CONDITIONAL CORRELATION ANALYSIS

A thesis submitted in partial fulfillment of the requirements for the degree of Master of Science

By

SANJEEV BHATTA
B.E., Computer Engineering, Tribhuvan University, 2012

2017
Wright State University
I HEREBY RECOMMEND THAT THE THESIS PREPARED UNDER MY SUPERVISION BY Sanjeev Bhatta ENTITLED Conditional Correlation Analysis BE ACCEPTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF Master of Science.

Guozhu Dong, Ph.D.
Thesis Director

Mateen M. Rizki, Ph.D.
Chair, Department of Computer Science and Engineering

Committee on Final Examination

Guozhu Dong, Ph.D.

Keke Chen, Ph.D.

Derek Doran, Ph.D.

Robert E. W. Fyffe, Ph.D.
Vice President for Research and Dean of the Graduate School
Correlation analysis is a frequently used statistical measure to examine the relationship among variables in different practical applications. However, the traditional correlation analysis uses an overly simplistic method to do so. It measures how two variables are related in an application by examining only their relationship in the entire underlying data space. As a result, traditional correlation analysis may miss a strong correlation between those variables especially when that relationship exists in the small subpopulation of the larger data space. This is no longer acceptable and may lose a fair share of information in this era of Big Data which often contains highly diverse nature of data where data can differ in a noticeable manner within the same application.

To remedy this situation, we are introducing a new approach called Conditional Correlation Analysis (CCR) in this thesis. Instead of computing the correlation among variables in the entire data space, this approach first divides the entire data space into multiple subpopulations using patterns. It then computes the correlation for each subpopulation and identifies the subpopulation which is highly different (in term of correlation strength) from the global population. Moreover, we introduce the concepts of CCRs and the ways to mine those CCRs, provides measures to evaluate the unusualness of CCRs and gives experiments to evaluate and illustrate the CCR approach in financial and medical applications.

**Keywords:** subpopulation, conditional correlation, big data, patterns, unusualness
# Contents

1 Introduction .................................................. 1

2 Related Work .................................................. 4
   2.1 Piecewise Linear Regression ................................. 4
   2.2 Pattern Aided Regression and CPXR .......................... 5
   2.3 Multiresolution Correlation Analysis ......................... 5

3 Preliminaries .................................................. 6
   3.1 Pearson Correlation Coefficient .............................. 6
   3.2 Simple Linear Regression Model ............................. 7
   3.3 Pattern .................................................. 7
   3.4 Discretization Method (Binning) .............................. 8
      3.4.1 Equi-Width Discretization .............................. 8
      3.4.2 Entropy Based Discretization ............................ 8
   3.5 Contrast Pattern Aided Regression ......................... 9

4 Conditional Correlation Analysis: Concepts and Measures .... 11
   4.1 CCR Concept ............................................. 11
   4.2 CCR Example ............................................. 13
   4.3 Mining CCRs using CPXR .................................... 15

5 Experimental Results ........................................... 16
   5.1 Data Sets and CPXR Parameter Settings ..................... 16
   5.2 Example of strong CCRs and analysis ....................... 18
   5.3 Different datasets often have CCRs of different strengths . 22
   5.4 Example of evenly distributed CCRs ........................ 27
   5.5 Advantages of CCR approach and CCR Mining ............... 31
5.6 Computation time ...................................................... 32

6 Applications of CCR .................................................. 33
  6.1 Application on Financial Dataset .................................. 33
  6.2 Application on Medical Dataset .................................. 34

7 Conclusion .............................................................. 35
  7.1 Summary .............................................................. 35
  7.2 Future Work .......................................................... 35

Bibliography .............................................................. 37
List of Figures

5.1 Global GoldMA365 vs LumberMA365 (Table 5.3), CC = 0.0429 . . . . . . . . . . . 20
5.2 Subpop 2 GoldMA365 vs LumberMA365 (Table 5.3), CC = 0.5254 . . . . . . . . . 21
5.3 Subpop 3 GoldMA365 vs LumberMA365 (Table 5.3), CC = 0.6626 . . . . . . . . . 21
5.4 Subpop 4 GoldMA365 vs LumberMA365 (Table 5.3), CC = 0.7568 . . . . . . . . . 22
5.5 Global Correlation: G628 vs G1362 for Colon, CC = 0.3363 . . . . . . . . . . . 25
5.6 Subpop 1 Correlation: G628 vs G1362 for Colon, CC = 0.5073 . . . . . . . . . . . 26
5.7 Subpop 2 Correlation: G628 vs G1362 for Colon, CC = 0.5604 . . . . . . . . . . . 26
5.8 Subpop 3 Correlation: G628 vs G1362 for Colon, CC = 0.3992 . . . . . . . . . . . 27
5.9 Global GoldMA365 vs 2YrTreaMA365 (Table 5.5, CC = -0.5256 . . . . . . . . . 28
5.10 Subpopu 1 GoldMA365 vs 2YrTreaMA365 (Table 5.5, CC = -0.5914 . . . . . . . . 28
5.11 Subpopu 2 GoldMA365 vs 2YrTreaMA365 (Table 5.5, CC = -0.7654 . . . . . . . . 29
5.12 Subpopu 3 GoldMA365 vs 2YrTreaMA365 (Table 5.5, CC = -0.5266 . . . . . . . . 29
5.13 Subpopu 4 GoldMA365 vs 2YrTreaMA365 (Table 5.5, CC = -0.6203 . . . . . . . . 30
5.14 Subpopu 5 GoldMA365 vs 2YrTreaMA365 (Table 5.5, CC = -0.6819 . . . . . . . . 30
5.15 Subpopu 6 GoldMA365 vs 2YrTreaMA365 (Table 5.5, CC = -0.5556 . . . . . . . . 31
List of Tables

1.1 CCRs: Gold vs SPY (Fin0717) ........................................... 3
4.1 CCRs: Gold vs EWJ Fin0717 (See Chapter 5) ......................... 14
5.1 Data Sources for Financial Data ........................................ 17
5.2 Description of Datasets Used .......................................... 17
5.3 CCRs: GoldMA365 vs LumberMA365 (FinMVAD) ..................... 19
5.4 Ratios (Subpop/Global) for CCRs in Table 5.3 ......................... 20
5.5 SCRs: GoldMA365 vs 2YrTreaMA365 (FinMVAD) .................... 23
5.6 Ratios (Subpop/Global) for CCRs in Table 5.5 ......................... 24
5.7 Ratios (Subpop/Global) for Fin0717 .................................. 24
5.8 CCRs: G1362 vs G628 for Colon ..................................... 25
5.9 Ratios (Subpop/Global) for CCRs in Table 5.8 ......................... 25
A lot of people have their great contribution to make this thesis possible. I would like to thanks my advisor Dr. Guozhu Dong for his guidance to make this thesis a reality. His immense knowledge, patience, and motivation helped me a lot for my research and my Master’s study. I could not have imagined a better mentor than him for my Master’s research.

I would also like to thank my committee members, Dr. Keke Chen and Dr. Derek Doran for their support and guidance towards the completion of this thesis. Their productive comments and suggestions helped me in different ways.

I also thank my colleagues at Kno.e.sis center and Wright State University for their support and suggestions.

Finally, I must express my very profound gratitude to my family for providing me with unfailing support and continuous encouragement throughout my years of study and through the process of researching and writing this thesis. This accomplishment would not have been possible without them.

Thank you.
1

Introduction

Correlation analysis is a frequently used approach to analyze the interdependence among variables in several practical applications including business, government, bioinformatics, medicine, finance, and investment management [Chunhachinda et al. 1997], [Strong 2008], [Elton et al. 2009], [Markowitz 1952], [Rogel-Salazar and Tella 2015], [Sandoval and Franca 2012], [Antonakakis and FILIS 2013], [Schwert 1989]. It shows how the behavior of one variable changes with the change in another variable. Strongly correlated variables mean that the behavior of those two variables is closely related. No correlation between the variables shows they do not change with respect to one another. This is how the traditional correlation analysis can be interpreted. However, traditional correlation analysis uses an overly simplistic approach— it studies how two given variables are correlated in an application by examining only their relationship in the entire underlying data space.

As a result, traditional correlation analysis may miss strong correlations, especially when such correlations exist only locally (over several subsets of the entire data space satisfying various conditions). To fill the gap associated with such weakness in traditional correlation analysis, we propose Conditional Correlation Analysis (CCR) approach in this thesis. The CCR approach proposed here is also motivated by the following: In this era of Big Data, we have a very large volume of high-dimensional and highly diverse data that are changing at a very high speed. It is now both possible and necessary to identify various different subpopulations within the dataset where given pairs of variables correlate differently, and analyze how correlation relationship evolves over time.

At the conceptual level, the CCR approach can be used to identify a small number of logical subpopulations, each of which being a pattern-defined subset of the whole data space, such that
correlation of two given variables in each subset is very unusual (e.g. very different from the global
correlation over the whole space of all the possible data) and the correlation relationships in the
different subsets are also very different from each other.

Importantly, the logical subpopulations and the discovery of these subpopulations are obtained
automatically based on the data and they may be unknown prior to the CCR analysis.

After introducing the concepts, this thesis discusses how to use Contrast Pattern Aided Regres-
sion (CPXR) [Dong and Taslimitehrani 2015] to identify strong and unusual conditional correlations.

The conditional correlation concept introduced in this thesis is very different from traditional
conditional correlation analysis [Baba et al. 2004], [Lawrance 1976], which is concerned with condi-
tional correlations of two variables given another variable.

The CCR approach has the capability to uncover several interesting types of conditional corre-
lations, including:

- It can identify logical subpopulation where two variables $x$ and $y$ have strong correlations but
  the global correlation between $x$ and $y$ is very weak.
- It can identify logical subpopulation where two variables $x$ and $y$ have strong correlations, the
  global correlation between $x$ and $y$ is also very strong, but the local correlations and the global
  correlation are very different (e.g. having very different slopes).
- It can identify multiple logical subsets where two variables $x$ and $y$ have strong correlations, and
  those correlations are very different from each other.

Such findings are useful for decision making in many applications, including business, finance, gov-
ernment, medicine, and investment management.

Table 1.1 illustrates conditional correlations in terms of the correlation between Gold price and
SPY price for the period between 2007 and 2017.

Correlation coefficients shown in the table represents correlation coefficients for five different
subpopulations of the entire dataset for the given period determined by the respective patterns
shown in the second column. Another correlation coefficient is for the default model, which is also
the subpopulation of the global data space. The last row in the table gives the correlation coefficient
for the global population.

The correlation coefficient for Gold and SPY in the entire period (last row of the table) is 0.0896,
which is only about 10% of the correlation value of 0.8711 in subpopulation M2. We also see that
the correlation coefficients for M2 and M3 have large comparable magnitude but they have opposite
signs (so one is a strong positive correlation and the other is a strongly negative one).
Using the CCR approach as example, we will argue and illustrate that Big Data and big knowledge analysis has both more challenges and opportunities. Moreover, the conditional correlation based approach is likely a very useful approach to uncover useful knowledge and to untangle the complexity of knowledge in the Big Data era.

In the rest of this thesis, Chapter 2 gives an overview of the related work. All the required preliminaries are presented in Chapter 3. Chapter 4 defines the concepts of conditional correlation analysis and associated measures, as well as how to use CPXR to mine such correlations. Chapter 5 reports experimental findings on several applications. Chapter 6 presents the application of CCR and Chapter 7 offers concluding remarks.

Table 1.1: CCRs: Gold vs SPY (Fin0717)

<table>
<thead>
<tr>
<th>Popu</th>
<th>Pattern $P_i$ and LR Models $M_i$</th>
<th>CC</th>
</tr>
</thead>
<tbody>
<tr>
<td>M1</td>
<td>$P_1: 103.2293 &lt;= SPY &lt; 148.8261$ [M_1: Gold = 2.2475 + 1.1337 * SPY]</td>
<td>0.3954</td>
</tr>
<tr>
<td>M2</td>
<td>$P_2: 31.1272 &lt;= EWJ &lt; 38.0454$ [M_2: Gold = -27.4408 + 1.5112 * SPY]</td>
<td>0.8711</td>
</tr>
<tr>
<td>M3</td>
<td>$P_3: 1.1211 &lt;= 10YrTrea &lt; 2.2422$ [&amp; 0.0000 &lt;= 2YrTrea &lt; 0.9859 [&amp; 5.7842 &lt;= NaturalGas &lt; 131.4251] [M_3: Gold = 226.2973 + -0.5373 * SPY]</td>
<td>-0.9044</td>
</tr>
<tr>
<td>M4</td>
<td>$P_4: 79.2249 &lt;= Dollar &lt; 87.3010$ [M_4: Gold = 94.7121 + 0.2616 * SPY]</td>
<td>0.3415</td>
</tr>
<tr>
<td>M5</td>
<td>$P_5: 0.0000 &lt;= 2YrTrea &lt; 0.9859$ [&amp; 52.3947 &lt;= VTI &lt; 76.1176 [&amp; 5.7842 &lt;= NaturalGas &lt; 131.4251] [M_5: Gold = 121.2390 + 0.2758 * SPY]</td>
<td>0.2360</td>
</tr>
<tr>
<td>Default</td>
<td>$P_d: not(P1 \lor P2 \lor P3 \lor P4 \lor P5)$ [M_d: Gold = 69.0615 + 0.2024 * SPY]</td>
<td>0.8402</td>
</tr>
<tr>
<td>Global</td>
<td>$M_g: Gold = 115.0426 + 0.0490 * SPY$</td>
<td>0.0896</td>
</tr>
</tbody>
</table>
2

Related Work

Our research focuses mainly on correlation analysis and related to the areas of regression and patterns which are the very broad topic and unrealistic to discuss here in detail. Therefore, we will have a closer look on only those fields that are very much related to our research. We are discussing Piecewise Linear Regression, Pattern Aided Regression Model and CPXR, and Multiresolution Correlation Analysis in this section.

2.1 Piecewise Linear Regression

The importance of effectively modeling different subpopulations with different regression models has been widely recognized, as is confirmed by the popularity of the piecewise linear regression model (PLR) [Mcgee and Carleton 1970], and the popularity of studies (e.g. [Cook and Weisberg 1983]) on heteroscedasticity in regression and correlation analysis. To the best of our knowledge, there are no algorithms that build PLR models with conditions on more than one variable except for CPXR [Dong and Taslimitehrani 2015] which we have used as the major algorithm to define patterns to identify potential subpopulation in our CCR approach.

Piecewise linear regression method or segmented linear regression method partitions predictor variable into several different intervals. Separate line of fit is calculated for each segment and the line segments are joined to each other at points called breakpoints.

CCR computed from Pattern Aided Regression (PXR) Models [Dong and Taslimitehrani 2015] are more general than PLR models. During condition formation, PLR model uses only one predictor variable to define intervals. This is not the case in CCR which uses multiple attributes for the formation of the conditions. PLR method can overall define the idea of conditional correlation analysis but CCR method using PXR is highly superior to the PLR method.
2.2 Pattern Aided Regression and CPXR

This work is closely related to Pattern Aided Regression and the CPXR algorithm [Dong and Taslimitehrani 2015]. However, this thesis is new in multiple fronts, including introducing the conditional correlation analysis problem, in the concepts for conditional correlation analysis, and in using the concepts and approach of this thesis to illustrate the challenges, opportunities, and potential approaches for Big Data analysis.

CPXR method builds the regression model on the multiple attributes. However, CCR makes use of the CPXR method by only taking two attributes for the construction of the regression model in each subpopulation. Both CCR and CPXR work the same way for pattern identification process.

2.3 Multiresolution Correlation Analysis

The Multiresolution Correlation Analysis (MCA) method [Feigelman et al. 2014] uses a single conditional variable to determine multiple subpopulations in the dataset where two variables (between whom the correlation is being identified) are highly correlated. It also uses the approach of visually identifying subpopulations based on the local pairwise correlation.

This method shows the importance of identifying the potential subpopulation in a dataset to better understand their relationship but this method is only suitable for the low dimensional data as it uses a single variable as a condition to identify potential subpopulations. It also uses the graphical method to visually identify those subpopulations.

Our approach is best suited in identifying the best subpopulation in highly diverse and high dimensional dataset using human understandable patterns. MCA uses a sorting variable as the conditional variable to define the subpopulation. CCR uses all the attributes (except the response variable) in the dataset as the conditional variables to define patterns in the dataset.

CCR uses the surprisingness measures to compare the correlation coefficients in each subpopulation with the global population to identify the most unusual subpopulation. MCA approach does not propose any measures to further compare the correlation of the subpopulation with the global population.
3 Preliminaries

In this section, we discuss following topics required for the CCR analysis.

- Pearson Correlation Coefficient
- Simple Linear Regression Model
- Patterns
- Discretization Method (Binning)
- Contrast Pattern Aided Regression Method

Throughout this thesis, we will use “variables” and “attributes” as synonyms.

3.1 Pearson Correlation Coefficient

Given a set $S = \{(x_1, y_1), ..., (x_n, y_n)\}$ of $n$ pairs of $x$ and $y$ values, the Pearson correlation coefficient is defined as

$$r_{x,y,S} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2} \sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}},$$

where $\bar{x} = \frac{\sum_{i=1}^{n}x_i}{n}$ and $\bar{y} = \frac{\sum_{i=1}^{n}y_i}{n}.$

The Pearson correlation coefficient is one of the most commonly used measures for evaluating the correlation between variables. The coefficient $r_{x,y,S}$ varies in a numeric range between $-1$ and $+1.$ A larger magnitude indicates a strong correlation, which can be strong positive correlation or strong negative correlation. A value of zero suggests that there is no correlation between the two variables.

1We add the three subscripts $x, y$ and $S$ for use later.
3.2 Simple Linear Regression Model

The simple linear regression model is most popular numerical prediction modeling method. It deals with modeling the effect of a single predictor variable on a response variable in the given application. The predictor variable is also called independent, regressors, or explanatory variable and the response variable is also called dependent variable.

Given a simple linear regression model \( f(x) = \alpha + \beta x \) of \( y \) based on \( x \), the \( R^2 \) of \( f \) on \( S = \{(x_1, y_1), ..., (x_n, y_n)\} \) is defined by

\[
R^2 = 1 - \frac{\sum_{i=1}^{n}(f(x_i) - y_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}.
\]

It turns out that the \( R^2 \) of the linear regression of \( y \) based on \( x \) is always the squared value of \( r_{x,y,S} \). The sign of \( r_{x,y,S} \) (positive or negative), is the same as the sign of the slope (the coefficient \( \beta \) of \( x \)) in the regression function. The magnitude of \( r_{x,y,S} \) is also strongly related to the magnitude of the slope \( \beta \). Below we often use the slope and the \( R^2 \) of linear regression models to refer to the strength of the correlation.

3.3 Pattern

We now define “patterns,” which will be used to define logical subsets in the data.

A pattern (also called as a condition) \( P \) is the conjunction of a small number of single-attribute conditions; a single-attribute condition has the form “\( A = a \)” if \( A \) is a categorical attribute and \( a \) is a value of \( A \), or the form “\( b \leq B < c \)” or “\( b \leq B \)” or “\( B < c \)” if \( B \) is a numerical attribute and \( b \) and \( c \) are values of \( B \). An example pattern is “\( \text{gender} = \text{female} \) & \( 30 \leq \text{age} < 40 \)”. A tuple \( t \) is said to satisfy a pattern \( P \) if \( t \) satisfies all of the single-attribute conditions in \( P \). (The notion of “\( t \) satisfies a single-attribute condition” is defined naturally.) Given a dataset \( D \), the matching dataset of \( P \) is defined as \( \text{mds}(P) = \{s \mid s \in D \text{ and } s \text{ satisfies } P\} \).

We now give some background information on patterns. Given a dataset \( D \), the \( \text{mds}() \) of patterns partition the set of all possible patterns into equivalence classes. Two patterns are in the same equivalence class iff they have the same \( \text{mds} \). An equivalence class of such patterns has just one maximal pattern (not contained in any other pattern in the equivalence class), and it typically contains multiple minimal “generator” patterns (not containing any other patterns in the equivalence class). All patterns in one equivalence class have the same “meaning” in conditional correlation analysis; we can pick any of them to represent the equivalence class. We often pick the minimal ones since they are concise and are easier to understand.
3.4 Discretization Method (Binning)

CCR method requires the transformation of numerical predictor variables to categorical counterparts (bins). We used equi-width and entropy based discretization methods for our thesis.

3.4.1 Equi-Width Discretization

Equi-width discretization divides each predictor variable \( X \) into \( k \) intervals of equal size. The width \( w \) of the interval is defined by,

\[
w = \frac{\text{max}(A) - \text{min}(A)}{k}
\]

where \( k \) is given by the user.

The interval boundaries are then given as:

\[
[\text{min}(A), \text{min}(A) + w), [\text{min}(A) + w, \text{min}(A) + 2w), \ldots, [\text{min}(A) + (k-1)w, \text{max}(A)]
\]

**Example:** Let’s assume \( A = \{1, 4, 8, 12, 16, 18, 24, 27, 32\} \) is a set of numerical values and \( k = 4 \).

Equi-width intervals are \([0, 8), [8, 16), [16, 24), [24, 32]\) and then,

1, 4 ∈ \( Bin_1 = [0, 8) \)
8, 12 ∈ \( Bin_2 = [8, 16) \)
16, 18 ∈ \( Bin_3 = [16, 24) \)
24, 27, 32 ∈ \( Bin_4 = [24, 32] \).

3.4.2 Entropy Based Discretization

Entropy based discretization [Fayyad and Irani 1993] is used to discretize the range of continuous-valued attribute into multiple intervals. This method is an iterative method where data are split multiple times where class variable is used to find the best splits to make the bins as pure as possible. This implies that the majority of the values in a bin have the same class label. The Entropy based discretization method is characterized by finding the splits with maximum information gain. Assuming \( A \) is a numerical variable with a binary class label, entropy based discretization can be summarized as:

1. Calculate entropy for the class label using,
3.5. CONTRAST PATTERN AIDED REGRESSION

\[ Entropy(A) = -p \log p - n \log n \]

where \( p = \frac{c_0}{A} \) and \( n = \frac{c_1}{A} \). \( c_0 \) and \( c_1 \) are number of instances with class labels 0 and 1, and \( A \) is the total number of instances.

2. Given a split point \( v \), calculate entropy for the class label using,

\[ Entropy(A, v) = \frac{A_1}{A} * Entropy(A_1) + \frac{A_2}{A} * Entropy(A_2) \]

where \( A_1 \) is the number of instances of \( A \) in \([\text{min}(A), v)\) and \( A_2 \) is the number of instances of \( A \) in \([v, \text{max}(A)]\).

3. Given a split point \( v \), calculate information gain for the class label using,

\[ Information Gain(A, v) = Entropy(A) - Entropy(A, v) \]

4. The goal is to get the split point with maximal information gain. The best split is found after examining all possible split points.

3.5 Contrast Pattern Aided Regression

The contrast pattern aided regression (CPXR) method, introduced in [Dong and Taslimitehrani 2015], is a new, robust pattern-based method for building prediction models. CPXR has the ability to deal with highly complicated and diverse predictor-response relationships, giving high prediction accuracies. The prediction models returned by CPXR are also easier to understand than those returned by artificial neural network and ensemble regression methods.

CPXR builds pattern-aided regression models (PXR) of the form,

\[ PM = ((P_1, f_1, w_1), (P_2, f_2, w_2), f_d), \]

where each \( P_i \) is a pattern, and each \( f_i \) and \( w_i \) are respectively the local multiple linear regression model and the weighting factor associated with \( P_i \). \( f_d \) is the default model, which is the regression model developed to characterize those samples that do not match any of those \( k \) patterns. For each instance \( X \), let \( \pi_X \) denote the set of patterns in \( \{P_1, ..., P_k\} \) matching \( X \). The predicted value for \( X \) by the regression function of \( PM \) is given by the weighted average of the predicted values of the local prediction models \( f_i \) such that \( P_i \in \pi_X \).

At a very high level, CPXR starts with a baseline regression model \( f_0 \) built on the original training dataset \( D_t \) using the standard multiple linear regression technique. The training database \( D_t \) is then divided into two subsets (to be viewed as two classes): \( LE \) (large errors) and \( SE \) (small
3.5. CONTRAST PATTERN AIDED REGRESSION

errors), using some cut point selected so that the size of $LE$ is a desired proportion of $D_t$. In order to
discretize input variables and define items, an entropy-based binning method is used. Then, CPXR
mines all contrast patterns of the $LE$ class, namely those patterns that match more samples in $LE$
than in $SE$. Such patterns are used since they are likely to capture subgroups of data where $f_0$
makes large prediction errors. Several filters are also used to remove certain patterns, which are very
similar to others. Then a local multiple linear regression model $f_i$ is built for each remaining contrast
pattern $P_i$. Those patterns and local multiple linear regression models, which do not improve the
accuracy of predictions, are removed at this step. CPXR then applies a double (nested) loop to
search for an optimal set $\{P_1, \ldots, P_k\}$. In these loops, CPXR adds a pattern to the set, or replaces
a pattern in the set by another one to minimize the prediction errors of the PXR model. CPXR
terminates when pattern addition or pattern replacement does not improve the accuracy of the PXR
model significantly.

The main idea of CPXR is to use a pattern, conjunction of several conditions on a small number
of predictor variables, as a logical characterization of a subgroup of data, and a local regression model
(corresponded to pattern) as a behavioral characterization of the predictor-response relationship for
data instances of that subgroup of data. What makes CPXR a powerful technique is it can pair
a pattern and a local regression model to represent a specific predictor-response relationship for a
subgroup of data. It also has the flexibility in pairing multiple patterns and local regression models
to represent distinct predictor-response relationships for multiple subgroups of data. Another reason
why CPXR substantially outperforms most of other prediction methods is that it uses an effective
mechanism to select a highly collaborative set of a small number of patterns to maximize their overall
combined prediction accuracy. (See [Dong and Taslimitehrani 2015] for more technical details.)
We discussed all the preliminaries required for CCR approach in chapter 3. In this chapter, we first present a simple example to introduce the main idea of this thesis. We then present the concept of Conditional Correlation Analysis (CCR) approach and the process to mine those CCRs from a given dataset.

4.1 CCR Concept

Following definition needs to be introduced to introduce the CCR concept.

Definition: A conditional correlation analysis (CCR) for two variables $x$ and $y$ over a dataset $D$ (representing an application) is given by a pattern $P$ (over the attributes of $D$), together with a linear regression model of the form $y = \alpha_P + \beta_P \cdot x$ that describes the correlation strength between $x$ and $y$ on the local models $mds(P)$.

Remark 1: We note that the correlation for the local models $mds(P)$ is a localized correlation, as it is concerned with the correlation of $x$ and $y$ only for data in $mds(P)$.

Remark 2: Using patterns to define local models has several advantages.

(1) Pattern defined local models are logical since they are defined using simple human-understandable conditions. They are more informative and more understandable than approaches that use data points as centers or representatives. For example, the local models defined by

"gender = female & 30 \leq age < 40"

is easy to explain.

(2) We can control the complexity of the patterns by selecting the shorter ones and also selecting
4.1. **CCR CONCEPT**

those that contain highly distinctive attributes.

(3) Searching over the set of potentially useful patterns is often more efficient than searching over the set of all possible clusterings, especially when contrast patterns [Dong and Bailey 2013] linked with the search objectives are used.

(4) Using pattern defined local models can lead to more robust measures.\(^1\)

The goodness of a CCR given by a pattern \(P\) depends on the strength of the correlation on 
\(mds(P)\) and how this localized correlation differs from the global correlation on the entirety of \(D\).  
Such goodness can be identified by comparing the localized regression model \(y = \alpha_P + \beta_P x\) over \(mds(P)\) and the global regression model \(y = \alpha + \beta x\) over \(D\).

Specifically, the surprisingness of the CCR given by \(P\) can be defined by considering two factors:  
The difference between their Pearson correlation coefficient values \((r_{x,y,D} vs r_{x,y,mds(P)})\), and the difference between their slopes \((\beta vs \beta_P)\).

There are two ways to compute the differences, one in terms of the ratio operator and the other in terms of the − operator.  
(1) The ratio based definition is helpful for visualizing the differences (including the difference in signs).  
The ratio of the CCR over \(P\) on the Pearson coefficient is given by \(ratio_r(P) = \frac{r_{x,y,mds(P)}}{r_{x,y,D}}\), and the ratio of the CCR on the slope is given by \(ratio_s(P) = \frac{\beta_P}{\beta}\).

(2) The − based definitions are:  
The difference of the CCR over \(P\) on the Pearson coefficient is given by \(diff_r(P) = \left| r_{x,y,mds(P)} - r_{x,y,D} \right|\), and the difference of the CCR on the slope is given by \(diff_s(P) = \beta_P - \beta\).

To avoid division by 0, we add 0.1 if the divisor is positive, or −0.1 if it is negative, before computing the ratio.

There are several possible interesting situations:

(1) \(surp_r(P)\) is negative. This indicates that the two correlations have different polarities (one being positive and the other being negative).

(2) \(|surp_r(P)|\) is very large or small. This indicates that the two correlations differ significantly with respect to their correlation strength.

(3) \(surp_s(P)\) is negative. This indicates that the two slopes have different signs.

(4) \(|surp_s(P)|\) is large. This indicates that the two slopes differ significantly in magnitude.

We note that the relationship between \(\alpha\) and \(\alpha_P\) is strongly related to the relationship between \(\beta\) and \(\beta_P\). Hence we will not introduce a measure on the difference between \(\alpha\) and \(\alpha_P\).

\(^1\)It may be possible that contrived data can be constructed where pattern-defined subpopulations do not have very unusual correlations but there are clusters having very unusual correlations.
4.2 CCR Example

To illustrate the CCR concept, we present a very simple example. Table 4.1 represents a CCR model for Gold and EWJ taken from financial data from 2007 to 2017.

A CCR model is defined by a set of patterns and a local regression model along with the correlation value between variables. The first column in the table represents different subpopulations within the dataset. Subpopulations are named from M1 to M7. Default represents the subpopulation determined after excluding all subpopulations (from M1 to M7) from the global population. Global represents the global population containing the complete dataset. Each population is determined by a set of patterns. The set can contain one or more patterns which identify the matching dataset for each subpopulation. A regression model is computed for each subpopulation which is later used to compute the correlation coefficient. Patterns and LR Models in Table 4.1 represents all the patterns and regression model for all subpopulations, default subpopulation and global population. The corresponding CC in the third column provides the correlation coefficient for all the populations.

The correlation value for global population is nearly zero which is not the case for subpopulations M2, M3, M4 and M6 (see Table 4.1). Gold and EWJ seems to be strongly correlated in those subpopulations. The default subpopulation also shows a stronger correlation than the global population but the strength is increased in the opposite polarity. CCR can extract these different behaviors in correlation between the variables in the dataset.
Table 4.1: CCRs: Gold vs EWJ Fin0717 (See Chapter 5)

<table>
<thead>
<tr>
<th>Models</th>
<th>Pattern $P_i$ and LR Models $M_i$</th>
<th>CC</th>
</tr>
</thead>
<tbody>
<tr>
<td>M1</td>
<td>$P_1: 79.2249 &lt;= Dollar &lt; 87.3010$</td>
<td>0.0428</td>
</tr>
<tr>
<td></td>
<td>$M_1: \text{Gold} = 120.0372 + 0.2182 * EWJ$</td>
<td></td>
</tr>
<tr>
<td>M2</td>
<td>$P_2: 52.3947 &lt;= VTI &lt; 76.1176$</td>
<td>-0.8404</td>
</tr>
<tr>
<td></td>
<td>$M_2: \text{Gold} = 384.4031 + -6.4409 * EWJ$</td>
<td></td>
</tr>
<tr>
<td>M3</td>
<td>$P_3: 103.2293 &lt;= SPY &lt; 148.8261$</td>
<td>-0.5755</td>
</tr>
<tr>
<td></td>
<td>$&amp; 0.0000 &lt;= 2YrTrea &lt; 0.9859$</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$&amp; 5.7842 &lt;= NaturalGas &lt; 131.4251$</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$&amp; 52.3947 &lt;= VTI &lt; 76.1176$</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$M_3: \text{Gold} = 263.6948 + -2.9840 * EWJ$</td>
<td></td>
</tr>
<tr>
<td>M4</td>
<td>$P_4: 36.9241 &lt;= Milk &lt; 46.7761$</td>
<td>-0.6872</td>
</tr>
<tr>
<td></td>
<td>$M_4: \text{Gold} = 285.3994 + -3.9741 * EWJ$</td>
<td></td>
</tr>
<tr>
<td>M5</td>
<td>$P_5: 31.1272 &lt;= EWJ &lt; 38.0454$</td>
<td>0.2074</td>
</tr>
<tr>
<td></td>
<td>$M_5: \text{Gold} = -4.9947 + 3.9664 * EWJ$</td>
<td></td>
</tr>
<tr>
<td>M6</td>
<td>$P_6: 36.9241 &lt;= Milk &lt; 46.7761$</td>
<td>-0.7686</td>
</tr>
<tr>
<td></td>
<td>$&amp; 79.2249 &lt;= Dollar &lt; 87.3010$</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$&amp; 52.3947 &lt;= VTI &lt; 76.1176$</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$&amp; 103.2293 &lt;= SPY &lt; 148.8261$</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$M_6: \text{Gold} = 471.2247 + -9.0693 * EWJ$</td>
<td></td>
</tr>
<tr>
<td>M7</td>
<td>$P_7: 1.1211 &lt;= 10YrTrea &lt; 2.2422$</td>
<td>0.1684</td>
</tr>
<tr>
<td></td>
<td>$&amp; 0.0000 &lt;= 2YrTrea &lt; 0.9859$</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$&amp; 5.7842 &lt;= NaturalGas &lt; 131.4251$</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$&amp; 31.1272 &lt;= EWJ &lt; 38.0454$</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$M_7: \text{Gold} = 133.9279 + 0.8287 * EWJ$</td>
<td></td>
</tr>
<tr>
<td>Default</td>
<td>$P_d: not(P1 \lor P2 \lor P3 \lor P4 \lor P5 \lor P6 \lor P7)$</td>
<td>0.5594</td>
</tr>
<tr>
<td></td>
<td>$M_d: \text{Gold} = 59.1174 + 1.1864 * EWJ$</td>
<td></td>
</tr>
<tr>
<td>Global</td>
<td>$M_g: \text{Gold} = 149.8587 + -0.6865 * EWJ$</td>
<td>-0.1664</td>
</tr>
</tbody>
</table>
4.3 Mining CCRs using CPXR

Given a dataset \( D \), we wish to mine a succinct set of CCRs that have highly unusual local subpopulation correlations. In this thesis, we utilize CPXR to discover, illustrate, and analyze CCRs\(^2\).

We make use of the Contrast Pattern Aided Regression (CPXR) method to mine CCRs. CPXR uses a multiple numbers of predictor variables to predict single response variable \( y \) and build a multiple regression model of the response variable \( y \).

To mine CCRs for two variables \( x \) and \( y \), we use a modified CPXR as follows. For each pattern \( P \), instead of building a multiple regression model of \( y \), we build a simple linear regression of \( y \) using the \( x \). The other parts of CPXR where all the variables except the response variable contribute for the pattern construction remain unchanged. Finally, for all the subpopulations defined by the patterns, a local regression model is constructed which is further used for computation of correlation on each subpopulation.

\(^2\)Other algorithms can also be designed to mine CCRs directly.
5

Experimental Results

This section discusses the results and findings of Conditional Correlation Analysis (CCR) performed in four different datasets. We used financial (Fin0717, FinMVAD, FinLagD) and medical (Colon) data (see Section 5.1) for our experiment. First, the findings indicate that highly unusual CCRs exists (when the global correlation is very low). Moreover, we present some of the CCRs discovered by our CPXR-based method and analyze them concerning unusualness measures. The computation time for CCR is also computed in the experiment.

It should be noted that we could not compare the method introduced in this thesis for CCR against others as this is the first and novel method considering the CCR approach.

5.1 Data Sets and CPXR Parameter Settings

We report on four datasets\(^1\) (see Table 5.2 for statistics).

1) Colon. This came from the microarray dataset for Colon Cancer [Alon et al. 1999]. The original dataset contained 2000 genes and 62 data tuples with two classes. We used the entropy based feature selection method [Fayyad and Irani 1993] to select 15 genes with highest information gain for use in the experiment. The 15 genes are (we used the column number in the original dataset as the gene number): G1362, G1361, G543, G1625, G1381, G1593, G1232, G1915, G628, G149, G1240, G1117, G840, G1730, G1953.

2) Fin0717. We collected the daily prices for the following 15 asset classes (Gold, Dollar, 2YrTres-
ury, 5YrTres-
y, 10YrTres-
y, CrudeOil, NaturalGas, RedWheat, Milk, Lumber, SPY, EWJ, AUD, VTI, EWS) between 03/08/2007 and 03/08/2017. Sources for attributes are provided in Ta-
ble 5.1. Most were taken from the Yahoo website, the dollar index was from Investing website and Treasury rates were from Treasury website.

\(^1\)We note that it can happen for datasets to have no highly unusual subpopulations.
## 5.1. DATA SETS AND CPXR PARAMETER SETTINGS

Table 5.1: Data Sources for Financial Data

<table>
<thead>
<tr>
<th>SN</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td><a href="http://finance.yahoo.com">http://finance.yahoo.com</a></td>
</tr>
<tr>
<td>2</td>
<td><a href="https://www.investing.com">https://www.investing.com</a></td>
</tr>
<tr>
<td>3</td>
<td><a href="https://www.treasury.gov">https://www.treasury.gov</a></td>
</tr>
</tbody>
</table>

For the next two datasets, we use $X(t)$ to denote the closing price of asset $X$ on day $t$.

3) FinMVAD. Given a period $p$ (which we fixed at 365 days) and an asset symbol $X$, we use a new attribute, denoted as $X_{dmvap}$, to represent the difference of $X$ relative to $p$-day MVA: For each day $t \geq p$, let $X_{dmvap}(t) = (X(t) - X_{mvap}(t-1))/X_{mvap}(t-1)$, where $X_{mvap}(t-1)$ denotes the $p$-period moving average of $X$ between $[t-p, t-1]$. (The resulting table after this transformation contains $p$ less rows than the original data.)

4) FinLagD. Given a period $p$ (which we fixed at 365 days) and an asset symbol $X$, we use a new attribute, denoted as $X_{dp}$, to represent the difference of $X$ relative to $X$’s value $p$-day ago: For each day $t > p$, let $X_{dp}(t) = (X(t) - X(t-p))/X(t-p)$. (The resulting table also contains $p$ less rows than the original data.)

Table 5.2: Description of Datasets Used

<table>
<thead>
<tr>
<th>Dataset</th>
<th>#attributes</th>
<th>#instances</th>
</tr>
</thead>
<tbody>
<tr>
<td>Colon</td>
<td>15</td>
<td>62</td>
</tr>
<tr>
<td>Fin0717</td>
<td>15</td>
<td>2347</td>
</tr>
<tr>
<td>FinMVAD</td>
<td>15</td>
<td>1982</td>
</tr>
<tr>
<td>FinLagD</td>
<td>15</td>
<td>1982</td>
</tr>
</tbody>
</table>

We used simple linear regression for the CPXR method. CPXR has two parameters (alpha for dividing data into LE and SE, and minsup for mining contrast patterns whose support in LE is at least minsup). The following parameter values were used: alpha = 0.5 and minsup=0.04 for Fin0717 (the original finance data without transformation); alpha=0.5 and minsup=0.03 for FinMVAD (finance data with 365 days moving average); and alpha=0.2 and minsup=0.03 for Colon.
5.2 Example of strong CCRs and analysis

We now show that quite often we can find strong CCRs with very unusual measure values.

Table 5.3 presents CCRs for GoldMA365 vs LumberMA365 found from FinMVAD (obtained using 356 days MVA transformation). We see that subpopulations M3 and M4 have $CC = 0.66$ and $CC = 0.76$ respectively, which are very different from that of the global correlation with $CC = 0.04$; the ratios (see Table 5.4) on $CC$ for M3 and M4 are 15 and 17 respectively; the ratios on the regression slope are 43 and 50 respectively. Figure 5.1 shows their global correlation for all data, Figure 5.3 shows their correlation for subpopulation 3, and Figure 5.4 shows their correlation for subpopulation 4. Table 5.4 presents the ratios for $CC$ and regression slope of the subpopulations vs the global correlation.
5.2. **EXAMPLE OF STRONG CCRS AND ANALYSIS**

Table 5.3: CCRs: GoldMA365 vs LumberMA365 (FinMVAD)

<table>
<thead>
<tr>
<th>Popu</th>
<th>Pattern $P_i$ and LR Models $M_i$</th>
<th>$CC$</th>
</tr>
</thead>
<tbody>
<tr>
<td>M1</td>
<td>$P_1: \ -0.5873 &lt; 10Y &lt; -0.1736$ $&amp;  \ 0.3622 &lt; CO &lt; 0.1256$ $&amp;  \ 0.1435 &lt; Mi &lt; 0.1719$ $&amp;  \ 0.2852 &lt; Lu &lt; -0.0144$ $&amp;  \ 0.2784 &lt; Na &lt; 0.3019$</td>
<td>0.2068</td>
</tr>
<tr>
<td></td>
<td>$M_1: \ Go = 0.1973 + 0.2198 \ast Lu$</td>
<td></td>
</tr>
<tr>
<td>M2</td>
<td>$P_2: \ -0.1736 &lt; 10 &lt; 0.24007$ $&amp;  \ 0.1167 &lt; EJ &lt; 0.0601$</td>
<td>0.5254</td>
</tr>
<tr>
<td></td>
<td>$M_2: \ Go = 0.0120 + 0.7255 \ast Lu$</td>
<td></td>
</tr>
<tr>
<td>M3</td>
<td>$P_3: \ -0.1167 &lt; EJ &lt; 0.0601$ $&amp;  \ 0.0542 &lt; Re &lt; 0.1451$ $&amp;  \ 0.2852 &lt; Lu &lt; -0.0144$</td>
<td>0.6626</td>
</tr>
<tr>
<td></td>
<td>$M_3: \ Go = 0.1275 + 1.1619 \ast Lu$</td>
<td></td>
</tr>
<tr>
<td>M4</td>
<td>$P_4: \ -0.5873 &lt; 10 &lt; -0.1736$ $&amp;  \ 0.3622 &lt; CO &lt; 0.1256$ $&amp;  \ 0.0542 &lt; Re &lt; 0.1451$ $&amp;  \ 0.3259 &lt; 2Y &lt; 0.3490$</td>
<td>0.7568</td>
</tr>
<tr>
<td></td>
<td>$M_4: \ Go = 0.3746 + 1.3437 \ast Lu$</td>
<td></td>
</tr>
<tr>
<td>Default</td>
<td>$P_d: \ not(P_1 \lor P_2 \lor P_3 \lor P_4)$</td>
<td>0.4919</td>
</tr>
<tr>
<td></td>
<td>$M_d: \ Go = -0.1232 + 0.2643 \ast Lu$</td>
<td></td>
</tr>
<tr>
<td>Global</td>
<td>$M_g: \ Go = -0.0211 + 0.0267 \ast Lu$</td>
<td>0.0429</td>
</tr>
</tbody>
</table>

Table 5.4: Ratios (Subpop/Global) for CCRs in Table 5.3

<table>
<thead>
<tr>
<th>Popu</th>
<th>Ratio CC</th>
<th>Ratio R²</th>
<th>Ratio α</th>
<th>Ratio β</th>
</tr>
</thead>
<tbody>
<tr>
<td>M1</td>
<td>4.8196</td>
<td>23.2288</td>
<td>-9.3136</td>
<td>8.2169</td>
</tr>
<tr>
<td>M2</td>
<td>12.2416</td>
<td>149.8574</td>
<td>-0.5694</td>
<td>27.1187</td>
</tr>
<tr>
<td>M3</td>
<td>15.4354</td>
<td>238.2533</td>
<td>-6.0211</td>
<td>43.4319</td>
</tr>
<tr>
<td>M4</td>
<td>17.6300</td>
<td>310.8185</td>
<td>-17.6837</td>
<td>50.2265</td>
</tr>
<tr>
<td>Default</td>
<td>11.4599</td>
<td>0.2420</td>
<td>5.8175</td>
<td>9.8808</td>
</tr>
</tbody>
</table>

Figure 5.1: Global GoldMA365 vs LumberMA365 (Table 5.3), CC = 0.0429
5.2. EXAMPLE OF STRONG CCRS AND ANALYSIS

Figure 5.2: Subpop 2 GoldMA365 vs LumberMA365 (Table 5.3), CC = 0.5254

Figure 5.3: Subpop 3 GoldMA365 vs LumberMA365 (Table 5.3), CC = 0.6626
5.3 Different datasets often have CCRs of different strengths

CCRs found from Colon for G1362 vs G628 are given in Table 5.8. The ratios for $CC$ and slope are given in Table 5.9. Among all subpopulations listed, the largest ratio on $CC$ is 1.67, and the largest ratio of the slope is 3.3. These are weaker than that for GoldMA365 vs LumberMA365 obtained from FinMVAD.

CCRs found from FinMVAD for GoldMA365 vs 2YrTreaMA365 are weaker than those found for Gold and Lumber, as shown in Table 5.5. Here, the most unusual CCR has a $CC = -0.77$, while the global correlation has a $CC = -0.53$. Tables 5.6 shows that the unusualness measure is much smaller than that for GoldMA365 vs LumberMA365 obtained from FinMVAD.

CCRs found from FinLagD for Gold vs Dollar are also weaker than those found for Gold and Lumber. Indeed, the most unusual Pearson correlation coefficients for the subpopulations we discovered were $CC = 0.05$ and $CC = -0.24$. These are not very different than the global correlation whose CC is $-0.19$. We omit the details here.

It can also happen that some datasets have no meaningful CCRs at all.
Table 5.5: SCRs: GoldMA365 vs 2YrTreaMA365 (FinMVAD)

<table>
<thead>
<tr>
<th>Popu</th>
<th>Pattern $P_i$ and LR Models $M_i$</th>
<th>CC</th>
</tr>
</thead>
</table>
| M1   | $P_1: \: -0.3622 \: \leq \: CO \: < \: 0.1256$  
      | $M_1: \: Go = -0.0110 + -0.2173 \: \times \: 2Y$ | -0.5914 |
| M2   | $P_2: \: -0.2852 \: \leq \: Lu \: < \: -0.0144$  
      | & $0.3259 \: \leq \: 2Y \: < \: 0.3490$  
      | & $0.3622 \: \leq \: CO \: < \: 0.1256$ | -0.7654 |
|      | $M_2: \: Go = -0.0062 + -0.6035 \: \times \: 2Y$ |    |
| M3   | $P_3: \: -0.2937 \: \leq \: EJ \: < \: -0.1167$  
      | & $0.2852 \: \leq \: Lu \: < \: -0.0144$  
      | & $0.3622 \: \leq \: CO \: < \: 0.1256$  
      | & $0.5873 \: \leq \: 10Y \: < \: -0.1736$  
      | & $0.1435 \: \leq \: Mi \: < \: 0.1719$  
      | & $0.2784 \: \leq \: Na \: < \: 0.3019$ | -0.5266 |
|      | $M_3: \: Go = 0.1164 + -0.3989 \: \times \: 2Y$ |    |
| M4   | $P_4: \: -0.1736 \: \leq \: 10Y \: < \: 0.2400$  
      | & $0.3622 \: \leq \: CO \: < \: 0.1256$ | -0.6203 |
|      | $M_4: \: Go = -0.0626 + -0.2059 \: \times \: 2Y$ |    |
| M5   | $P_5: \: -0.1167 \: \leq \: EJ \: < \: 0.0601$  
      | & $0.4417 \: \leq \: 5Y \: < \: 0.1174$  
      | & $0.3622 \: \leq \: CO \: < \: 0.1256$  
      | & $0.3259 \: \leq \: 2Y \: < \: 0.3490$ | -0.6819 |
|      | $M_5: \: Go = -0.0950 + -0.4525 \: \times \: 2Y$ |    |
| M6   | $P_6: \: -0.1435 \: \leq \: Mi \: < \: 0.1719$  
      | & $0.0542 \: \leq \: Re \: < \: 0.1451$ | -0.5556 |
|      | $M_6: \: Go = 0.0277 + -0.1521 \: \times \: 2Y$ |    |
| Default | $P_d: \: not(P1 \lor P2 \lor P3 \lor P4 \lor P5 \lor P6)$ | -0.6338 |
| Global | $M_d: \: Go = 0.0301 + -0.1072 \: \times \: 2Y$ | -0.5256 |

Abbrevs: CO – CrudeOilMA365, EJ – EWJMA365, Go – GoldMA365,  
Lu – LumberMA365, Mi – MilkMA365, Na – NaturalGasMA365,  
Re – RedWheatMA365, 2Y – 2YrTreaMA365, 5Y – 5YrTreaMA365,  
10Y – 10YrTreaMA365
Table 5.6: Ratios (Subpop/Global) for CCRs in Table 5.5

<table>
<thead>
<tr>
<th>Popu</th>
<th>Ratio CC</th>
<th>Ratio $R^2$</th>
<th>Ratio $\alpha$</th>
<th>Ratio $\beta$</th>
</tr>
</thead>
<tbody>
<tr>
<td>M1</td>
<td>1.1251</td>
<td>1.2659</td>
<td>4.1233</td>
<td>1.5385</td>
</tr>
<tr>
<td>M2</td>
<td>1.4561</td>
<td>2.1204</td>
<td>2.3383</td>
<td>4.2714</td>
</tr>
<tr>
<td>M3</td>
<td>1.0019</td>
<td>1.0038</td>
<td>-43.2760</td>
<td>2.8237</td>
</tr>
<tr>
<td>M4</td>
<td>1.1801</td>
<td>1.3928</td>
<td>23.3001</td>
<td>1.4572</td>
</tr>
<tr>
<td>M5</td>
<td>1.2972</td>
<td>1.6827</td>
<td>35.3193</td>
<td>3.2031</td>
</tr>
<tr>
<td>M6</td>
<td>1.0569</td>
<td>1.1172</td>
<td>-10.3309</td>
<td>1.0764</td>
</tr>
<tr>
<td>Default</td>
<td>1.2056</td>
<td>0.4017</td>
<td>-11.2097</td>
<td>0.7594</td>
</tr>
</tbody>
</table>

Table 5.7: Ratios (Subpop/Global) for Fin0717

<table>
<thead>
<tr>
<th>Popu</th>
<th>Ratio CC</th>
<th>Ratio $R^2$</th>
<th>Ratio $\alpha$</th>
<th>Ratio $\beta$</th>
</tr>
</thead>
<tbody>
<tr>
<td>M1</td>
<td>4.4113</td>
<td>19.4603</td>
<td>0.0195</td>
<td>23.1358</td>
</tr>
<tr>
<td>M2</td>
<td>9.7175</td>
<td>94.4298</td>
<td>-0.2385</td>
<td>30.8401</td>
</tr>
<tr>
<td>M3</td>
<td>-10.0898</td>
<td>101.8042</td>
<td>1.9670</td>
<td>-10.9645</td>
</tr>
<tr>
<td>M4</td>
<td>3.8099</td>
<td>14.5153</td>
<td>0.8232</td>
<td>5.3396</td>
</tr>
<tr>
<td>M5</td>
<td>2.6329</td>
<td>6.9323</td>
<td>1.0538</td>
<td>5.6286</td>
</tr>
<tr>
<td>Default</td>
<td>9.3727</td>
<td>0.7059</td>
<td>0.6003</td>
<td>4.1304</td>
</tr>
</tbody>
</table>
5.3. **DIFFERENT DATASETS OFTEN HAVE CCRS OF DIFFERENT STRENGTHS**

Table 5.8: CCRs: G1362 vs G628 for Colon

<table>
<thead>
<tr>
<th>Popu</th>
<th>Pattern $P_i$ and LR Models $M_i$</th>
<th>CC</th>
</tr>
</thead>
<tbody>
<tr>
<td>M1</td>
<td>$P_1: 49.3506 \leq G_{1362} &lt; 309.7854$</td>
<td>$M_1: G_{1362} = 23.4166 + 0.4341 \times G_{628}$</td>
</tr>
<tr>
<td>M2</td>
<td>$P_2: 10.3896 \leq G_{841} &lt; 360.6447$</td>
<td>$M_2: G_{1362} = 29.7957 + 0.3830 \times G_{628}$</td>
</tr>
<tr>
<td>M3</td>
<td>$P_3: 20.6792 \leq G_{1594} &lt; 88.08192$</td>
<td>$M_3: G_{1362} = 69.2597 + 0.1048 \times G_{628}$</td>
</tr>
<tr>
<td>Default</td>
<td>$P_d: not(P_1 \lor P_2 \lor P_3)$</td>
<td>$M_d: G_{1362} = 149.5552 + -0.0327 \times G_{628}$</td>
</tr>
<tr>
<td>Global</td>
<td>$M_g: G_{1362} = 85.3893 + 0.1314 \times G_{628}$</td>
<td>0.3363</td>
</tr>
</tbody>
</table>

Table 5.9: Ratios (Subpop/Global) for CCRs in Table 5.8

<table>
<thead>
<tr>
<th>Popu</th>
<th>Ratio CC</th>
<th>Ratio $R^2$</th>
<th>Ratio $\alpha$</th>
<th>Ratio $\beta$</th>
</tr>
</thead>
<tbody>
<tr>
<td>M1</td>
<td>1.5083</td>
<td>2.2751</td>
<td>0.2742</td>
<td>3.3029</td>
</tr>
<tr>
<td>M2</td>
<td>1.6662</td>
<td>2.7764</td>
<td>0.3489</td>
<td>2.9145</td>
</tr>
<tr>
<td>M3</td>
<td>1.1868</td>
<td>1.4085</td>
<td>0.8111</td>
<td>0.7978</td>
</tr>
<tr>
<td>Default</td>
<td>-0.6811</td>
<td>0.0524</td>
<td>1.7514</td>
<td>-0.2490</td>
</tr>
</tbody>
</table>

Figure 5.5: Global Correlation: G628 vs G1362 for Colon, CC = 0.3363
5.3. DIFFERENT DATASETS OFTEN HAVE CCRS OF DIFFERENT STRENGTHS

Figure 5.6: Subpop 1 Correlation: G628 vs G1362 for Colon, CC = 0.5073

![Figure 5.6](image1)

Figure 5.7: Subpop 2 Correlation: G628 vs G1362 for Colon, CC = 0.5604

![Figure 5.7](image2)
5.4 Example of evenly distributed CCRs

Some dataset doesn’t contain any kind of surprisingness with respect to the correlation coefficient in its subpopulations (identified by CCR).

On the basis of correlation coefficient, we can see an even distribution of subpopulations within the dataset which can also be a crucial factor in decision making for applications in finance and medicine. Table 5.5 represents such dataset where global correlation and correlation in all subpopulations identified by CCR are nearly equal. Under the given conditions, the relation between GoldMA365 and 2YrTreaMA365 shows the similar behavior for the span of ten years (2007-2017).

Correlation in the global population is -0.53 which is nearly same for all subpopulations from M1 to M6 (see Table 5.5). CCR always keeps looking for an unusual relationship among the variables in the dataset but might end up discovering nothing interesting in their relationship.
5.4. EXAMPLE OF EVENLY DISTRIBUTED CCRS

Figure 5.9: Global GoldMA365 vs 2YrTreaMA365 (Table 5.5, CC = -0.5256)

Figure 5.10: Subpopu 1 GoldMA365 vs 2YrTreaMA365 (Table 5.5, CC = -0.5914)
Figure 5.11: Subpopu 2 GoldMA365 vs 2YrTreaMA365 (Table 5.5, CC = -0.7654)

Figure 5.12: Subpopu 3 GoldMA365 vs 2YrTreaMA365 (Table 5.5, CC = -0.5266)
5.4. EXAMPLE OF EVENLY DISTRIBUTED CCRS

Figure 5.13: Subpopu 4 GoldMA365 vs 2YrTreaMA365 (Table 5.5, CC = -0.6203)

Figure 5.14: Subpopu 5 GoldMA365 vs 2YrTreaMA365 (Table 5.5, CC = -0.6819)
5.5 Advantages of CCR approach and CCR Mining

Experiments reported above show that strong correlation in subpopulations between two variables can be found while the global correlation between the two variables is very weak. This is demonstrated from Table 1.1. New knowledge such as such strong correlations in subpopulations can be used to better understand the relationship between the two variables and to help gain better results in applications.

Remark: From Table 1.1, we can see the following: the CC for subpopulation M2 is a very high 0.87 indicating the correlation between Gold and SPY is a very strongly positive one in this
subpopulation, and the $CC$ for subpopulation $M3$ is a highly negative $-0.90$ indicating the correlation between Gold and SPY is a very strongly positive one in this subpopulation. We also see that the global $CC$ is a very low $0.09$. This is an example where a highly positive correlation in a subpopulation cancels out a highly negative correlation in another subpopulation – there is no strong positive or strong negative correlation at the global level.

## 5.6 Computation time

We tested the CCR approach on a six processor machine (with a 1.6 GHz CPU, 4 GB of RAM), concerning running time. For the four datasets, we considered (having 15 attributes and between 62 and 2347 rows (data instances)), the running time is between 4 seconds and 122 seconds.
6

Applications of CCR

Conditional Correlation Analysis (CCR) is applicable in wide variety of applications including finance, medicine, business, economy, environmental sciences, and government. In this section, we discuss some of the applications of our proposed approach. It can study the relationship among variables for a dataset in deeper level by examining them in different subpopulations. It looks for the unusual relationship among variables in subpopulations and find the interesting subpopulation which can help for the analysis of the application. It can also be used to define the heterogeneity of subpopulations in the dataset.

6.1 Application on Financial Dataset

The stock market, also known as the equity market, is a vital component of a free market economy. It provides company for an access to capital in exchange for ownership given to investors. A person trying to invest in the stock market considers different factor before investing money in the market. Correlation is one of the major factors that plays an important role in making the decision to do so. CCR can help to determine the relationship among different factors that play a key role in the stock market. It can help determine the relation that was never discovered by the traditional correlation methods.

The application of CCR in different financial datasets (Fin0717, FinMVAD) (see Section 5.1) shows that CCR is capable of uncovering the strong relationship between financial variables in subpopulations of the dataset. Examples used in the experimental section shows that interesting correlation within the subpopulation exist even if the correlation in the global population is very low. These facts were never considered while using the traditional correlation analysis method.

CCR can help an investor make a right decision towards his/her investment by giving the proper and complete information about the change in variables within the subpopulation.
6.2 Application on Medical Dataset

We used a microarray gene sequence dataset to show the power of CCR in the field of medicine. Many another dataset from the medical applications can use CCR to determine the strength of the relationship among different variables in the subpopulations.

The global population might not give the true picture of the relationship between the variables. The conditions generated by CCR divides the dataset into different subpopulations and determine the relationship among those variables which can sometimes be very different from as presented by the global population.

The application of CCR in gene expression dataset (Colon) (see Section 5.1) shows that correlation between certain gene values are stronger in some subpopulations than in the global population (see Table 5.8).

Table 5.8 also shows that $G_{1362}$ and $G_{628}$ in the default population is negatively correlated which is in contrast to the global population. More interesting behavior can be uncovered in some other datasets in the medical application. This kind of information can provide completely new perspective to see the relationship between different variables.
7

Conclusion

In this chapter, we summarize the findings and discuss the future works for this thesis.

7.1 Summary

Correlation plays an important part in our day to day life. Identifying highly correlated variables and understanding their relationship to determine the trends can help uncover various interesting relations for many applications.

This thesis introduced the Conditional Correlation Analysis (CCR) approach. It discussed how to mine such CCRs, provided measures to evaluate the unusualness of CCRs, and gave experiments to evaluate and illustrate the CCR approach. Finally, it used the CCR example to illustrate the challenges and opportunities, as well as possible knowledge types and models for the new requirements, for the era of Big Data.

The subpopulation approach taken in this thesis may have big potential for modeling and computing other kinds of knowledge types concerning subpopulations. One major challenge for discovering such subpopulation based structural models is that there are a huge number of possible subpopulations that may need to be considered.

Many challenging issues remain open, including how to mine the most unusual local model’s correlations, with or without constraints. The concepts and approach can also be extended to many other knowledge types for use in the era of Big Data.

7.2 Future Work

In this work, we identified the potential subpopulation with the unusual correlation from the global population and described some of the surprisingness measures.
7.2. **FUTURE WORK**

Studying the behavior of different subpopulations can be taken to next level using CCR. An index to determine the population heterogeneity can be identified using the CCR approach. This can be achieved by comparing all the subpopulations with each other to determine how they are distributed in the global population.

CCR approach also can be used as a prediction modeling tool to further predict the relationship among different variables in the application by studying their behavior in subpopulations for a given span of time. Regression model determined for each subpopulation can be used to predict the relationship between variables for the future.
References


7.2. FUTURE WORK


