of ROWK clustering in the case of unequal relative class sizes. For brevity, only Case study 3 is examined. The results for other cases (unreported) are similar to those of Case 3.

All parameters of Case 3 remain unchanged except that the number of members in classes is not equal. Particularly, *Class* ξ_1^0 has the largest size (1400 members) while *Class* ξ_2^0 has the smallest (600 members). The other three classes have 1000 members. Figure 4.16 displays the three class memberships (i.e. *Class* ξ_1^0, *Class* ξ_2^0 *and Class* ξ_4^0) by the first two canonical variables derived from CDA. It is basically similar to Figure 4.11b except that the size of *Class* ξ_1^0's dominates the other classes. Given the increase in size of *Class* ξ_1^0 and the large overlap between *Class* ξ_2^0 and *Class* ξ_4^0, there is tendency for only two clusters to be found in ROWK clustering; one has the majority of its membership in *Class* ξ_1^0 and the other contains members of two other classes. Figure 4.17 confirms this. Using modified BIC (unreported) and the kneel point in Figure 4.17, the optimal number of clusters are identified as two. Accordingly, as can be seen in Table 4-18, the final set of optimal weights is found to be $\{w_1, w_2, w_3, w_4, w_5\}$ = {0.2343, 0.2297, 0.536, 0, 0} which is almost similar to the results with equal membership size (i.e. {0.266, 0.232, 0.502, 0, 0}). Hence, the three channels that contribute to ROWK's superior performance are robust to the relative size of classes.

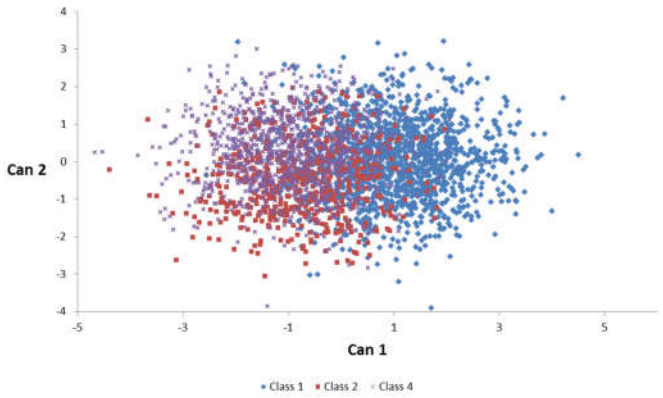


Figure 4.16- Three Unequal-sized Class (Case 3) Membership by the First Two Canonical Variables Derived from CDA.

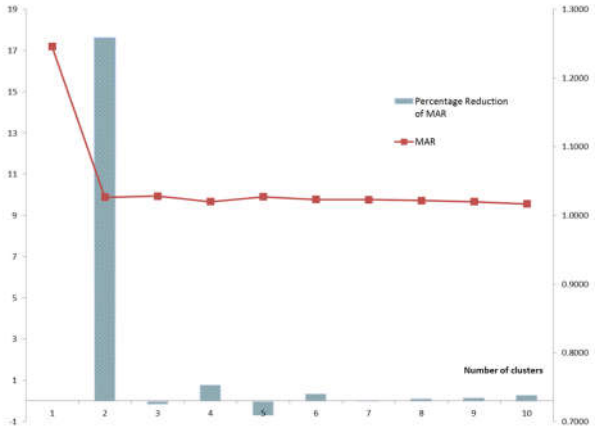


Figure 4.17- ROWK Results at Each Number of Clusters

Table 4-18: Optimal Weights by ROWK Clustering-Unequal Class Size (Case3)

ROWK Results at Each Number of Cluster K'						
Number of clusters (k)	MAR	Z1	Z2	Z3	Z4	Z5
1	1.2455					
2	**1.0261**	**0.2343**	**0.2297**	**0.536**	**0**	**0**
3	1.0278	0.2702	0.2189	0.5109	0	0
4	1.0197	0.0958	0.0958	0.8083	0	0
5	1.0268	0.2998	0.0936	0.6066	0	0
6	1.0231	0.1997	0.1021	0.6982	0	0
7	1.0226	0.1741	0.089	0.7368	0	0
8	1.0215	0.2329	0.0827	0.6844	0	0
9	1.0197	0.2036	0.0723	0.7241	0	0
10	1.0166	0.1807	0.0744	0.7449	0	0

Table 4-19 presents the performance of ROWK clustering relative to other methods. Briefly, for all indicators (*MAR, MSR*, class purity, *R-square* and *vs_IDEAL*), ROWK clustering performance is superior to those running regressions without cluster identification (*ALL*), using clusters identified through standardized (*STD_K*) and unstandardized K-means clustering (*UNSTD_K*).

Table 4-19: Performance of ROWK (Case 3 with Unequal Class Size)

"DIF" denotes differences between ROWK clustering and its benchmarks Paired t-tests are used to test for mean differences. T-statistics are presented in brackets. IDEAL denotes the case of 100% correctly-assigned members; *ALL* denotes the case of running regressions without clustering; *ROWK, STD_K* and *UNSTD_K* denote the cases of running regression within each cluster found in ROWK clustering, standardized K-means clustering and unstandardized K-means clustering respectively. $p^{ver3}_{overall\_3class}$ denotes the purity index version 3 based on three classes (*Class* ξ_1^0 *, Class* ξ_2^0 *and Class*$\xi\ _4^0$*). Var Y* denotes total variance of dependent variables. R_sq denotes the index of model fitness *(=1-MSR/Var Y)*; *vs.IDEAL* denotes how good of a procedure's *MAR* relative to that of ideal case, computed as $100*(MAR_{all} - MAR_i)/(MAR_{all}-MAR_{ideal})$.

*, **, *** denote significance at 10%, 5% and 1%, respectively.

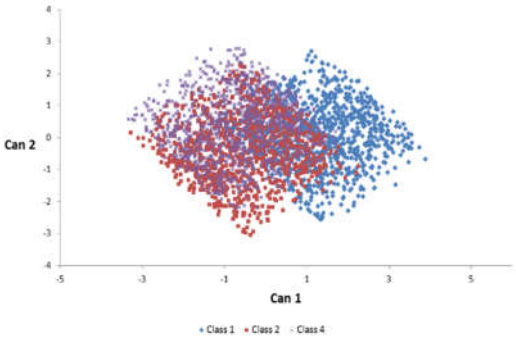
	ROWK	IDEAL	ALL	STD_K	UNSTD_K
W1	**0.2343**	_	_	_	_
W2	**0.2297**	_	_	_	_
W3	**0.536**	_	_	_	_
W4	**0**	_	_	_	_
W5	**0**	_	_	_	_
MAR	1.0261	0.8194	1.2455	1.0354	1.0827
DIF_MAR	_	-0.2067***	0.2193***	0.0092*	0.0565***
		(-5.28)	(6.77)	(1.33)	(2.54)
MSR	1.7991	1.0432	2.4584	1.8436	1.9856
DIF_MSR	_	-0.7559***	0.6593***	0.0445*	0.1865**
		(-4.58)	*(3.48)*	*(1.48)*	*(2.11)*
$p^{ver3}_{overall\_3class}$	0.618	1	0.3689	0.6163	0.5857
R_SQ (%)	0.5463	0.7369	0.38	0.535	0.4992
vs. IDEAL	0.4658	1	0	0.4344	0.3341
N	5000	5000	5000	5000	5000

4.5.3 Different Feature Distributions

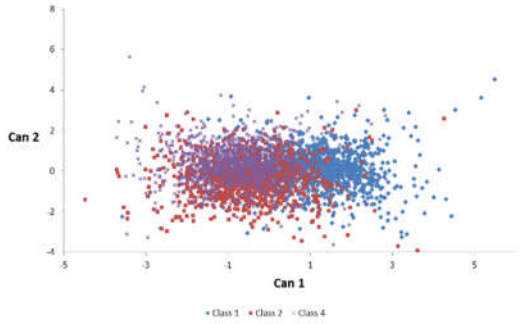
For the three aforementioned case studies, class members are generated to have multivariate normal distributions centered at the corresponding class centroid. However, real data does not always follow a normal distribution (Tanioka & Yadohisa, 2012). This section explores the sensitivity of ROWK performance to other types of class member distributions. Log-normal, Student t and uniform distributions, which are common distributions observed in financial data, are examined (Hürlimann, 2001). Only results for Case 3 are presented. The results for Case 1 and Case 2 are similar and therefore not reported for the sake of brevity.

Figure 4.18 graphically presents three-class memberships (*Class* ξ_1*, Class* ξ_2 *and Class* ξ_4) in Case 3 with uniform, Student-t and log-normal class distributions,

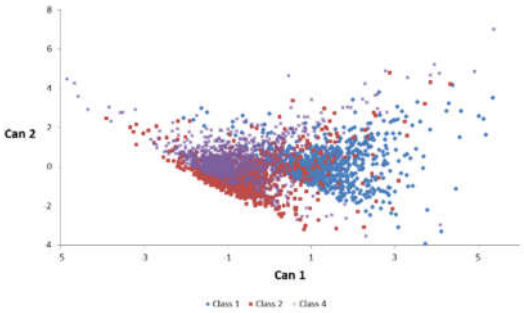
respectively. See Appendix Figures B7a, b and c for the distributions of features. Panel A of Table 4-20 presents the optimal set of weights determined by ROWK clustering for different types of class distribution. For all examined distributions, while the most relevant generated feature, z_3, received the highest weight, the random noise feature z_5 and the irrelevant feature z_4 are eliminated. Furthermore, Panel B shows that for all examined indicators, ROWK clustering achieves better performance relative to other benchmarks. In summary, this evidence supports the robustness of ROWK clustering to the type of class distribution.



a, Uniform Distribution



b, Student-t Distribution



c, Log-normal Distribution

Figure 4.18- Three-Class Membership by the First Two Canonical Variables Derived from CDA – Case 3 with Various Types of Class Distribution

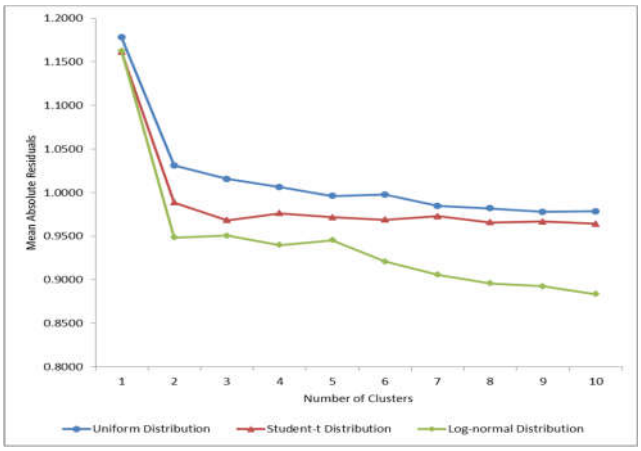


Figure 4.19- ROWK Results at Each Number of Clusters (Case 3 with Various Types of Class Membership Distribution)

Table 4-20: Performance of ROWK (Case 3 with Various Types of Class Distribution)

"*DIF*" denotes differences between ROWK clustering and its benchmarks Paired t-tests are used to test for mean differences. T-statistics are presented in brackets. *IDEAL* denotes the case of 100% correctly-assigned members; *ALL* denotes the case of running regressions without clustering; *ROWK, STD_K* and *UNSTD_K* denote the cases of running regressions within each cluster found in ROWK clustering, standardized K-means clustering and unstandardized K-means clustering respectively. $p^{ver3}_{overall\_3class}$ denotes the purity index version 3 based on three classes (*Class* ξ_1^0 *, Class* ξ_2^0 *and Class*ξ $^0_4$). *, **, *** denote significance at 10%, 5% and 1%, respectively.

Panel A: ROWK Clustering Optimal Weights					
	W_1	W_2	W_3	W_4	W_5
Uniform	0.1346	0.1703	0.6951	0	0
Student-t	0.3117	0.1616	0.5267	0	0
Log-normal	0.1991	0.0046	0.7963	0	0
	ROWK	IDEAL	ALL	STD_K	UNSTD_K
Panel B: Uniform					
MAR	1.0154	0.7875	1.1779	1.0252	1.0806
DIF_MAR	_	-0.228***	0.1625***	0.0098*	0.0651***
		(-4.53)	*(3.61)*	*(1.51)*	*(2.43)*
MSR	1.7616	0.957	2.2539	1.791	1.9525
DIF_MSR	_	-0.8045***	0.4924***	0.0295*	0.191***
		(-7.17)	*(5.77)*	*(1.35)*	*(2.68)*
$p^{ver3}_{overall\_3class}$	0.55412	1	0.33333	0.53915	0.4987
Panel C: Student-t					
MAR	0.9681	0.7878	1.1614	0.9946	1.026
DIF_MAR	_	-0.1803***	0.1932***	0.0265**	0.0578***
		(-418)	*(4.01)*	*(1.98)*	*(2.90)*
MSR	1.6091	0.9585	2.201	1.7018	1.7848
DIF_MSR	_	-0.6506***	0.5919***	0.0928**	0.1758***
		(-6.17)	*(5.31)*	*(1.79)*	*(2.79)*
$p^{ver3}_{overall\_3class}$	0.6554	1	0.3333	0.5502	0.5208

Table 4-20 (cont.) Panel D: Log-normal					
MAR	0.9483	0.7879	1.1625	0.9604	0.9738
DIF_MAR	_	-0.1604***	0.2142***	0.0121*	0.0255**
		(-3.68)	*(4.71)*	*(1.43)*	*(1.82)*
MSR	1.4837	0.9581	2.2119	1.539	1.5799
DIF_MSR	_	-0.5257***	0.7282***	0.0553*	0.0962**
		(-5.11)	*(5.67)*	*(1.35)*	*(2.01)*
$p^{ver3}_{overall\_3class}$	0.5818	1	0.3333	0.5716	0.5591
N	5000	5000	5000	5000	5000

4.5.4 Different Types of Standardization

For the above three case studies, when dealing with the problem of HGSC, traditional methods are chosen as benchmarks. The first benchmark runs the regression without clustering (*ALL*). The second and third and final benchmarks run regressions for each cluster found by unstandardized K-means clustering (*UNSTD_K*), standardized K-means clustering (*STD_K*), and weighted K-means clustering (*WK*) developed by Huang et al (2008), respectively. This thesis examines another five different methods of standardization following Tanioka & Yadohisa (2012). Their formulas are as follows:

(4.2)

$$Z_v^0 = Z_v$$

$$Z_v^1 = (Z_v - \overline{Z_v})/std(Z_v)$$

$$Z_v^2 = Z_v/\max(Z_v)$$

$$Z_v^3 = Z_v/(\max(Z_v) - \min(Z_v)$$

$$Z_v^4 = Z_v/(N * \overline{Z_v})$$

$$Z_v^5 = rank(Z_v) \text{ and}$$

$$Z_v^6 = Z_v/\left(Z_v^{q(97.5)} - Z_v^{q(2.5)}\right)$$

where $Z_v^{q(x)}$ is the x*th* percentile. Z_v^0 is the original (unstandardized) feature, Z_v^1 is the z-score of a feature employed in *STD_K*, and $z_v^s, s = 2, \dots, 6$ are five other examined methods of standardization. For the sake of brevity, the results of this section are not reported and are available upon request. Consistent with Tanioka & Yadohisa (2012), the

results show that of these other standardization methods, the ranking method z_v^5 is the most effective. More importantly, with respect to the problem of HGSC, ROWK clustering consistently outperforms all the above examined standardization methods. Nevertheless, this raises a question of whether ROWK's clustering performance could be further enhanced by ranking features to deal with the outlier effect and non-normal class distributions. We leave it to future studies to shed more light on this.

4.5.5 ROWK Clustering and Factor Analysis

Case study 2 above examined the effect of multicollinearity on the performance of ROWK clustering relative to its benchmarks. Extraction methods (i.e. FA) are widely used to deal with the problem of correlated features in cluster analysis application studies (e.g. Ando & Bai, 2016; Di Cimbrini, 2015; Li & Li, 2008; Mohd-Rahim, et al., 2014; Nimtrakoon &Tayles, 2015). Accordingly, this section compares the effectiveness of ROWK clustering with FA in addressing the problem of multicollinearity. Case study 2 is used to shed light on this issue. The following tests are more conservative in that data in Case study 2 are generated to favor the performance of factor analysis. Specifically, in Case study 2, z_3 and z_4 are generated to be highly within-class correlated (ρ=0.537). Given z_3 is a highly relevant feature, the first component/factor of FA will capture this correlation and contain information of z_3, the most relevant feature.

Principal axis factor analysis with Varimax rotation is used to extract factors from five features. Figure 4.20 graphs the eigenvalue against the number of factors. Only the first two factors have positive eigenvalues, therefore two factors are extracted, accounting for about 37% variation of all features. This is reasonable since only z_3 and z_4 are highly correlated.

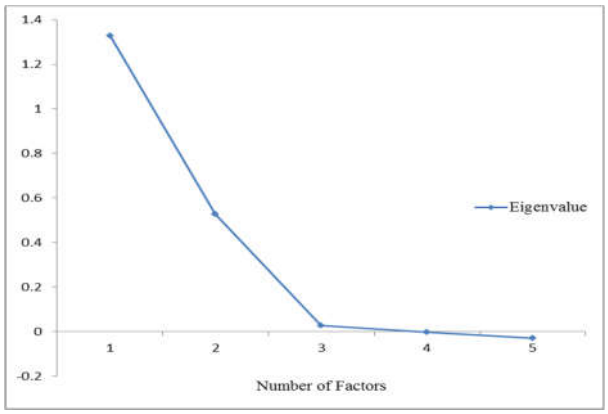


Figure 4.20- Eigenvalues of the Reduced Correlation Matrix

Table 4-21 presents the rotated factor patterns, which are the correlations between the variable and the factor. As expected, z_3 and z_4 are merely embedded in Factor 1, while Factor 2 only comprises z_1. Noticeably, z_2 and z_5 have little correlation with the extracted factors; therefore they have no effect on the factor scores that are used in later clustering. Given that the data in Case study 2 are generated to favor the performance of factor analysis, it is expected that the performance of K-means clustering using the extracted factor scores (*FACTOR* for short) is superior to those of either *UNSTD_K* or *STD_K*. Table 4-22 supports this statement. *FACTOR* has better results for *MAR, MSR*, class purity, *R_sq* relative to those of *UNSTD_K* or *STD_K*. However, *ROWK* still modestly outperformances *FACTOR* at the 10% level of significance. This thesis also explores the case of conducting ROWK clustering using factor scores (*ROWK_FACTOR* for short). *ROWK_FACTOR* outperforms *FACTOR*, but the results are statistically no different from those of *ROWK*. An unreported test (available upon request) shows that when the features in Case study 2 are changed in such a way to reduce the degree of conservatism (i.e. z_4 is highly correlated to z_2, not z_3), then the first factor comprises z_4 and z_2, while the second and third are merely z_1 and z_3 respectively. Hence, FA places more weight on z_4 and z_2, and lower weights on the highly relevant features Z_3 and Z_1. Consequently, the performance of *FACTOR* is inferior to that of *UNSTD_K* or *STD_K*, and even poorer relative to that of ROWK clustering.

Table 4-21: Rotated Factor Pattern		
	Factor1	Factor2
Z1	0.0492	0.7126
Z2	0.1765	0.0954
Z3	0.9872	-0.1720
Z4	0.5490	-0.0036
Z5	-0.0004	0.0133

In summary, the above evidence is consistent with Witten & Tibshirani (2010), who contend that there is no guarantee that these components contain the target signal that the researcher is seeking to identify using clustering. Consequently, the performance of cluster analysis using extracted components is inferior to those using original data if the relevance of these components for the identification of cluster structures is not in line with the extracting methods (e.g. z_4 are highly correlated to z_2, not z_3). In contrast, ROWK clustering as introduced in this thesis employs external criterion from regression analysis (i.e. *MAR*) to guide the clustering process to identify and reduce the weight of irrelevant correlated features, thereby mitigating the problem of multicollinearity.

Table 4-22: Performance of ROWK vs. Factor Analysis (Case 2)

"*DIF*" denotes differences between ROWK clustering and its benchmarks. Paired t-tests are used to test for mean differences. T-statistics are presented in brackets. IDEAL denotes the case of 100% correctly-assigned members; *ROWK, STD_K* and *UNSTD_K* denote the cases of running regressions within each cluster found in ROWK clustering, standardized K-means clustering and unstandardized K-means clustering respectively. *FACTOR and ROWK_FACTOR* denotes K-means and ROWK clustering using the extracted factor scores, respectively. $p^{ver3}_{overall\_3class}$ denotes the purity index version 3 based on three classes (*Class* ξ^0_1 *, Class* ξ^0_2 *and Class*$\xi^{\ 0}_4$). *Var Y* denotes total variance of dependent variables. *R_sq* denotes the index of model fitness *(=1-MSR/Var Y*); *vs.IDEAL* denotes how good of a procedure's *MAR* relative to that of ideal case, computed as $100*(MAR_{all} - MAR_i)/(MAR_{all}-MAR_{ideal})$.
*, **, *** denote significance at 10%, 5% and 1%, respectively.

	ROWK	STD_K	UNSTD_K	FACTOR	ROWK_FACTOR
W1	**0.1806**	_	_	_	W_{F1}= 0.6923
W2	**0.0684**	_	_	_	W_{F1}= 0.3077
W3	**0.7447**	_	_	_	_
W4	**0.0064**	_	_	_	_
W5	**0**	_	_	_	_
MAR	**0.8313**	0.8615	0.8637	0.8383	0.8324
DIF_MAR	_	0.0296***	0.0318***	*0.007**	*0.001*
		(2.78)	*(2.99)*	*(1.35)*	*(0.48)*
MSR	**1.1113**	1.2082	1.2141	1.1378	1.1160
DIF_MSR	_	0.097***	0.103***	0.0265*	*0.0047*
		(4.32)	*(3.18)*	*(1.53)*	*(0.69)*
$p^{ver3}_{overall\_3class}$	**0.623**	0.471	0.515	0.584	0.6035
R_SQ	**0.4844**	0.4393	0.4365	0.4721	0.4822
vs. IDEAL	**0.7124**	0.5358	0.5227	67.42	0.7094
N	**5000**	5000	5000	5000	5000

4.6 SUMMARY

This chapter presents the empirical results for the first research aim. Specifically, this study examines those factors that affect the performance of clustering with respect to the precision of cluster recognition and regression estimates. The four examined factors are class (true cluster) density, class centroid distance, the degree of heterogeneity in regression coefficients and multicollinearity. Consistent with hypothesis H1a (1b), this study finds that the higher the class density (the distances between class centres), the greater the precision for identification of class membership when running regressions within classes. Additionally, the positive relationship between class density (distances between class centres) and clustering precision is stronger when distances between class centres (class densities) are lower. Consistent with hypothesis H1c, the evidence presented supports the proposition that *MARs* decline significantly when (1) distances

between class centres increase, (2) class densities increase and (3) differences in regression coefficients between classes are larger. Furthermore, in support of hypothesis H2 and consistent with Sambandam (2003), evidence is presented of the significantly negative influence of multicollinearity on cluster analysis performance.

This thesis also conducts empirical tests regarding the impact of standardization on the performance of K-means clustering via different contexts. Consistent with hypothesis H3, relative to using unstandardized features, standardization of features results in significantly larger (lower) *MARs*. This is observed when a feature's weight results from differences in distances between class centres (class densities) measured by the feature relative to those measured by other features. This raises concerns regarding the effectiveness of standard K-means clustering whereby cluster features are routinely standardized before running cluster analysis.

This section presents simulated case studies to test hypothesis H4, which posits three channels through which ROWK improves the performance of cluster analysis with respect to the HSGC issue. As hypothesized, this study finds significant improvements of ROWK relative to K-means clustering and weighted K-means clustering through three channels. The channels attributable to these improvements are confirmed. Specifically, ROWK places more (less) weight on more (less) relevant features (Case 1); reduces the influence of multicollinearity by reducing the weights of irrelevant features which are highly correlated with relevant features (Case 2) and captures relevance not only by its contribution to cluster recognition but also by regression estimation (Case 3).

Several robustness tests are conducted. Notably, out-of-sample results remain unchanged, ruling out concerns of data mining. Further, the performance of ROWK is found to be robust to different distributions of cluster features, i.e. student, log-normal or uniform distributions, which are prevalent distributions of financial data. ROWK is also documented to be robust to the existence of unequal classes, which is a challenge for standard K-means clustering. Finally, the performance of FA is examined when features are correlated. Consistent with Witten & Tibshirani (2010), the performance of CA using the extracted components is worse than those using original data if the relevance of these components to the identification of cluster structures is not in line with the extracting

method. In contrast, a novel feature of ROWK clustering is to employ external criterion from regression analysis (i.e. *MAR*) to guide the clustering process. An important outcome is the identification and reduction of the weight of irrelevant correlated features, thereby mitigating the problem of multicollinearity.

CHAPTER 5

EMPIRICAL RESULTS OF ROWK CLUSTERING ON EARNINGS PERSISTENCE

5.1 INTRODUCTION

This chapter presents the empirical results for the second research aim, the first application of the proposed ROWK clustering method, to address the issue of HGSC on earnings persistence. Specifically, it discusses the findings pursuant to hypotheses H5 to H9 to answer the question whether ROWK clustering can help to reveal currently unclear earnings persistence patterns.

First, Section 5.2 presents descriptive statistics of variables of earnings persistence and cluster features. Section 5.3 contains the main results of the chapter. It begins with Section 5.3.1 with the results of ROWK clustering for the earning persistence model. Optimal weights, cluster membership and the number of clusters are identified. Consistent with hypothesis H5, there is evidence that features contribute differently to the identification of clusters. Earnings volatility and the absolute level of accruals are found to be the most dominant features of ROWK clustering, which is consistent with contemporary literature on earnings persistence.

Next, Section 5.3.2 displays the results of earnings persistence patterns. The patterns support hypotheses H5 to H7, revealing that firms exhibit different earnings persistence between ROWK clusters, and that ROWK clustering can lead to better earnings predictability. Section 5.3.3 explores the characteristics of each cluster, revealing several noticeable distinguishable characteristics between clusters. The findings on the heterogeneities of conservatism, earnings management, firm life cycles and industry classification across clusters reported in Section 5.3.4 are consistent with hypothesis H8.

The evidence of whether analysts' forecasts incorporate information of earnings patterns identified by ROWK clustering is presented in Section 5.4. Inconsistent with hypothesis

H9, analysts are found to overestimate future earnings, yet they show a partial understanding of earnings persistence patterns. Notably, the evidence reveals that analysts only incorporate earnings persistence patterns in the short-term, and ignore the impact on long-term future earnings.

Section 5.5 provides numerous robustness tests that support the thesis findings. Out-of-sample results are similar to those from in-sample, providing evidence to support the superior performance of ROWK clustering, ruling out potential concerns of data mining. Further, the performance of ROWK clustering is not sensitive to the inclusion of interaction terms on the earning persistence model, a different proxy for earnings, and long-term future earnings. The chapter is summarised in Section 5.6.

5.2 DESCRIPTIVE STATISTICS, AND TEST SPECIFICATION

As discussed in Section 3.3.2, the main results focus on the 1988-2004 period in a replication of Dichev & Tang (2009) who document the dominance of earnings volatility in distinguishing earnings persistence. The thesis sample is then randomly divided into two equal sub-samples. The first sample (9,794 observations) is for a comprehensive exploratory analysis of the application of ROWK clustering on the earnings persistence model. The second sample (10,070) is used for out-of-sample robustness tests. The differences in the thesis samples and those of Dichev and Tang mainly come from the thesis restriction requiring sample data to be available for 17 clustering features. As a robustness test, the thesis also conducted out-of-sample tests which show that the results are not sensitive to different samples. See Appendix Table C1 for the derivation of the sample and descriptive statistics for the complete sample covering the 1988-2011 period.

Table 5-1 presents the derivation of the sample and descriptive statistics for the 1988-2004 sample. The first seventeen variables are clustering features. Note that the table presents descriptive statistics of original values of the variables winsorized at 1% top and bottom (i.e. not the rank-transformed values). The results are in line with other studies that examine similar variables and time periods. As stated previously, the thesis compares its descriptive statistics with other studies, notably Dichev & Tang (2009) in order to

assure that any differences in findings are due to the new ROWK methodology and not due to major differences in samples. Means of deflated cash flows from operation (*OCF_DEF*, 9.05%), earnings (*IBC_DEF*, 3.15%) and accruals (*ACC_DEF*, -5.9%) are consistent with Dichev & Tang (2009) who document the corresponding values as 8.5%, 3.1% and -5.5% respectively. The mean of profit margin (*PM*) is 10.5%, which is slightly higher than the mean of 1978-1996 period (9.2%) reported by Fairfield & Yohn (2001). In contrast, the mean of asset turnover (*ATO*) of 1.81 is much lower than the mean of 2.25 reported in Fairfield & Yohn (2001). Both the means and medians of change in profit margin (*ΔPM*) and change in asset turnover (*ΔATO*) are positive, however are not significantly different from zero. Financial leverage (*FLEV*) and operating leverage (*OLLEV*) are similar at around 30%. Means of sales growth (*SALES_GR*), net borrowing costs (*NBC*) and dividend payout (*DIV*) are 9.2%, 8.1% and 0.34, respectively. The sample means of firm size (log of total assets, *SIZE*) and age (*AGE*) are 7.1 and 24.5 respectively. The sample mean of earnings volatility (*VOL_IBC_DEF*) is 3.89%, which is slightly lower than the 4.00% reported by Dichev & Tang (2009).

There are significant reductions in sample size when variables *NBC* and *DIV* are added. The final sample size is 19,864, reducing by nearly 4% to 19,179 when *NBC* is included, and by nearly 6% to 18,804 when *DIV* is added. To overcome this issue, new measures of *NBC* and *DIV* are estimated as discussed previously in Section 3.3.2.

The distribution of examined variables is the main concern of this thesis. Figure 5.1 presents the distributions of the clustering features[59]. Similar to Dichev & Tang (2009), the distribution of earnings volatility is bounded at zero on the left and is heavily right skewed. Note that all variables are winsorized at the 1st and 99th percentiles. Heavily right skewed distributions are also observed for several other features, such as asset turnover, current ratio, capital expenditure, and operating leverage. To avoid potential problems associated with skewness and (possible) outliers of feature distributions, the original cluster features are next ranked as follows. The first version ranks by 100th percentiles as in Tanioka & Yadohisa (2012). The second version is a normal-distribution ranking

[59] This figure is the same as Figure 3.2.

where cluster features are first ranked and then transformed into normal distributions. This aligns well with the original data for our features which tend to display relatively normal distributions if skewness is removed. For brevity, only the second version is presented here. The results remain unchanged if version 1 is used and are available upon request from the author. Unless specifically mentioned as an original feature, further discussion of cluster features refer to the rank-transformed version 2.

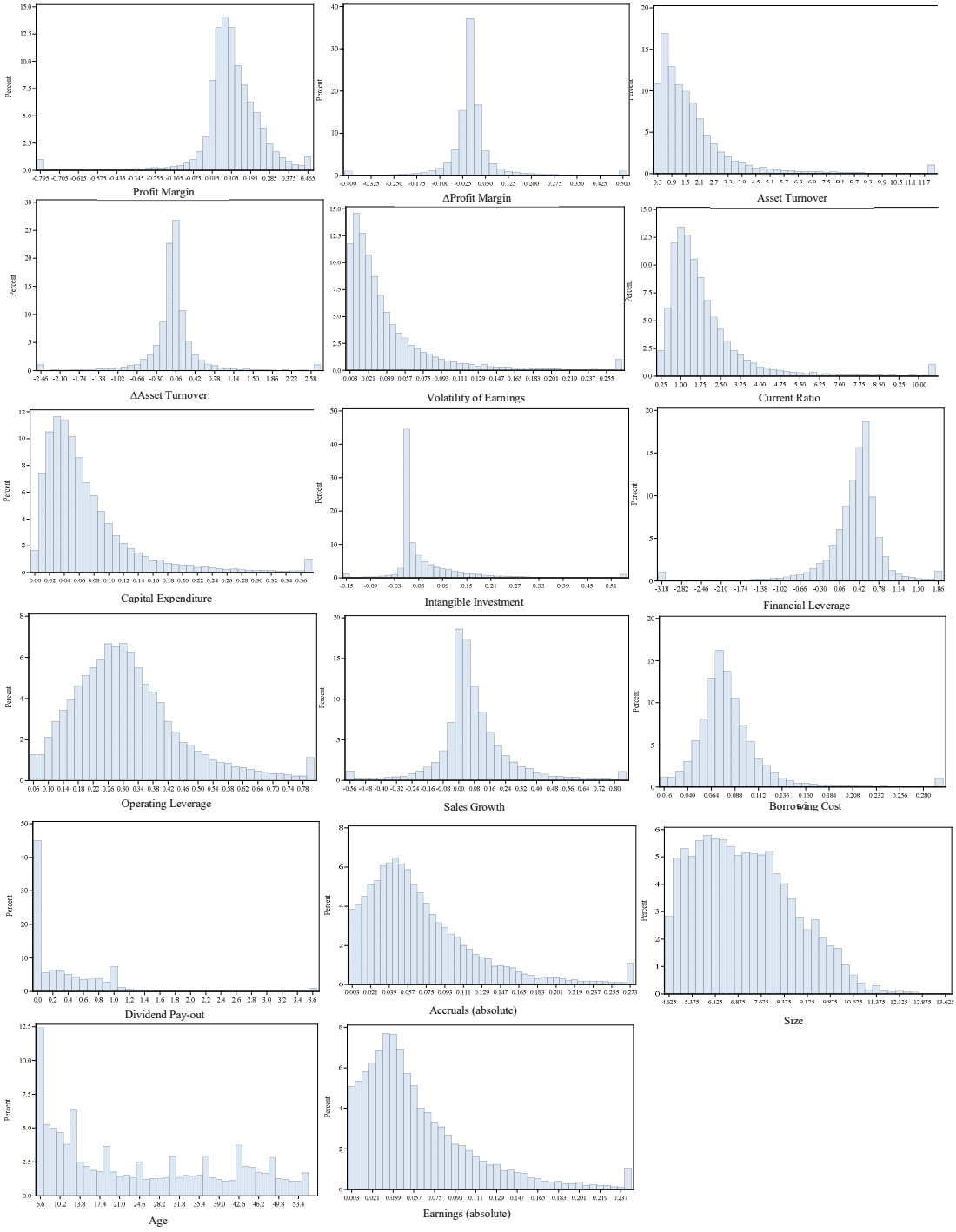


Figure 5.1- Distribution of Cluster Features Winsorized at 1% Top and Bottom

Table 5-1: Derivation of the Sample and Descriptive Statistics

Panel A: Derivation of the Sample

COMPUSTAT firm-years from 1984–2004 with 12/31 fiscal year-end	124,107
Firm-years with available deflated earnings, cash flows and accruals	79,818
Exclusion of financial firms (SIC code from 6000 to 6999)	68,183
Firm-years with assets greater than $100 million	37,329
Firm-years with available data on earnings volatility and cash flow volatility (based on the most recent 5 years)	22,591
Firm-years remaining after truncating the top and bottom 1% on deflated earnings, accruals and cash flows	21,377
Firm-years with available data of cluster features	19,864
Final sample	***19,864***

Panel B: Descriptive Statistics

No.	Variables	N	Mean	Median	Std.Dev	Min	Max
1	PM	19,864	0.1051	0.1006	0.1462	-0.7132	0.4660
2	ΔPM	19,864	0.0028	0.0014	0.0865	-0.3661	0.4435
3	ATO	19,864	1.8153	1.3142	1.8345	0.1837	11.8835
4	ΔATO	19,864	0.0085	0.0085	0.5561	-2.4386	2.6682
5	VOL_IBC	19,864	0.0389	0.0241	0.0451	0.0016	0.2667
6	CR	19,864	1.8949	1.4538	1.6053	0.2666	10.5015
7	CAPX_DEF	19,864	0.0700	0.0520	0.0632	0.0035	0.3597
8	INTAN_INV_DEF	19,864	0.0448	0.0060	0.0961	-0.1394	0.5344
9	FLEV	19,864	0.3027	0.4152	0.6258	-3.2388	1.8185
10	OLLEV	19,864	0.3154	0.2953	0.1494	0.0623	0.8397
11	SALE_GR	19,864	0.0923	0.0407	0.2753	-0.4960	1.5234
12	NBC	19,179	0.0817	0.0760	0.0370	0.0175	0.2983
13	DIV	18,804	0.3453	0.0984	0.5805	0.0000	3.8275
14	AB_ACC_DEF	19,864	0.0687	0.0560	0.0527	0.0015	0.2747
15	SIZE	19,864	7.0908	6.9367	1.5640	4.6531	10.9200
16	AGE	19,864	24.5382	21.0000	15.3050	6.0000	54.0000
17	ABS_IBC_DEF	19,864	0.0601	0.0467	0.0487	0.0013	0.2427
18	IBC_DEF	19,864	0.0315	0.0375	0.0728	-0.4099	0.2495
19	ACC_DEF	19,864	-0.0590	-0.0527	0.0656	-0.3978	0.1689
20	VOL_OCF	19,864	0.0400	0.0314	0.0314	0.0044	0.1768
21	OCF_DEF	19,864	0.0905	0.0874	0.0732	-0.1998	0.3352

PM, ΔPM denote profit margin and the change in profit margin; ATO, ΔATO represent asset turnover and the change of asset turnover; VOL_IBC, VOL_OCF denote volatility of earnings and volatility of operating cash flows measured as standard deviation of the most recent five years; CR is current ratio; CAPX_DEF and INTAN_INV_DEF are measures of deflated capital expenditure and deflated investment in intangible assets; FLEV and OLLEV denote financial leverage and operating leverage; SALE_GR is sales growth; NBC is net borrowing cost; ACC_DEF and AB_ACC_DEF denote deflated accruals and absolute value of deflated accruals; DIV, SIZE and AGE represent dividend payout, log of total assets and firm age respectively; IBC_DEF and ABS_IBC_DEF denote deflated earnings and absolute value of deflated earnings; OCF is operating cash flows. See Section 2.5.2 for measurement details of the variables.

Table 5-2 exhibits the correlation coefficients between clustering features. See Appendix Table C2 for the correlation coefficients between clustering features for the complete sample covering the 1988-2011 period. Over 96% (131 out of 136) of the correlation

coefficients are significant at the 10% level, and nine pairs of clustering features have their absolute value of correlation over 0.3. This multicollinearity issue challenges the performance of standard clustering techniques, raising the need for the new ROWK clustering method proposed in this thesis. No correlation coefficients are higher than 0.7, so all cluster features are used to execute the ROWK clustering procedure.

Table 5-2: Correlations of Clustering Features

		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
1	PM	1	0.18	-0.44	-0.03	-0.35	-0.20	0.16	-0.07	0.06	-0.15	0.01	-0.13	-0.27	-0.10	0.27	0.11	0.16
2	ΔPM		1	0.05	0.22	0.08	0.03	-0.03	0.02	-0.02	0.03	0.18	0.04	-0.24	-0.03	-0.03	-0.03	0.10
3	ATO			1	0.14	0.18	0.27	-0.10	0.28	-0.29	0.53	0.31	0.16	-0.24	0.01	-0.23	0.01	0.22
4	ΔATO				1	0.04	-0.03	-0.09	-0.08	0.07	0.12	0.33	0.12	0.01	0.11	0.00	0.03	0.03
5	VOL_IBC					1	0.28	-0.06	0.16	-0.17	-0.03	0.01	0.11	-0.03	0.24	-0.32	-0.28	0.18
6	CR						1	-0.19	0.26	-0.50	-0.15	0.13	0.02	-0.25	-0.09	-0.38	-0.17	0.18
7	CAPX_DEF							1	-0.07	-0.05	-0.12	0.13	-0.05	-0.09	0.23	0.06	-0.04	0.11
8	INTAN_INV_DEF								1	-0.27	0.07	0.25	-0.03	-0.19	-0.05	-0.05	-0.07	0.18
9	FLEV									1	-0.09	-0.14	-0.05	0.27	0.01	0.24	0.13	-0.34
10	OLLEV										1	-0.03	0.14	0.07	-0.05	0.22	0.28	0.05
11	SALE_GR											1	0.03	-0.36	-0.05	-0.13	-0.13	0.19
12	NBC												1	-0.01	0.04	-0.07	-0.03	0.01
13	DIV													1	0.15	0.20	0.18	-0.24
14	AB_ACC_DEF														1	-0.07	-0.13	0.12
15	SIZE															1	0.42	-0.13
16	AGE																1	-0.10
17	ABS_IBC_DEF																	1

PM, ΔPM denote profit margin and the change in profit margin; ATO, ΔATO represent asset turnover and the change of asset turnover; VOL_IBC, VOL_OCF denote volatility of earnings and volatility of operating cash flows measured as standard deviation of the most recent five years; CR is current ratio; CAPX_DEF and INTAN_INV_DEF are measures of deflated capital expenditure and deflated investment in intangible assets; FLEV and OLLEV denote financial leverage and operating leverage; SALE_GR is sales growth; NBC is net borrowing cost; ACC_DEF and AB_ACC_DEF denote deflated accruals and absolute value of deflated accruals; DIV, SIZE and AGE represent dividend payout, log of total assets and firm age respectively; IBC_DEF and ABS_IBC_DEF denote deflated earnings and absolute value of deflated earnings; OCF is operating cash flows. See Section 2.5.2 for measurement details of the variables. All correlations are significant at 10% level, except those shaded in blue. The yellow-shaded squares highlight correlations over +0.3 or less than -0.3.

5.3 ROWK CLUSTERING AND EARNINGS PERSISTENCE PATTERNS

This section presents the results from the application of ROWK clustering for the earnings persistence model. After pre-processing variables of the earnings persistence model and all clustering features, ROWK clustering is executed.

5.3.1 ROWK Optimal Weights

First, all cluster features are ranked by their mean of squared residuals (*MSR*) measured by running K-means clustering for each feature one at a time. The act of transforming clustering features into their rank version eliminates concerns of noise of data observed in the original data. Hence, the thesis employs the *MSRs* as the criterion to guide the process of identifying optimal weights of features. The use of *MARs* does not change any results. Figure 5.2 displays clustering features' *MSRs* in order when the number of clusters is set to 8. *AB_ACC_DEF* achieves the lowest *MSR* among features, while *VOL_IBC* attains the second lowest, consistent with previous studies on earnings persistence (e.g. Dichev & Tang, 2009; Sloan, 1996). It is also observed that *INTAN_INV_DEF* alone can distinguish earnings persistence, resulting in low *MSRs* (ranking the 5th lowest among 17 examined features). This is consistent with Kwon & Yin (2015) who observe a difference in earnings persistence between high tech and low tech firms.

Interestingly, *PM* has the third lowest *MSR*, lower than that of *ΔPM* (ranking 9th) and *ΔATO* (ranking 6th). It means that *PM* is better than the latter variables in identifying the patterns of firms' earnings persistence. This is in contrast to the findings of Fairfield & Yohn (2001), Amir et al. (2011), and Bauman (2014) who document the superiority of *ΔPM* and *ΔATO* relative to their levels in providing information on earnings persistence and future earnings. One possible explanation for this inconsistency is the high correlation between *PM* and the two lowest *MSR* features, i.e. *VOL_IBC* (ρ = -0.35) and *AB_ACC_DEF* (ρ = -0.1). The corresponding correlations with *VOL_IBC* and *AB_ACC_DEF* in the case of *ATO*, *ΔPM* and *ΔATO* are (0.18, 0.01), (0.08; -0.03), (0.04, 0.11), respectively. Although at this stage we have no definitive conclusion, it is expected that if the superiority of *PM* is purely due to its high correlation with *VOL_IBC* and

AB_ACC_DEF, then ROWK clustering will correctly recognize it and will not assign a large weight to *PM*.

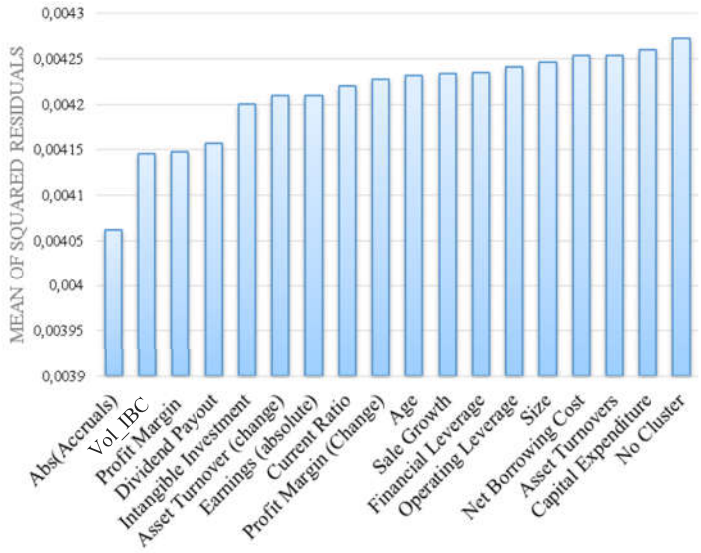


Figure 5.2- MSRs for Each Feature at 8 clusters

After ranking clustering features based on their *MSRs* when running K-means clustering using one feature at the time, the optimal weights are found by executing ROWKs' algorithm as described in Step 4, Section 3.2.1.1.2. Figure 5.3 shows the graphs of *MSRs* and the modified Bayesian criteria (modified *BIC*, see Equation 3.6) by the number of clusters. Both knee points of *MSRs* and the modified *BIC* lead to the same optimal number of clusters at 8. Therefore, firms are optimally divided into eight clusters.

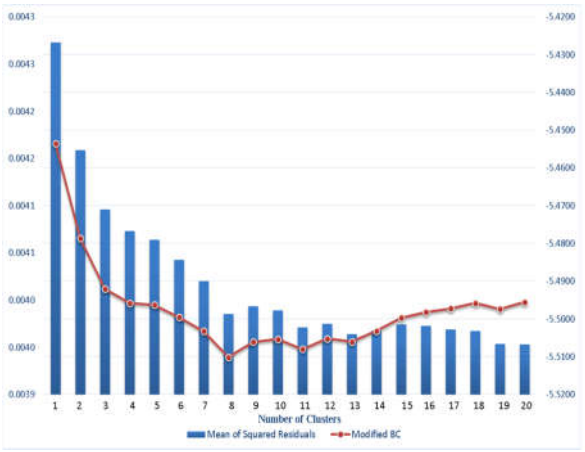


Figure 5.3- MSRs and Modified BIC by the Number of Clusters

Table 5-3 presents the results of ROWK optimal weights for each number of clusters. See Appendix Table C3 for the stepwise results of ROWK clustering when the number of clusters is set as the optimal, i.e. 8. There are several important findings. First, across different examined numbers of clusters (2 to 20), the ROWK clustering procedure consistently assigns non-zero weights to the following five features: *ΔPM; ΔATO; VOL_IBC; INTAN_INV_DEF* and *AB_ACC_DEF*. Meanwhile, the other thirteen features consistently receive zero-weights across different numbers of clusters. Second, ROWK clustering assigns the highest weight to *AB_ACC_DEF* (0.393). This is consistent with Sloan (1996) who documents lower persistence of the accruals component of current earnings. He also observes that the firm stock price acts as if investors do not fully understand the lower persistence of accruals, and "fixate" on earnings. Subsequently, researchers have refined research designs over the last decade and focused upon testing types/components of accruals that are less persistent, the negative relationship between accruals and stock returns, and the explanation of this negative relationship (Richardson et al., 2010)[60] [61]. ROWK also places an equally high weight on *VOL_IBC*, (i.e. 0.393). This is consistent with Dichev & Tang (2009). They find that among other examined variables, earnings volatility is superior in distinguishing earnings persistence across partitioned groups, and consequently attains better earnings predictability.

This thesis does not aim to dig deeply into the theoretical background of the mechanisms of how accruals and earnings volatility impact on earnings persistence and earnings predictability. Rather, this study considers whether ROWK clustering can be used as an exploratory method to discover relevant features that contribute most to distinguish unknown clusters, if any. In this regard, the highest weights that ROWK clustering

[60] Richardson, Sloan, Soliman & Tuna (2005) extend the work of Sloan (1996) by examining the connection between accrual reliability and earnings persistence. They find that accruals that are less reliable result in lower earnings persistence. Dechow, Richardson & Sloan (2008) further examine the persistence and pricing of cash components of earnings. While the higher persistence of cash flows mainly stems from equity components, the other components (the change in cash balance and issuances/distributions to debt) have persistence as low as those of accruals.

[61] The accruals anomaly (the negative relationship between the accruals component of earnings and abnormal future returns) has received significant attention from researchers. Explanations consider investor attributes such as behavioural biases (Hirshleifer, 2001; Subrahmanyam, 2007), limited attention (Hirshleifer, Hou, Teoh & Zhang, 2004), and misunderstanding diminishing returns to new investments (Dechow et al., 2008). Competing explanations consider earnings management and accounting distortions (Dechow & Dichev 2002, Richardson, Sloan, Soliman & Tuna, 2006).

assigns to *AB_ACC_DEF* and *VOL_IBC* (both features account for up to nearly 80% of the weights of all seventeen examined features) provide compelling evidence of the effectiveness of ROWK as an exploratory tool to explore the unknown patterns underlying examined regression models.

Table 5-3: ROWK Optimal Feature Weights Across Different Numbers of Clusters

Features/No. of Clusters	1	2	3	4	5	6	7	**8**	9	10
PM	_	0	0	0	0	0	0	**0**	0	0
ΔPM	_	0.013	0.016	0.02	0.034	0.015	0.017	**0.027**	0.046	0.004
ATO	_	0	0	0	0	0	0	**0**	0	0
ΔATO	_	0.069	0.088	0.11	0.104	0.129	0.12	**0.119**	0.107	0.112
VOL_IBC	_	0.215	0.474	0.341	0.344	0.428	0.397	**0.393**	0.43	0.449
CR	_	0	0	0	0	0	0	**0**	0	0
CAPX_DEF	_	0	0	0	0	0	0	**0**	0	0
INTAN_INV_DEF	_	0.041	0.131	0.165	0.173	0.074	0.069	**0.068**	0.062	0.064
FLEV	_	0	0	0	0	0	0	**0**	0	0
OLLEV	_	0	0	0	0	0	0	**0**	0	0
SALE_GR	_	0	0	0	0	0	0	**0**	0	0
NBC	_	0	0	0	0	0	0	**0**	0	0
DIV	_	0	0	0	0	0	0	**0**	0	0
AB_ACC_DEF	_	0.663	0.291	0.364	0.344	0.353	0.397	**0.393**	0.355	0.371
SIZE	_	0	0	0	0	0	0	**0**	0	0
AGE	_	0	0	0	0	0	0	**0**	0	0
ABS_IBC_DEF	_	0	0	0	0	0	0	**0**	0	0
MSRs(x100)	0.4273	0.4159	0.4096	0.4073	0.4064	0.4043	0.402	**0.3985**	0.3994	0.3989
Features/No. of Clusters	11	12	13	14	15	16	17	18	19	20
PM	0	0	0	0	0	0	0	0	0	0
ΔPM	0.025	0.049	0.049	0.039	0.039	0.033	0.034	0.041	0.06	0.009
ATO	0	0	0	0	0	0	0	0	0	0
ΔATO	0.11	0.114	0.114	0.091	0.091	0.091	0.056	0.067	0.1	0.073
VOL_IBC	0.439	0.457	0.457	0.363	0.363	0.365	0.38	0.453	0.187	0.441
CR	0	0	0	0	0	0	0	0	0	0
CAPX_DEF	0	0	0	0	0	0	0	0	0	0
INTAN_INV_DEF	0.063	0.002	0.002	0	0	0	0	0	0	0.001
FLEV	0	0	0	0	0	0	0	0	0	0
OLLEV	0	0	0	0	0	0	0	0	0	0
SALE_GR	0	0	0	0	0	0	0	0	0	0
NBC	0	0	0	0	0	0	0	0	0	0
DIV	0	0	0	0	0	0	0	0	0	0
AB_ACC_DEF	0.363	0.377	0.377	0.507	0.507	0.51	0.53	0.439	0.653	0.477
SIZE	0	0	0	0	0	0	0	0	0	0
AGE	0	0	0	0	0	0	0	0	0	0
ABS_IBC_DEF	0	0	0	0	0	0	0	0	0	0
MSRs(x100)	0.3971	0.3975	0.3964	0.3968	0.3974	0.3973	0.3969	0.3967	0.3953	0.3953

PM, ΔPM denote profit margin and the change in profit margin; ATO, ΔATO represent asset turnover and the change of asset turnover; VOL_IBC, VOL_OCF denote volatility of earnings and volatility of operating cash flows measured as standard deviation of the most recent five years; CR is current ratio; CAPX_DEF and INTAN_INV_DEF are measures of deflated capital expenditure and deflated investment in intangible assets; FLEV and OLLEV denote financial leverage and operating leverage; SALE_GR is sales growth; NBC is net borrowing cost; ACC_DEF and AB_ACC_DEF denote deflated accruals and absolute value of deflated accruals; DIV, SIZE and AGE represent dividend payout, log of total assets and firm age respectively; IBC_DEF and ABS_IBC_DEF denote deflated earnings and absolute value of deflated earnings; OCF is operating cash flows. See Section 2.5.2 for measurement details of the variables. MSR(x100) denotes 100 multiplied by MSR when running the earnings persistence regression within each clusters.

The third important finding identifies three other features that have consistently non-zero weights: *ΔPM, ΔATO* and *INTAN_INV_DEF*, with the corresponding weights of 0.027; 0.119 and 0.068, respectively. Counter to the findings of Fairfield & Yohn (2001), Bauman (2014) and Amir et al. (2011), *PM* is superior to *ΔPM* and *ΔATO* in identifying the patterns of firms' earnings persistence. The contrasting findings are explained by the correlation of *PM* with the two most important features (i.e. *AB_ACC_DEF* and *VOL_IBC*). As illustrated in Chapter 4 using simulated data, one of the mechanisms contributing to the superior performance of ROWK clustering over standard clustering techniques is its effectiveness in reducing the impact of multicollinearity. As expected, when zero weight is placed on *PM*, ROWK clustering assigns considerable weight to each of *ΔATO* and *ΔPM*. ROWK also assigns non-zero weight to *INTAN_INV_DEF*, which is consistent with Kwon & Yin (2015) who find a difference in earnings persistence between high and low tech firms.

Finally, features exhibit significant differences in their contributions to identify clusters. Table 5-4 compares *MSRs* derived from different cluster identifications, i.e. *ROWK, EQUAL_5, EQUAL_17* and *ALL*. *ROWK* denotes running the earnings persistence model regression within each cluster identified by executing ROWK clustering. *EQUAL_5 (EQUAL_17)* indicates the case of running the earnings persistence regression within each cluster identified by executing K-means clustering with equal weights for only five non-zero-weight ROWK features (all 17 clustering features). *ALL* represents the case of running the earnings persistence regression without considering the existence of clusters, i.e. using the full sample as a single cluster. As can be seen from Table 5-4, the *MSR* of *ROWK* is significantly lower (at the 1% level) than those of *EQUAL_5, EQUAL_17* and *ALL*. This is consistent with hypothesis H5, which posits that feature weights identified by ROWK clustering are not equal. This finding is important because financial applications of K-Means clustering where features are treated equally do not guarantee that the underlying clusters, if any, can be revealed since it is likely that some features are more relevant while others are either less relevant or even noise. In the thesis' context, even in the case of employing all 17 examined features with equal weights, the reduction of *MSR* as compared to the case of *ALL* is less than half that of ROWK clustering which employs only five relevant features (-0.000106 vs. -0.000288).

Table 5-4: ROWKs' Optimal Weights vs. Equal Weights

	ROWK	EQUAL_17	EQUAL_5	ALL
PM	**0**	0.059	0	_
ΔPM	**0.027**	0.059	0.2	_
ATO	**0**	0.059	0	_
ΔATO	**0.119**	0.059	0.2	_
VOL_IBC	**0.393**	0.059	0.2	_
CR	**0**	0.059	0	_
CAPX_DEF	**0**	0.059	0	_
INTAN_INV_DEF	**0.068**	0.059	0.2	_
FLEV	**0**	0.059	0	_
OLLEV	**0**	0.059	0	_
SALE_GR	**0**	0.059	0	_
NBC	**0**	0.059	0	_
DIV	**0**	0.059	0	_
AB_ACC_DEF	**0.393**	0.059	0.2	_
SIZE	**0**	0.059	0	_
AGE	**0**	0.059	0	_
ABS_IBC_DEF	**0**	0.059	0	_
MSR	**0.003985**	0.004167	0.004123	0.004273
Dif_MSR	_	0.000182	0.000138	0.000288
		(3.31)	*(2.81)*	*(3.83)*

PM, ΔPM denote profit margin and the change in profit margin; ATO, ΔATO represent asset turnover and the change of asset turnover; VOL_IBC, VOL_OCF denote volatility of earnings and volatility of operating cash flows measured as standard deviation of the most recent five years; CR is current ratio; CAPX_DEF and INTAN_ INV_DEF are measures of deflated capital expenditure and deflated investment in intangible assets; FLEV and OLLEV denote financial leverage and operating leverage; SALE_GR is sales growth; NBC is net borrowing cost; ACC_DEF and AB_ACC_DEF denote deflated accruals and absolute value of deflated accruals; DIV, SIZE and AGE represent dividend payout, log of total assets and firm age respectively; IBC_DEF and ABS_IBC_DEF denote deflated earnings and absolute value of deflated earnings; OCF is operating cash flows. See Section 2.5.2 for measurement details of variables. MSR denotes the mean squared residuals when running the earnings persistence regression within each cluster. ROWK denotes the case of ROWK optimal weights; EQUAL_5 (EQUAL_17) denotes the case of K-Means clustering with equal weights of only five (all 17 clustering features) non-zero-weight ROWK features. ALL indicates the case of treating all observations as a single cluster.

5.3.2 Earnings Persistence Patterns

5.3.2.1 HGSC in the Earnings Persistence Model

Table 5-5 displays results for the earnings persistence regression across different examined methods. It shows the intercepts, persistence coefficients and adjusted R^2 of the regression of future earnings (next year) on current earnings. These results provide evidence about the economic and statistical significance of the hypothesis H6 proposition that firms exhibit different earnings persistence between ROWK clusters. Regression results for the full sample in Panel A document an intercept coefficient of 0.0087,

persistence coefficients of 0.65 and adjusted R^2 of 0.35. These results are in line with existing studies on earnings persistence (e.g. Dichev & Tang, 2009).

Panel B of Table 5-5 presents results when the earnings persistence regression is run within each cluster found by ROWK clustering. As discussed above, firms are assigned to eight clusters. For the purpose of illustration, clusters are ordered based on persistence coefficients from highest to lowest. At this stage, no specific names are assigned to clusters. Instead, the name is based on the assignment of the SAS program. An examination of Panel B reveals that there is strong evidence of differences in intercept and earnings persistence coefficients between clusters. The difference between the highest intercept coefficient (Cluster 2) and the lowest (Cluster 5) is nearly 5%, which is highly significant (p value <0.0001). Earnings persistence is highest at Clusters 6 and 7 (0.96). Cluster 2 has the lowest earnings persistence of 0.494. The difference between the highest and lowest persistence is 0.467, which is highly significant (p value < 0.0001). The above evidence is consistent with hypothesis H6 predicting that there are differences in earnings persistence and intercepts between ROWK clusters. In addition, these differences seem large in magnitude and suggest that cluster membership is economically important.

Next, earnings persistence conditioned on each feature that has non-zero ROWK optimal weights is examined. Specifically, firms are grouped onto octiles formed on each feature. Then, the earnings persistence regression is run for each group. These results provide a benchmark to assess the economic magnitude of ROWK clustering to distinguish earnings persistence. Panel C of Table 5-5 illustrates Dichev & Tang (2009)'s result on our sample by conditioning on *VOL_IBC*, one of the two features receiving the highest weight. Earnings persistence monotonically reduces from Octile 1 to Octile 8 of *VOL_IBC*. However, the difference between the highest and lowest persistence coefficients conditional on *VOL_IBC* is 0.44, significantly lower than those conditional on ROWK's clusters (i.e. 0.47). Furthermore, intercept coefficients exhibit small differences across octiles of *VOL_IBC*. The difference between the highest and lowest intercepts is only 1.1% and is not significantly different from zero. This evidence is consistent with results reported by Dichev & Tang (2009). Panels D, E, F and G display

the results of the earnings persistence model conditional on *ABS_ACC_DEF, ΔPM, ΔATO* and *INTAN_INV_DEF*. The results are in line with other studies that examine similar variables and time periods. Generally, partitioning the sample based on a single feature results in lower differences in coefficients of both intercept and earnings persistence across groups than those using ROWK clustering. These results are consistent with the first part of hypothesis H7 predicting that ROWK clustering results in larger differences in earnings persistence between clusters than a single-feature partitioning technique.

Clusters identified by ROWK clustering also display distinguishable earnings predictability as can be seen from the last column of Table 5-5. The adjusted R^2 derived from running the earnings persistence regression for firms in Cluster 6 is 56.3%, which is double that of Cluster 2 (28.3%). The difference of adjusted R^2 (i.e. 28%) between these clusters is highly significant at 1% level based on the bootstrapping method discussed on Chapter 3. Only *VOL_IBC* has a higher corresponding difference of adjusted R^2 (39%) than those of ROWK. The other features have differences of adjusted R^2 no more than 14%. As regard to total predictability (*MSR* for all clusters), the regression of earnings persistence within each cluster found by ROWK clustering achieves the lowest *MSR* at 0.003985, which is significantly lower than those conditional on each feature. The above evidence supports the second part of hypothesis H7 predicting that ROWK clustering results in higher earnings predictability than a single variable cluster partitioning technique.

Table 5-5: Results for the Earnings Persistence Regression

$$Earnings_{i,t+1} = \mu_{\zeta_k^0} + Earnings_{i,t}\theta_{\zeta_k^0} + u_{i,t}$$

Panel A: Regression Result for the Full Sample				
	N	$\mu_{\zeta_k^0}$	$\theta_{\zeta_k^0}$	Adj. R^2
Full sample	9794	0.0087	0.6518	0.3479
Panel B: Regression Results by Clusters Identified by ROWK Clustering				
Clusters	N	$\mu_{\zeta_k^0}$	$\theta_{\zeta_k^0}$	Adj. R^2
Cluster 8	1351	-0.001	**0.961**	**0.563**
Cluster 7	1129	-0.012	0.960	0.413
Cluster 6	1755	-0.002	0.868	0.433
Cluster 5	851	**-0.026**	0.834	0.339
Cluster 4	1704	0.010	0.831	0.443
Cluster 3	1147	0.003	0.817	0.392
Cluster 2	975	-0.004	0.620	0.290
Cluster 1	882	**0.024**	**0.494**	**0.283**
Difference (H-L)		0.04967	0.46753	0.28
P-value on Difference		*<0.0001*	*<0.0001*	*<0.0001*
Total MSR (*100)				0.3985
Adj. R2				0.3904
Panel C: Regression Results by Octiles of Earnings Volatility				
Octiles by VOL_IBC	N	$\mu_{\zeta_k^0}$	$\theta_{\zeta_k^0}$	Adj. R^2
Octile 1	1219	0.000	**0.942**	**0.651**
Octile 2	1225	0.000	0.902	0.412
Octile 3	1226	0.001	0.883	0.460
Octile 4	1224	0.000	0.839	0.358
Octile 5	1229	0.003	0.825	0.435
Octile 6	1223	**0.007**	0.679	0.294
Octile 7	1228	0.007	0.617	0.300
Octile 8	1220	**-0.004**	**0.500**	**0.255**
Difference (H-L)		0.011	0.442	0.396
P-value on Difference		*0.15*	*<0.001*	*<0.0001*
Difference MSR vs. ROWK (*100)				0.0215
P-value on Difference				*<0.001*
Difference of Adj. R^2 vs. ROWK (p-value)				-0.0263 (<0.0001)
Panel D: Regression Results by Octiles of Absolute Value of Accruals				
Octiles by ABS_ACC_DEF	N	$\mu_{\zeta_k^0}$	$\theta_{\zeta_k^0}$	Adj. R^2
Octile 1	1219	**-0.018**	**0.988**	**0.465**
Octile 2	1225	-0.008	0.815	0.314
Octile 3	1226	-0.001	0.769	0.329
Octile 4	1224	-0.002	0.886	0.436
Octile 5	1229	0.005	0.805	0.398
Octile 6	1223	0.008	0.726	0.369
Octile 7	1228	0.009	0.710	0.336
Octile 8	1220	**0.022**	**0.524**	**0.329**
Difference (H-L)		0.040	0.464	0.136
P-value on Difference		*<0.001*	*<0.001*	*0.045*
Difference MSR vs. ROWK (*100)				0.0105
P-value on Difference				*<0.01*
Difference of Adj. R^2 vs. ROWK (p-value)				-0.0152(<0.001)

Table 5-5 (cont.) Panel E: Regression Results by Octiles of Change in Profit Margin

Octiles by ΔPM	N	$\mu_{\zeta_k^0}$	$\theta_{\zeta_k^0}$	Adj. R^2
Octile 1	1219	**0.001**	0.615	0.317
Octile 2	1225	0.006	0.609	0.269
Octile 3	1226	0.009	**0.594**	**0.279**
Octile 4	1224	0.011	0.639	0.294
Octile 5	1229	0.007	**0.771**	**0.394**
Octile 6	1223	**0.015**	0.646	0.335
Octile 7	1228	0.013	0.639	0.295
Octile 8	1220	0.005	0.628	0.316
Difference (H-L)		0.014	0.177	0.115
P-value on Difference		*0.068*	*0.074*	*0.062*
Difference MSR vs. ROWK (*100)				0.0259
P-value on Difference				*<0.001*
Difference of Adj. R^2 vs. ROWK (p-value)				-0.0393 (<0.0001)

Panel F: Regression Results by Octiles of Change in Asset Turnover

Octiles by ΔATO	N	$\mu_{\zeta_k^0}$	$\theta_{\zeta_k^0}$	Adj. R^2
Octile 1	1219	**-0.002**	0.612	0.285
Octile 2	1225	0.002	0.698	0.363
Octile 3	1226	0.004	0.707	0.388
Octile 4	1224	0.005	0.715	0.356
Octile 5	1229	0.008	**0.736**	**0.397**
Octile 6	1223	0.011	0.680	0.383
Octile 7	1228	0.017	0.635	0.364
Octile 8	1220	**0.019**	**0.577**	**0.337**
Difference (H-L)		0.021	0.159	0.060
P-value on Difference		*0.017*	*0.058*	*0.096*
Difference MSR vs. ROWK (*100)				0.0231
P-value on Difference				*<0.001*
Difference of Adj. R^2 vs. ROWK (p-value)				-0.0349 (<0.001)

Panel G: Regression Results by Octiles of Intangible Investments

Octiles by INTAN_INV	N	$\mu_{\zeta_k^0}$	$\theta_{\zeta_k^0}$	Adj. R^2
Octile 1	1150	0.008	**0.490**	**0.241**
Octile 2	552	0.010	0.711	0.356
Octile 3	2401	**0.015**	0.542	0.240
Octile 4	799	0.003	**0.807**	**0.375**
Octile 5	1221	0.010	0.685	0.343
Octile 6	1223	0.007	0.709	0.364
Octile 7	1228	0.003	0.712	0.367
Octile 8	1220	**0.001**	0.697	0.433
Difference (H-L)		0.014	0.317	0.134
P-value on Difference		*0.089*	*<0.01*	*0.074*
Difference MSR vs. ROWK (*100)				0.0216
P-value on Difference				*<0.001*
Difference of Adj. R^2 vs. ROWK (p-value)				-0.0269 (<0.0001)

Earnings is defined as earnings before extraordinary item deflated by average total assets. ABS_ACC_DEF is the absolute amount of accruals. VOL_IBC_DEF is defined as the firm-specific standard deviation of earnings over the most recent 5 years. INTAN_ INV_DEF denotes deflated investment in intangible assets; PM, ΔPM denote profit margin and the change in profit margin; ATO, ΔATO represent asset turnover and the change in asset turnover. *Difference (H-L)* indicate the difference between highest value and lowest value. *Difference MSR vs. ROWK (*100)* denotes the 100-time difference between the mean squared residuals derived from running the earnings persistence model within each octile of examined features and within each cluster identified by ROWK clustering. The p-value for the difference in the intercepts and persistence coefficients is derived from a t-test. The p-value for the difference in the Adj_R^2 between the highest value and lowest value is derived from a bootstrap test (see Chapter 3.2.2 for full details). *Difference of Adj. R^2 vs. ROWK* denotes the difference between adjusted R^2 from running the earnings

persistence model within each octile of examined features and within each cluster identified by ROWK clustering. The p-value for the difference in the Adj_R^2 is derived from Vuong's test. The bold numbers of intercept and persistence coefficients indicate the extreme values (highest or lowest). The bold numbers of adjusted R^2 are in accordance with the bold persistence coefficients.

5.3.2.2 ROWK and Non-linearity

An important finding from Panels C to G of Table 5-5 concerns the pattern of intercept and persistence coefficients. Persistence across octiles of *ABS_ACC_DEF, ΔPM, ΔATO* and *INTAN_INV_DEF* do not exhibit monotonically decreasing patterns as in *VOL_IBC*. For example, firms with the lowest octiles of *ABS_ACC_DEF* have very high earnings persistence (0.988). Then the persistence coefficient drops to 0.769 at Octile 3 before jumping to 0.89 at Octile 4. From Octile 4, the persistence coefficient monotonically reduces to 0.524 at Octile 8. Notably, the persistence patterns in case of *ΔATO* and *INTAN_INV_DEF* display a U-shape. This raises some concerns regarding the veracity of including interaction terms in regression models to account for the effects of partitioning variables.

In contrast, ROWK clustering could be used to explore these non-linear patterns. To explore this possibility, this thesis further examines the performance of ROWK with the inclusion of interaction terms[62]. Particularly, the five cluster features that have non-zero weights after running ROWK clustering are interacted with earnings and are added to the original earnings persistence model, resulting in Equation 5.1.

(5.1)

$$\begin{aligned} Earnings_{i,t+1} = {} & \mu_{\xi_k^0} + \theta_{\xi_k^0} Earnings_{i,t} + \gamma_{1,\xi_k^0} VOL_{i,t} + \gamma_{2,\xi_k^0} \Delta PM_{i,t} + \gamma_{3,\xi_k^0} \Delta ATO_{i,t} \\ & + \gamma_{4,\xi_k^0} INTAN\_INV\_DEF_{i,t} + \gamma_{5,\xi_k^0} ABS\_ACC\_DEF_{i,t} + \pi_{1,\xi_k^0} VOL_{i,t} * Earnings_{i,t} \\ & + \pi_{2,\xi_k^0} \Delta PM_{i,t} * Earnings_{i,t} + \pi_{3,\xi_k^0} \Delta ATO_{i,t} * Earnings_{i,t} \\ & + \pi_{4,\xi_k^0} INTAN\_INV\_DEF_{i,t} * Earnings_{i,t} + \pi_{5,\xi_k^0} ABS\_ACC\_DEF_{i,t} * Earnings_{i,t} \\ & + u_{i,t}, i \end{aligned}$$

where *VOL, ΔPM, ΔATO, INTAN_INV_DEF* and *ABS_ACC_DEF* are the five cluster features that have non-zero weights after running the original ROWK clustering from

[62] The inclusion of interaction terms is a standard treatment to account for linear or monotonic effects of variables on the relation of causal variables and an outcome (Puhani, 2012).

Equation 3.18. Then, ROWK clustering is executed using the earnings persistence model following Equation 5.1. The list of cluster features is the same as in the case without interaction terms. For brevity, the results of ROWK execution are not reported and are available upon request.

ROWK clustering using Equation 5.1 identifies three features as having non-zero weights, i.e. *ΔATO*, value of accruals (*ACC_DEF*) and *PM*[63]. The corresponding optimal number of clusters is two, i.e. *Cluster Int_1* and *Cluster Int_2*. The optimal weights corresponding to these features are ($w_{\Delta ATO}$, w_{ACC}, w_{PM}) = (0.349, 0.242, 0.409). The highest weight placed upon *PM* is consistent with the findings of Amir et al. (2011). Previous studies on the prediction of future earnings document that while *ΔPM* does help to predict future earnings, the value of *PM* itself does not (e.g. Fairfield & Yohn, 2001). However, Amir et al. (2011) find that while the unconditional persistence of *PM* is lower than that of *ATO*, the conditional persistence of *PM* is higher than that of *ATO*. Consequently, the market's reaction to *PM* is stronger than to *ATO*. This thesis attributes the higher contribution of *PM* to explain earnings persistence to the ability of *PM* to identify non-linear effects that are not addressed by the inclusion of interaction terms. In this case, ROWK clustering as an exploration technique of both linear and non-linear underlying patterns is successful in correctly identifying the relevance of *PM* and assigning the highest weight to it.

Table 5-6 displays the regression results with the inclusion of interaction terms for the full sample and for each cluster identified by ROWK clustering. As expected, for the full sample, the inclusion of interaction terms increases the adjusted R-squared from 34.79% to 38.39%. Most of the added variables are significant. However, for the interaction terms, only the interactions with earnings volatility (i.e. *VOL_IBC_DEF*Earnings*) and accruals (*ACC_DEF*Earnings*) are significant, while those of change in profit margin (*ΔPM*Earnings*), change in asset turnover (*ΔATO*Earnings*) and intangible investment (*INTAN_INV_DEF*Earnings*) are not significant. This result is expected since previous results displayed non-monotonic patterns of slope coefficients. This raises doubts on the

[63] When including interaction terms, the value of accruals contributes to ROWK clustering better than the absolute value of accruals.

appropriateness of using of interaction terms to account for non-monotonic or non-linear effects of (moderator) variables on the relationship between casual variables and the outcome.

In contrast, ROWK clustering successfully identifies these non-monotonic patterns. It turns out that the full sample comprises two latent clusters where the effects of the interaction terms with *ΔPM* and *ΔATO* are opposite, causing the coefficients of *ΔPM*Earnings* and *ΔATO*Earnings* to be insignificant for the full sample regression. For example, the coefficient of *ΔPM*Earnings* (0.141) is insignificantly positive for *Cluster Int_1*, but becomes highly significantly negative (-0.75) for *Cluster Int_2*. This is consistent with findings of Bauman (2014) who addresses the puzzle of persistence of *ΔPM* in forecasting future profitability by partitioning firms based on the direction of *ΔPM*. The same pattern is also observed in the case of *ΔATO*Earnings*.

This thesis does not strive to develop any theory to explain why firms in these two clusters exhibit different coefficients. This is beyond the scope of the thesis and can be the focus of following research. Instead, these results provide compelling evidence that ROWK clustering could be an effective method to explore non-monotonic or non-linear relationships that cannot be captured by the inclusion of interaction terms.

Table 5-6: ROWK Clustering with the Inclusion of Interaction Terms

	Full Sample		ROWK Clustering			
			Cluster Int_1		Cluster Int_2	
	(1)	(2)	(3)	(4)	(5)	(6)
	Parameter Estimate	Pr > |t|	Parameter Estimate	Pr > |t|	Parameter Estimate	Pr > |t|
Intercept	**-0.006**	*<.0001*	**-0.01**	*<.0001*	**-0.005**	*0.0043*
Earnings	**0.872**	*<.0001*	**0.882**	*<.0001*	**0.898**	*<.0001*
ΔPM	0.012	*0.1236*	0.013	*0.3502*	**0.065**	*<.0001*
ΔATO	**0.004**	*0.0004*	**0.004**	*0.0296*	-0.004	*0.1684*
VOL_IBC_DEF	**-0.087**	*<.0001*	**-0.066**	*<.0001*	-0.017	*0.4154*
ACC_DEF	**0.185**	*<.0001*	**0.194**	*<.0001*	**0.219**	*<.0001*
INTAN_ INV_DEF	**-0.026**	*<.0001*	**-0.031**	*0.0084*	-0.008	*0.4451*
ΔPM *Earnings	0.006	*0.9186*	0.141	*0.1236*	**-0.751**	*<.0001*
ΔATO *Earnings	0.015	*0.1466*	**-0.032**	*0.0305*	**0.149**	*<.0001*
VOL_IBC_DEF *Earnings	**-0.496**	*<.0001*	-0.2	*0.2092*	**-1.32**	*<.0001*
ACC_DEF *Earnings	**-1.095**	*<.0001*	**-1.079**	*<.0001*	**-1.723**	*<.0001*
INTAN_INV_DEF *Earnings	0.095	*0.1546*	0.098	*0.3369*	-0.018	*0.8854*
N	9794		4838		4956	
Adj. R_square	0.3839		0.3481		0.3064	
Corrected MSR (*100)	0.4032		0.396			
Dif_MSR (*100)	0.0071 *(p-value =0.061)*					
F test	5.78 *(p-value <0.01)*					

ROWK clustering is run using the earnings persistence with interaction terms model. Two clusters are identified, i.e. Cluster Int_1 and Cluster Int_2. Earnings is defined as earnings before extraordinary item deflated by the average total assets. ACC_DEF is the amount of deflated accruals. VOL_IBC_DEF is defined as the firm-specific standard deviation of earnings over the most recent 5 years. INTAN_ INV_DEF denotes deflated investment in intangible assets; PM, ΔPM denote profit margin and the change in profit margin; ATO, ΔATO represent asset turnover and the change in asset turnover. *"Corrected MSR (*100)"* denotes 100 times the mean squared residuals adjusted by its corresponding degree of freedom. "*Dif_MSR (*100)"* indicates the difference of *"Corrected MSR (*100)"* between the full sample and ROWK clustering. F_test is the F-statistic drawn from the F-test with the null hypothesis of equal coefficients between Cluster Int_1 and Cluster Int_2. Bold values indicate significant coefficients at or below the 10% level.

5.3.2.3 Long-term Analysis

We further examine patterns of earnings persistence and the intercepts across ROWK clusters over the next five years. This serves two purposes. First, we wish to see whether the clusters found by ROWK clustering in the persistence model exhibit heterogeneous-group coefficients transitorily or over the long term. If ROWK clustering is effective over the long term, then we expect to see less or delayed convergence of earnings. Second, given that firm value depends more on permanent changes in earnings streams rather than on transitory shocks (Pimentel & De Aguiar, 2016), evidence of long-term effects of cluster membership on earnings persistence patterns would imply an important role for ROWK clustering in firm valuation.

Table 5-7 presents the results of the earnings persistence model across ROWK clusters when one-year-ahead earnings ($Earnings_{t+1}$) is replaced by subsequent years up to five years ($Earnings_{t+2}$,…,$Earnings_{t+5}$). Full sample results are shown in Panel A. Consistent with previous studies on long-term earnings persistence, earnings persistence reduces monotonically through time, starting at 0.65 for the next year earnings and declining to only 0.33 for the next five years of earnings. Consequently, the predictive power of earnings declines sharply from 34.8% to only 7.8% for longer prediction horizons.

For parsimony, this thesis presents comparisons between *Clusters 1* and *8* (Panels B and C), and *Clusters 4* and *5* (Panel D and E). Even a cursory scrutiny of these four panels unveils considerable differences in the long-run earnings patterns of the underlying clusters. In Panel B, firms in *Cluster 1* display a quick deterioration of persistence (0.49–0.16) and Adj. R^2 (0.28–0.04) over the 5-year predictive horizon. In contrast, Panel C reveals that firms in *Cluster 8* continue to exhibit robust earnings persistence (0.96-0.58) and predictive power (0.56-0.19) over the entire 5-year horizon. In the same manner, the differences between *Cluster 4* and *5* intercepts remain stable even after 5 years, hovering around 3%. These results suggests that ROWK clustering may be a powerful tool to identify the different patterns of intercepts and earnings persistence in the long-run.

Table 5-7: ROWK Clustering and Long-term Earnings Patterns

	N	$\mu_{\xi_k^0}$	$\theta_{\xi_k^0}$	Adj. R^2
Panel A: Regression results for the full sample				
Earnings (t+1)=α+βEarnings(t)	9794	0.009	0.652	0.348
Earnings (t+2)=α+βEarnings(t)	9012	0.017	0.479	0.204
Earnings (t+3)=α+βEarnings(t)	8445	0.02	0.412	0.147
Earnings (t+4)=α+βEarnings(t)	7929	0.018	0.379	0.105
Earnings (t+5)=α+βEarnings(t)	7455	0.019	0.33	0.078
Panel B: Regression results for Cluster 1				
Earnings (t+1)=α+βEarnings(t)	882	0.024	0.494	0.283
Earnings (t+2)=α+βEarnings(t)	754	0.027	0.278	0.118
Earnings (t+3)=α+βEarnings(t)	695	0.028	0.245	0.081
Earnings (t+4)=α+βEarnings(t)	628	0.029	0.229	0.072
Earnings (t+5)=α+βEarnings(t)	572	0.024	0.163	0.038
Panel C: Regression results for Cluster 8				
Earnings (t+1)=α+βEarnings(t)	1351	-0.001	0.961	0.563
Earnings (t+2)=α+βEarnings(t)	1296	0.003	0.851	0.453
Earnings (t+3)=α+βEarnings(t)	1244	0.003	0.827	0.313
Earnings (t+4)=α+βEarnings(t)	1193	0.002	0.763	0.251
Earnings (t+5)=α+βEarnings(t)	1150	0.009	0.579	0.186
Panel D: Regression results for Cluster 4				
Earnings (t+1)=α+βEarnings(t)	1704	0.010	0.831	0.443
Earnings (t+2)=α+βEarnings(t)	1597	0.014	0.717	0.275
Earnings (t+3)=α+βEarnings(t)	1499	0.017	0.633	0.227
Earnings (t+4)=α+βEarnings(t)	1406	0.020	0.555	0.146
Earnings (t+5)=α+βEarnings(t)	1320	0.015	0.551	0.126
Panel E: Regression results for Cluster 5				
Earnings (t+1)=α+βEarnings(t)	851	-0.026	0.834	0.339
Earnings (t+2)=α+βEarnings(t)	753	-0.011	0.642	0.237
Earnings (t+3)=α+βEarnings(t)	700	-0.010	0.469	0.116
Earnings (t+4)=α+βEarnings(t)	646	-0.01	0.499	0.113
Earnings (t+5)=α+βEarnings(t)	598	-0.014	0.518	0.113

5.3.2.4 Graphs of Mean Reversion across ROWK Clusters

5.3.2.4.1 Cluster 1 vs. Cluster 8

Figure 5.4 presents a graphical view of the results in Panel B and C of Table 5-7. The graphs use the same scale to track the evolution of median deflated earnings over the next five years for the full sample (Figure 5.4a), for firms in *Cluster 1* (Figure 5.4b) and firms in *Cluster 8* (Figure 5.4c). Figure 5.4a shows the expected mean reversion. The difference in current deflated earnings between firms in the top and bottom earnings quintiles is 15%, reducing to only 5.3% by the end of the firth year. Consistent with results in Panel B of Table 5-7 the second graph demonstrates much quicker mean reversion for firms in *Cluster 1*. The range of deflated earnings shrinks dramatically from 33% in year *t* to only

an eighth (4.4%) in *t+5*. In contrast, the third graph reveals that firms in *Cluster 8* display high inertia of earnings flows. The range of current deflated earnings is 4.7% and only reduces modestly to 3.5% after five years. These graphs depict clean, simple and long-lasting different patterns of mean reversion between ROWK clusters.

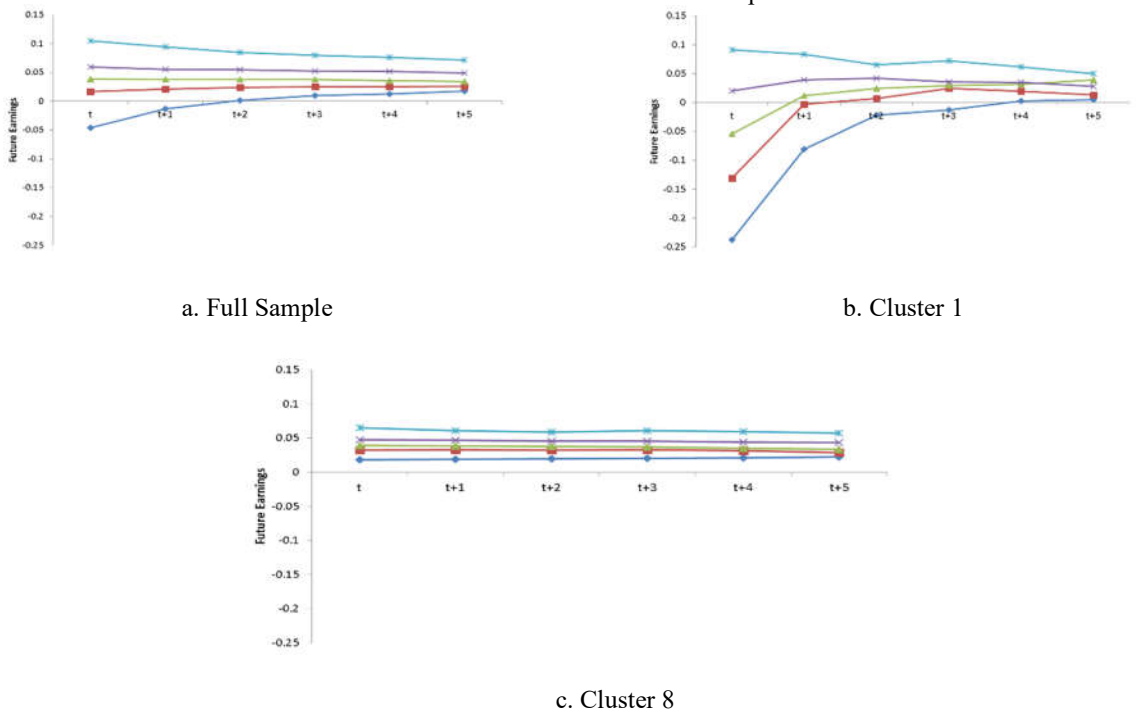
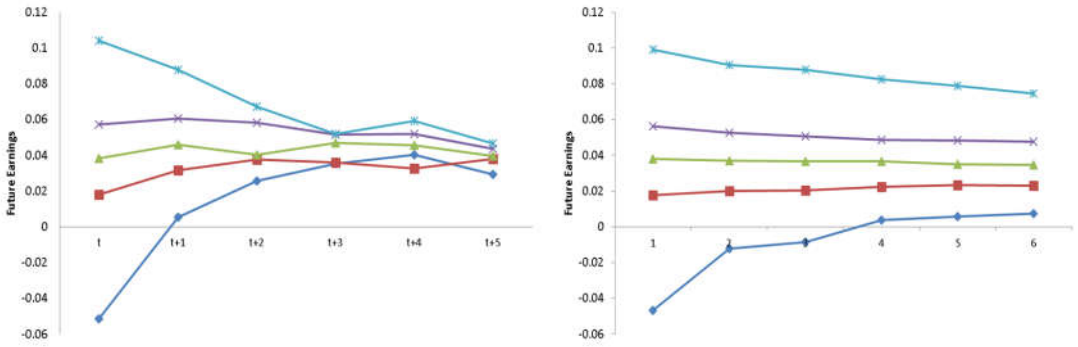


Figure 5.4- Mean Reversion of 5-year Future Earnings. In Fig a, the full sample is sorted into five quintiles by value of current earnings. Then, the graph plots the median current earnings by future earnings for each quintile. Fig b and Fig c are similarly constructed, but replacing the full sample by firms only in *Cluster 1* and *Cluster 8*, respectively.

Figure 5.4 also reveal the potential effects of earnings level on persistence patterns. The range of median deflated earnings between the top and bottom quintiles for firms in *Cluster 1* is 33%, being much higher than the 4.7% observed for firms in Cluster 8. Recall from Figure 5.2 that absolute earnings (*ABS_IBC*) ranks sixth among 17 features in achieving low *MSR* when running K-means using a single feature. Yet ROWK clustering assigns zero weight to *ABS_IBC*. One explanation for this divergence could potentially be explained by the correlation between *ABS_IBC* and the two highest-weighted features, i.e. *VOL_IBC* and *ABS_ACC*.

To support this argument, this thesis takes a further step to control for earnings using the two-pass sorting procedure. Specifically, firms are sorted annually into 20 portfolios based on their current earnings. *Portfolio 1* corresponds to the portfolio of the lowest current earnings, and *Portfolio 20* corresponds to that of the highest current earnings. Next, ROWK clustering is executed within each portfolio. Combining the lowest earnings persistence cluster from *Portfolios 1* to *4* produces Quintile 1 (the lowest earnings) of *Cluster 1*. Combining the lowest earnings persistence cluster from Portfolios 5 to 7 produces Quintile 2 of *Cluster 1,* and so on. The same procedure is repeated to generate quintiles of firms in *Cluster 8*.

Figure 5.5 presents the results from this sorting procedure for firms in *Clusters 1* (Figure 5.5a) and *8* (Figure 5.5b). Given the purpose of the two-pass sorting procedure, consistent with expectations, the ranges of current earnings between the top and bottom quintiles are similar between firms in *Clusters 1* and *8*. The range for firms in *Cluster 1* is 15.5%, only slightly higher than the range for firms in *Cluster 8* (14.6%). The sorting procedure successfully reduces the range of current earnings between *Clusters 1* and *8*. Therefore, differences in the reversion of earnings between *Clusters 1* and *8* now purely result from differences in the characteristics of these two clusters, not from the dispersion in current earnings. Clear evidence of different earnings persistence across graphs is depicted, with firms in *Cluster 1* reverting more quickly. While the range in current median earnings of 15.5% declines dramatically to 1.7% in the five subsequent years for firms in *Cluster 1*, the corresponding percentages are 14.6% and 6.7% for *Cluster 8*. In summary, this evidence provides graphic support for the Hypothesis H6 prediction that there are differences of earnings persistence and intercepts between ROWK clusters, lasting up to five years beyond the cluster formation period.



a. Cluster 1 Controlled By Current Earnings b. Cluster 8 Controlled By Current Earnings

Figure 5.5- Mean Reversion of 5-year Future Earnings Controlling for the Dispersion of Current Earnings. Fig a (b) plots mean reversion of 5-year future earnings of firms in Cluster 1(8) controlling for current earnings by the two-pass sorting procedure. Firms are sorted annually into 20 portfolios based on their current earnings. Then, ROWK clustering is executed within each portfolio. Combining the lowest earnings persistence cluster from Portfolios 1 to 4 results in Quintile 1 (low earnings) of Cluster 1. Combining the lowest earnings persistence cluster from Portfolios 5 to 7 produces Quintile 2 of Cluster 1 and so on (Fig a). The same procedure is repeated to generate quintiles of the firms in Cluster 8 (Fig b)

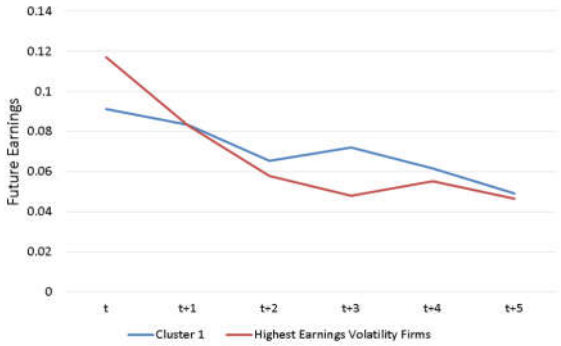
5.3.2.4.2 Cluster 1 vs. the Portfolio of Firms with Highest Earnings Volatility

Earnings volatility is documented as a dominant factor affecting earnings persistence and earnings predictability (Dichev & Tang, 2009). This is consistent with our findings that that ROWK assigns the highest weight to *VOL_IBC*. However, ROWK clustering, being a tool to coordinate relevant information from all features, can identify incremental or superior persistence patterns than those from a single feature (in this case *VOL_IBC*). Panels B and C of Table 5-5 reveal a noticeable difference between firms in *Cluster 1* and firms in the highest earnings volatility portfolio (*OCTILE_8_VOL*). Firms in *Cluster 1* and *OCTILE_8_VOL* are similar in that they display extremely low persistence in earnings. However, while firms in *OCTILE_8_VOL* display an insignificant negative intercept (-0.4%), firms in *Cluster 1* have the highest intercept (2.4%).

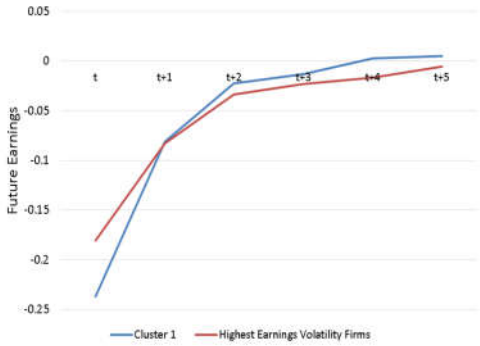
This section provides graphical evidence with respect to this difference. Figure 5.6a plots earnings streams of the top current earnings quintiles of firms in *Cluster 1* and firms in *OCTILE_8_VOL*. In year t, firms in *OCTILE_8_VOL* have 2.57% higher median earnings than those of firms in *Cluster 1*. However, after one year, the positive intercept coefficient

causes firms in *Cluster 1* to display less mean-reversion than those in *OCTILE_8_VOL*. Consequently, these firms have equal median earnings after one year at around 8.3%. Firms in *OCTILE_8_VOL* then have lower median earnings starting from year *t+2*, with the differences of median earnings ranging from -0.8% at year *t+2* to -2.4% in year *t*+3. A similar pattern is also observed for the bottom quintile of earnings as graphed by Figure 5.6b.

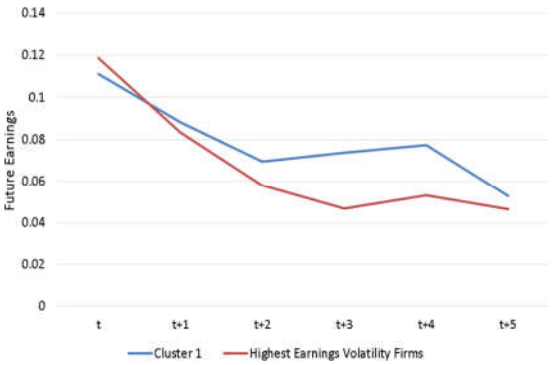
Figure 5.6c and Figure 5.6d display results of Figure 5.6a and Figure 5.6b using the previously discussed two-pass sorting procedure to control for dispersion of current earnings. The results remain unchanged. These findings provide supporting evidence that ROWK clustering, which utilises relevant information from all features, is able to identify incremental or superior persistence patterns relative to methods using any single feature.



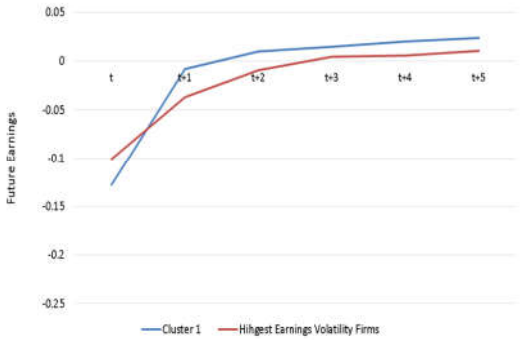
a) Top Quintile of Current Earnings



b) Bottom Quintile of Current Earnings



c) Controlled by Current Earnings-
Top Quintile of Current Earnings



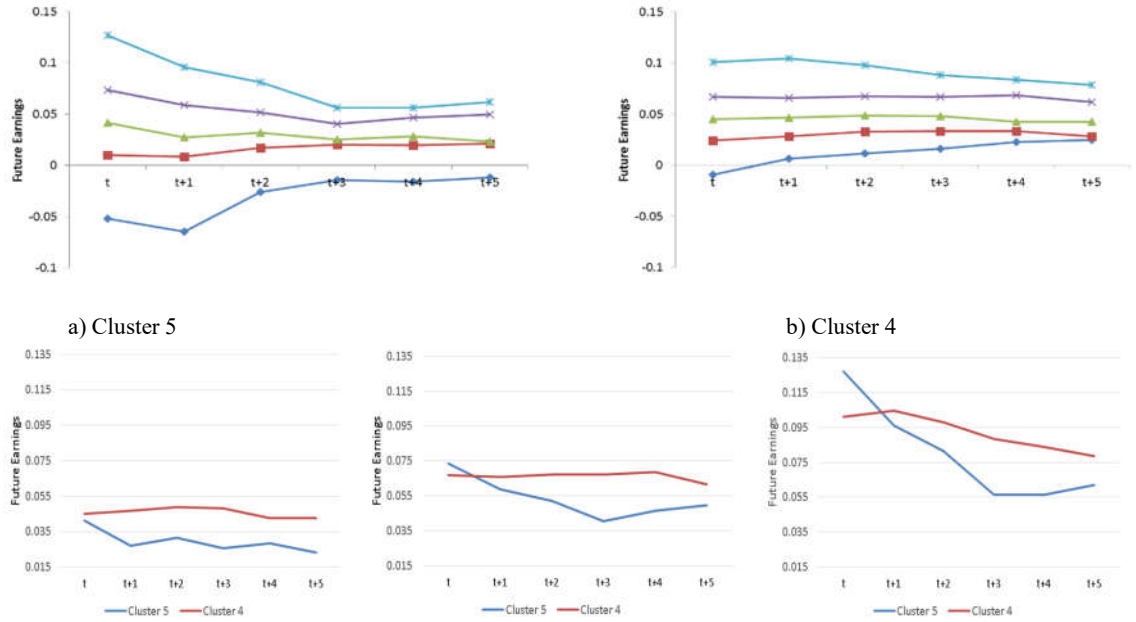
d) Controlled by Current Earnings-
Bottom Quintile of Current Earnings

Figure 5.6- Mean Reversion of 5-year Future Earnings- Cluster 1 vs. Highest Earnings Volatility Firms. Fig a (b) plots mean reversion of 5-year future earnings of the top (bottom) current earnings quintile of Cluster 1 firms and firms in the highest earnings volatility octiles. Figs c and d replicate Figs a and b respectively with a two-pass sorting procedure to control for dispersion of current earnings. Firms are sorted annually into 20 portfolios based on their current earnings. Then, ROWK clustering is executed within each portfolio. Combining the lowest earnings persistence cluster from Portfolios 1 to 4 (17 to 20) results in the bottom (top) earnings quintile of Cluster 1. For the highest earnings volatility firms, for each of 20 portfolios, firms are further divided into octiles of earnings volatility. Combining the highest earnings volatility octiles from Portfolios 1 to 4 (17 to 20) produces the bottom (top) earnings quintile of the highest earnings volatility firms.

5.3.2.4.3 Cluster 4 vs. Cluster 5

Table 5-7 not only reveals a difference in earnings persistence between ROWK's clusters, but also exhibits heterogeneities of intercepts across ROWK clusters. Firms in *Clusters 4* and *5* exhibit similar persistence, but they significantly differ with respect to the intercept coefficients, μ. While firms in *Cluster 4* have intercept coefficients of 1%, the corresponding percentage for firms in *Cluster 5* is -2.6%. Figure 5.7 displays this difference graphically. Figure 5.7a and Figure 5.7b plot mean reversion of 5-year future earnings of firms in these clusters. It is clear that even firms in these clusters share the same persistence, with firms in *Cluster 5* converging more quickly for the first three years.

Graphs Figure **5.7**c, Figure **5.7**d and Figure **5.7**e directly compare earnings patterns between firms in these clusters within the third, fourth and fifth earnings octiles. The first and second octiles are not presented as their current earnings are dissimilar. For all of these graphs, it is apparent that firms in *Cluster 4* have stable earnings while firms in *Cluster 5* experience strong declines in earnings up to next three years. The ranges of median earnings between firms in *Cluster 4* and *Cluster 5* in year *t* are 0.38%, -0.67% and -2.59% at the third, fourth and fifth earnings octiles, respectively. These ranges increase dramatically and reach a maximum in year *t+3* with the corresponding values of 2.5%, 2.64% and 3.17% at the third, fourth and fifth earnings octiles, respectively. In summary, these results provide compelling evidence of consistently different earnings patterns between firms in *Cluster 4* and *Cluster 5*, even when these firms have similar earnings persistence.



a) Cluster 5

b) Cluster 4

c) Cluster 4 vs. Cluster 5- Octile 3

d) Cluster 4 vs. Cluster 5- Octile 4

e) Cluster 4 vs. Cluster 5- Octile 5

Figure 5.7- Mean Reversion of 5-year Future Earnings- Cluster 4 vs. Cluster 5. Fig a (b) plots mean reversion of 5-year future earnings of firms in *Cluster 4* and *Cluster 5*. Figs c, d and e plot comparable earnings quintiles, i.e. the third, fourth and fifth quintiles respectively, between firms in *Cluster 4* and *Cluster 5.*

5.3.3 Cluster Description

5.3.3.1 Cluster Characteristics

In the previous section, the names of clusters were assigned by order of earnings persistence. This section takes a closer look at the characteristics of each cluster found by ROWK clustering. The results are presented in Table 5-8 that displays medians of the original features, organised in columns according to earnings persistence[64]. The far left column of the table describes *Cluster 8*, which includes the highest persistence firms, and the far right column describes *Cluster 1*, the lowest persistence cluster.

[64] The use of median is to address the issue of non-normal distributions and high skewness of original features. However, the results are unchanged when means are used.

The first finding from Table 5-8 reports the significance of the Kruskal-Wallis test. All 17 examined features have less than 0.0001 p-value, rejecting the null hypothesis that median values of features are equal across ROWK clusters. This fact demonstrates that when a feature stands alone, it can be useful to distinguish clusters. However, when all features are used, ROWK clustering can identify those features that are relevant and those that are not. The mechanism underlying this is the second channel of ROWK as discussed in Chapter 4. Particularly, ROWK is designed to address the problem of high correlation between features. Accordingly, among 17 features, only 5 features jointly contribute to clustering. The other irrelevant 12 features, which are correlated with these 5 features, receive zero weights by ROWK clustering.

Table 5-9 is a transformation of Table 5-8 and demonstrates distinguishing characteristics of each cluster. As expected, two outstanding clusters, i.e. *Clusters 1* and *8*, have special characteristics. C*luster* 8 comprises firms with the highest earnings persistence. These firms have extremely low *VOL_IBC* (0.4%), asset efficiency (*ATO* is only at 0.674), and financial flexibility (low *CR* and high *FLEV*). Firms in these clusters are large and well-established (*SIZE* and *AGE* are $2.2 billion and 36, respectively).

In contrast, firms in *Cluster 1* have extremely high *VOL_IBC* and *ABS_ACC_DEF*. Looking more closely at accruals, the high value of *ABS_ACC_DEF* comes from the low value of accruals (*ACC_DEF* is -16.6%)[65]. These firms are small, young, and invest less in intangible assets. An important feature of these firms is that they experience poor operating performance with very low *PM* (3.7%) and low *ATO* (1.29). However, positive annual *ΔPM* (0.7%) and *ΔATO* (0.15) indicate that these firms are progressing. This explains why these firms have low earnings persistence and high intercepts. The characteristics of firms in *Cluster 8* behave like mature firms, while firms in *Cluster 1* have features of firms in the life cycle stages of introduction and growth (Dickinson, 2011). This will be explored in the next section.

[65] Recall that ABS_ACC denotes the absolute value of accruals. The high value of ABS_ACC could fall into the case of extremely high value of accruals or extremely low value of accruals.

Other clusters also display unique patterns. Firms in *Cluster 7*, which have persistence as high as firms in *Cluster 8* but with low intercepts, have the highest *ACC_DEF*. Firms in *Cluster 6* display middle ranges of all features. Firms in *Cluster 2*, which have the second lowest earnings persistence, experience large reductions in *PM* and *ATO* (*ΔPM* is -1.9% and *ΔATO* is -0.27). They have high *INTAN_INV_DEF* (3.2%) and high *VOL_IBC* (5.6%). In contrast, firms in Cluster 3 have outstanding performance with high *SALE_GR* (10%) and strong improvement in asset turnover (*ΔATO* = 0.21) and profit margin (*ΔPM*= 1.8%).

An important finding comes from comparing firms in *Cluster 5* and *Cluster 4*, both of which have the same earnings persistence (around 0.8) but different intercepts (-2.6% vs. 1% respectively). These firms differ in several aspects. While firms in *Cluster 5* have high *VOL_IBC* (std is 6.8%) and high *ACC_DEF* (-1.7%), firms in Cluster 4 have stable earnings streams (std of *VOL_IBC* is 1.7%) and low *ACC_DEF* (-9.5%). Firms in *Cluster 5* are young and small, while firms in *Cluster 4* are in the mid-range of age and size. In the forecast analyst section, we will examine whether analysts incorporate into their forecasts the difference in intercepts of firms in *Clusters 4* and *5*. In the same manner, we also test whether analyst forecasts distinguish the difference in earnings persistence between firms in *Clusters 1* and *8.*

Table 5-8: Descriptions of ROWK's Clusters- Median of Original Features

	Clus. 8	Clus. 7	Clus. 6	Clus. 5	Clus. 4	Clus. 3	Clus. 2	Clus. 1	K-W Test
$\mu_{\zeta_k^0}$	-0.001	-0.012	-0.002	-0.026	0.010	0.003	-0.004	0.024	
$\theta_{\zeta_k^0}$	0.961	0.96	0.868	0.834	0.831	0.817	0.62	0.494	
Adj. R^2	0.563	0.413	0.433	0.339	0.443	0.392	0.29	0.283	
N	1351	1129	1755	851	1704	1147	975	882	
PM	0.178	0.098	0.101	0.085	0.112	0.078	0.069	0.037	<.0001
ΔPM	-0.001	0.001	-0.001	0.006	0.002	0.018	-0.019	0.007	<.0001
ATO	0.674	1.476	1.329	1.467	1.347	1.745	1.349	1.295	<.0001
ΔATO	0.007	-0.001	-0.025	-0.068	0.039	0.213	-0.27	0.148	<.0001
VOL_IBC	0.004	0.016	0.019	0.068	0.017	0.046	0.056	0.079	<.0001
CR	0.973	1.644	1.547	2.007	1.271	1.662	1.771	1.49	<.0001
CAPX_DEF	0.053	0.042	0.049	0.041	0.078	0.044	0.058	0.054	<.0001
INTAN_INV_DEF	0	0.009	0.011	0.044	0.005	0.01	0.032	0.001	<.0001
FLEV	0.529	0.367	0.414	0.241	0.41	0.386	0.255	0.374	<.0001
OLLEV	0.3	0.294	0.292	0.284	0.287	0.3	0.255	0.266	<.0001
SALE_GR	0.022	0.062	0.037	0.054	0.057	0.099	0.013	0.049	<.0001
NBC	0.941	0.989	0.986	0.997	0.989	1.02	0.99	1.014	<.0001
DIV	1.468	0.972	1.075	0.779	1.045	0.947	1.186	2.804	<.0001
AB_ACC_DEF	0.047	0.013	0.045	0.022	0.097	0.059	0.091	0.166	<.0001
SIZE	2297	930	1433	396	1253	735	486	476	<.0001
AGE	36	24	25	13	20	16	12	12	<.0001
ABS_IBC_DEF	0.039	0.045	0.042	0.056	0.047	0.052	0.055	0.091	<.0001
ACC_DEF	-0.046	-0.008	-0.043	-0.017	-0.095	-0.057	-0.086	-0.166	<.0001

PM, ΔPM denote profit margin and the change in profit margin; ATO, ΔATO represent asset turnover and the change in asset turnover; VOL_IBC, VOL_OCF denote volatility of earnings and volatility of operating cash flows measured as standard deviation of the most recent five years; CR is current ratio; CAPX_DEF and INTAN_ INV_DEF are measures of deflated capital expenditure and deflated investment in intangible assets; FLEV and OLLEV denote financial leverage and operating leverage; SALE_GR is sales growth; NBC is net borrowing cost; ACC_DEF and AB_ACC_DEF denote deflated accruals and absolute value of deflated accruals; DIV, SIZE and AGE represent dividend payout, log of total assets and firm age respectively; IBC_DEF and ABS_IBC_DEF denote deflated earnings and absolute value of deflated earnings; OCF is operating cash flows. See Section 2.5.2 for details of formulae for the variables. $\mu_{\zeta_k^0}$, $\theta_{\zeta_k^0}$ and Adj. R^2 denote estimations intercept, earnings persistence and adjusted R^2 derived from running the earnings persistence regression (Equation 3.18) within each ROWK cluster. K-W Test is the p-value of the Kruskal-Wallis test.

Table 5-9: Characteristics of ROWK's Clusters

	Clus. 8	Clus. 7	Clus. 6	Clus. 5	Clus. 4	Clus. 3	Clus. 2	Clus. 1
$\mu_{\zeta_k^0}$	-	L	-	L	H	-	-	HG
$\theta_{\zeta_k^0}$	HG	H	-	-	-	-	L	LW
N	1351	1129	1755	851	1704	1147	975	882
PM	HG	-	-	-	-	-	-	LW
ΔPM	-	-	-	-	-	HG	LW	-
ATO	LW	-	-	-	-	H	-	-
ΔATO	-	-	-	L	-	HG	LW	H
VOL_IBC	LW	-	-	H	L	-	H	HG
CR	LW	-	-	HG	-	-	-	-
CAPX_DEF	-	-	-	-	H	-	-	-
INTAN_INV_DEF	LW	-	-	HG			H	L
FLEV	HG	-	-	LW	-	-	-	-
OLLEV	-	-	-	-	-	-	-	-
SALE_GR	-	-	-	-	-	HG	L	-
NBC	LW	-	-	-	-	-	-	-
DIV	-	-	-	LW	-	-	-	HG
AB_ACC_DEF	-	LW	-	L	H		H	HG
SIZE	HG	-	-	L	-	-	-	L
AGE	HG	-	-	L	-	-	-	L
ABS_IBC_DEF	-	-	-	-	-	-	-	-
ACC_DEF	-	HG	-	H	L	-	L	LW

PM, ΔPM denote profit margin and the change in profit margin; ATO, ΔATO represent asset turnover and the change in asset turnover; VOL_IBC, VOL_OCF denote volatility of earnings and volatility of operating cash flows measured as
standard deviation of the most recent five years; CR is current ratio; CAPX_DEF and INTAN_ INV_DEF are measures of deflated capital expenditure and deflated investment in intangible assets; FLEV and OLLEV denote financial leverage and operating leverage; SALE_GR is sales growth; NBC is net borrowing cost; ACC_DEF and AB_ACC_DEF denote deflated accruals and absolute value of deflated accruals; DIV, SIZE and AGE represent dividend payout, log of total assets and firm age respectively; IBC_DEF and ABS_IBC_DEF denote deflated earnings and absolute value of deflated earnings; OCF is operating cash flows. See Section 2.5.2 for details of formulae for the variables. $\mu_{\zeta_k^0}$, $\theta_{\zeta_k^0}$ and Adj. R^2 denote estimations intercept, earnings persistence and adjusted R^2 derived from running the earnings persistence regression (Equation 3.18) within each ROWK cluster. L (H) indicates low (high value); LW (HG) indicates the lowest (highest) value; '-' indicates middle values.

Figure 5.8 graphs ROWK cluster memberships by the first two canonical variables (which account for 93.5% of total eigenvalues) derived from CDA with five non-zero-weighted features and cluster membership identified via ROWK clustering. Each circle represents a cluster. The centre of a circle is indicated by the corresponding cluster centroid. The radius of a circle indicates the mean distance between cluster members and the cluster centroid. As expected, *Cluster* 1's radius is the biggest while *Cluster 8*'s circle is the smallest. Members of these two clusters are also far apart in distance. It is also observed that the order of circles based on the first canonical variable (*Can 1*) is in line

with earnings persistence. In contrast, the differences in intercept are explained more by the second canonical variable (*Can 2*). Results (unreported) show that *VOL_IBC* and *ABS_ACC_DEF* are the main components of these two canonical variables, consistent with the highest weights that ROWK clustering assigns to these features.

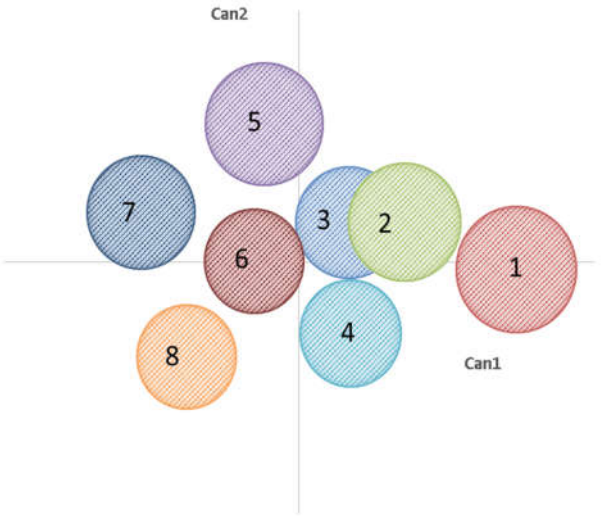


Figure 5.8- Cluster Location by the First Two Canonical Variables Derived from CDA with Cluster Membership Identified via ROWK Clustering

5.3.3.2 Survival Rate and Transition Analysis

5.3.3.2.1 Transition Analysis

Panel A of Table 5-10 examines the transition of firm-observations from a cluster to another in five subsequent years. Blue-bolded values denote the highest proportions, while the red-bolded values indicate the lowest proportions. Several noteworthy findings are observed from this analysis. First, firms in *Cluster 8* are quite stable with nearly 60% of them remaining in *Cluster 8* over 5 years, while firms in other clusters, specifically *Cluster 1* (17.57%), *Cluster 2* (12.91%), *Cluster 3* (15.46%) and *Cluster 5* (12.73%) exhibit a high tendency to move to other clusters. Second, firms in clusters with low earnings persistence (*Cluster 1, 2* and *3*) are more likely to continue to stay in low earnings persistence clusters. In the same manner, firms in clusters with high earnings persistence (*Cluster 6, 7* and *8*) tend to remain in high earnings persistence clusters even

after 5-years' time. For example, 83% firms in *Cluster 8* remain in *Clusters 6, 7*, and *8* over 5 years. Finally, though they share the same earnings persistence (0.83), transition patterns of firms in *Cluster 5* are similar to those of lower earnings persistence clusters, while firms in *Cluster 4* display similar transition patterns with those of higher earnings persistence clusters.

5.3.3.2.2 Survivorship

Survivorship of firms across clusters is examined due to two reasons. First, given the distinguishable patterns of earnings and cluster characteristics, firms across clusters should exhibit different rates of survival. Examining survivorship of firms across ROWK clusters sheds more light on whether and to what extent firms across clusters exhibit differences in survival rates. Second, survivorship bias is an important issue in inter-temporal analyses (Dickinson, 2011). Understanding firm survival rates helps to measure the impact of survivorship bias on previous results of inter-temporal analyses.

Survivorship analysis is presented in Panel B of Table 5-10. Almost 78% of firms survive five years ahead in the pooled sample. Acquisition and mergers, transition to private ownership, or bankruptcy are possible reasons for sample attrition. The last two columns of Panel B show the proportion of firms within a cluster that delist due to merger or performance-related reasons in the year prior to delisting. For example, delisting of firms in *Cluster 1* are due to mergers and acquisitions (76.55%) and performance (6.4%). P-values report *Z*-statistics (two-tailed) on tests of whether delisting proportions between firms in *Clusters 1* and *8* significantly differ.

As expected, firms with lower persistence have lower rates of survival. Five years after the formation period, nearly one third of firms in *Cluster 1* delisted. In contrast, the corresponding fraction of *Cluster 8* firms that delisted is less than one sixth. The differences in survival rates between firms in these clusters are significant at the 5% level from year *t+3*. Despite sharing the same persistence, firms in *Cluster 5* have slightly lower survival rates relative to those of firms in *Cluster 4*. This is to be expected since earnings trends for firms in *Cluster 5* are downward at an annual rate of -2.6%, while they are upward at an annual rate of 1% for firms in *Cluster 4*.

Regarding the reasons for delisting, there is no clear pattern for firms delisted due to mergers and acquisitions, which account for 62% of firms in *Cluster 8* to 78% of firms in *Cluster 5*. Unsurprisingly, firms in *Cluster 5* have the highest proportion of acquisition and mergers since they are small, have high earnings volatility and experience reductions in efficiency of assets management. Consequently, they have the lowest intercepts (-2.6%). Performance-related delistings reveal more obvious patterns. Firms with lower earnings persistence tend to have a higher proportion of delisting because of performance related reasons (e.g. 6.4% for firms in *Cluster 1* vs. 2.7% for firms in *Cluster 8*).

Table 5-10: Transition Analysis and Survival Rates of ROWKs' Cluster Membership (%)

Panel A: Evolution of Firms Across ROWKs' Cluster

Formation Time	Future Period	t+1	t+2	t+3	t+4	t+5	Formation Time	Future Period	t+1	t+2	t+3	t+4	t+5
C. 1	C. 1	**32.42**	**25.31**	**20.8**	**20.71**	17.57	C. 2	C. 1	18	17.08	15.78	13.21	11.26
	C. 2	15.97	20.92	20.8	18.45	16.32		C. 2	**25.2**	16.26	14.75	15.28	12.91
	C. 3	26.45	20.92	18.67	12.94	11.72		C. 3	21.07	**25.78**	**21.31**	**19.95**	18.87
	C. 4	5.81	9.41	12.53	14.56	**20.5**		C. 4	8.67	14.45	14.55	17.36	**19.21**
	C. 5	15.81	16.95	18.13	20.06	10.88		C. 5	15.47	12.32	11.27	10.88	9.6
	C. 6	2.1	3.97	4.8	7.12	15.48		C. 6	6.8	7.39	13.52	14.25	15.23
	C. 7	1.45	2.3	4	5.18	5.86		C. 7	4.67	6.4	7.58	6.74	9.93
	C. 8	**0**	**0.21**	**0.27**	**0.97**	**1.67**		C. 8	**0.13**	**0.33**	**1.23**	**2.33**	**2.98**
C. 3	C. 1	9.92	8.54	6.94	7.23	7.99	C. 4	C. 1	3.86	4.97	5.74	5.56	6.3
	C. 2	13.52	15.13	14.04	9.3	12.37		C. 2	3	4.88	5.85	5.93	6.14
	C. 3	**28.68**	17.79	17.77	16.32	15.46		C. 3	6	5.31	6.05	7.2	7.53
	C. 4	8.94	12.61	16.07	19.42	14.43		C. 4	**49.46**	**42.33**	**39.69**	**38.13**	**33.03**
	C. 5	16.14	13.73	11.34	6.82	7.99		C. 5	**0.57**	**1.89**	**2.26**	**2.27**	**2.76**
	C. 6	14.29	**19.75**	**19.29**	**23.35**	**21.91**		C. 6	19.66	21.85	19.49	19.07	19.05
	C. 7	7.85	10.78	12.01	12.6	14.18		C. 7	9.86	9.34	9.03	8.84	10.14
	C. 8	**0.65**	**1.68**	**2.54**	**4.96**	**5.67**		C. 8	7.58	9.43	11.9	13.01	15.05
C. 5	C. 1	10.76	12.23	11.78	10.21	8.99	C. 6	C. 1	2.6	3.17	4.18	3.83	4.72
	C. 2	22.46	15.58	12.74	11.41	11.24		C. 2	3.63	**2.42**	3.09	3.96	**3.81**
	C. 3	17.63	20.71	**19.71**	**20.42**	**20.22**		C. 3	5.55	7.35	7.17	7.29	7.46
	C. 4	4.52	6.71	8.65	8.71	8.61		C. 4	21.23	20.05	20.42	19.28	16.89
	C. 5	**29.49**	**21.89**	16.83	11.11	12.73		C. 5	**2.12**	3.01	**2.59**	**3.83**	4.57
	C. 6	9.67	14.4	18.27	19.52	22.47		C. 6	**40.34**	**34.42**	**30.38**	**24.97**	**24.96**
	C. 7	5.15	8.48	11.3	15.92	11.99		C. 7	16.85	15.87	15.14	16.19	15.37
	C. 8	**0.31**	**0**	**0.72**	**2.7**	**3.75**		C. 8	7.67	13.7	17.03	20.64	22.22
C. 7	C. 1	4.39	3.66	3.73	5.2	3.74	C. 8	C. 1	0.34	0.6	0.8	**0.67**	**0.79**
	C. 2	3.32	3.14	**3.73**	**5.2**	4.74		C. 2	0.34	0.79	0.69	1.07	**0.79**
	C. 3	4.07	6.54	6.97	7.51	6.23		C. 3	0.09	0.99	1.14	1.2	1.75
	C. 4	12.1	13.73	13.45	13.68	16.71		C. 4	10.74	12.1	12.36	12.27	12.72
	C. 5	**1.82**	**3.14**	4.86	5.59	**4.49**		C. 5	**0**	**0.1**	**0.57**	0.8	0.95
	C. 6	30.62	**27.32**	**26.26**	**25.05**	**26.68**		C. 6	5.76	8.23	11.67	14.8	14.79
	C. 7	**30.51**	26.67	24.47	23.12	22.94		C. 7	9.11	10.62	10.53	9.73	11.13
	C. 8	13.17	15.82	16.53	14.64	14.46		C. 8	**73.63**	**66.57**	**62.24**	**59.47**	**57.07**

Table 5-10 (cont)
Panel B: Firms Survival Rates next Five Year after Formation Period

Formation Stage		Survival Rate (%)					Delisted Merger (%) (row)	Performance Related (%) (row)	Delisted Merger (%) (column)	Performance Related (%) (column)
Cluster	N (t=0)	t+1	t+2	t+3	t+4	t+5				
Pooled	9794	100	94.09	88.22	82.82	77.89	74.25	4.56		
C. 1	882	100	90.82	83.11	75.40	68.82	76.55	6.40	12.56	17.1
C. 2	975	100	92.72	85.85	79.08	73.44	75.60	5.60	12.02	14.51
C. 3	1147	100	94.42	86.57	81.08	74.98	75.00	6.52	13.16	18.65
C. 4	1704	100	94.31	88.73	83.63	78.58	73.00	3.29	16.24	11.92
C. 5	851	100	92.36	85.78	79.44	73.33	78.60	4.19	10.74	9.33
C. 6	1755	100	94.99	90.20	85.98	81.99	76.89	4.31	15.86	14.51
C. 7	1129	100	94.15	88.75	82.99	78.83	74.01	3.33	11.32	8.29
C. 8	1351	100	96.52	92.52	88.68	85.42	62.35	2.69	8.11	5.7
Differences (C1-C8)	*0*		*-5.7*	*-9.41*	*-13.28*	*-16.6*	*14.2*	*3.71*		
p-value of differences	*(0.50)*		*(0.04)*	*(0.011)*	*(<0.01)*	*(<0.001)*	*(<0.001)*	*(<0.01)*		
Differences (C4-C5)	*0*		*1.95*	*2.95*	*4.19*	*5.25*	*-5.6*	*-0.9*		
p-value of differences	*(0.50)*		*(0.21)*	*(0.09)*	*(0.031)*	*(0.017)*	*(0.0013)*	*(0.044)*		

All firms are grouped into 8 clusters after running ROWK clustering. Then in Panel A, the cluster transition of firms for the next five years is tracked. Blue-bolded values denote the highest proportions, while the red-bolded ones indicate the lowest proportions. In Panel B, delisting data is extracted from the Compustat database. 'Delisted Merger' or 'Performance Related' indicates firms that delist due to mergers (Compustat code, i.e. dlrsncd, as 01) or firm performance (Compustat code, i.e. dlrsncd, as 02 and 03). 'Differences (C1-C8)' and 'Differences (C1-C8)' denote differences of survival rates between firms in Clusters 1 and 8 and between firms in Clusters 4 and 5, respectively. 'p-value of differences' denotes p-values of the corresponding differences, calculated based on the Z-statistics from a test of equal survival rates between firms in these clusters. Bold numbers indicate a significant difference in survival rates at the 0.05 significance level (two-tailed).

5.3.4 Heterogeneities of Industry Classification, Firm Life Cycles, Earning Management and Conservatism across Clusters

Panels A, B, and C of Table 5-11 present the frequency of ROWK cluster membership by industry classification, stages of firm life cycle and earnings management, respectively. Panel D shows means and medians of proxies for conditional and unconditional accounting conservatism[66]. For all panels, the test statistics (*Chi-Square* statistics in Panels A, B and C, and *Kruskal-Wallis* statistics in Panel D) show that between ROWK clusters, firms exhibit significant heterogeneities in industry affiliation, stage of firm life cycle, earnings management, and magnitude of accounting report conservatism. This is consistent with the Hypothesis H8 proposal that clusters found in ROWK clustering exhibit heterogeneities with respect to the above attributes.

Particularly, 38% of firms with the lowest earning persistence (*Cluster 1*) are in the Energy and Business Equipment industries, while 1.25% of firms in *Cluster 1* belong to the Utility industry. In contrast, nearly two thirds of firms in *Cluster 8* are in the Utility industry. This is consistent with the findings of Dichev & Tang (2009) who observe that 54% firms in the top quintiles of earnings volatility are in the Utility industry[67]. Firms in *Cluster 3* are predominantly (21%) in the Business Equipment industry. In other clusters, there is no clear pattern, except that the majority of firms in the Energy (31%) and Telecommunication (40%) industries are members of *Cluster 4*.

Regarding stage of firm life cycle, as can be seen from Panel B of Table 5-11, over 80% of firms are in stages of growth and maturity. This is as expected since the thesis follows the work of Dickinson (2011) to identify stages of firm life cycle. In her study, 75% of firms in the sample covering the 1989-2005 period are growth or mature firms. Several notable results are found. Firms in *Cluster 8* are more likely to be growth and mature firms characterised with stable earnings streams, large size and older. In contrast, the proportions of firms in *Cluster 1* that fall within the stages of introduction

[66] Refer to Chapter 3, Section 3.2.2.3.1 for details of the industry classification scheme, identification of firm life cycle stages, recognition of earnings management, and measurement of degrees of accounting conservatism.

[67] As discussed in the previous section, firms in *Cluster 8* have the lowest earnings volatility, consequently there is much overlap between firms in *Cluster 8* and the lowest earnings volatility.

(9%), decline (4.8%) and shakeout (10%) are significantly higher than the all-cluster averages with the corresponding numbers as 5.6%, 2.5% and 7.5% respectively.

Firms in *Clusters 4* and *5* also have distinguishing life cycle patterns. While firms in *Cluster 4* tend to be in the mature stage (63%) and have low likelihood to experience the stages of decline (1.3%) and shakeout (3.9%), high proportions of firms in *Cluster 5* are struggling in the stages of Introduction (12.35%), Decline (6%) and Shakeout (11.53%). This may explain the observation that firms in *Clusters 4* and *5* are similar in earnings persistence (0.83), but dissimilar in intercepts, with *Cluster 4* having a positive intercept at 1% and *Cluster 5* having the lowest negative intercept at -2.6%.

Patterns of earnings management across ROWK clusters are displayed in Panel C. Firms are identified as downward earnings managed (*EM_DN*), upward earnings managed (*EM_UP*) and without earnings managed (*NO_EM*) using the method in a study by Jansen et al. (2012). Seventeen percent of firm-year observations in the sample upwardly manage earnings. The corresponding number for downward earnings management is similar at 18%. These numbers are in line with those reported by Jansen et al. (2012).

Given their poor performance, firms in *Cluster 1* are expected to manage earnings upward to beat analyst/market expectations. It turns out that over 25% of firms in *Cluster 1* manage their earnings downward, but fewer than 6% intentionally increase their earnings. At first glance, this seems surprising. However, firms that widely miss or fail to beat analysts' forecasts tend to manage earnings downward (Cao, Shaari, & Donnelly, 2018; Jansen et al., 2012). These firms have an incentive to "take a bath" and then exploit the reversal in the future. This is consistent with our evidence of a strong tendency for firms in *Cluster 1* to manage their earnings downward. Firms in *Cluster 3*, if managing earnings, also tend to manage earnings downward. However, firms in *Cluster 2* and *Cluster 5*, if managing earnings, tend to manage their earnings upward. The proportion of firm-year observations in *Clusters 2* and *5* that manage earnings upward is 8 and 4 times relative to those that manage earnings downward, respectively. Other clusters do not exhibit much difference in their tendencies to favour upward or downward earnings management.

The last panel presents the results of degrees of accounting conservatism across ROWK clusters. *U_CON (C_CON)* denotes unconditional (conditional) accounting conservatism measured following the work of Chen et al., (2014). It is observed that both the mean and median of *U_CON* monotonically reduce *from Cluster 1* to *Cluster 8*. In the case of conditional accounting conservatism, firms in *Cluster 1* still display the highest level of conservatism. This evidence provides an additional explanation of the tendency for high mean-reversion for firms in *Cluster 1*, since firms with high levels of both conditional and unconditional accounting conservatism tend to have lower earnings persistence (Chen et al., 2014; Penman & Zhang, 2002).

In summary, these findings support Hypothesis H8 predictions that the clusters found in ROWK clustering exhibit heterogeneities with respect to accounting conservatism, earnings management, firm life cycles and industry membership. In addition, it upholds the thesis's argument in favour of using financial ratios as the input list for ROWK clustering rather than directly incorporating direct proxies that contain measurement errors. In additional tests, the earnings persistence model is run for each industry (*REG_IND*), firm life cycle stage (*REG_LC*), category of earnings management (*REG_E*) and decile of *U_CON* and *C_CON*. The results, which are not reported but are available upon request, show that the *MSRs* of these cases are significantly larger compared to those demonstrated using ROWK clusters[68].

[68] We also examined whether or not industry has an incremental effect on the earnings persistence model when industry is controlled for using ROWK clustering. To do that, we added industry dummy variables to the earnings persistence model, and then ran the earnings persistence regression within each cluster. The results, which are not reported but are available upon request, show that the coefficients of the industry dummies become insignificant after controlling for ROWK clusters.

Table 5-11: Frequency of ROWK's Clusters by Industry

Panel A: Industry									
Industry Row Percentage % *Column Percentage %*	Cluster								
	1	2	3	4	5	6	7	8	Total
1.Consumer Non-Durables	5.16 *3.74*	8.29 *5.44*	12.21 *6.8*	18.31 *6.87*	7.04 *5.29*	23.63 *8.6*	17.21 *9.74*	8.14 *3.85*	6.52
2.Consumer Durables	5.41 *2.15*	10.54 *3.79*	16.52 *5.06*	14.81 *3.05*	11.4 *4.7*	21.65 *4.33*	14.25 *4.43*	5.41 *1.41*	3.58
3.Manufacturing	5.98 *10.54*	10.8 *17.23*	16.13 *21.88*	13.43 *12.27*	9.58 *17.51*	24.55 *21.77*	14.01 *19.31*	5.53 *6.37*	15.89
4. Energy	19.32 *16.67*	15.51 *12.1*	13.14 *8.72*	31.67 *14.14*	4.86 *4.35*	10.91 *4.73*	3.29 *2.21*	1.31 *0.74*	7.77
5.Chemicals	8.66 *3.51*	8.1 *2.97*	14.25 *4.45*	18.72 *3.93*	12.01 *5.05*	24.02 *4.9*	11.17 *3.54*	3.07 *0.81*	3.66
6.Business Equipment	20.11 *21.32*	21.93 *21.03*	14.55 *11.86*	8.66 *4.75*	15.94 *17.51*	9.84 *5.24*	7.49 *6.2*	1.5 *1.04*	9.55
7.Telecommunications	13.1 *8.73*	7.65 *4.62*	10.54 *5.41*	40.48 *13.97*	5.44 *3.76*	12.93 *4.33*	4.59 *2.39*	5.27 *2.29*	6
8. Utilities	0.74 *1.25*	0.8 *1.23*	1.61 *2.09*	9.17 *8.04*	0.87 *1.53*	16.93 *14.42*	11.78 *15.59*	58.1 *64.25*	15.25
9. Retails and Wholesales	7.44 *6.35*	10.36 *8*	8.9 *5.84*	18.86 *8.33*	6.11 *5.41*	19.79 *8.49*	18.59 *12.4*	9.96 *5.55*	7.69
10.Healthcare	8.19 *5.9*	13.86 *9.03*	14.17 *7.85*	9.29 *3.46*	22.2 *16.57*	16.38 *5.93*	12.44 *7*	3.46 *1.63*	6.48
12.Other	10.15 *19.84*	8.24 *14.56*	13.34 *20.05*	20.94 *21.19*	9.05 *18.33*	17.58 *17.26*	11.25 *17.18*	9.45 *12.07*	17.6
Total_% (N=9794)	9.01	9.96	11.71	17.4	8.69	17.92	11.53	13.79	100
	DF	Value	Prob						
Chi-Square	70	4460	<.0001						

Panel B: Firm Life Cycle									
Life Cycle Stages Row Percentage % *Column Percentage %*	Cluster								
	1	2	3	4	5	6	7	8	Total
Introduction	14.21 *8.9*	19.85 *11.21*	15.3 *7.34*	9.47 *3.05*	19.13 *12.35*	11.84 *3.71*	8.74 *4.28*	1.46 *0.59*	5.62
Growth	8.28 *32.31*	12.35 *43.42*	9.63 *28.76*	14.46 *29.01*	9.95 *40*	18.44 *35.94*	14.93 *45.45*	11.97 *30.36*	34.99
Mature	7.98 *43.95*	6.96 *34.57*	11.81 *49.83*	22.16 *62.77*	5.31 *30.12*	18.69 *51.45*	8.85 *38.06*	18.24 *65.33*	49.4
Decline	17 *4.79*	15.38 *3.91*	14.57 *3.15*	8.91 *1.29*	20.65 *6*	11.34 *1.6*	9.31 *2.05*	2.83 *0.52*	2.53
Shakeout	12.07 *10.05*	9.19 *6.89*	17.15 *10.93*	9.05 *3.88*	13.44 *11.53*	17.56 *7.3*	15.64 *10.16*	5.9 *3.19*	7.46
Total_% (N=9767)	9.01	9.96	11.71	17.4	8.69	17.92	11.53	13.79	100
	DF	Value	Prob						
Chi-Square	28	841	<.0001						

Table 5-11 (cont.) Panel C: Earnings Management									
Earnings Management Row Percentage % *Column Percentage %*	Cluster								
	1	2	3	4	5	6	7	8	Total %
EM_DN	12.73 *25.4*	1.88 *3.38*	12.91 *19.79*	21.03 *21.71*	3.75 *7.76*	15.41 *15.44*	11.09 *17.27*	21.21 *27.61*	17.96
EM_UP	3.07 *5.78*	16.38 *27.9*	1.87 *2.7*	14.03 *13.67*	13.67 *26.67*	22.64 *21.42*	14.09 *20.73*	14.27 *17.54*	16.96
No_EM	9.52 *68.82*	10.51 *68.72*	13.95 *77.51*	17.27 *64.61*	8.75 *65.57*	17.38 *63.13*	10.98 *62*	11.63 *54.85*	65.08
Total_% (N=9767)	9.01	9.96	11.71	17.4	8.69	17.92	11.53	13.79	100
	DF	Value	Prob						
Chi-Square	14	699	<.0001						
Panel D: Accounting Conservatism									
	Cluster								
U_CON	1	2	3	4	5	6	7	8	K-W
Mean	6.69	5.86	5.37	5.48	5.23	4.71	4.39	4.50	
Median	7	6	6	6	5	5	4	4	*<0.01*
N	855	946	1110	1638	813	1669	1075	1167	9273
C_CON	1	2	3	4	5	6	7	8	K-W
Mean	4.26	3.14	3.8	3.46	3.21	3.44	3.55	3.6	
Median	4	3	4	3	3	3	3	4	*0.03*
N	767	863	1026	1396	780	1486	982	812	8112

Industry Classification is based on the Fama-French 12 Industry classification scheme. Stages of firm life cycle are identified using the cash flow patterns as in Dickinson (2011). EM_DN, EM_UP and No_EM indicate downward, upward and no earnings management, identified following Jansen et al., 2012. U_CON and C_CON denote unconditional and conditional conservatism as estimated in Chen et al., (2014). Chi-square statistics test the null hypothesis of no association between the row and column variables in a two-way table. K-W denotes p-values of the Kruskal-Wallis test, with the null hypothesis of no differences in medians across clusters.

5.4 ANALYST FORECASTS

Previous results reported in this thesis demonstrate the success of ROWK clustering to identify the problem of HGSC in the earnings persistence model. To the extent that financial analysts are able to rationally utilise information from financial statements and are able to understand cluster patterns in earnings persistence, differences in regression coefficients across firms are expected to be embedded in analyst forecasts. On the other hand, a number of studies produce evidence of systematic biases in investors' and analysts' forecast errors, which suggests that investors and analysts do not fully impound the implications of existing information (Ball & Bartov, 1996; Dichev & Tang, 2009; Monte-Mor, et al., 2018; Ulupinar, 2018).

This section presents results from tests of Hypothesis H9 predicting that information from ROWK's cluster identification predicts analyst forecast errors. To do this, the thesis employs two methods. The first uses two-way sorting portfolios to control for the dispersion of current earnings. Specifically, firms are sorted annually into 20 portfolios by current earnings, and ROWK clustering is executed within each portfolio. Consequently, for each earnings portfolio, eight ROWK clusters are identified. Accordingly, the examined two-way sorting portfolios are generated as the intersection of clusters and earnings portfolios. Table 5-12 presents the construction of the examined portfolios.

Table 5-12: Two-Way Sorting Portfolio

Abbr.	Description	Formation
HE_HP	High Earnings-High Persistent Portfolio	Cluster 8 of earnings portfolios 17-20
HE_LP	High Earnings-Low Persistent Portfolio	Cluster 1 of earnings portfolios 17-20
LE_HP	Low Earnings-High Persistent Portfolio	Cluster 8 of earnings portfolios 1-4
LE_LP	Low Earnings-Low Persistent Portfolio	Cluster 1 of earnings portfolios 1-4
HE_HI	High Earnings-High Intercept Portfolio	Cluster 4 of earnings portfolios 17-20
HE_LI	High Earnings-Low Intercept Portfolio	Cluster 5 of earnings portfolios 17-20
LE_HI	Low Earnings-High Intercept Portfolio	Cluster 4 of earnings portfolios 1-4
LE_LI	Low Earnings-Low Intercept Portfolio	Cluster 5 of earnings portfolios 1-4

Table 5-13 presents the comparison of the realized earnings and analyst forecasts for up to three years between firms in *Clusters 1* and *8* (Panels A and B) and between firms in *Clusters 4* and *5* (Panels C and D). Realized earnings is as reported in *I/B/E/S*. Analyst forecast errors for year *t+i*, ($FE_{t+i}, i = 1,..,3$) are defined as the differences between the actual earnings for year *t+i* and the *first median* analyst forecast of year *t+i immediately* following the announcement of earnings for year *t+i-1*.

Panel A compares the realized earnings and analyst forecasts between *High Earnings-High Persistence Portfolio* (*HE_HP*) and *High Earnings-Low Persistence Portfolio* (*HE_LP*). The results confirm that the two-way sorting procedure works well in the sense that both the mean and median current realized earnings for *HE_HP* and *HE_LP* are nearly the same. Therefore, any deviations in future profitability can be fully attributed to differences in the persistence patterns between firms in *Cluster 1* and *Cluster 8*.

The next columns and rows in Panel A provide the properties of the realized earnings and analysts' forecasts for these two portfolios at each of the next three years. An examination of the mean and median reveals that analyst forecasts correctly anticipate that firms in *HE_LP* will experience faster mean reversion in future earnings than firms in *HE_HP*. However, this correction only exists in the short term, i.e. at time *t+1*, while in the long run there is no evidence that analysts incorporate the difference in earnings between these portfolios. Particularly, the predicted divergence in profitability across *HE_HP* and *HE_LP* portfolios at time *t* is 1.2% at the mean and median, which is similar to actual divergence with the corresponding numbers as 1.6% and 1.1%. Nevertheless, at times *t+2* and *t+3*, the predicted divergences in profitability across *HE_HP* and *HE_LP* portfolios are significantly lower than those of actual diversions. The same results are observed in Panel C that compares earnings streams between *HE_HI* and *HE_LI* (i.e. *Cluster 4* vs. *Cluster 5*).

Thus, these results provide evidence that analyst forecasts correctly anticipate that firms in the low earnings persistence cluster will experience faster mean reversion in future earnings than firms in the high earnings persistence cluster. However, this correction only exists in the short term, i.e. at time t+1, while in the long run there is no evidence that analysts incorporate the differences in earnings between these portfolios. Hence, these results partly support hypothesis H9 in the sense that information from ROWK's cluster identification partially predicts analyst forecast errors in the short-term, and strongly predicts analyst forecast errors in the long-term.

Panels C and D of Table 5-13 provide the analysis for the lowest current realized earnings portfolios. The results from these panels reveal some important findings. In the case of current realized earnings portfolios, it seems that analysts' forecasts incorporate the information from different earnings patterns between clusters not only precisely (i.e. assign the correct sign of the divergence), but also excessively (i.e. exaggerate the magnitude of the divergence). Second, analysts' forecasts exhibit more (upward) biases for firms in the lowest quintiles of earnings. These biases could obscure the results as to whether or not analysts can understand different earnings patterns across clusters.

Table 5-13: Analyst Earnings Forecasts vs. Actual Earnings												
Time	T			T+1			T+2			T+3		
	N	Mean	Med	N	Mean	Med	N	Mean	Med	N	Mean	Med
Panel A: Cluster 1 vs. Cluster 8- High Current Earnings												
Realized earnings												
HE_HP	249	0.138	0.131	248	0.132	0.126	221	0.127	0.126	199	0.122	0.122
HE_LP	182	0.139	0.128	182	0.116	0.115	157	0.107	0.105	139	0.104	0.1
Difference		*-0.001*	*0.002*		*0.016*	*0.011*		*0.02*	*0.02*		*0.018*	*0.022*
Analyst Forecasts												
HE_HP				249	0.133	0.126	221	0.135	0.126	180	0.139	0.133
HE_LP				182	0.121	0.114	157	0.127	0.117	130	0.129	0.122
Difference					*0.012*	*0.012*		*0.008*	*0.01*		*0.01*	*0.011*
p-value [A]					*0.09*	*0.35*		*0.005*	*0.003*		*0.039*	*0.001*
Panel B: Cluster 1 vs. Cluster 8- Low Current Earnings												
Realized earnings												
LE_HP	240	-0.008	0.005	237	0.002	0.012	212	0.009	0.016	187	0.015	0.018
LE_LP	176	-0.015	0.007	174	0.012	0.016	159	0.03	0.028	148	0.038	0.039
Difference		*0.007*	*-0.002*		*-0.01*	*-0.004*		*-0.021*	*-0.012*		*-0.022*	*-0.021*
Analyst Forecasts												
LE_HP				240	0.006	0.013	210	0.016	0.017	150	0.025	0.022
LE_LP				176	0.021	0.023	157	0.047	0.039	129	0.068	0.05
Difference					*-0.014*	*-0.01*		*-0.031*	*-0.022*		*-0.043*	*-0.029*
p-value [A]					*0.13*	*0.04*		*<0.01*	*<0.01*		*<0.01*	*0.07*
Panel C: Cluster 4 vs. Cluster 5- High Current Earnings												
Realized earnings												
HE_HI	348	0.141	0.133	346	0.131	0.129	314	0.126	0.124	286	0.121	0.12
HE_LI	171	0.14	0.126	171	0.116	0.117	158	0.098	0.098	147	0.093	0.084
Difference		*0.002*	*0.007*		*0.016*	*0.012*		*0.028*	*0.026*		*0.029*	*0.036*
Analyst Forecasts												
HE_HI				348	0.132	0.125	311	0.133	0.128	260	0.133	0.127
HE_LI				171	0.123	0.121	159	0.13	0.127	132	0.13	0.126
Difference					*0.009*	*0.004*		*0.003*	*0.002*		*0.003*	*0.001*
p-value [A]					*0.05*	*0.04*		*<0.01*	*<0.01*		*<0.01*	*<0.01*
Panel D: Cluster 4 vs. Cluster 5- Low Current Earnings												
Realized earnings												
LE_HI	345	-0.012	0.005	334	0.005	0.014	308	0.018	0.023	275	0.023	0.027
LE_LI	176	-0.015	0.005	176	0.003	0.015	158	0.023	0.028	143	0.036	0.039
Difference		*0.004*	*0*		*0.002*	*-0.001*		*-0.004*	*-0.005*		*-0.013*	*-0.012*
Analyst Forecasts												
LE_HI				344	0.009	0.016	311	0.025	0.024	233	0.04	0.033
LE_LI				176	0.018	0.022	157	0.044	0.041	124	0.06	0.053
Difference					*-0.009*	*-0.007*		*-0.019*	*-0.017*		*-0.02*	*-0.02*
p-value [A]					*<0.01*	*0.07*		*<0.01*	*<0.01*		*<0.01*	*<0.01*

[A] p-value on tests of the difference between forecast and realized earnings differences (t-tests for differences in means and Wilcoxon test for differences in medians).

The two portfolios are constructed in the following way. In the first step, the full sample is sorted into 20 portfolios based on the value of I/B/E/S realized current earnings. In the second step, ROWK clustering is executed within each portfolio. Then for each earnings portfolio, eight ROWK clusters are identified. Accordingly, the examined two-way sorting portfolios are generated as the intersection of clusters and earnings portfolios. See Table 5.10 for the descriptions and formations of examined two-way sorting portfolios. Realized earnings for t+1 is the realized earnings for year t+1 as reported in I/B/E/S. Analyst forecast errors for year t+1, (FE_{t+1}) are defined as the differences between the actual earnings for year t+1 and the *first* median analyst forecast of year t+1 immediately following the announcement of earnings for year t. Analyst forecast errors for year t+2,…, t+5 are constructed in the same manner as t+1.

To resolve this issue, this thesis take the further step of using the analysts' forecast error model as follows:

(5.2)

$$FE_{t+i} = \alpha_0 + \alpha_1 LOW\_INT + \alpha_2 FE_t + \alpha_3 LOW\_INT * FE_t + u_t$$

(5.3)

$$FE_{t+i} = \beta_0 + \beta_1 LOW\_P + \beta_2 FE_t + \beta_3 LOW\_P * FE_t + u_t$$

where *LOW_INT (LOW_P)* is a dummy variable, which equals 1 if a firm is in *Cluster 5 (Cluster 1)* and 0 if a firm is in *Cluster 4 (Cluster 8). FE* denotes forecast error. To control for positive auto-correlations of analyst forecast errors, it is necessary to add earnings surprises at time *t* (FE_t) to the predictive model of future forecast errors (Abarbanell & Bernard, 1992; Dickinson, 2011). Analyst forecast error for year *t* (FE_t) is measured as the difference between the actual earnings for year *t* and the *last* median analyst forecast for year *t prior* to the announcement of earnings for year *t*. Analyst forecast errors for year *t+1* (FE_{t+1}) are defined as the differences between the *actual* earnings for year *t+1* and the *first* median analyst forecast of year *t+1 immediately* following the announcement of earnings for year *t*. Analyst forecast errors for year *t+2,..., t+5* are constructed in the same manner as for *t+1*. The coefficients of interest are α_1 and α_3 for Equation 5.2 and β_1 and β_3 for Equation 5.3.

Table 5-14 presents the regression results. Columns two to four display results of Equation 5.2 when the dependent variables are analyst forecast errors for year *t+1, t+2* and *t+3* respectively. Columns six to eight present results of Equation 5.3. As expected, the slope coefficients on FE_t are positive and significant for all portfolios, ranging from 0.38 (0.4636-0.0878) for firms in *Cluster 8* to 0.63 (0.4549+0.1726) for firms in *Cluster 4*. More importantly, information from cluster memberships is able to predict 1-year ahead analyst forecasts errors since all coefficients of interest, i.e. α_1 and α_3 (for Equation 5.2) and β_1 and β_3 (for Equation 5.3) are significant when the dependent variable is FE_{t+1}. The significantly negative coefficients of *LOW_P* and *LOW_P*FE* are in line with findings of Dichev & Tang (2009).

Consistent with the thesis findings from the two-way sorting portfolios, analysts seem to ignore the impact of cluster patterns on long-run future earnings. The magnitudes of the coefficients for *LOW_INT* and *LOW_P* are even higher in the long run. Particularly, firms in *Cluster 5* (*LOW_INT=1*) have 0.77% lower median analyst forecast errors for 1-year ahead earnings relative to firms in *Cluster 4*, but this dispersion widens to -2.4% and -2.1% for 2-year and 3-year ahead earnings, respectively. The corresponding numbers for *LOW_P* are -0.2%, -0.9% and -1.3% for 1-year, 2-year and 3-year ahead earnings, respectively. This evidence supports the conclusion that analysts *partly* understand the implications of ROWK cluster information for short-term future earnings, and *ignore* the implications for long-term future earnings. As a result, conditioning on such information permits the identification of reliable and economically important patterns in analyst forecast errors.

To further examine whether the information from ROWK clustering provides additional information after controlling for earnings volatility, the best factor driving earnings persistence as found in the study of Dichev & Tang (2009), the analysts' forecast error model is next run within the first (lowest) and firth (highest) quintiles of earnings volatility. Since only firms in clusters 1,2,3 and 5 belong to the highest earnings volatility portfolio, the model is as follows:

(5.4)

$$FE_{t+i} = \gamma_0 + \gamma_1 FE_t + \gamma_2 C1 + \gamma_3 C1 * FE_t + \gamma_4 C2 + \gamma_5 C2 * FE_t + \gamma_6 C3 + \gamma_7 C3 * FE_t + u_t$$

where C_i (i=1,2,3) is a dummy variable, which equals 1 if a firm is in *Cluster i* and 0 otherwise. In this model, the base group is Cluster 5, i.e. when C_i=0 (i=1,2,3).

In the same manner, since only firms in Clusters 4,6,7 and 8 belong to the lowest earnings volatility portfolio, the model is as follows:

(5.5)

$$FE_{t+i} = \pi_0 + \pi_1 FE_t + \pi_2 C4 + \pi_3 C4 * FE_t + \pi_4 C7 + \pi_5 C7 * FE_t + \pi_6 C8 + \pi_7 C8 * FE_t + u_t$$

where C_i (i=4,7,8) is a dummy variable, which equals 1 if a firm is in *Cluster i* and 0 otherwise. In this model, the base group is Cluster 6, i.e. when C_i=0 (i=4,7,8).

The results are presented in Panel B of Table 5.14. Briefly, several dummy variables are highly significant, and the F-tests (all dummy and interaction variables equal 0) are also highly significant for most cases (except equation 5.4 with FE_{t+3}). This evidence gives further support to the previous finding that ROWK clustering provides incremental information after controlling for the best factor driving earnings persistence, i.e. earnings volatility.

Table 5-14: Analyst Forecast Errors Conditional on ROWK's Clusters

Panel A: Without Controlling for Earnings Volatility

	Equation 5.1				Equation 5.2		
Variable	FE_{t+1}	FE_{t+2}	FE_{t+3}	Variable	FE_{t+1}	FE_{t+2}	FE_{t+3}
Intercept	-0.0009	**-0.0064**	**-0.0091**	Intercept	-0.0010	**-0.0033**	**-0.0047**
	(0.141)	*(0.0006)*	*(<0.0001)*		*(0.1435)*	*(0.0036)*	*(0.0026)*
LOW_INT	**-0.0077**	**-0.0242**	**-0.0209**	LOW_P	**-0.0022**	**-0.0091**	**-0.0132**
	(<0.0001)	*(<0.0001)*	*(<0.0001)*		*(0.0469)*	*(<0.0001)*	*(<0.001)*
FE_t	**0.4549**	**0.8382**	**0.5541**	FE_t	**0.4636**	0.2499	0.2752
	(<0.0001)	*(0.0073)*	*(0.0305)*		*(<0.0001)*	*(0.0971)*	*(0.259)*
LOW_INT*	**0.1726**	0.2381	-0.1750	LOW_P*	**-0.0878**	0.0234	0.1267
FE_t	*0.013*	*(0.3231)*	*(0.3503)*	FE_t	*(0.0002)*	*(0.5841)*	*(0.1566)*
N	2479	2220	1783		2094	1903	1485
Adj.R^2	0.0508	0.0337	0.0362		0.0151	0.0148	0.0248
Firm FE	Y	Y	Y		Y	Y	Y
Year FE	Y	Y	Y		Y	Y	Y

Panel B: Controlling for Earnings Volatility

	Highest Earnings Volatility Firms				Lowest Earnings Volatility Firms		
Variable	FE_{t+1}	FE_{t+2}	FE_{t+3}	Variable	FE_{t+1}	FE_{t+2}	FE_{t+3}
Intercept	**-0.0095**	**-0.024**	**-0.03**	Intercept	**-0.0037**	**-0.0082**	**-0.0118**
	(<.0001)	*(<.0001)*	*(<.0001)*		*(0.0002)*	*(<.0001)*	*(<.0001)*
FE_t	**0.6943**	0.3166	0.3684	FE_t	0.0156	0.3472	**-2.011**
	(<.0001)	*(0.3101)*	*(0.3917)*		*(0.9618)*	*(0.5576)*	*(0.0631)*
C3	**0.0032**	0.0077	0.0015	C4	0.0016	0.0015	0.0002
	(0.2246)	*(0.1013)*	*(0.8146)*		*(0.1788)*	*(0.4778)*	*(0.9414)*
C3*FE	-0.2380	0.1002	-0.7273	C4*FE	**0.9013**	**1.2696**	**4.0757**
	(0.3432)	*(0.8235)*	*(0.2577)*		*(0.0127)*	*(0.054)*	*(0.0004)*
C1	**0.0072**	**0.008**	0.0089	C8	**0.0026**	**0.0045**	**0.0069**
	(0.0056)	*(0.082)*	(0.1566)		*(0.0116)*	*(0.0147)*	*(0.0124)*
C1*FE	**-0.5162**	-0.1528	0.0667	C8*FE	**0.6574**	**0.4265**	**2.9271**
	(0.0142)	*(0.6863)*	*(0.8984)*		*(0.0624)*	*(0.5064)*	*(0.0109)*
C2	**0.0043**	**0.0096**	0.009	C7	0.0001	**0.0037**	0.0014
	(0.0809)	*(0.032)*	*(0.1391)*		*(0.9112)*	*(0.0888)*	*(0.6759)*
C2*FE	-0.0396	0.2576	-0.577	C7*FE	**1.4443**	0.0173	2.0629
	(0.8625)	*(0.5371)*	*(0.3226)*		*(0.0003)*	*(0.981)*	*(0.1141)*
N	1865	1609	1237	N	1871	1741	1411
Adj.R^2	0.0231	0.0037	0.0002	Adj.R^2	0.0575	0.0259	0.0353
Pr > F	<.0001	0.0717	0.4097	Pr > F	<.0001	<.0001	<.0001
Firm FE	Y	Y	Y	Firm FE	Y	Y	Y
Year FE	Y	Y	Y	Year FE	Y	Y	Y

5.5 ROBUSTNESS TESTS

This section presents robustness tests of the previous findings regarding the superior performance of ROWK clustering to address the problem of HGSC in the earnings persistence model.

5.5.1 Out-of-sample Results

5.5.1.1 Earnings Persistence

All of the previous results are based upon a sample that randomly draws from the full 1988-2004 period data sample. This sample is called the '*training sample'*, while the remaining random sample is called the '*testing sample'*. Now, the same optimal weights that are found by ROWK clustering in the training sample are applied to the testing sample.

Table 5-15 presents the results of earnings persistence patterns for the testing sample. Consistent with the in-sample results, ROWK clustering continues to exhibit superior performance to identify heterogeneities on coefficients of the earnings persistence model. The differences between the highest and lowest values of intercepts, persistence coefficients and adjusted R-squared are statistically significant at levels of 3.2%, 44.9% and 33.0%, respectively. The *MSRs* for ROWK clustering are significantly lower than those using either the full sample or a single feature to partition the sample. Briefly, the results from ROWK clustering reveal its potential ability to identify heterogeneities of coefficients for the earnings persistence model for both in-sample and out-of-sample data.

Table 5-15: Out-of-sample Results for the Earnings Persistence Regression:
$Earnings_{i,t+1} = \mu_{\zeta_k^0} + Earnings_{i,t}\theta_{\zeta_k^0} + u_{i,t}$

Panel A: Regression Result for the Full Sample				
	N	$\mu_{\zeta_k^0}$	$\theta_{\zeta_k^0}$	Adj. R^2
Full sample	10070	0.009	0.681	0.393
Difference MSR vs. ROWK (*100)				*0.021*
P-value on Difference				*(<0.0001)*
Panel B: Regression Results by Clusters Identified by ROWK Clustering				
Clusters	N	$\mu_{\zeta_k^0}$	$\theta_{\zeta_k^0}$	Adj. R^2
Cluster 8	1155	-0.002	0.956	0.626
Cluster 7	1388	-0.007	0.902	0.460
Cluster 6	1916	0.001	0.896	0.478
Cluster 5	819	-0.014	0.821	0.357
Cluster 4	1322	0.011	0.822	0.479
Cluster 3	1253	0.002	0.826	0.435
Cluster 2	1248	-0.008	0.778	0.369
Cluster 1	969	0.018	0.507	0.297
Difference (H-L)		*0.032*	*0.449*	*0.330*
P-value on Difference		*(<0.0001)*	*(<0.0001)*	*(<0.0001)*
Total MSR (*100)				0.354
Panel C: Regression Results by Octiles of Earnings Volatility				
Octiles by VOL_IBC	N	$\mu_{\zeta_k^0}$	$\theta_{\zeta_k^0}$	Adj. R^2
Octile 1	1252	-0.002	0.985	0.678
Octile 2	1261	-0.001	0.924	0.621
Octile 3	1261	0.002	0.876	0.471
Octile 4	1259	-0.001	0.882	0.407
Octile 5	1262	0.001	0.825	0.402
Octile 6	1261	0.003	0.793	0.406
Octile 7	1261	0.006	0.642	0.358
Octile 8	1253	0.002	0.513	0.273
Difference (H-L)		*0.0083*	*0.473*	*0.404*
P-value on Difference		*(0.27)*	*(<0.0001)*	*(<0.0001)*
Difference MSR vs. ROWK (*100)				*0.009*
P-value on Difference				*(0.031)*
Panel D: Regression Results by Octiles of Absolute Value of Accruals				
Octiles by ABS_ACC_DEF	N	$\mu_{\zeta_k^0}$	$\theta_{\zeta_k^0}$	Adj. R^2
Octile 1	1252	-0.008	0.897	0.418
Octile 2	1261	-0.004	0.858	0.426
Octile 3	1262	-0.004	0.863	0.427
Octile 4	1258	-0.002	0.833	0.365
Octile 5	1262	0.002	0.835	0.459
Octile 6	1261	0.001	0.839	0.442
Octile 7	1261	0.007	0.763	0.448
Octile 8	1253	0.023	0.539	0.356
Difference (H-L)		*0.0311*	0.358	0.062
P-value on Difference		*(<0.0001)*	*(<0.001)*	*(0.037)*
Difference MSR vs. ROWK (*100)				*0.005*
P-value on Difference				*(0.067)*

Optimal weights found by ROWK clustering on the training sample are applied to the testing sample. *Difference (H-L)* indicates the difference between the highest and lowest values. *Difference MSR vs. ROWK (*100)* denotes the 100-times difference between the mean squared residuals derived from running the earnings persistence model within each octile of examined features and within each cluster identified by ROWK clustering. The p-value for the difference in the intercepts and persistence coefficients is derived from a t-test. The p-value for the difference in the Adj_R^2 is derived from a bootstrap test (see Chapter 3.2.2 for full details). The bold numbers of intercept and persistence coefficients indicate the extreme values (highest or lowest). The bold numbers of adjusted R^2 are in accordance with the bold persistence coefficients

5.5.1.2 Interaction Terms

For the sake of brevity, the out-of-sample results from ROWK clustering when interaction terms are included in the earnings persistence model are not reported and are available upon request. Briefly, all main results remain unchanged.

5.5.1.3 Out-of-sample Earnings Prediction

Earnings predictability influences equity valuation. Prior to 2000, fundamental analysis or equity valuation, with its aim to identify elements of financial statements that are relevant to assess firm value, attempts without theoretical guidance to select accounting variables or their ratios to predict earnings and consequently stock returns (Richardson et al., 2010). It results not only in concerns regarding the in-sample estimations, but also displays low out-of-sample predictive power (e.g. Lipe, 1986; Ou & Penman, 1989).

The superiority of ROWK clustering to identify heterogeneities on coefficients of the earnings persistence model suggests that it could help investors to predict future earnings more precisely. This section examines the out-of-sample performance of ROWK clustering to forecast one-year-ahead earnings. Different benchmark techniques are used to make comparisons with ROWK clustering. Please refer to Table 3-3 for the descriptions of the examined benchmarks.

Each year *t*, ROWK clustering is run using the previous five-year data (*t-5* to *t-1*). The use of previous data ensures that the prediction is based on available data. Regression estimates for each cluster, ROWK cluster membership and their corresponding centroids are saved. Then each firm in year *t* is assigned to the cluster that has the lowest distance to that firm. To predict firm earnings in year *t+1*, the estimations from the earnings persistence model are applied to firms in year *t* based on their cluster membership. Forecasts are produced by a rolling-forward estimation, starting from 1994 (with 1988-1993 estimation periods) to 2011.

The results of earnings predictions are presented in Figure 5.9. The thesis relies on mean and median absolute forecast errors (*MEAN_AFE* and *MED_AFE*) to indicate forecasting accuracy. Forecast error is the difference between the actual earnings minus the forecast based on the predictive models. For brevity, only *MEAN_AFE*

results are presented. The results are unchanged using medians and are available upon request. The examined period covers two financial crises, i.e. Dot Com Bubble and 2008-Financial Crisis.

Several observations are made. First, for all models, the *MEAN_AFEs* (an inverted indicator of earnings predictability) tend to increase over time. They start at approximately 3% during the beginning of the 1990s, and hover around 4% by the end the examined period, i.e. 2011. Second, *MEAN_AFEs* increase dramatically during the crises, particularly during the 2008 Financial Crisis to over 6%. More importantly, ROWK clustering outperforms all benchmark models in predicting one-year-ahead earnings. *MEAN_AFEs* of ROWK clustering are 4.15%, which is significantly lower than those of *ALL* (4.35%), *VOL* (4.24%), *ACC* (4.28%), *E_MOD* (4.25%), *K_MEAN_NON_ZERO* (4.3%) and *WK* (4.29%). Unreported results (available upon request) show that although the ROWK optimal weights fluctuate over time, the relative ordering of weights remain unchanged. *VOL_IBC* and *ABS_ACC_DEF* are the two most relevant features and are consistently assigned the highest weights by ROWK clustering.

As discussed in Section 5.3.2.2 (ROWK and Non-linearity), there is evidence of non-monotonic effects of moderator variables. As evidenced by different coefficients of interaction terms across ROWK clusters, this study observed that ROWK clustering using the earnings persistence model with interactions terms was able to identify these effects. To further investigate this issue, the thesis examines earnings predictability of models with interaction terms.

Figure 5.10 displays these results. *ALL* and *ALL_INT* denote results from regressions of the earnings persistence models without and with interaction terms, respectively. *ROWK* and *ROWK_INT* denotes regressions of the earnings persistence models for each cluster found by ROWK clustering executed based on earnings persistence without and with interaction terms, respectively. *K_MEAN_INT_NON_ZERO* denotes results from running K_means with equal weights for only relevant features using the earnings persistence model with interaction terms. Several observations are notable. First, the inclusion of interaction terms only modestly increases the precision of earnings forecasts. *MEAN_AFE* of *ALL_INT* is 4.3%, being slightly lower than 4.35% for the case of *ALL*. Second, ROWK clustering outperforms other models on

earnings predictability, including the models with interaction terms (i.e. *ALL_INT* and *K_MEAN_INT_NON_ZERO*). This evidence supports the proposition that ROWK clustering is a potentially effective method to cope with the non-monotonic effects of moderator variables. Finally, the *MEAN_AFE* of *ROWK* and *ROWK_INT* are similar, indicating that the performance of ROWK clustering is less affected by the inclusion of interaction terms.

To finalize this section, this thesis quantitatively investigates those channels that contribute to the outperformance of ROWK clustering. The first channel is that ROWK clustering addresses the problem of HGSC by considering the effect of moderator variables. It is analogous to including interaction terms in the regression model. This channel improves model precision by 0.051%, measured by the difference of *MEAN_AFEs* between the models of *ALL* and *ALL_INT*. The second channel is the employment of the clustering technique to account for non-monotonic effects that cannot be addressed through the inclusion of interaction terms. This channel improves model precision by 0.01%, measured by the difference of *MEAN_AFEs* between the cases of *ALL_INT* and *K_MEAN_INT_NON_ZERO*. The final channel arises from the process by which ROWK clustering connects the clustering technique and the regression model, correctly identifies degrees of contribution of features, and consequently places correct weights to features. As expected, this channel makes the greatest contribution to the superior performance of ROWK clustering. This channel improves model precision by 0.134%, measured by the differences in *MEAN_AFEs* between the cases of *K_MEAN_INT_NON_ZERO* and *ROWK*. This is not surprising since cluster features make different contributions to earnings persistence patterns as documented in previous results.

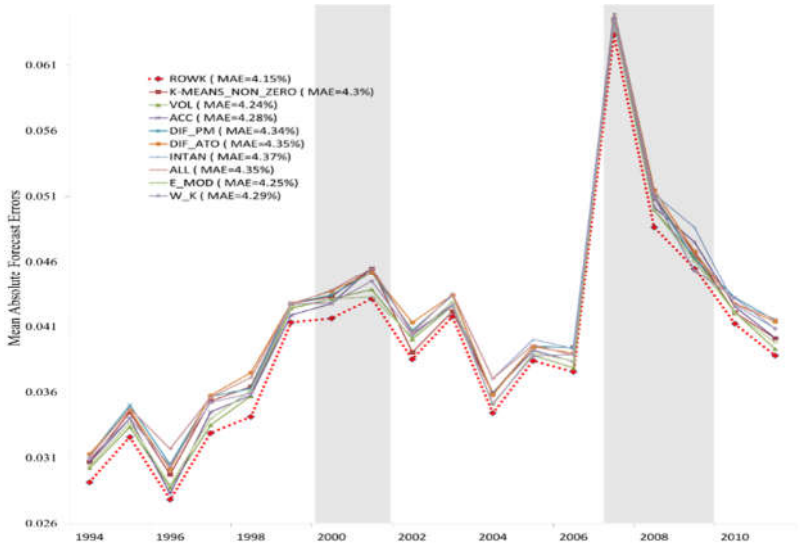


Figure 5.9- One-year-ahead Earnings Forecast Errors by Different Prediction Models

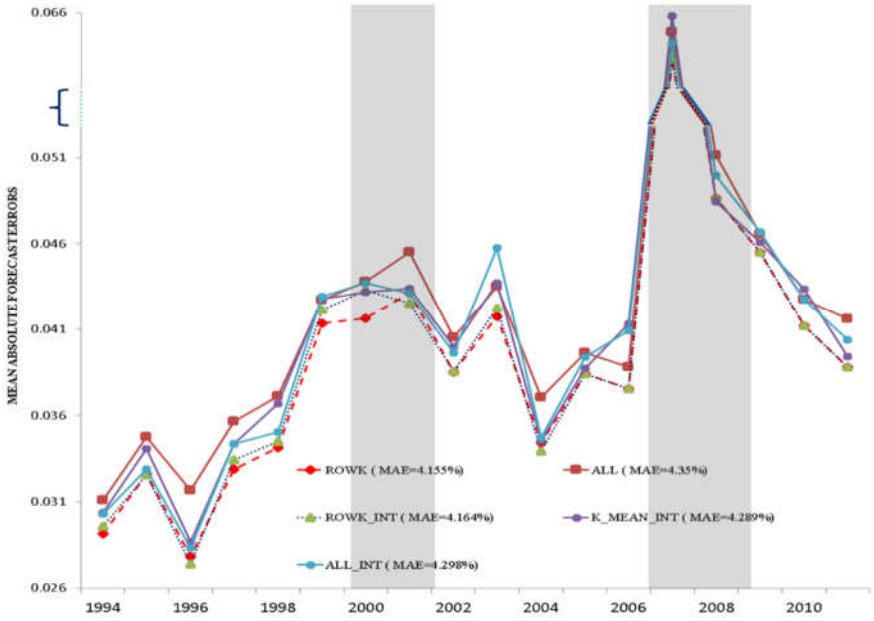


Figure 5.10- One-year-ahead Earnings Forecast Errors- Interaction Terms

5.5.2 Other Robustness Tests

Recent studies of earnings predictability tend to use return on net operating assets (*RNOA*) as the measure of firm earnings (e.g. Amor-Tapia & Tascón Fernández, 2014; Bauman, 2014; Fairfield & Yohn, 2001). Operating earnings is studied in prior research on earnings persistence and earnings forecasts which consider the role of operating activities in creating value (Fairfield & Yohn, 2001). Following the study of Dichev & Tang (2009), all results in this chapter are based on deflated earnings measured as income before extraordinary items divided by average total assets

(*IBC_DEF*). To assess whether the results are sensitive to a different measure of earnings, we repeat our tests replacing *IBC_DEF* by *RNOA*. All results, which are not reported and are available upon request, remain materially unchanged.

Section 5.3.4 documents a significant drop in the sample size (22%) due to firms failing to survive over the five-year period. Furthermore, patterns of survivor rates differ across clusters. Hence, the long-term analysis could be biased. To reconcile this issue, the thesis employs two robustness tests. For the first test, only firms that survive up to five years since the formation period are included. For the second test, if firms delist, their next (missing) earnings will be assigned. Particularly, for firms that are delisted for performance reasons (i.e. bankruptcy), the next year of earnings is coded as zero, and the stock annual returns are set to -100%. For firms that delist due to mergers or acquisitions, the next year of earnings are coded as 30% of previous earnings and stock returns are treated as -30%. All results, which are not reported and are available upon request, remain materially unchanged.

5.6 SUMMARY

This chapter presents the empirical results relating to the second thesis research aim to address the issue of HGSC on earnings persistence. Results are reported from the very first application of ROWK, the newly proposed clustering method introduced in this thesis. Specifically, the chapter discusses the findings pursuant to hypotheses H5 and H9 to assess whether ROWK clustering can help to reveal the earnings persistence patterns that would otherwise remain unclear.

ROWK clustering is implemented for the earnings persistence model with 17 examined features. The results show that only five features are relevant for clustering and assigned non-zero weights. These features include earnings volatility (*VOL_IBC_DEF*), absolute value of accruals (*ABS_ACC_DEF*), intangible investment (*INTAN_INV_DEF*), change in profit margin (*ΔPM*) and change in asset turnover (*ΔATO*). These results are consistent with those of several studies on earnings persistence and predictability (e.g. Amir et al., 2011; Bauman, 2014; Dichev & Tang, 2009 etc.). Furthermore, ROWK clustering assigns different optimal weights to features, consistent with Hypothesis H5 predicting that firms exhibit different earnings persistence between ROWK clusters.

Examining the earnings persistence patterns across ROWK clusters also reveals that, consistent with Hypothesis H6, there are differences in earnings persistence and intercepts between ROWK clusters. Importantly, these differences seem large in magnitude and suggest that cluster membership is economically important. Moreover, partitioning the sample based on a single feature results in lower differences in coefficients of both intercepts and earnings persistence across groups relative to those of ROWK clustering. These results are consistent with the Hypothesis H7 prediction that ROWK clustering results in larger differences in earnings persistence between clusters than a single-variable cluster partitioning technique.

The thesis further investigates the characteristics of each clusters found by ROWK clustering. The results exhibit distinguishing characteristics of each cluster, especially between firms in *Clusters 4* and *5* (same earnings persistence but different intercepts), and firms in *Clusters 1* and *8* (low persistence vs. high persistence). Consistent with Hypothesis H8, firms also exhibit distinguishable patterns of survivorship, transition rates, industry classification, earnings management, stages of firm life cycle and accounting conservatism across clusters.

To the extent that financial analysts are able to rationally utilise information from financial statements and are able to understand cluster patterns in earnings persistence, differences in regression coefficients across firms are expected to be implicitly embedded in analyst forecasts. Using two-way sorting portfolios to control for dispersion in current actual earnings, this thesis documents that analysts' forecasts only partially incorporate the information from cluster patterns in the short run, while ignoring the impact of these patterns on long-term future earnings. This lends partial support to Hypothesis H9 which proposes that information from ROWK's cluster identification predicts analyst forecast errors. The results from estimating the analyst forecast error model further support this statement.

The chapter concludes with numerous robustness tests. First, all results remain unchanged for out-of-sample data. The results are also insensitive to the treatment of survivorship bias and different proxies for earnings. Second, differences of earnings patterns across clusters persist for up to five years after the formation period. Third, the results support the existence of non-monotonic effects of moderator variables (i.e. features) on the earnings persistence model that cannot be addressed by the inclusion

of interaction terms. Further, the results show that ROWK clustering successfully addresses these non-monotonic effects by grouping firms into similar clusters. Finally, channels contributing to the performance of ROWK clustering are identified. As expected, the channel relating to the correction of feature weight assignment contributes the most to the superior performance of ROWK clustering. Accounting for the full 0.14% reduction in *MEAN_AFE* resulting from ROWK clustering relative to the earnings persistence model for the full sample (*ALL*), nearly 95% of the reduction (0.134%) is achieved through the channel of more precise assignment of feature weights, while the remaining 5% arises from other channels.

The next chapter, Chapter 6, concludes this thesis. In a nutshell, it summarises the research aims, hypotheses and methodology followed by a synopsis of the thesis' findings. Next, the contributions and limitations of the study are discussed. The thesis concludes with suggestions for future research and extensions.

CHAPTER 6

CONCLUSION

6.1 INTRODUCTION

This final chapter summarises and discusses the thesis's findings. In particular, Section 6.2 reviews the two thesis aims, the nine associated hypotheses, methodology and empirical results. Next, the contributions of the thesis are presented in Section 6.3. Section 6.4 identifies the limitations of the thesis that guide suggestions for further research presented in final section, Section 6.5.

6.2 REVIEW OF THESIS AIMS, RESEARCH QUESTIONS, HYPOTHESES, METHODOLOGY AND MAJOR FINDINGS

This thesis investigates a solution to the prevailing problem of HGSC, and then relates it to an application within the finance domain. In order to achieve this objective, two thesis aims are established and were achieved through the thesis. The first research aim of this thesis is to develop a new clustering method that can be applied in financial research to address the problem of HGSC. The second aim of the thesis is to apply the newly proposed ROWK clustering method to mitigate problems of HGSC in earnings persistence models. To achieve these aims, nine associated hypotheses, driven by existing research gaps were examined. The following two sub-sections (6.2.1, 6.2.2) summarise associated hypotheses, methodology, and major findings relating to each thesis aim.

6.2.1 The first thesis aim: To develop a new clustering method that can be applied in financial research to address the problem of HGSC

The first thesis aim is to propose a new clustering method that is able to address the shortcomings of current clustering techniques in dealing with the issue of HGSC. To achieve the first research aim, a thorough review of literature is conducted (Chapter 2). The literature of HGSC reveals evidence of low predictive power and instability

of estimates in regression analysis (e.g. Lipe, 1986; Nissim & Penman, 2001; Ou & Penman, 1989). Examining causes and developing methods to deal with these unsuccessful empirical results should be a matter of urgency. It could be that the relationship between predictors and outcome variables is non-linear, or there are violations of the coefficients' homogeneity assumption that is usually taken for granted in quantitative research (e.g. Fu, 2010; Giroud & Mueller, 2011; Tapia & Tascón Fernández, 2014). These issues encourage an urgent call for industry-specific or firm-specific models, careful econometrics and prudent partitioning of the data (Nissim & Penman, 2001). However, approaches to partitioning data that financial researchers employ to address the problem of HGSC have tended to be simplistic, such as by dividing a whole sample into different quantiles of certain firm features at which they expect to observe different relationships (e.g. Dichev & Tang, 2009; Dickinson, 2011; Little et al., 2009).

CA, in the form of unsupervised classification assigning objects into unlabeled classes, could be a potential solution. However, most studies have neglected to recognize the potential usefulness of firm partitioning using CA to address the HGSC issue (e.g. Goldfeld & Quandt, 1972; Lin & Ng, 2012). Furthermore, CA and particularly, K-means clustering have inherent shortcomings, especially those relating to multicollinearity and feature weightings, causing incorrect classification of objects (Amorim & Mirkin, 2012; Sambandam, 2003;). These findings reveal the need for a new clustering method that tackles the shortcomings of current clustering techniques to deal with the issue of HGSC. ROWK clustering, a novel clustering technique proposed in this thesis, is such a technique.

Accordingly, four related hypotheses are developed. The first three hypotheses examine factors that affect the performance of CA with respect to the precision of cluster recognition and regression estimations. The findings provide a guide for researchers to consider the feasibility of using CA to solve the problem of HGSC. The fourth hypothesis investigates three channels through which ROWK improves the performance of cluster analysis with respect to the HSGC problem. Table 6-1 summarises the first thesis aim, its associated hypotheses, methodology and findings.

6.2.2 The second thesis aim: To apply the newly proposed method, i.e. ROWK clustering, to mitigate problems of HGSC in earnings persistence models.

Given the theoretical background and empirical evidence of the existence of HGSC in earnings persistence models, the thesis employs earnings persistence models as a potential candidate to apply ROWK clustering. The second thesis aim is addressed by the next five hypotheses H5 to H9. Hypotheses H5 to H8 examine the performance of ROWK clustering to identify the heterogeneities of coefficients of earnings persistence models. The final hypothesis queries whether analysts understand the earnings persistence patterns embedded by ROWK clustering and incorporate these in their earnings forecasts. Table 6-2 summarises the second thesis aim, its associated hypotheses, methodology and findings.

Table 6-1: Summary for the First Thesis Aim

This table summarises the first thesis aim, its associated hypotheses, empirical specifications used to test these hypotheses and the findings.

Thesis Aim 1: To develop a new clustering method that can be applied in financial research to address the problem of HGSC

Hypotheses	**Methodology**	**Results**
H1a: *The higher the clusters' density, the greater the precision for identification of the true clusters when running regressions within clusters. This positive relationship between cluster density and clustering precision is stronger when distances between clusters' centres are lower.* H1b: *The greater the distance between clusters' centres, the greater the precision for identification of the true clusters when running regressions within clusters. This positive relationship between distances of clusters' centres and clustering precision is stronger when clusters' densities are lower.* H1c: *The MAR/MSR when running regressions within clusters is lower when (1) the distances between clusters' centres are greater, or (2) the clusters' densities are higher, or (3) the differences of regression coefficients between clusters are larger.* H2*: When features are highly correlated, the precision for identification of the true clusters is lower, and the MAR/MSR when running regressions within clusters is*	The econometric framework for the proposed new clustering method, ROWK, is introduced. ROWK clustering includes four procedural steps. First, the regression model is identified. Second, features are selected. Then data is pre-processed before being used in the final step to find the optimal weights. Four hypotheses are examined via simulated data sets with different sets of parameters. Hypothesis H1a relates to distances between classes proxied by λ, the parameter that represents the extent of differences between *Class* ξ_1^0 and other classes. Hypothesis H1b relates to the densities of classes as measured by θ. For testing hypothesis H2, all input parameters are the same as in the testing of Hypothesis 1, except for the *within* covariance matrix of features which are generated to differ from the diagonal matrix. For testing	***Consistent with hypothesis H1a (1b***), this study finds that the higher the class density (the distances between class centres), the greater the precision for identification of class membership when running regressions within classes. Additionally, the positive relationship between class density (distances between class centres) and clustering precision is stronger when distances between class centres (class densities) are lower. ***Consistent with hypothesis H1c***, the evidence presented supports the proposition that the mean absolute residuals (*MAR*) decline significantly when (1) distances between class centres increase, (2) class densities increase and (3) differences in regression coefficients between classes are larger. Furthermore, in ***support of hypothesis H2*** and consistent with Sambandam (2003), evidence is presented of the significantly negative influence of multicollinearity on cluster analysis performance. ***Consistent with hypothesis H3***, relative to using unstandardized features, standardization of features is revealed to result in significantly larger (lower) *MARs*.

higher, as compared to the case of low or uncorrelated features. H3a: *When a feature's weight results from differences of clusters' densities measured by the feature relative to those measured by other features, standardization improves the performance of clustering as compared to those of unstandardized features.* H3b*: When a feature's weight results from differences of distances between clusters' centres measured by the feature relative to those measured by other features, standardization decreases the performance of clustering as compared to those of unstandardized features.*	hypothesis H3, two scenarios of feature weights are generated. 100 simulated data samples are generated for each case. K-means clustering is run for these 100 samples, and the average results are presented. T-tests are used to test for the significance of differences	This is observed when a feature's weight results from differences in distances between class centres (class densities) measured by the feature relative to those measured by other features. This raises concerns regarding the effectiveness of standard K-means clustering whereby cluster features are routinely standardized before running cluster analysis. These findings provide a useful guide for researchers to assess the feasibility of conducting cluster analysis to address the HGSC problem. A clear firm pattern indicated by high cluster density, and/or large centroid distances and/or huge discrepancies of regression estimations between clusters are signals of the success of using cluster analysis to address the HGSC problem.
H4: *When features have different degrees of contribution to cluster identification and regression estimation, ROWK outperforms generic K-means (both standardized and un-standardized) with regard to the precision of cluster recognition and regression estimation. The mechanisms underlying the outperformance of ROWK are through these channels. Specifically, ROWK is hypothesized to:* *a, Place more (less) weight on more (less) relevant features.*	To test this hypothesis, three simulated cases are investigated. Each case sheds light on each channel through which ROWK improves the performance of cluster analysis with respect to the HSGC problem. Case 1 includes a simple set of simulated data with uncorrelated features. Case 2 employs the same data as Case 1, but with correlated features. Case 3 analyses a situation where features'	***As hypothesized***, this study finds significant improvements of ROWK relative to K-means clustering and WK through three channels. The channels attributable to these improvements are confirmed. Specifically, ROWK places more (less) weight on more (less) relevant features (Case study 1); reduces the influence of multicollinearity by reducing the weights of irrelevant features which are highly correlated with relevant features (Case study 2) and captures relevance not only by its contribution to cluster

b Reduce the influence of the multicollinearity problem by reducing the weights of irrelevant features which are highly correlated with relevant features. *c Capture relevance not only by its contribution to cluster recognition but also by regression estimation.*	weights come from two sources, i.e. contributions to cluster recognition and to regression estimations. Robustness tests include out-of-sample tests, different distributions of cluster features, unequal class sizes, and with transformed features by factor analysis.	recognition but also by regression estimation (Case study 3). Results are robust to several robustness tests, including out-of-sample tests, different distributions of cluster features, unequal class sizes, and with transformed features by factor analysis.

Table 6-2: Summary for the Second Thesis Aim

This table summarises the second thesis aim, its associated hypotheses, empirical specifications used to test these hypotheses and the findings.

Thesis Aim 2: To apply the newly proposed method, i.e. ROWK clustering, to mitigate problems of HGSC in earnings persistence models.

Hypotheses	**Methodology**	**Results**
H5*: Feature weights identified by ROWK clustering are not equal.*	The research design to apply ROWK clustering to the earnings persistence model is presented. It includes four steps. The first step introduces the regression model of earnings persistence. The second step presents the list of features that are used in clustering. Pre-processing of clustering data is discussed in the third step. The final step executes ROWK clustering to identify three important aspects, i.e. the optimal number of clusters, the optimal cluster weights, and cluster membership on earnings persistence To test hypothesis H5, the *MSR* in the case of equal weights (*MSR_EQW*) is computed. Then we compare the *MSR* of the optimal weights (*MSR_ROWK*) to those of equal weights. If H5	Seventeen clustering features are investigated. Among of them, only 5 features are relevant and receive non-zero weights after running ROWK clustering. These features include earnings volatility (*VOL_IBC_DEF*), absolute value of accruals (*ABS_ACC_DEF*), intangible investment (*INTAN_INV_DEF*), change in profit margin (*ΔPM*) and change in asset turnover (*ΔATO*). Furthermore, ROWK clustering assigns different optimal weights to features, ***consistent with Hypothesis H5*** predicting that firms exhibit different earnings persistence between ROWK clusters. The highest weights are

	is supported, then a significant difference between *MSR_EQW* and *MSR_ROWK* is expected.	assigned to earnings volatility and absolute value accruals, together accounting for nearly 80% of the total weight of all 17 features.
H6*: Firms exhibit different earnings persistence between ROWK clusters*	To test hypothesis H6, the earnings persistence model is added with dummy intercept and slope variables receiving a value equal to 1 for the observations in the lowest earnings persistence cluster. A t-test is used to test for differences in earnings persistence across clusters. This thesis follows the bootstrapping approach employed in Dichev & Tang (2009) to compare Adj R^2 between clusters.	***Consistent with Hypothesis H6***, there are differences in earnings persistence and intercepts between ROWK clusters. Importantly, these differences seem large in magnitude and suggest that cluster membership is economically important. The difference between the highest and lowest intercept coefficients is nearly 5%, which is highly significant (p value <0.0001). The difference between the highest and lowest persistence is 0.467, which is also highly significant (p value < 0.0001).
H7*: ROWK clustering results in larger differences in earnings persistence between clusters and lower earnings prediction errors than a single variable cluster partitioning technique.*	Several benchmark techniques are used to compare with ROWK clustering. The first benchmark model is to run the earnings persistence model for the full sample. The second benchmark model is standard K-means using all listed features. The third benchmark model is again K-means clustering but with only listed features that have non-zero ROWK optimal weights. The fourth benchmark model involves normal partitions using only one cluster feature. The final benchmark model is WK as developed by Huang et al. (2008).	Partitioning the sample based on a single feature results in lower coefficients' differences of both intercept and earnings persistence across groups than those of ROWK clustering. As regard to earnings predictability, regression of earnings persistence within each cluster found by ROWK clustering achieves the lowest *MSR* at 0.00385, which is significantly lower than those conditional on each feature, ***consistent with hypothesis H7***.
H8: *The clusters found in ROWK clustering exhibit heterogeneities with respect to accounting conservatism,*	Fama-French 12-industry classifications are used to divide firms into different industries. The thesis follows Dickinson (2011) to identify firm life cycles using information from cash flow statements. Following the work of Jansen et al., 2012, earnings	***Consistent with hypothesis H8***, the clusters found in ROWK clustering exhibit heterogeneities with respect to accounting conservatism, earnings management, firm life cycles and industry membership. In addition, it upholds

earnings management, firm life cycles and industry membership.	management is identified using the signs of ΔPM and ΔATO. This thesis follows Khan & Watts (2009) to estimate conditional conservatism (*C_CON*). Following Givoly & Hayn (2000) and Chen et al., (2014), unconditional conservatism (*U_CON*) is measured as negative cumulative non-operating accruals.	the thesis's argument in favour of using financial ratios as the input list for ROWK clustering rather than directly incorporating the direct proxies that contain measurement errors. In additional tests, the earnings persistence model is run for each industry firm's life cycle stage, category of earnings management and decile of *U_CON* and *C_CON*. The results show that the *MSR*s of these cases are significantly larger compared to those experienced using ROWK clusters.
H9: Information from ROWK's cluster identification predicts analyst forecast errors.	Two approaches are presented to test this hypothesis. The first approach uses portfolio sorting, and the second approach employs a model of analyst forecast errors. Two-way portfolio sorting procedure is employed to control for dispersion of earnings. The analyst forecast errors model takes account of serial correlation of analyst forecast errors.	Evidence from two-way portfolio sorting shows that analysts' forecasts only partially incorporate the information from cluster patterns in the short run, while ignoring impacts of these patterns on long-term future earnings. Results from the analysis of the analyst forecast errors model uphold the findings from the two-way sorting portfolios. As a result, conditioning on such information permits the identification of reliable and economically important patterns in analyst forecast errors. These results ***partly support hypothesis H9*** in the sense that information from ROWK's cluster identification only modestly predicts analyst forecast errors in the short-term, but strongly predicts analyst forecast errors in the long-term.

6.3 CONTRIBUTIONS OF THE THESIS

This study contributes to both CA and the financial literature in several important ways. While the consequences resulting from the violation of the constant-coefficients assumption are recognized by researchers, existing studies pay little attention to the development of techniques to solve it (Richardson et al., 2010). This thesis is the first study to systematically apply CA to address the problem of HGSC within financial research. "Systematically" in this context means that the thesis does not simply apply standard techniques of CA to group firms. Instead, the thesis comprehensively examines factors impacting the performance of cluster analysis to address the problem of HGSC. Then it discusses shortcomings of CA and proposes a new method to cope with these drawbacks. Finally, it illustrates the utility of this proposed method by examining its performance using both simulated and real data.

Note that CA was first introduced and developed in the natural sciences from the need to classify data into homogeneous objects. It has been applied more recently in business to market segmentation studies (Dolnicar, 2002). Yet, very few studies apply CA to finance. Even in those finance studies that do apply cluster analysis, most of them only employ K-means clustering as a supplemental component. This thesis pioneers the application of CA to group homogeneous firms, and mitigate the problem of HGSC. Therefore, it provides a guide for future researchers in the finance discipline to apply CA to their field of study.

This thesis take a further step of being the first study to mitigate the inherent drawbacks of CA that have not been sufficiently recognized and adjusted for in much of the past research (e.g. Epure et al., 2011, Lee et al., 2004, Li & Li, 2008). One advantage is that by using the new technique developed in this thesis, researchers can gain more precise coefficient estimates in predictive models. In order to achieve this objective, the thesis proposes a novel clustering technique, called Regression-Oriented-Weighted-K_means clustering (ROWK). It combines K-means clustering and regression analysis. On the one hand, the K-means algorithm iteratively assigns similar observations into clusters. On the

other hand, the *MAR/MSR* from running regressions is used to guide the process of weight adjustment in clustering and mitigate the problem of multicollinearity. Accordingly, the thesis introduces academic researchers to a useful new tool that employs cluster analysis to address the problem of HGSC.

This thesis also contributes to research in the finance discipline by introducing a standard procedure to apply CA to solve financial problems. Four steps of executing ROWK clustering along with the algorithm of finding optimal weights are comprehensively presented, facilitating future applications of ROWK clustering to financial research. The proposed method has the advantage of being easy to understand and execute using typical statistical analysis programs. Hence the thesis equips researchers with a powerful tool to enhance regression results whenever there are indications of heterogeneous coefficients, which are frequently problematic in financial research.

Despite recent efforts to address the issue of heterogeneous parameters, most studies focus merely on the regression side, ignoring underlying reasons for the problem, i.e. cluster patterns (e.g. Ando & Bai, 2016; Lin & Ng, 2012). To find the optimal weights of clustering features, the ROWK procedure proposed in this study helps to distinguish those factors that are essential to identity cluster patterns. It equips researchers with a powerful tool to empirically explore which features are more important. For example, this thesis determines that, of the examined factors, earnings volatility and accruals are the most relevant to distinguish patterns of earnings persistence.

This thesis is also the first study to investigate (though simulated data) the factors that impact upon the performance of CA in order to resolve the breach of the homogeneous coefficients assumption. This study provides a novel guide for researchers to consider the feasibility of adopting ROWK clustering. Density or compactness within clusters, distance between cluster centres, and the level of correlation between variables are found to be among the factors that influence the performance of CA in general and ROWK in particular to address problems of heterogeneous coefficients across groups of firms.

Additionally, this thesis makes an original contribution with respect to the application of CA (more precisely ROWK clustering) to identify patterns of earnings persistence in

business firms. The thesis provides evidence of HGSC on earnings persistence, and shows the usefulness of using information from clusters identified by ROWK clustering to predict analyst forecast errors. The thesis's findings will be of particular interest to both academic researchers and investors who have concerns surrounding earnings forecasting. Earnings forecasts are contextual, and CA reveals underlying firm clusters at which different contexts are likely to be observed. Consequently, it helps researchers and investors/analysts to achieve higher accuracy in earnings forecasts. The results reported in this thesis confirm that the precision of earnings forecasts can be significantly improved by incorporating information from ROWK clustering.

The implications of the findings are of widespread importance to future research that is associated with formation of portfolios or identification of benchmark portfolios. A conventional practice in quantitative business research is the assignment of firms into portfolios by quantiles of an examined variable such as firm size or market to book ratio. The thesis reveals the superiority of CA as a system of portfolio formation because it takes into account variables' distributions and is able to deal with several examined variables simultaneously.

Finally, given that ROWK clustering successfully addresses the issue of HGSC in both simulated and real data, the thesis creates promising opportunities for future studies to apply ROWK clustering in other examined models when there is concern of HGSC. Some suggestions for potential areas that could benefit from the application of ROWK clustering will be discussed later in Section 6.5.

6.4 LIMITATIONS OF THE THESIS

Not surprisingly, given the aim of developing a new method of clustering to deal with the problem of HGSC, the thesis acknowledges several limitations relating to its methodology. Particularly, to simplify the econometric framework of ROWK clustering, regression models are assumed to suffer only from the problem of HGSC, while other econometric problems, including endogeneity, are assumed to already be resolved. It is important to recognise that ROWK clustering ***is not*** designed to cope with the problem

of endogeneity. The problem of endogeneity must be addressed before conducting ROWK clustering. This requirement is able to be addressed through the abundant literature on solutions to endogeneity. In the case when the endogeneity problem is not resolved, CA in general and ROWK clustering in particular do not guarantee to correctly identify the data patterns. The problem of endogeneity is only mitigated when the effects of the omitted variable (that causes endogeneity) on the examined variables are also one of the causes of HGSC. For example, a well-known example is the relationship between education and income. This relationship cannot be correctly empirically estimated due to omitted variable bias, i.e. ability. However, if the clusters identified through ROWK exhibit heterogeneities on ability, running a regression within each cluster can reduce the bias caused by the omitted variable of ability.

Further methodological issues are encountered in the algorithm to find optimal weights of features. As discussed in Section 3.2.1.1.2, in the ROWK procedure proposed in this thesis, finding solutions for an optimization problem is extremely challenging. While ROWK aims to find the optimal features' weights, the optimization function relates to the regression criterion, i.e. *MAR/MSR*. Consequently, the feature weights are not embedded in the optimization function, making it impossible to employ (numerical) derivatives to form new guesses for the parameter value. To tackle this problem, the thesis introduces an algorithm as detailed in Section 3.2.1.1.2. Even though the proposed optimization algorithm performs well in the empirical result, it is an extremely time consuming process. Finding new guesses, identifying initial set of weights, and the unknown optimal number of clusters are reasons that cause ROWK's algorithm to repeat the process many times, resulting in huge amounts of time to reach to the end of the process, specifically when the number of cluster feature is high. Particularly for the application part in the thesis, 17 cluster features are used. The time to run ROWK clustering with these 17 features for the earnings persistence model took several hours. Nevertheless, with the rapid advancement of technology enabling more powerful computing power, the severity of this issue can be expected to reduce measurably.

Regarding data, the scope of the thesis focuses exclusively on simulation data to examine the performance of ROWK clustering (Chapter 4). A clearer picture may result from

using various real data sets to examine the performance of ROWK clustering, rather than using simulated data sets. However, given the time constraints of producing a thesis and the data availability limitations of real data sets, the use of simulated data sets is chosen. Simulated data sets have the advantage of flexibility of parameter adjustment, providing a feasible way to investigate the three channels through which ROWK improves the performance of CA with respect to the HSGC problem.

For the application of ROWK clustering to earnings persistence (Chapter 5), the thesis conducts the analysis using US data: Financial statement data from COMPUSTAT, stock data from CRSP and analyst forecast data from Thomson Reuter (IBES). The main limitation is that only annual financial statement data is available. Thus, estimated models using annual data do not capture any changes in cluster patterns (if any) within a given year. Hence, the results could improve further by achieving lower *MAR/MSR* or more finely distinguished regression coefficients across clusters when higher frequency data sets (monthly or quarterly data etc.) are used.

The next sub-section presents several potential extensions to the work presented in this thesis.

6.5 SUGGESTIONS FOR FUTURE RESEARCH

Several of the limitations of the thesis discussed in the previous section give rise to suggestions as to how the research might be extended in the future. These suggestions focus on improved methodology, the data and potential applications of ROWK clustering.

With respect to methodology, it would be useful to investigate whether other clustering techniques could replace K-means in the ROWK clustering. Recall that ROWK clustering employs K-means with different feature weights to minimise the *MAR/MSR* of the regression equation. The reason for choosing K-means clustering to group firms is discussed in Chapter 3 (Methodology). However, the choice of using other clustering techniques is possible. Hierarchical clustering, a method of CA which seeks to build a hierarchy of clusters, is an example. Another potential clustering technique to replace K-means is expectation–maximisation clustering (*EM clustering*) which employs an *EM*

algorithm to group subjects[69]. The crucial point is that while the chosen clustering technique is used to group firms, the weight of features for clustering needs to be identified by the regression criteria (e.g. to minimise the regression *MAR/MSR*).

Another methodological suggestion for future research stems from the optimized criterion. In Chapter 4 of the thesis, feature weights are identified to minimise the regression *MAR* to mitigate the effect of outliers or noise features of simulated data. However, the thesis does not rule out the possibility of using other regression criteria. For example, when there is no concern of effects from outliers, *MSR* could be a more appropriate regression criterion as in the case of applying ROWK clustering to earnings persistence described in Chapter 5. Additionally, when the ultimate objective is to distinguish the regression's coefficient estimations across clusters, maximising mean squared distances between coefficient estimations across clusters is an appropriate choice.

In terms of the database itself, since firms can move into different clusters within a year, annual data may not be optimal. Thus, this study could be broadened and strengthened by employing quarterly data. The results of earnings persistence patterns could also be broadened using global data instead of US data. Whether the earnings persistence pattern identified by ROWK clustering is consistent across different countries is an open question that may be addressed in future work. Moreover, it is also possible to add more features to run ROWK clustering. For example, firms display more earnings persistence when their boards are dedicated to advising, compared with those whose boards are not focused on advising (Hsu & Hu, 2016). Thus, adding a feature to proxy for the degree that a firm's board is dedicated to advising could enhance the performance of ROWK clustering.

The problem of HGSC in regression estimation is well documented in finance research (Lin & Ng, 2012). Given the superior performance of ROWK to address the problem of HGSC on simulated data and earnings persistence, future finance research may benefit from the application of the ROWK procedure whenever there are suspicions that

[69] See Murtagh & Contreras (2012) for a review of hierarchical clustering. See Yang, Lai, & Lin (2012) for a review of EM clustering.

regression coefficients are group-specific. Some potential applications of ROWK in finance are proposed next.

Research efforts to develop a robust approach to measure firm life cycle stages are sparse and constrained to simple identities (Dickinson, 2011). For example, firms falling into the same phase of life cycle are likely to have the same age and size, hence these are common proxies for life cycles. Dickinson (2011) further employs cash flow patterns to proxy for firm life cycle. She conjectures that firms exhibit some similarities as they evolve across their life cycles. Moreover, regression estimations relating to earnings forecasts based on firm life cycles exhibit group-specific coefficients. Therefore, cluster analysis, and specifically ROWK may be a potential solution to identify the stages within firm life cycles.

Discretionary accruals estimates could also be improved by the use of ROWK. Researchers typically estimate discretionary accruals by running a regression of non-discretionary accruals within each industry. However, this practice generally leads to imprecise estimates due to the different relations between the dependent variable and its determinants within each industry (Fairfield et al., 2003). Therefore, ROWK can be used to combine the model of non-discretionary accruals and weighted K-means to identify clusters where the coefficients in the regression model are homogeneous.

Capital expenditure is a widely examined topic in the literature (Titman, Wei, & Xie, 2004). When a firm increases its capital investment, the market may interpret the information as favorable or unfavorable. Evidence from empirical research strongly upholds the negative long-term stock reactions to capital investments. For example, when firms issue equity to finance their investments, their stock returns are generally negative (e.g. Loughran & Ritter, 1995). On the other hand, during a stock repurchase, which is an indicator of decreased investment, firms' stock prices increase significantly (e.g. Ikenberry, Lakonishok, & Vermaelen, 1995). Numerous studies document a negative relationship between a firm's capital expenditures and long-run stock returns (e.g. Fairfield et al., 2003; Kogan & Papanikolaou, 2013; Ozdagli, 2012; Titman et al., 2004),

or a negative relationship between a firm's capital expenditure and future profitability (e.g. Bauman, 2014; Dickinson & Sommers, 2012; Fairfield et al., 2003; Sunder, 1980).

One explanation of the negative relationship between increased investment and future earnings is diminishing marginal returns on investments, arising because more profitable investments tend to be exploited before less profitable investments (Fairfield et al., 2003). However, several characteristics are useful to assess whether a firm investment is successful. Firms characterized by different stages of life cycle, industries, levels of free cash flows, financial leverage, types of control (i.e. manager, owner), and growth opportunities should show differences in the performance of their investments (Kogan & Papanikolaou, 2013; Giroud & Mueller, 2011; Perfect, Peterson, & Peterson, 1995; Titman et al., 2004)[70]. Cluster analysis, which aims at placing observations into different clusters such that observations in the same cluster are homogeneous to each other but are different from ones in other groups (Fred & Jain, 2005), could be employed to identify firm clusters whereby firms in different clusters exhibit different investment performances.

[70] According to Titman et al. (2004) firms with higher investment expenditures are likely to have greater investment opportunities. This is consistent with the findings of Kogan & Papanikolaou (2013) who further explain that when a given firm implements an investment project the market will make an upward revision of the firm's growth opportunities. The effect of increased investment is stronger if there is evidence that firms are underinvesting, such as firms with marginal Tobin's q greater than one (Perfect, et al., 1995). There is empirical evidence supporting this "good" story of investments. This evidence mainly focuses on announcements of major capital investments (e.g. Vogt, 1997), so it is likely that these findings are biased because firms have tendencies to announce publicly only favorable investment expenditures. Moreover, observations of higher short-term stock prices are usually associated with higher capital investments, which could be explained from the fact that firms find it easier to increase investment expenditures when their stock prices are high (Titman et al., 2004).

In contrast, there are also numerous reasons why increased investment expenditures are considered to be unfavorable. Firms with higher free cash flows have greater chances for their managers to spend too much money on value-destroying corporate acquisitions (Giroud & Mueller, 2011). This is consistent with the literature that shows firms experience negative benchmark-adjusted long-run returns (i.e. the Fama-French three factors and the Carhart momentum factor) when they substantially increase their capital investment (Titman et al., 2004).

The above section reveals considerable opportunity for the application of ROWK clustering to solve current problems in finance. The opportunities may even extend to other disciplines. We leave this to researchers to explore further.

APPENDIX

APPENDIX A: METHODOLOGY

A1: θ_k^2 IS THE EXPECTATION OF THE MEAN SQUARED DISTANCES BETWEEN MEMBERS OF A CLASS k TO ITS CENTRE

Proof:

$$\sum_{i\in\xi_k^0}\sum_{v=1}^{V}\left(z_{i,v}^{k}-c_{k,v}\right)^2 = \sum_{i\in\xi_k^0}\sum_{v=1}^{V}\in_{i,v}^{k\ 2}$$

$$= \sum_{v=1}^{V}\sum_{i\in\xi_k^0}\in_{i,v}^{k\ 2}$$

Take the expectation of each side of the equation:

$$E\left[\sum_{i\in\xi_k^0}\sum_{v=1}^{V}\left(z_{i,v}^{k}-c_{k,v}\right)^2\right] = E\left[\sum_{v=1}^{V}\sum_{i\in\xi_k^0}\in_{i,v}^{k\ 2}\right]$$

$$= \sum_{v=1}^{V}E\left[\sum_{i\in\xi_k^0}\in_{i,v}^{k\ 2}\right]$$

$$= \sum_{v=1}^{V}N_k^0\sigma_{k,v}^2$$

$$= N_k^0\sum_{v=1}^{V}w_{k,den\_v}^2\frac{\theta_k^2}{\sum_{v=1}^{V}w_{k,den\_v}^2}$$

$$= N_k^0\theta_k^2$$

Then $\theta_k^2 = \frac{E\left[\sum_{i\in\xi_k^0}\sum_{v=1}^{V}\left(z_{i,v}^{k}-c_{k,v}\right)^2\right]}{N_k^0}$ (Q.E.D.)

A2: THE MODEL OF EARNINGS PREDICTION

The testing of Hypothesis H7 compares the performance of ROWK clustering with respect to earnings persistence and earnings predictability to that of other benchmark techniques. To test for earnings predictability, the thesis also includes an earnings prediction model that incorporates many earnings predictors covered by recent studies of earnings predictability. Specifically, the model of one-year-ahead earnings is as follows:

(6.1)

$$\Delta\text{Earnings}_{i,t+1} = \alpha + \beta_1 \text{Earnings}_{i,t} + \beta_2 \Delta\text{Earnings}_{i,t} + \beta_3 \Delta\text{COA}_{i,t} + \beta_4 \Delta\text{NCOA}_{i,t} + \beta_5 \Delta\text{ATO}_{i,t} + \beta_6 \Delta\text{PM}_{i,t} + \beta_7 EM\_UP_{i,t} + \beta_8 \Delta\text{PM}_{i,t} * EM\_UP_{i,t} + \beta_9 EM\_DN_{i,t} + \beta_{10} \Delta\text{PM}_{i,t} * EM\_DN_{i,t} + \varepsilon_{i,t+1}$$

where:

$\text{EM\_UP}_{i,t} = 1$ if $\Delta\text{PM}_{i,t} > 0$ and $\Delta\text{ATO}_{i,t} < 0$ and $EM\_DN_{i,t} \neq 1$, and 0 otherwise, and

$\text{EM\_DN}_{i,t} = 1$ if $\Delta\text{PM}_{i,t} < 0$ and $\Delta\text{ATO}_{i,t} > 0$ and $EM\_UP_{i,t} \neq 1$, and 0 otherwise.

It is well-accepted that earnings mean-revert (Bauman, 2014; Fairfield & Yohn, 2001), so it is expected that $\beta_1 < 0$. $\Delta\text{Earnings}_{i,t}$ is included in the predictive model to examine the autocorrelation of ΔEarnings beyond that captured by mean reversion (Bauman, 2014). Growth in net operating assets (ΔNOA_t) captures the effect of investments in future profitability (Dickinson & Sommers, 2012). As ΔNOA_t is considered a broad measure of accruals, it captures the less persistent accruals component of earnings as documented in Sloan (1996) and Richardson et al., (2010)[71]. ΔNOA_t could then be further divided into two components, i.e. change in current operating accruals (ΔCOA_t) and change in noncurrent operating accruals ($\Delta NCOA_t$) which, according to Bauman (2014) have different effects on future profitability. ΔNOA_t or its components (i.e. ΔCOA_t, $\Delta NCOA_t$) represent the effect of investments or accruals on future profitability. The accruals component of earnings has less persistence than cash flow components as documented by Sloan (1996). In addition, investment negatively impacts firm profitability due to diminishing marginal returns to new investment (Fairfield et al., 2003), and the over-investment problem (Cooper et al., 2008). Therefore, it is expected that $\beta_3 < 0$ and $\beta_4 < 0$.

To capture the incremental information of the **Δ**Earnings decomposition documented in Fairfield & Yohn (2001), change in asset turnover (**Δ**ATO) and change in profit margin (**Δ**PM) are included. However, instead of the **Δ**ATO and **Δ**PM defined in Fairfield &

[71] Richardson et al.(2010) demonstrate the link between the accruals measure in Sloan (1996) and the broad measure of accruals (**Δ**NOA) as follows:

ΔNOA = Accruals – ΔTax Payable + Depreciation + Δ Net Non-Current Operating Assets

Yohn (2001), this thesis uses the first difference of ATO and PM to avoid multicollinearity with **Δ**RNOA[72]. According to Soliman (2008), this helps to avoid the potential omitted variable bias due to the omission of **Δ**RNOA.

Fairfield & Yohn (2001) find that an increase in ATO that denotes an increase in the productivity of a firm's assets also signifies an increase in future profitability. Therefore, it is expected that $\beta_5 > 0$. Furthermore, they also find that **Δ**PM alone reveals no further information about future profitability. They argue that there are two opposite sources underpinning **Δ**PM, which are changes in operating efficiency and earnings management. While an increase (decrease) in operating efficiency tends to increase (decrease) future profitability, an increase of profit margin resulting from upward earnings management could result in a reversed direction. Therefore, the failure of **Δ**PM to predict changes in RNOA as has been documented in Fairfield & Yohn (2001) should be explained by the failure to disentangle the effects of operating efficiency and earnings managements.

To address this issue, this thesis employs an earnings management diagnostic introduced by Jansen et al. (2012) who argue that the opposite signs of coefficients for **Δ**ATO and **Δ**PM are a signal of earnings management. In equation 6.1, we define *EM_UP (EM_DN)* = 1 when there is upward (downward) earnings and 0 otherwise. The proxy for earnings management will be discussed further in the next section. By this construction, β_7 and β_9 represent the fixed reversal effects of upward and downward earnings management, respectively. Therefore, it is expected that $\beta_7 < 0$ and $\beta_9 > 0$. β_6 denotes the effect of changes in PM due to changes in pure operating efficiency, thus it is expected that $\beta_6 > 0$. β_8 (β_{10}) represents the influence of upward (downward) earnings on the effect of **Δ**PM. Thus, it is expected that $\beta_8 < 0$ and $\beta_{10} < 0$. It is also expected that $(\beta_6 + \beta_8) < 0$ and $(\beta_6 + \beta_{10}) < 0$.

A3: INDUSTRY CLASSIFICATION

[72] Fairfield and Yohn (2001) define the change in assets turnover and the change in profit margin as follows:

$$The\ change\ in\ assets\ turnover = \Delta ATO_t * PM_{t-1}$$
$$The\ change\ in\ profit\ margin = \Delta PM_t * ATO_{t-1}$$

They argue that these measures allow for cross-sectional differences in the mix of ATO and PM.

The thesis chooses Fama-French 12-industry classifications to divide firms into different industries. Table Appendix A1 presents details of the Fama-French 12-industry classification.

Table A1: Fama-French Twelve Industry Classification

Abbreviation	Description
NoDur	Consumer NonDurables -Food, Tobacco, Textiles, Apparel, Leather, Toys
Durbl	Consumer Durables - Cars, TVs, Furniture, Household Appliances
Manuf	Manufacturing -Machinery, Trucks, Planes, Off Furn, Paper, Com Printing
Enrgy	Oil, Gas, and Coal Extraction and Products
Chems	Chemicals and Allied Products
BusEq	Business Equipment - Computers, Software, and Electronic Equipment
Telcm	Telephone and Television Transmission
Utils	Utilities
Shops	Wholesale, Retail, and Some Services (Laundries, Repair Shops)
Hlth	Healthcare, Medical Equipment, and Drugs
Fin	Finance
Other	Other - Mines, Constr, BldMt, Trans, Hotels, Bus Serv, Entertainment

A4: FIRM LIFE CYCLES

To identify cycles of firm life for the test of hypothesis H7, the thesis follows Dickinson (2011), who identifies firm life cycles using information from cash flow statements. Using the sign of three net cash flow activities (operating, investing, and financing), firms are assigned into one of the five stages, i.e. introduction, growth, mature, shake-out and decline. Life cycles as proxied by cash flow patterns are predicted to identify differential profitability persistence between clusters. Table A2 presents in detail the life cycle classification using cash flow patterns. For example, a firm falls in the growth stage if its signs of operating CF, investing CF and Financing CF are negative, negative, positive (-,-,+) respectively.

Table A2: Firm Life Cycles Identified using Cash Flow Patterns (Dickinson, 2011)

Signs/Stages	Introduction	Growth	Mature	Shake-out	Shake out	Shake-out	Decline	Decline
Operating CFs	-	+	+	-	+	+	-	-
Investing CFs	-	-	-	-	+	+	+	+
Financing CFs	+	+	-	-	+	-	+	-

APPENDIX B: ROWK AND THE PROBLEM OF HGSC- SIMULATION RESULTS

B1: CDA WITH TRUE MEMBERSHIP- CASE 1

Table B1 presents the results from running CDA for observations in Case 1 with true membership. It means that observation identification is already known in advanced. CDA is performed on the list of features with the knowledge of true class membership to derive canonical variables. PROC SCANDIC in the SAS program is used to run CDA. It can be seen that the first canonical variable (Can 1) accounts for over 90% of total eigenvalues. Expectedly, when looking at total-sample standardized canonical coefficients, Z_3 contributes most to the variance of Can 1.

Table B1

Canonical Discriminant Analysis (Figure 4.2-Case 1)

Eigenvalues of Inv(E)*H = CanRsq/(1-CanRsq)				Test of H0: The canonical correlations in the current row and all that follow are zero				
Can	Eigen-value	Propor-tion	Cumula-tive	Likeli-hood Ratio	Approx-imate F Value	Num DF	Den DF	Pr > F
1	5.4698	0.9333	0.9333	0.1074	794.45	20	16554	<.0001
2	0.1967	0.0336	0.9668	0.6946	162.55	12	13208	<.0001
3	0.1249	0.0213	0.9881	0.8312	161.2	6	9986	<.0001
4	0.0695	0.0119	1	0.935	173.45	2	4994	<.0001

Total-Sample Standardized Canonical Coefficients				
Var	Can1	Can2	Can3	Can4
Z1	-0.0276	0.9203	-0.0404	-0.5648
Z2	0.1241	0.3081	0.9529	0.3884
Z3	2.4292	-0.1866	0.0177	-0.4873
Z4	0.2298	0.4632	-0.5532	0.8492
Z5	0.0183	-0.0578	0.0234	0.025

Class Means on Canonical Variables				
Class	Can1	Can2	Can3	Can4
1	4.0002	-0.1068	-0.1711	0.2325
2	-1.8873	0.3835	0.3656	0.3257
3	1.0326	0.5206	0.0832	-0.4055
4	-0.6387	-0.7414	0.3171	-0.1495
5	-2.5068	-0.0559	-0.5947	-0.0032

N		5000	DF Total	4999
Variables		5	DF Within Classes	4995
Classes		5	DF Between Classes	4
Wilks' Lambda	Pr > F	<.0001		
Pillai's Trace	Pr > F	<.0001		
Hotelling-Lawley Trace	Pr > F	<.0001		
Roy's Greatest Root	Pr > F	<.0001		

B2: FREQUENCY OF CLASS MEMBERSHIP BY CLUSTER (CASE 1)

Table B2 presents the frequency of class membership by ROWK clusters for firms in Case 1. The label p_i^{verj} denotes purity version *j* for cluster/class *i*. With respect to class purity_version3, 71.3% of members are correctly assigned. Results remain unchanged for purity versions 1 and 3. In contrast, only 57.48% of members are correctly assigned when using conventional *UNSTD_K*. Performance of *STD_K* is even poorer with only 49.56% members precisely assigned. This is consistent with the fact that the standard

deviation of z_3 is slightly larger than those of other features (i.e. the different weights of features come mainly from differences in distances between class centres).

Table B2

Frequency of Class Membership by Cluster (Case 1)

Panel A: ROWK Clustering								
Class/Clus	1	2	3	4	5	# of Obs	p_i^{ver2}	p_i^{ver3}
1	0	946	1	0	53	1000	0.946	0.94223
2	272	0	559	133	36	1000	0.559	0.55844
3	5	57	68	98	772	1000	0.772	0.75024
4	83	1	169	593	154	1000	0.593	0.65237
5	697	0	204	85	14	1000	0.697	0.65941
# of obs.	1057	1004	1001	909	1029	5000		
p_i^{ver1}	0.65941	0.94223	0.55844	0.65237	0.75024			
$p_{overall}^{ver1}$	0.7134							
$p_{overall}^{ver2}$	0.7134							
$p_{overall}^{ver3}$	0.71254							
Panel B: Unstandardized K-means Clustering								
Class/Clus	1	2	3	4	5	# of Obs	p_i^{ver2}	p_i^{ver3}
1	1	1	14	955	29	1000	0.955	0.81904
2	201	421	170	3	205	1000	0.421	0.46987
3	112	103	114	190	481	1000	0.481	0.51334
4	148	209	493	18	132	1000	0.493	0.51461
5	581	162	167	0	90	1000	0.581	0.55705
# of Obs.	1043	896	958	1166	937	5000		
p_i^{ver1}	0.55705	0.46987	0.51461	0.81904	0.51334			
$p_{overall}^{ver1}$	0.5862							
$p_{overall}^{ver2}$	0.5862							
$p_{overall}^{ver3}$	0.5748							

Table B2 (cont)

Panel B: Standardized K-means Clustering								
Class/Clus	1	2	3	4	5	# of Obs	p_i^{ver2}	p_i^{ver3}
1	10	915	29	16	30	1000	0.915	0.81478
2	155	14	219	296	316	1000	0.316	0.32016
3	141	163	130	137	429	1000	0.429	0.43465
4	149	31	321	398	101	1000	0.398	0.39641
5	477	0	255	157	111	1000	0.477	0.5118
# of Obs.	932	1123	954	1004	987	5000		
p_i^{ver1}	0.5118	0.81478	0.33648	0.39641	0.43465			
$p_{overall}^{ver1}$	0.508							
$p_{overall}^{ver2}$	0.507							
$p_{overall}^{ver3}$	0.4956							

B3: CDA WITH TRUE MEMBERSHIP- CASE 1

Table B3 presents the results from running CDA for observations in Case 2 with true membership. It means that observation identification is already known in advance. CDA is performed on the list of features with the knowledge of true class membership to derive canonical variables. PROC SCANDIC in the SAS program is used to run CDA. Similar to those of Case 1, the first canonical variable (Can 1) accounts for over 90% of total eigenvalues. Expectedly, when looking at total-sample standardized canonical coefficients, z_3 contributes most to the variance of Can 1.

Table B3
Canonical Discriminant Analysis (Figure 4.8-Case 2)

Eigenvalues of Inv(E)*H = CanRsq/(1-CanRsq)				Test of H0: The canonical correlations in the current row and all that follow are zero				
Can	Eigen-value	Propor-tion	Cumula-tive	Likeli-hood Ratio	Approx-imate F Value	Num DF	Den DF	Pr > F
1	6.3762	0.9384	0.9384	0.0918	872.46	20	16554	<.0001
2	0.2071	0.0305	0.9689	0.6776	174.38	12	13208	<.0001
3	0.1243	0.0183	0.9872	0.8180	175.85	6	9986	<.0001
4	0.0873	0.0128	1	0.9197	217.97	2	4994	<.0001

Total-Sample Standardized Canonical Coefficients				
Variable	**Can1**	**Can2**	**Can3**	**Can4**
Z1	-0.026	0.8175	0.1269	-0.7006
Z2	0.1343	0.1117	0.9763	0.4404
Z3	2.9614	-0.2429	-0.0322	-0.5145
Z4	-0.5869	0.7634	-0.4419	0.8926
Z5	0.0172	-0.0498	0.0023	0.0394

Table B3(cont)
Class Means on Canonical Variables

Class	**Can1**	**Can2**	**Can3**	**Can4**
1	4.1398	0.0434	-0.2045	0.2899
2	-2.2159	0.305	0.4367	0.3294
3	1.1964	0.435	0.1668	-0.4795
4	-0.2371	-0.8626	0.1754	-0.1132
5	-2.8832	0.0793	-0.5744	-0.0266

N		5000	DF Total	4999
Variables		5	DF Within Classes	4995
Classes		5	DF Between Classes	4
Wilks' Lambda	Pr > F	<.0001		
Pillai's Trace	Pr > F	<.0001		
Hotelling-Lawley Trace	Pr > F	<.0001		
Roy's Greatest Root	Pr > F	<.0001		

B4: FREQUENCY OF CLASS MEMBERSHIP BY CLUSTER (CASE 2)

Table B4 presents the frequency of class membership by ROWK clusters for firms in Case 2. The label p_i^{verj} denotes purity version j for cluster/class i. Regarding cluster purity version 3, 60% of observations are correctly assigned. The corresponding values for *STD_K* and *UNSTD_K* are lower at 43.7% and 47.9% respectively. Results remain unchanged for purity versions 1 and 3. This evidence lends further support for hypothesis H4b. By mitigating the problem of multicollinearity, ROWK clustering improves the precision of cluster identification, and consequently improves the results of regression estimations.

Table B4
Frequency of Class Membership by Cluster (Case 2)

Panel A: ROWK Clustering							
Class/Clus	1	2	3	4	# of Obs	p_i^{ver2}	p_i^{ver3}
1	45	953	0	2	1000	0.953	0.8999
2	79	0	672	249	1000	0.672	0.4274
3	790	96	17	97	1000	0.79	0.6700
4	226	10	137	627	1000	0.627	0.5268
5	39	0	746	215	1000	0.746	0.4745
# of Obs.	1179	1059	1572	1190	5000		
p_i^{ver1}	0.6700	0.8999	0.4745	0.5268	0		
$p_{overall}^{ver1}$	0.6232						
$p_{overall}^{ver2}$	0.7576						
$p_{overall}^{ver3}$	0.5998						
Panel B: Unstandardized Clustering							
Class/Clus	1	2	3	4	# of Obs	p_i^{ver2}	p_i^{ver3}
1	932	13	38	17	1000	0.932	0.7259
2	23	213	433	331	1000	0.433	0.3407
3	256	175	469	100	1000	0.469	0.3690
4	68	197	170	565	1000	0.565	0.4564
5	5	609	161	225	1000	0.609	0.5046
# of Obs.	1284	1207	1271	1238	5000		
p_i^{ver1}	0.7259	0.5046	0.3690	0.4564			
$p_{overall}^{ver1}$	0.515						
$p_{overall}^{ver2}$	0.6016						
$p_{overall}^{ver3}$	0.4793						
Panel C: Standardized Clustering							
Class/Clus	1	2	3	4	# of Obs	p_i^{ver2}	p_i^{ver3}
1	60	883	29	28	1000	0.883	0.72735
2	364	33	225	378	1000	0.378	0.29647
3	454	222	153	171	1000	0.454	0.38312
4	134	63	353	450	1000	0.45	0.35294
5	173	13	566	248	1000	0.566	0.42685

# of Obs.	1185	1214	1326	1275	5000
p_i^{ver1}	0.38312	0.72735	0.42685	0.35294	
$p_{overall}^{ver1}$	0.4706				
$p_{overall}^{ver2}$	0.5462				
$p_{overall}^{ver3}$	0.4374				

B5: CDA WITH TRUE MEMBERSHIP- CASE 3

Table B5 presents the result of running CDA for observations in Case 3 with all five features and five classes. The CDA is performed on the list of five features with the knowledge of true class membership to derive canonical variables. PROC SCANDIC in the SAS program is used to run CDA. Contrary to the results of Cases 1 and 2, the first canonical variable (Can 1) only accounts for over 40% of total eigenvalues. The second canonical variable (Can 2) also contributes to nearly the same proportion as in Can 1. This is reasonable since both z_3 and z_4 are built to distinguish class membership. When looking at total-sample standardized canonical coefficients, z_3 contributes most to the variance of Can 2, while z_4 is the dominant component of Can 1. Expectedly, Can 1 is solely computed to identify membership of *Class* ξ_3 and *Class* ξ_5 while Can 2 is computed to identify membership of *Class* ξ_1, *Class* ξ_2 and *Class* ξ_4.

Table B5
Canonical Discriminant Analysis (Figure 4.11a- Case 3)

Eigenvalues of Inv(E)*H = CanRsq/(1-CanRsq)				Test of H0: The canonical correlations in the current row and all that follow are zero				
Can	Eigen-value	Propor-tion	Cumula-tive	Likeli-hood Ratio	Approx-imate F Value	Num DF	Den DF	Pr > F
1	0.4284	0.498	0.498	0.4820	203.72	20	16554	<.0001
2	0.3775	0.4389	0.9368	0.6885	166.75	12	13208	<.0001
3	0.0539	0.0627	0.9995	0.9484	44.64	6	9986	<.0001
4	0.0004	0.0005	1	0.9996	0.99	2	4994	0.3707

Total-Sample Standardized Canonical Coefficients				
Variable	Can1	Can2	Can3	Can4
Z1	-0.0412	0.5255	-0.7145	0.4491
Z2	0.0356	-0.3301	0.3978	0.5254
Z3	-0.0592	0.7984	0.7974	-0.0235
Z4	1.1912	0.0843	-0.0199	0.0221
Z5	0.0216	0.0309	-0.092	-0.7854

Class Means on Canonical Variables				
Class	Can1	Can2	Can3	Can4
1	-0.0948	1.1155	0.0279	0.0163
2	0.0388	-0.467	-0.3945	0.0145
3	-1.0422	-0.0895	0.0197	-0.0239
4	0.0801	-0.6394	0.3355	0.018
5	1.0182	0.0803	0.0114	-0.0249

N	5000	DF Total	4999
Variables	5	DF Within Classes	4995
Classes	5	DF Between Classes	4

Wilks' Lambda	Pr > F	<.0001
Pillai's Trace	Pr > F	<.0001
Hotelling-Lawley Trace	Pr > F	<.0001
Roy's Greatest Root	Pr > F	<.0001

B6: FREQUENCY OF CLASS MEMBERSHIP BY CLUSTER (CASE 3)

Table B6 presents the frequency of *three* class membership by ROWK clusters for firms in Case 3. The label p_i^{verj} denotes purity version j for cluster/class i. Regarding cluster validation, if *purity_ver3* indices are computed using the *five-class* patterns, only 31% of observations are precisely assigned by ROWK, which is not much different from those of *STD_K* and *UNSTD_K*. This is rational since ROWK clustering does not aim to discover the *true* class patterns (i.e. five classes). Its ultimate goal is to explore class patterns that reduce the problem of HGSC, here *Class* ξ_1, *Class* ξ_2 and *Class* ξ_4. Consistent with our expectations, if only these class members are considered, 62.47% of observations are correctly assigned by ROWK, which is higher than the classifications

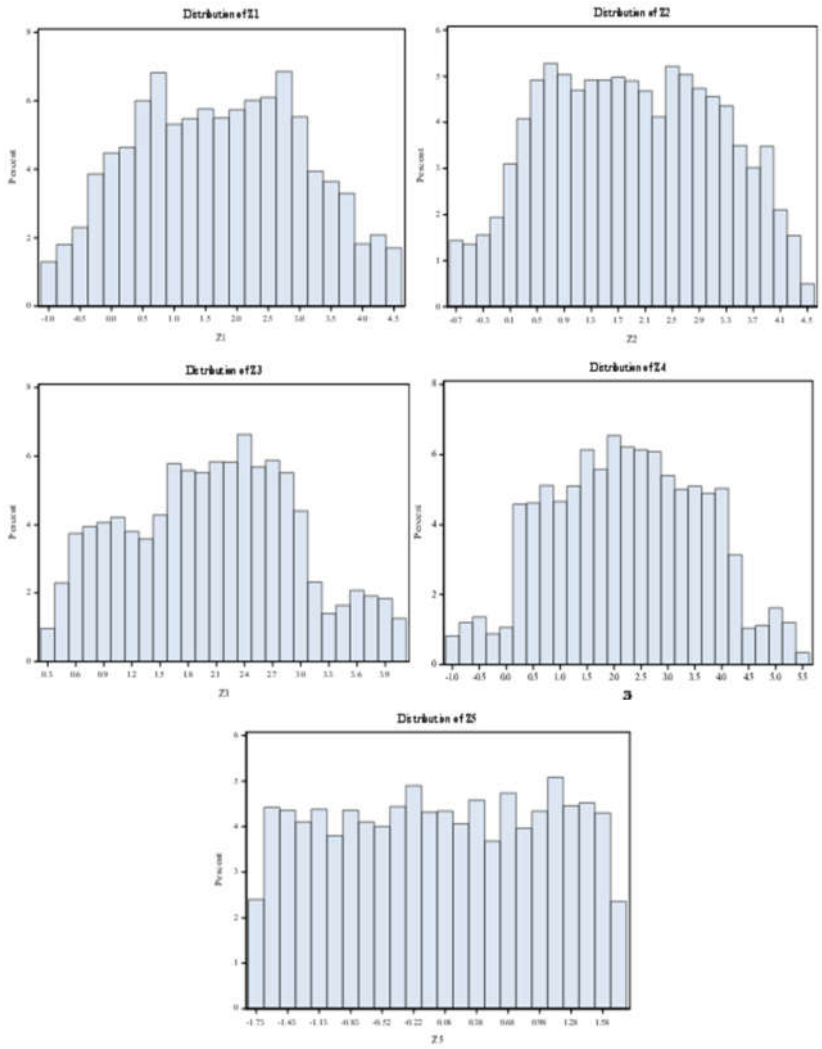
using *STD_K* (54.27%), *UNSTD_K* (48.91%) and *WK* (48.15%).This evidence gives further support for hypothesis H4c. By assigning a weight to a feature based on its contribution to both class recognitions and regression estimations, ROWK clustering improves the results of regression estimations.

Table B6
Frequency of Class Membership by Cluster (Case 3)

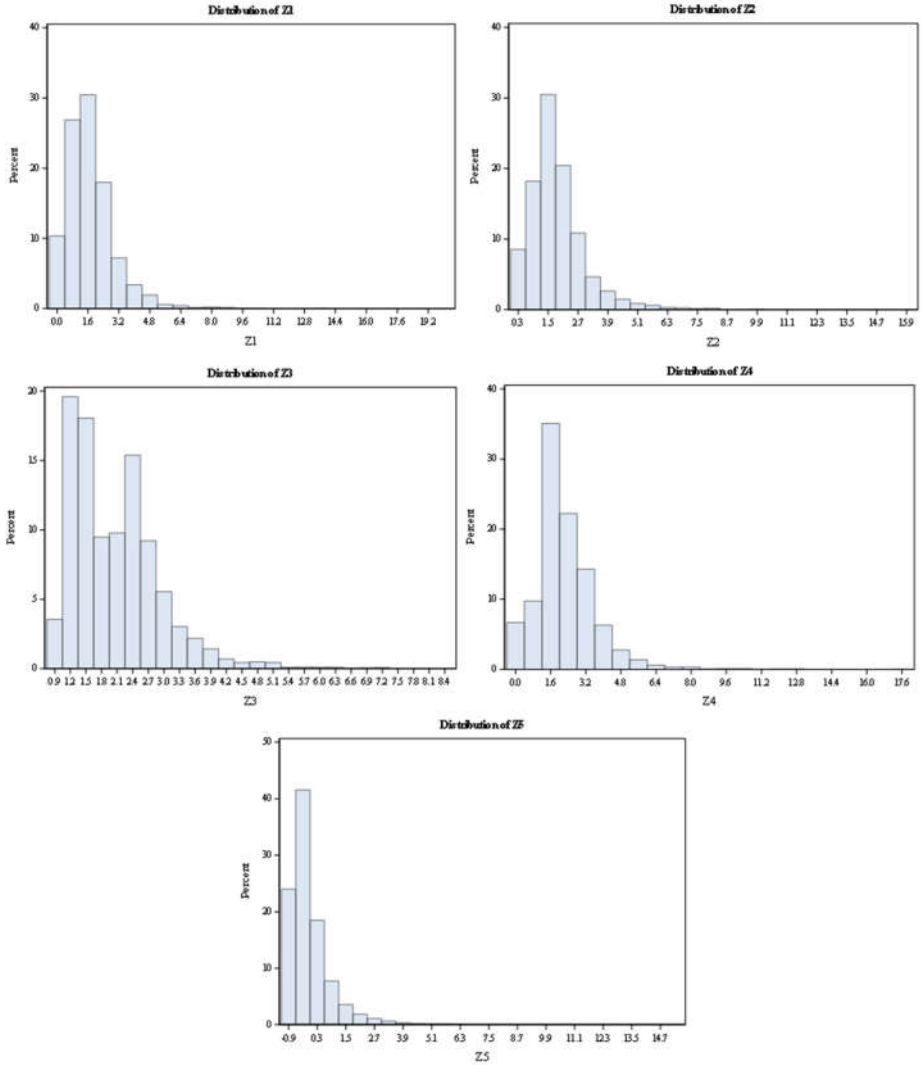
Panel A: ROWK Clustering						
Class/Clus	1	2	3	No. Obs	p_i^{ver2}	p_i^{ver3}
1	1269	241	157	1667	0.7613	0.7658
2	224	885	558	1667	0.5309	0.5331
4	164	534	968	1666	0.5810	0.5752
No. Obs.	1657	1660	1683	5000		
p_i^{ver1}	0.7658	0.5331	0.5752			
$p_{overall}^{ver1}$			0.6244			
$p_{overall}^{ver2}$			0.6244			
$p_{overall}^{ver3}$			0.6247			
Panel B: Un-Standardized K-means Clustering						
Class/Clus	1	2	3	# of Obs	p_i^{ver2}	p_i^{ver3}
1	306	1075	286	1667	0.6449	0.6384
2	613	420	634	1667	0.3803	0.3794
4	726	189	751	1666	0.4508	0.4494
# of Obs.	1645	1684	1671	5000		
p_i^{ver1}	0.44134	0.63836	0.44943			
$p_{overall}^{ver1}$			0.5104			
$p_{overall}^{ver2}$			0.4920			
$p_{overall}^{ver3}$			0.4891			
Panel C: Standardized K-means Clustering						
Class/Clus	1	2	3	# of Obs	p_i^{ver2}	p_i^{ver3}
1	238	1250	179	1667	0.7499	0.7340
2	678	274	715	1667	0.4289	0.4228
4	690	179	797	1666	0.4784	0.4713
# of Obs.	1606	1703	1691	5000		
p_i^{ver1}	0.42964	0.734	0.47132			
$p_{overall}^{ver1}$			0.5474			
$p_{overall}^{ver2}$			0.5524			
$p_{overall}^{ver3}$			0.5427			

B7: DISTRIBUTION OF FEATURES (CASE 3 WITH VARIOUS DISTRIBUTIONs OF CLASS MEMBERSHIP)

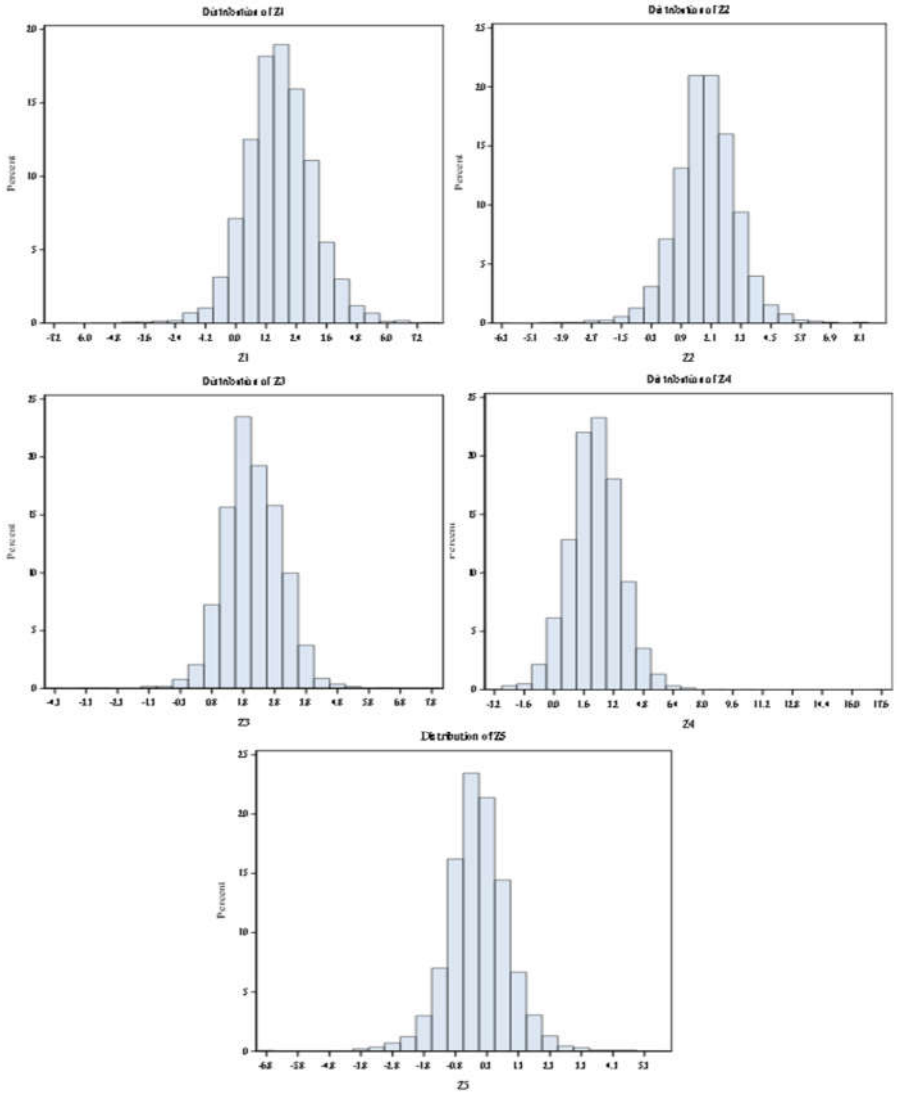
Figures B1a, b, c presents distributions of five cluster features for Case 3 with Uniform, Log-normal and Student-t Distribution of class membership, respectively.



a. Distribution of Features (Case 3 with Uniform Distribution of Class Membership)



b. Distribution of Features (Case 3 with Log-normal Distribution of Class Membership)



c. Distribution of Features (Case 3 with Student-t Distribution of Class Membership)

Figure B1- Distribution of Features (Case 3 with Various Distribution of Class Membership)

APPENDIX C: EMPERICAL RESULTS OF ROWK CLUSTERING ON EARNINGS PERSISTENCE

C1: DERIVATION OF THE SAMPLE AND DESCRIPTIVE STATISTICS FOR THE COMPLETE SAMPLE (1988-2011)

Table C1 presents the derivation of the sample and descriptive statistics for the complete sample covering the 1988-2011 period. There are a total of 33,686 firm-year observations. Of the 21 variables examined, the first seventeen are clustering features. Note that the table presents descriptive statistics of original values of the variables winsorized at 1% top and bottom (i.e. not the rank-transformed values). The results of the complete sample are similar to those of the investigated sample (1988-2004). Means of deflated cash flows from operations (*OCF_DEF*, 9.29%), earnings (*IBC_DEF*, 3.27%) and accruals (*ACC_DEF*, -6.01%) are again consistent with Dichev & Tang (2009) who document the corresponding values as 8.5%, 3.1% and -5.5% respectively. The mean of profit margin (*PM*) is 10.43%, which is slightly higher than the mean of the 1978-1996 period (9.2%) reported by Fairfield & Yohn (2001). In contrast, the mean of asset turnover (*ATO*) of 1.84 is much lower than the mean of 2.25 reported in Fairfield & Yohn (2001). Both the means and medians of change in profit margin (*ΔPM*) and change in asset turnover (*ΔATO*) are not significantly different from zero. Financial leverage (*FLEV*) in the later period reduces moderately, causing the value of FLEV for the 1988-2011 falls at 0.24, which is lower to that of operating leverage (*OLLEV, 032*). Means of sales growth (*SALES_GR*), net borrowing costs (*NBC*) and dividend payout (*DIV*) are 9.15%, 7.97 % and 0.33, which are unchanged from the 1988-2004 period. The sample means of firm size (log of total assets, *SIZE*) and age (*AGE*) are 7.3 and 24.2 respectively. The sample mean of earnings volatility (*VOL_IBC_DEF*) is 4.3%, which is significantly higher than that of the 1988-2004 period (3.9%). In summary, the results of the complete sample are very similar to those of the investigated sample (1988-2004).

Table C1

Derivation of the Sample and Descriptive Statistics for the Complete Sample

Panel A: Derivation of the sample

COMPUSTAT firm-years over 1988–2011 with 12/31 fiscal year-end	170.347
Firm-years with available deflated earnings, cash flows and accruals	118,953
Exclusion of financial firms (SIC code from 6000 to 6999)	98,287
Firm-years with assets greater than $100 millions	57,032
Firm-years with available data on earnings volatility and cash flow volatility (based on the most recent 5 years)	38,506
Firm-years remaining after truncating the top and bottom 1% on deflated earnings, accruals and cash flows	36,467
Firm-years with available data of cluster features	33,686
Final sample	*33,686*

Panel B: Descriptive statistics

No.	Variables	N	Mean	Median	Std.Dev	Min	Max
1	PM	33686	0.1043	0.1021	0.1597	-0.8122	0.4768
2	ΔPM	33684	0.0023	0.0013	0.0943	-0.4141	0.4730
3	ATO	33686	1.8429	1.2969	1.9480	0.1640	12.6808
4	ΔATO	33686	-0.0022	0.0078	0.5961	-2.8048	2.7540
5	VOL_IBC	33686	0.0430	0.0260	0.0498	0.0017	0.2907
6	CR	33686	1.9705	1.5068	1.6549	0.2790	10.5886
7	CAPX_DEF	33686	0.0689	0.0493	0.0667	0.0030	0.3762
8	INTAN_INT_DEF	33686	0.0475	0.0078	0.1018	-0.1548	0.5671
9	FLEV	33686	0.2447	0.3742	0.6695	-3.4023	1.7853
10	OLLEV	33686	0.3188	0.2968	0.1534	0.0592	0.8493
11	SALE_GR	33686	0.0915	0.0441	0.2744	-0.5034	1.4917
12	NBC	32101	0.0797	0.0719	0.0475	0.0137	0.3938
13	DIV	31934	0.3337	0.0572	0.6009	0.0000	3.9953
14	AB_ACC_DEF	33686	0.0693	0.0555	0.0551	0.0014	0.2891
15	SIZE	33686	7.2610	7.1058	1.6327	4.6636	11.4207
16	AGE	33686	24.2159	19.0000	15.9132	6.0000	60.0000
17	ABS_IBC_DEF	33686	0.0636	0.0486	0.0524	0.0014	0.2607
18	IBC_DEF	33686	0.0327	0.0380	0.0777	-0.4370	0.2695
19	ACC_DEF	33686	-0.0601	-0.0523	0.0675	-0.4174	0.1676
20	VOL_OCF	33686	0.0406	0.0316	0.0322	0.0045	0.1825
21	OCF_DEF	33686	0.0929	0.0883	0.0753	-0.2256	0.3470

PM, ΔPM denote profit margin and the change in profit margin; ATO, ΔATO represent asset turnover and the change of asset turnover; VOL_IBC, VOL_OCF denote volatility of earnings and volatility of operating cash flows measured as standard deviation of the most recent five years; CR is current ratio; CAPX_DEF and INTAN_INT_DEF are measures of deflated capital expenditure and deflated investment in intangible assets; FLEV and OLLEV denote financial leverage and operating leverage; SALE_GR is sales growth; NBC is net borrowing cost; ACC_DEF and AB_ACC_DEF denote deflated accruals and absolute value of deflated accruals; DIV, SIZE and AGE represent dividend payout, log of total assets and firm age respectively; IBC_DEF and ABS_IBC_DEF denote deflated earnings and absolute value of deflated earnings; OCF_DEF is deflated operating cash flows.

C2: CORRELATIONS OF CLUSTERING FEATURES FOR THE COMPLETE SAMPLE (1988-2011)

Table Table 5-2C2 exhibits the correlation coefficients between clustering features of the complete sample covering the 1988-2011 period. Again, the results are similar to those of the examined period (1988-2004). Over 96% (131 out of 136) of correlation coefficients are significant at the 10% level, and 14 pairs of clustering features have their absolute value of correlation over 0.3. This multicollinearity issue challenges the performance of standard clustering techniques, raising the need for the new ROWK clustering method proposed in this thesis. No correlation coefficients are higher than 0.7, so all cluster features are used to execute the ROWK clustering procedure.

Table C2: Correlations of Clustering Features for the Complete Sample (1988-2011)

		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
1	PM	1	0.18	-0.41	-0.03	-0.32	-0.16	0.18	-0.06	0.07	-0.15	0.04	-0.13	-0.30	-0.11	0.30	0.08	0.18
2	ΔPM		1	0.05	0.24	0.05	0.02	-0.02	0.02	-0.01	0.04	0.21	0.03	-0.23	-0.04	-0.02	-0.01	0.09
3	ATO			1	0.14	0.16	0.26	-0.15	0.28	-0.30	0.55	0.30	0.15	-0.23	0.00	-0.23	0.03	0.23
4	ΔATO				1	0.05	-0.02	-0.07	-0.09	0.05	0.11	0.35	0.11	0.00	0.10	0.00	0.02	0.04
5	VOL_IBC					1	0.28	-0.07	0.13	-0.19	-0.04	0.02	0.12	-0.02	0.25	-0.33	-0.28	0.20
6	CR						1	-0.22	0.24	-0.51	-0.14	0.11	0.01	-0.24	-0.10	-0.37	-0.13	0.19
7	CAPX_DEF							1	-0.14	0.02	-0.12	0.11	-0.03	-0.06	0.21	0.09	-0.03	0.08
8	INTAN_INT_DEF								1	-0.28	0.07	0.25	-0.05	-0.21	-0.06	-0.07	-0.07	0.16
9	FLEV									1	-0.11	-0.14	-0.05	0.28	0.02	0.26	0.11	-0.33
10	OLLEV										1	-0.01	0.13	0.03	-0.06	0.19	0.27	0.05
11	SALE_GR											1	0.04	-0.35	-0.06	-0.12	-0.12	0.20
12	NBC												1	0.00	0.04	-0.06	-0.01	-0.01
13	DIV													1	0.17	0.16	0.13	-0.25
14	AB_ACC_DEF														1	-0.09	-0.14	0.11
15	SIZE															1	0.37	-0.11
16	AGE																1	-0.08
17	ABS_IBC_DEF																	1

PM, ΔPM denote profit margin and the change in profit margin; ATO, ΔATO represent asset turnover and the change in asset turnover; VOL_IBC, VOL_OCF denote volatility of earnings and volatility of operating cash flows measured as standard deviation of the most recent five years; CR is current ratio; CAPX_DEF and INTAN_INT_DEF are measures of deflated capital expenditure and deflated investment in intangible assets; FLEV and OLLEV denote financial leverage and operating leverage; SALE_GR is sales growth; NBC is net borrowing cost; ACC_DEF and AB_ACC_DEF denote deflated accruals and absolute value of deflated accruals; DIV, SIZE and AGE represent dividend payout, log of total assets and firm age respectively; IBC_DEF and ABS_IBC_DEF denote deflated earnings and absolute value of deflated earnings; OCF is operating cash flows. See Section 2.5.2 for details of formulae for the variables. All correlations are significant at 10% level, except the blue-shaded ones. The yellow-shaded squares highlight those with correlations over 0.3.

C3: STEPWISE RESULTS OF ROWK CLUSTERING FOR THE CASE OF OPTIMAL NUMBER OF CLUSTERS

Appendix Table C3 presents the stepwise results of ROWK clustering when the number of clusters is set at to be the optimal, i.e. 8. In Stepwise 1, only AB_ACC_DEF (the lowest-MSR feature) is used. The corresponding MSR is 0.004062. In Stepwise 2, the second-lowest-MSR feature, i.e. VOL_IBC is added into the feature input list. The resulting set of weights found at the end of this stepwise is $\{w_{\text{AB\_ACC\_DEF}}, w_{\text{VOL\_IBC}}\} = \{0.59, 0.41\}$. The corresponding MSR reduces to 0.004035. The procedure continues until all features are added (Stepwise 17). Note that a feature is excluded when the inclusion of this feature does not reduce the MSR. As a result, the optimal set of weights is found at Stepwise 9 with the corresponding sets of weights as:

$\{w_{\text{AB\_ACC\_DEF}}, w_{\text{VOL\_IBC}}, w_{\text{INTAN\_INV\_DEF}}, w_{\Delta\text{ATO}}, w_{\Delta\text{PM}}\} = \{0.393, 0.393, 0.068, 0.119, 0.27\}$.

Table C3: Stepwise Results of ROWK Clustering for the Case of Optimal Number of Clusters ($K' = 8$)

Features/Stepwise	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
AB_ACC_DEF	1	0.59	0.55	0.395	0.478	0.404	0.399	0.402	**0.393**	0.386	0.389	0.386	0.391	0.391	0.386	0.39	0.385
VOL_IBC	0	0.41	0.382	0.569	0.402	0.404	0.399	0.402	**0.393**	0.386	0.389	0.386	0.391	0.391	0.386	0.39	0.385
PM	0	0	0.067	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0
DIV	0	0	0	0.036	0	0	0	0	**0**	0	0	0	0	0	0	0	0
INTAN_INT_DEF	0	0	0	0	0.12	0.07	0.069	0.07	**0.068**	0.067	0.068	0.067	0.068	0.068	0.067	0.068	0.067
ΔATO	0	0	0	0	0	0.122	0.121	0.122	**0.119**	0.117	0.118	0.117	0.118	0.118	0.117	0.118	0.116
ABS_IBC_DEF	0	0	0	0	0	0	0.013	0	**0**	0	0	0	0	0	0	0	0
CR	0	0	0	0	0	0	0	0.004	**0**	0	0	0	0	0	0	0	0
ΔPM	0	0	0	0	0	0	0	0	**0.027**	0.027	0.027	0.027	0.027	0.027	0.027	0.027	0.027
AGE	0	0	0	0	0	0	0	0	**0**	0.017	0	0	0	0	0	0	0
SALE_GR	0	0	0	0	0	0	0	0	**0**	0	0.009	0	0	0	0	0	0
FLEV	0	0	0	0	0	0	0	0	**0**	0	0	0.019	0	0	0	0	0
OLLEV	0	0	0	0	0	0	0	0	**0**	0	0	0	0.005	0	0	0	0
SIZE	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0.005	0	0	0
NBC	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0.019	0	0
ATO	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0.007	0
CAPX_DEF	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0.02
MSRs (x100)	0.4062	0.4035	0.4049	0.4046	0.4032	0.3994	0.4002	0.4002	**0.3985**	0.4015	0.4	0.4018	0.4021	0.4005	0.4022	0.3991	0.3999

PM, ΔPM denote profit margin and the change in profit margin; ATO, ΔATO represent asset turnover and the change in asset turnover; VOL_IBC, VOL_OCF denote volatility of earnings and volatility of operating cash flows measured as standard deviation of the most recent five years; CR is current ratio; CAPX_DEF and INTAN_INT_DEF are measures of deflated capital expenditure and deflated investment in intangible assets; FLEV and OLLEV denote financial leverage and operating leverage; SALE_GR is sales growth; NBC is net borrowing cost; ACC_DEF and AB_ACC_DEF denote deflated accruals and absolute value of deflated accruals; DIV, SIZE and AGE represent dividend payout, log of total assets and firm age respectively; IBC_DEF and ABS_IBC_DEF denote deflated earnings and absolute value of deflated earnings; OCF is operating cash flows. See Section 2.5.2 for details of formulae for the variables. MSR(x100) denotes the 100-time of mean squared residuals when running the earnings persistence regression within each cluster.

REFERENCES

Abarbanell, J., Bernard, V., 1992. Tests of analysts' overreaction/underreaction to earnings information as an explanation for anomalous stock price behaviour. *Journal of Finance*, 47, 1181–1207.

Alexander, G. J., & Peterson, M. A. (2007). An analysis of trade-size clustering and its relation to stealth trading. *Journal of Financial Economics*, 84(2), 435-471.

Amaran, S., Sahinidis, N. V., Sharda, B., & Bury, S. J. (2016). Simulation optimization: A review of algorithms and applications. *Annals of Operations Research*, 240(1), 351-380.

Amir, E., Kama, I., & Livnat, J. (2011). Conditional versus unconditional persistence of RNOA components: implications for valuation. *Review of Accounting Studies*, 16(2), 302-327.

Amorim, R. C., & Mirkin, B. (2012). Minkowski metric, feature weighting and anomalous cluster initializing in K-means clustering. *Pattern Recognition*, 45(3), 1061-1075.

Amor-Tapia, B., & Tascón Fernández, M. T. (2014). Estimation of future levels and changes in profitability: The Effect of the Relative Position of the Firm in Its Industry and the Operating-Financing Disaggregation. *Revista de Contabilidad*, 17(1), 30-46.

Ando, T., & Bai, J. S. (2016). Panel data models with grouped factor structure under unknown group membership. *Journal of Applied Econometrics*, 31(1), 163-191.

Anthony, J., & K. Ramesh. (1992). Association between accounting performance measures and stock prices. *Journal of Accounting & Economics*. 15 (2-3), 203–227.

Art, D, Gnanadesikan, R, & Kettenring, R. (1982). Data-based metrics for cluster analysis. *Utilitas Mathematica*, 21A, 75–99.

Arthur, D., Vassilvitskii, S., and Siam/Acm. (2007). *K-means plus plus: The Advantages of Careful Seeding*. Philadelphia: Siam.

Ball, R., & Bartov, E. (1996). How naive is the stock market's use of earnings information? *Journal of Accounting & Economics*, 21(3), 319-337.

Bartholdy, J., & Feng, T. Y. (2013). The quality of securities firms' earnings forecasts and stock recommendations: Do informational advantages, reputation and experience matter in china? *Pacific-Basin Finance Journal*, 24, 66-88.

Basu, S. (1997). The conservatism principle and the asymmetric timeliness of earnings. *Journal of Accounting & Economics*, 24(1), 3-37.

Bauman, Mark P. (2014). Forecasting operating profitability with Dupont analysis: further evidence. *Review of Accounting and Finance*, 13(22), 191-205.

Beh, A., & Bruyere, B. L. (2007). Segmentation by visitor motivation in three Kenyan national reserves. *Tourism Management*, 28(6), 1464-1471.

Bernard, V. L., & Skinner, D. J. (1996). What motivates managers' choice of discretionary accruals? *Journal of Accounting & Economics*, 22(1-3), 313-325.

Bishop, Y. M. M. (1969). Full contingency tables, logits, and split contingency tables. biometrics, 25(2), 383-399.

Bock, H. H. (1985). On some significance tests in cluster analysis. *Journal of Classification*, 2(1), 77-108.

Bradshaw, M. T., Drake, M. S., Myers, J. N., & Myers, L. A. (2012). A re-examination of analysts' superiority over time-series forecasts of annual earnings. *Review of Accounting Studies*, 17(4), 944-968.

Brusco, M. J., & Cradit, J. D. (2001). A variable-selection heuristic for K-means clustering. *Psychometrika*, 66(2), 249-270.

Burnside, C. (1996). Production function regressions, returns to scale, and externalities. *Journal of Monetary Economics*, 37(2), 177-201.

Campbell, John Y., & Shiller, Robert J. (1988). The dividend-price ratio and expectations of future dividends and discount factors. *Review of Financial Studies*, 1(3), 195-228.

Cao, T. Y., Shaari, H., & Donnelly, R. (2018). Impairment reversals: unbiased reporting or earnings management. *International Journal of Accounting and Information Management*, 26(2), 245-271.

Cha, S., McCleary, K. W., & Uysal, M. (1995). Travel motivations of Japanese overseas travellers: a factor-cluster segmentation approach. *Journal of Travel Research*, 34(1), 33-39.

Chen, L. H., Folsom, D. M., Paek, W., & Sami, H. (2014). Accounting conservatism, earnings persistence, and pricing multiples on earnings. *Accounting Horizons*, 28(2), 233-260.

Chen, Sheng-Syan, Ho, Lan-Chih, & Shih, Yi-Cheng. (2007). Intra-industry effects of corporate capital investment announcements. *Financial Management*, 36(2), 125-145.

Chiang, Mark Ming-Tso, & Mirkin, Boris. (2010). Intelligent choice of the number of clusters in K-means clustering: an experimental study with different cluster spreads. *Journal of Classification,* 27(1), 3-40.

Cohen, D. A., & Zarowin, P. (2010). Accrual-based and real earnings management activities around seasoned equity offerings. *Journal of Accounting & Economics*, 50(1), 2-19.

Cooper M.C., & Milligan G.W. (1988). The effect of measurement error on determining the number of clusters in cluster analysis. In: Gaul W., Schader M. (eds) *Data, Expert Knowledge and Decisions*. Springer, Berlin, Heidelberg.

Cooper, M. J., Gulen, H., & Schill, M. J. (2008). Asset growth and the cross-section of stock returns. *Journal of Finance*, 63(4), 1609-1651.

Dechow, P. M., Richardson, S. A., & Tuna, I. (2003). Why are earnings kinky? An examination of the earnings management explanation. *Review of Accounting Studies*. 8 (2), 355–84.

Dechow, P. M., Ge, W., & Schrand, C. (2010). Understanding earnings quality: A review of the proxies, their determinants, and their consequences. *Journal of Accounting and Economics*, 50, 344–401.

Dechow, P. M., Richardson, S. A., & Sloan, R. G. (2008). The persistence and pricing of the cash component of earnings. *Journal of Accounting Research*, 46(3), 537-566.

Desai, H., Hogan, C. E., & Wilkins, M. S. (2006). The reputational penalty for aggressive accounting: earnings restatements and management turnover. *Accounting Review*, 81(1), 83-112.

Desarbo, W. S., Carroll, J. D., Clark, L. A., & Green, P. E. (1984). Synthesized clustering-a method for amalgamating alternative clustering bases with different weighting of variables. *Psychometrika*, 49(1), 57-78.

Di Cimbrini, T. (2015). Welfare or politics? The identity of Italian mutual aid societies as revealed by a latent class cluster analysis of their annual reports. *Accounting History*, 20(3), 310-341.

Dichev, I.D., & Tang, V.W., (2008). Matching and the changing properties of accounting earnings over the last 40 years. *The Accounting Review*, 83, 1–36.

Dichev, I. D., & Tang, V. W. (2009). Earnings volatility and earnings predictability. *Journal of Accounting and Economics*, 47(1-2), 160-181.

Dickinson, V. (2011). Cash flow patterns as a proxy for firm life cycle. *Accounting Review*, 86(6), 1969-1994.

Dickinson, V., & Sommers, G. A. (2012). Which competitive efforts lead to future abnormal economic rents? Using accounting ratios to assess competitive advantage. *Journal of Business Finance & Accounting*, 39(3-4), 360-398.

Dolnicar, S. (2002). A review of unquestioned standards in using cluster analysis for data-driven market segmentation. *CD Conference Proceedings of the Australian and New Zealand Marketing Academy Conference 2002.*

Efron, B., & Stein, C. (1981). The jackknife estimate of variance. *Annals of Statistics*, 9(3), 586-596.

Epure, M., Kerstens, K., & Prior, D. (2011). Bank productivity and performance groups: a decomposition approach based upon the Luenberger productivity indicator. *European Journal of Operational Research*, 211(3), 630-641.

Fairfield, P. M., Whisenant, J. S., & Yohn, T. L. (2003). Accrued earnings and growth: implications for future profitability and market mispricing. *Accounting Review*, 78(1), 353-371.

Fairfield, P., & Yohn, T. L. (2001). Using asset turnover and profit margin to forecast changes in profitability. *Review of Accounting Studies*, 6, 371-385.

Fama, E. F., & French, K. R. (1993). Common risk-factors in the returns on stocks and bonds. *Journal of Financial Economics*, 33(1), 3-56.

Fama, E. F., & French, K. R. (1997). Industry costs of equity, *Journal of Financial Economics*, 43, 153-193.

Feltham, G. A., & Ohlson, J. A. (1995). Valuation and clean surplus accounting for operating and financial activities. *Contemporary Accounting Research*, 11(2), 689-731.

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7(2), 179–188.

Fred, A. L. N., & Jain, A. K. (2005). Combining multiple clusterings using evidence accumulation. *Ieee Transactions on Pattern Analysis and Machine Intelligence*, 27(6), 835-850.

Fu, F. (2010). Overinvestment and the operating performance of SEO firms. *Financial Management*, 39(1), 249-272.

Garla, S., Chakraborty, G., & Gaeth, G. (2012). Comparison of K-means, normal mixtures and probabilistic-d clustering for b2b segmentation using customer' perception. *Data Mining and Text Analytics*: *SAS Global Forum 2012.*

Giroud, X., & Mueller, H. M. (2011). Corporate governance, product market competition, and equity prices. *Journal of Finance*, 66(2), 563-600.

Givoly, D., & Hayn, C. (2000). The changing time-series properties of earnings, cash flows, and accruals: Has financial reporting become more conservative? *Journal of Accounting and Economics,* 29, 287–320.

Gnanadesikan, R., Kettenring, J. R., & Tsao, S. L. (1995). Weighting and selection of variables for cluster analysis. *Journal of Classification*, 12(1), 113-136.

Goldfeld, S.M, & Quandt, R. E. (1972), *Nonlinear Methods in Econometrics*. North Holland Press, ch.9.

Goldfeld, S.M, & Quandt, R. E. (1973). The estimation of structural shifts by switching regressions, *Annals of Economic and Social Measurement*, Volume 2, Number 4 (pp. 475-485).

Gungor, Z., & Unler, A. (2008). K-harmonic means data clustering with Tabu-search method. *Applied Mathematical Modelling*, 32(6), 1115-1125.

Gupta, M. C., & Huefner, R. J. (1972). Cluster analysis study of financial ratios and industry characteristics. *Journal of Accounting Research*, 10(1), 77-95.

Hirshleifer, D. (2001). Investor psychology and asset pricing. *Journal of Finance*, 56(4), 1533-1597.

Hirshleifer, D., Hou, K. W., Teoh, S. H., & Zhang, Y. L. (2004). Do investors overvalue firms with bloated balance sheets? *Journal of Accounting & Economics*, 38(1-3), 297-331.

Hotelling, H. (1936). Relations between two sets of variates. *Biometrika*, 28, 321-377.

Hribar, P., & Collins, D. W. (2002). Errors in estimating accruals: implications for empirical research. *Journal of Accounting Research*, 40(1), 105-134.

Hsiao, C., & Tahmiscioglu, A. K. (1997). A panel analysis of liquidity constraints and firm investment. *Journal of the American Statistical Association*, 92(438), 455-465.

Hsu, P. H., & Hu, X. S. (2016). Advisory board and earnings persistence. *Journal of Accounting Auditing and Finance*, 31(1), 134-157.

Huang, J.Z., Xu, J., Ng, M., & Ye, Y. (2008). Weighting method for feature selection in K-means. in: H. Liu, H. Motoda (Eds.), *Computational Methods of Feature Selection*, Chapman & Hall/CRC, 193-209.

Hürlimann, W. (2001). Financial data analysis with two symmetric distributions. *ASTIN Bulletin*, 31(1), 187-211.

Hwang, J. S., & Hu, T. H. (2015). A stepwise regression algorithm for high-dimensional variable selection. *Journal of Statistical Computation and Simulation*, 85(9), 1793-1806.

Ikenberry, D., Lakonishok, J., & Vermaelen, T. (1995). Market underreaction to open market share repurchases. *Journal of Finance*, 50(3), 982-983.

Jansen, I. P., Ramnath, S., & Yohn, T. L. (2012). A diagnostic for earnings management using changes in asset turnover and profit margin. *Contemporary Accounting Research*, 29(1), 221-251.

Jensen, R. E. (1971). A cluster analysis study of financial performance of selected business firms. *The Accounting Review*, 46, 36-56.

Kelly, B., & Pruitt, S. (2013). Market expectations in the cross-section of present values. *Journal of Finance*, 68(5), 1721-1756.

Ketchen, D. J., & Shook, C. L. (1996). The application of cluster analysis in strategic management research: An analysis and critique. *Strategic Management Journal*, 17(6), 441-458.

Khan, M., & Watts, R. L. 2009. Estimation and empirical properties of a firm-year measure of accounting conservatism. *Journal of Accounting and Economics*. 48, 132–150.

Kim, H. D., Lee, D. H., Choe, H., & Seo, Il W. (2014). The evolution of cluster network structure and firm growth: a study of industrial software clusters. *Scientometrics,* 99(1), 77-95.

Kogan, L., & Papanikolaou, D. (2013). Firm characteristics and stock returns: The role of investment-specific shocks. *Review of Financial Studies*, 26(11), 2718-2759.

Kothari, S. P., Leone, A. J., & Wasley, C. E. (2005). Performance matched discretionary accrual measures. *Journal of Accounting & Economics*, 39(1), 163-197.

Kwon, S. S., & Yin, J. (2015). A comparison of earnings persistence in high-tech and non-high-tech firms. *Review of Quantitative Finance and Accounting*, 44(4), 645-668.

Lee, C. K., Lee, Y. K., & Wicks, B. E. (2004). Segmentation of festival motivation by nationality and satisfaction. *Tourism Management*, 25(1), 61-70.

Lettau, M., & Van Nieuwerburgh, S. (2008). Reconciling the return predictability evidence. *Review of Financial Studies*, 21(4), 1606-1652.

Li, F. (2011). Earnings quality based on corporate investment decisions. *Journal of Accounting Research*, 49(3), 721-752.

Li, Zhan-lei, & Li, Nan. (2008). Empirical research on herding behaviour of corporate financing decisions. In H. Lan (Ed.), *2008 International Conference on Management Science & Engineering* (pp. 1318-1324).

Lin, Chang-Ching, & Ng, Serena. (2012). Estimation of panel data models with parameter heterogeneity when group membership is unknown. *Journal of Econometric Methods*, 1(1), 42-55.

Lin, T. W., Yang, H. E., & Ieee. (2006). Female consumer behaviour in a banking environment. *2006 Ieee International Conference on Service Operations and Logistics, and Informatics*, 564.

Lipe, R. C. (1986). The information contained in the components of earnings. *Journal of Accounting Research*, 24, 37-64.

Little, P. L., Little, B. L., & Coffee, D. (2009). The Du Pont model: Evaluating alternative strategies in the retail industry. *Academy of Strategic Management Journal*, 8, 71-80.

Loughran, T., & Ritter, J. R. (1995). The new issues puzzle. *Journal of Finance*, 50(1), 23-51.

Maenpaa, K. (2006). Clustering the consumers on the basis of their perceptions of the Internet banking services. *Internet Research*, 16(3), 304-322.

Taboga, M. (2017). *Lectures on Probability Theory and Mathematical Statistics* (3rd ed) CreateSpace Independent Publishing Platform.

McNichols, M (2000), Research design issues in earnings management studies, *Journal of Accounting and Public Policy*, 19, (4-5), 313-345.

McNichols, M., Rajan, M. V., & Reichelstein, S. (2014). Conservatism correction for the market-to-book ratio and Tobin's q. *CESifo Working Paper*, 4626.

Mirkin, B. (2005). *Clustering for Data Mining: A Data Recovery Approach*. Boca Raton FL: Chapman and Hall/CRC.

Mohd-Rahim, F. A., Wang, C., Boussabaine, H., Abdul-Rahman, H., & Wood, L. C. (2014). Factor reduction and clustering for operational risk in software development. *Journal of Operational Risk*, 9(3), 53-88.

Monte-Mor, D. S., Galdi, F. C., & Costa, C. M. (2018). The role of accounting fundamentals and other information in analyst forecast errors. *International Finance*, 21(2), 175-194.

Murtagh, F., & Contreras, P. (2012). Algorithms for hierarchical clustering: An Overview. *Wiley Interdisciplinary Reviews-Data Mining and Knowledge Discovery*, 2(1), 86-97.

Ni, S. X. Y., Pearson, N. D., & Poteshman, A. M. (2005). Stock price clustering on option expiration dates. *Journal of Financial Economics*, 78(1), 49-87.

Nimtrakoon, S., & Tayles, M. (2015). Explaining management accounting practices and strategy in Thailand: A selection approach using cluster analysis. *Journal of Accounting in Emerging Economies*, 5(3), 269-298.

Ning, C., Xu, D. H., & Wirjanto, T. S. (2015). Is volatility clustering of asset returns asymmetric? *Journal of Banking & Finance*, 52, 62-76.

Nissim, D., & Penman, S. H. (2001). Ratio analysis and equity valuation: From research to practice. *Review of Accounting Studies*, 6, 109-154.

Nunes, P. M., Serrasqueiro, Z. S., & Leitao, J. (2010). Are there nonlinear relationships between the profitability of Portuguese service SME and its specific determinants? *Service Industries Journal*, 30(8), 1313-1341.

Ohlson, J. 1995. Earnings, book values, and dividends in equity valuation. *Contemporary Accounting Research*. 11, 661–687.

Ou, J. A., & Penman, S. H. (1989). Financial statement analysis and the prediction of stock return. *Journal of Accounting and Economics*, 11, 295-329.

Ozdagli, A. K. (2012). Financial leverage, corporate investment, and stock returns. *Review of Financial Studies*, 25(4), 1033-1069.

Park, H. (2011). A study of real estate customer segmentation using cluster analysis. *Journal of the Korean Data Analysis Society*, 13(4), 1811-1819.

Penman, S. 2007. *Financial Statement Analysis and Security Valuation*, 3rd ed. New York: McGraw-Hill/Irwin.

Penman, S. H., & Zhang, X. J. (2002). Accounting conservatism, the quality of earnings, and stock returns. *Accounting Review*, 77(2), 237-264.

Perfect, S. B., Peterson, D. R., & Peterson, P. P. (1995). Self-tender offers: The effects of free cash flow, cash flow signalling, and the measurement of Tobin's q. *Journal of Banking & Finance* 19, 1005-1023.

Pimentel, R. C., & De Aguiar, A. B. (2016). The role of earnings persistence in valuation accuracy and the time horizon. *Rae-Revista De Administracao De Empresas*, 56(1), 71-86.

Puhani, P. A. (2012). The treatment effect, the cross difference, and the interaction term in nonlinear "difference-in-differences" models. *Economics Letters*, 115(1), 85-87.

Qian, Y. (2006). K-means algorithm and its application for clustering companies listed in Zhejiang province. In A. Zanasi, C. A. Brebbia & N. F. F. Ebecken (Eds.), *Data Mining Vii: Data, Text and Web Mining and Their Business Applications* (Vol. 37, pp. 35-44).

Quandt, R. E. (1958). The estimation of the parameters of a linear-regression system obeying 2 separate regimes. *Journal of the American Statistical Association*, 53(284), 873-880.

Quandt, R. E. (1960). Tests of the hypothesis that a linear-regression system obeys 2 separate regimes. *Journal of the American Statistical Association*, 55(290), 324-330.

Quandt, R. E. (1972). A new approach to estimating switching regressions. *Journal of the American Statistical Association*, 67, 306-310.

Richardson, S. A., Sloan, R. G., Soliman, M. T., & Tuna, I. (2006). The implications of accounting distortions and growth for accruals and profitability. *Accounting Review*, 81(3), 713-743.

Richardson, S. A., Sloan, R. G., Soliman, M. T., & Tuna, I. (2005). Accrual reliability, earnings persistence and stock prices. *Journal of Accounting & Economics*, 39(3), 437-485.

Richardson, S., Tuna, I., & Wysocki, P. (2010). Accounting anomalies and fundamental analysis: a review of recent research advances. *Journal of Accounting & Economics*, 50(2-3), 410-454.

Sambandam, R. (2003). Cluster analysis gets complicated. *Marketing Research*, 15(1), 16-21.

SAS Institute Inc., Cary, NC. (2009). *Sas/Stat®9.2 User's Guide (Second ed.)*: SAS Institute Inc., Cary, NC.

Sloan, R. G. (1996). Do stock prices fully reflect information in accruals and cash flows about future earnings? *Accounting Review*, 71(3), 289-315.

Soliman, M. T. (2008). The use of Dupont analysis by market participants. *Accounting Review*, 83(3), 823-853.

Soukal, I., & Hedvicakova, M. (2012). Classification of the electronic retail core banking market consumers. *Proceedings of 30th International Conference Mathematical Methods in Economics*.

Steinley, D. (2004). Sandardizing variables in K-means clustering. In D. Banks, L. House, F. R. McMorris, P. Arabie & W. Gaul (Eds.), *Classification, Clustering, and Data Mining Applications*, 53-60.

Sun, W., Wang, J., & Fang, Y. (2012). Regularized K-means clustering of high-dimensional data and its asymptotic consistency. *Electronic Journal of Statistics*, 6, 148-167.

Sunder, S. (1980). Corporate capital-investment, accounting methods and earnings - A test of the control hypothesis. *Journal of Finance*, 35(2), 553-565.

Tan, P. N., Steinbach, M., & Kumar, V. (2005). *Introduction to Data Mining*, First Edition, Addison-Wesley Longman Publishing Co., Inc.

Tanioka, K., & Yadohisa, H. (2012). Effect of data standardization on the result of K-means clustering. In: Gaul W., Geyer-Schulz A., Schmidt-Thieme L., Kunze J. (eds*) Challenges at the Interface of Data Analysis, Computer Science, and Optimization. Studies in Classification, Data Analysis, and Knowledge Organisation*. Springer, Berlin, Heidelberg.

Ter Wal, Anne L. J. (2013). Cluster emergence and network evolution: A longitudinal analysis of the inventor network in Sophia-Antipolis. *Regional Studies*, 47(5), 651-668.

Fields, T. D., Lys, T. Z., & Vincent, L. (2001). Empirical research on accounting choice. *Journal of Accounting & Economics*, 31(1-3), 255-307.

Titman, S., Wei, K. C. J., & Xie, F. (2004). Capital investments and stock returns. *Journal of Financial and Quantitative Analysis*, 39(4), 677-700.

Ulupinar, B. (2018). The effect of managerial entrenchment on analyst bias. *Global Finance Journal*, 37, 25-38.

Vlckova, V., Lostakova, H., Patak, M., & Tanger. (2014). Selected problems arising within cluster analysis usage for market segmentation metal. *23rd International Conference on Metallurgy and Materials*, 1932-1939.

Vogt, S. C. (1997). Cash flow and capital spending: Evidence from capital expenditure announcements. *Financial Management*, 26(2), 44-57.

Vuolteenaho, T. (2002). What drives firm-level stock returns? *Journal of Finance*, 57(1), 233-264.

Vuong, Quang H. (1989). Likelihood Ratio Tests for Model Selection and non-nested Hypotheses, *Econometrica*, 57 (2),307–333.

Wang, P. (2013). The role of disaggregation of earnings in stock valuation and earnings forecasting. *Accounting and Business Research*, 43(5), 530-557.

Witten, D. M., & Tibshirani, R. (2010). A framework for feature selection in clustering. *Journal of the American Statistical Association*, 105(490), 713-726.

Wolk, H. I., Tearney, M. G., & Dodd, J. L. (2001). *Accounting Theory: A conceptual and institutional approach* (5th ed.). Cincinnati, OH: South-Western College Publishing.

Wong, M. A., & Lane, T. (1983). A kth nearest neighbour clustering procedure. *Journal of the Royal Statistical Society Series B-Methodological*, 45(3), 362-368.

Yang, M. S., Lai, C. Y., & Lin, C. Y. (2012). A robust em clustering algorithm for Gaussian mixture models. *Pattern Recognition*, 45(11), 3950-3961.

Zhao, G., & Maclean, A. L. (2000). A comparison of canonical discriminant analysis and principal component analysis for spectral transformation. *Photogrammetric Engineering and Remote Sensing*, 66(7), 841-847.