EFFICIENT DATA SPLITTING METHODS
FOR MACHINE LEARNING MODEL FITTING

by

Redouane Betrouni
A Dissertation
Submitted to the
Graduate Faculty
of
George Mason University
In Partial fulfillment of
The Requirements for the Degree
of
Doctor of Philosophy
Computational Sciences and Informatics

Committee:

_____	Dr. Jason M. Kinser, Committee Chair
_____	Dr. Edward J. Wegman, Committee Member
_____	Dr. Igor Griva, Committee Member
_____	Dr. Dhafer Marzougui, Committee Member
_____	Dr. Jason M. Kinser, Department Chair
_____	Dr. Donna M. Fox, Associate Dean, Office of Student Affairs & Special Programs, College of Science
_____	Dr. Fernando R. Miralles-Wilhelm, Dean, College of Science

Date: _____

Fall Semester 2021
George Mason University
Fairfax, VA

Efficient Data Splitting Methods for Machine Learning Model Fitting

A dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy at George Mason University

By

Redouane Betrouni
Master of Science
George Mason University, 2009
Bachelor of Science
University of Sciences and Technology Houari Boumediene, 1996

Director: Dr. Edward J. Wegman, Professor
Department of Computational and Data Sciences

Fall Semester 2021
George Mason University
Fairfax, VA

Dedication

bismi 'l-lāhi 'r-raḥmāni 'r-raḥīmi
بِسْمِ ٱللَّهِ ٱلرَّحْمَٰنِ ٱلرَّحِيمِ

To my loving Mother, my caring Father who sadly for me passed away before seeing this moment,
Wife and Daughters,
Brothers and Sisters.

Acknowledgments

First and foremost, I would like to thank my advisor and dissertation director Professor Edward J. Wegman for his contribution of time, indefectible help, and continuous support. Without his patience and perseverance during the many years I have been working with him, the development of this work certainly would not have matured to this point. I am very fortunate to have had the opportunity to work under his supervision. It has been an honor to be his forty-seventh Ph.D. student.

I would also like to express my sincere gratitude to Professors Jason Kinser, Igor Griva, and Dhafer Marzougui for taking an interest in my work, serving in my committee, and for their helpful and constructive comments. I am especially thankful to Jason Kinser for chairing my defense and to Igor Griva for being my examiner. I would like to thank our department administrator, Mrs. Karen Underwood for making all administrative business transparent to me, kept me organized and was always ready to help.

I gratefully acknowledge my employer, the U.S Census Bureau for the partial funding which made my Ph.D. work possible. I would like to particularly thank my supervisors Tim Fowler, Oliver Fisher, and Karen Battle for their continual support in my professional and academic development. Their encouragement and support allowed me the peace of mind to succeed in my research work. Working full-time while going to school at the same time can be challenging.

I give special thanks to my former Boss and mentor Dr. Aref Dajani, my colleague Dr. Brian Dumbacher, and fellow students William Ampeh, Dr. Yang Xu, and Dr. Sabyasachi Guharay being good buddies. I will certainly keep very good memories of the time we spent together discussing various research topics and other life concerns. Lastly but certainly not least, my deep and sincere gratitude to my family for their never-ending support and love; I am forever indebted to my mother Faouzia, who has always been encouraging me to finish my Ph.D. and graduate work, and to my father Mustapha, who sadly for me passed away before seeing this moment; He believed in my ability and always admired my determination to pursue my dream of education and my journey towards Ph.D. achievement. I cannot forget to mention an inspiration to my life; my uncle Abdelaziz who valued education and has always been cheering me on. I am very thankful to my elder brother Yacine, my younger brother Mohamed Elhadi, my spouse Souad and my daughters Rania, Lena, and Jennah, who have been supportive.

Table of Contents

List of Tables

List of Figures

Abstract

EFFICIENT DATA SPLITTING METHODS FOR MACHINE LEARNING MODEL FITTING

Redouane Betrouni, PhD

George Mason University, 2021

Dissertation Director: Dr. Edward J. Wegman

In this PhD dissertation, I developed a new sampling method which I named PCA-Systematic sampling as an improved stratified systematic sampling to optimally split data into training and testing subsets. This procedure will help machine learning algorithms avoid the classical mistake of overfitting. While it might be slightly more computationally expensive, it makes up for this apparent weakness by having a better estimate of test error and improving prediction accuracy. The dissertation provides computational and theoretical evidence to support the benefits of the new proposed sampling design over traditional approaches. Examples and mathematical evidence are presented to show how traditional splitting methods such as simple random sampling to partition data can distort relationship between important covariates and the variable of interest for the test dataset and as a consequence leads to either poor model construction or poor model fitting assessment.

In this dissertation, I create a sampling utility score index as a data quality control tool to assess data splits or sampling designs. This dissertation demonstrates the benefits of my sampling utility index as its mathematical property is derived and studied, sensitivity analysis is conducted to investigate how it behaves under different scenarios of sampling designs.

Finally, this dissertation contributes to the field of survey sampling and predictive modeling when the new developed methodology is implemented on three distinct publicly available datasets. I show in this dissertation how this new scheme of new sampling design developed and named PCA-Systematic can be used as an application on real surveys data like the Annual Survey of Public Employment and Payroll (ASPEP) and the American Housing Survey (AHS) data. I provide evidence of improvement in estimates with comparison to the traditional methods of systematic sampling. My novel PCA-Stratified-Systematic sampling method outperforms current and best state of the art sampling methods for the classification problem of Fisher IRIS data.

Chapter 1: Introduction

This chapter begins with an overview of data splitting. It includes the methodology and its purpose for both machine learning and statistical modeling. This chapter reveals a crucial area of improvement for data splitting. Treating data splitting as a typical sampling problem will make it more efficient. Using the theory of sampling and best practices to create the training data set will help mitigate overfitting and improve the fitted model's prediction accuracy. The typical use of simple random sampling for data splitting may distort important information among the variables for the training data, test data, or validation data (that is, the relationship between the target variable Y and the set of features X if random data splitting is conducted poorly). Consequently, this will affect the fitted model's quality, the tuned model, and the estimation of the selected model's test error. Thus, this chapter presents diverse sampling designs and how and when to pick one strategy over another, which essentially depends on the data structure.

This chapter describes Principal Component Analysis (PCA) at the end, a critical component for my novel PCA-Systematic sampling design.

1.1 Data Splitting

Data splitting is used in statistical analyses, mainly in applications of variance and bias estimation of estimators. For example, the Jackknife resampling technique, which was initially developed in 1949 by Quenouille, Maurice H, then later in 1956 enhanced to correct for bias [28, 29] is one variant of splitting data using either Delete-one Jackknife or Delete-k Jackknife [12]. These two methods of data splitting are operationally equivalent in terms of implementation to cross-validation using either Leave-One-Out Cross-Validation (LOOCV) or k-Fold Cross-Validation (k-fold CV), respectively [15]. In 1958 John Tukey made a

substantial adjustment to the technique, made it more generalizable, and coined the term Jackknife. The technique rendered variance estimation of estimators plausible even when a direct mathematical formula is hard to derive algebraically unless one uses numerical approximations [38]. The other primary application of data splitting is in machine learning for fitting models and error assessment of models. Data splitting is used as a strategy for algorithm selection and estimation of risk associated with selecting a particular algorithm [3]. There are two fundamental tasks needed while fitting a machine learning algorithm

- Model selection which involves estimating the performance of different candidate models in order to choose the best one [14]. Data splitting is one method that can be used to accomplish this task. The portion of the data that is used for this purpose is known as a validation set.

- Model assessment involves the task of estimating the prediction error (generalization error) of the selected final model in the step of model selection above [14].

There are several tools to perform this task, but one way to achieve this task is to use data splitting. This involves reserving a portion of data that is unseen by the fitted model (data that has not been used while fitting and selecting the model). This portion of data is known as test data. This task is done independently from the fit, which means it is done at the beginning before the model is fitted.

Data splitting is an essential tool needed to estimate the test error associated with fitting a given statistical learning model. The test data set can be used to evaluate model performance. Another essential purpose of data splitting is to determine the appropriate level of model complexity, such as determining the optimum number of parameters or tuning parameters in the case of fitting parametric models.

Data splitting in machine learning is necessary because data used in estimation or fitting a model cannot provide fair overall assessments of the sampling error inherent in the model. Hence, there is a need to separate the fitting from the error assessment. To meet this goal non-overlapping sets of the data are necessary.

Numerous papers have been published in the past to show data splitting as a method to determine the validity of statistical models. For example, Ronald D. Snee in 1977, in the context of multiple linear regression and ridge regression models, compared multiple validation methods with data splitting in which a portion of the data is used to estimate the model coefficients, and the remainder of the data is used to measure the prediction accuracy of the model [35]. He recommended using The DUPLEX algorithm, developed by R. W. Kennard [27, 35], for dividing the data into the estimation set and prediction set when there is no obvious variable such as time to use as a basis to split the data. [35]

In Ronald D. Snee work, data splitting based on the DUPLEX algorithm was compared with three methods of validation that were based on comparison of the model prediction of $\hat{y}$ and coefficients $\hat{\beta}_i$ with an assumed physical theoretical model, collection of new data to check model prediction and comparison of results with theoretical models and simulated data. [35]

Picard and Cook examined the principals of data splitting in the context of ordinary least regression (OLS) models [26]. They discussed the topics of how many observations n_V should be allocated for assessment and model validation. They established a mathematical framework that can quantify the trade-off between prediction and validation as a function of n_V and n_E, where n_E is the number of observations allocated for parameters estimation and model development. They were able to devise a useful criterion that can serve a device to be used as an objective function to be optimized. In their work, they mimicked the Cp statistic developed by C. L. Mallows [22] and using the integrated mean squared (IMSE) error of prediction of future observations [26]. Cp Statistic is an alternative criterion used to assess the fit of multiple linear regression model similar to the adjusted coefficient of determination R^2 as it addresses the issue of overfitting by penalizing models with large number of predictors p and favors less complex models.

In Picard and Cook work, the predicted value of an unseen future observation is computed using the fitted model [26].

LeBaron and Weigend found that the impact of model accuracy and performance can be significantly affected by the data splitting mechanism [19]. They evaluated data splits using residual bootstrapping, which assumes that the model is correctly specified and that the residuals are independent and identically distributed but not necessarily normally distributed. Their work was conducted in the context of Artificial Neural Network (ANN) models for financial time series model fitting using the New York Stock Exchange data. They found that factors needed to construct models such as feature selection, random weight initialization, and choice of the number of hidden units might not have the same impact the data splitting has on the fitted model's quality. Data splitting affect the variability in model performance substantially.

Having random noise in the training examples can lead to overfitting [24]. Overfitting can occur when a given machine learning model fits the training data so well, but the model performs poorly outside the training data; this can be caused if the model is to try to fit every training data, including noisy data. If the dataset split is poorly implemented, the data subsets will not sufficiently cover the data, and especially the variance will increase [24, 30].

There are different ways of splitting a dataset into training and validation sets:

- One time split into training and testing sets, this is known as the Holdout Method.
- Multiple times (Cross-validation), very similar to Jackknife operationally.
- 3-way splitting into training, testing, and validating, this is popular in ANN.
    1. The training set is basically to train the model; this is the first stage of the 2-way and 3-way splitting.
    2. The testing set is a reserved data portion to tune parameters for parametric learners.

3. The validation set is reserved for pure assessment for the model to be selected in the testing stage above; this portion of data is untouched by model fitting in the first stage or model pruning or tuning in the second stage.

When data exist with abundance, it is best to choose a 3-way split by randomly dividing the data into three parts to form the training, validation, and testing data portions. It is a difficult task, however, to achieve an optimum way to allocate sample sizes to the three parts as this depends on the signal-to-noise ratio that will end up in each portion [14].

There are other alternative methods to approximate the validation step of data splitting [14]. These methods use analytical expressions to be used for model selection and estimate the selected model's test error. These methods are based on the maximum likelihood concept. In a nutshell, the goal is to maximize the probability of the assumed model condition on the observed training data.

The mathematical expression penalizes having a large number of parameters to favor less complex models. These methods are considered in-sample prediction error estimates, which can be over-optimistic compared to the out-of-sample test error estimates that can be provided by data splitting.

These methods include but not limited to

1. Mallow's C_p metric was developed in 1973 by Colin Lingwood Mallows [22] for model selection and to evaluate multiple linear regression models that are fit using ordinary least squares as an objective function. The C_p metric is defined to be

$$C_p = \frac{1}{\hat{\sigma}^2} RSS_p - N + 2p$$

2. Akaike information criterion (AIC) a statistic developed by Hirotugu Akaike in 1974 [1]. The value of the AIC for the model is the following

$$AIC = 2\lambda - 2\log(\widehat{L})$$

where λ is the number of estimated parameters in the model and $\widehat{L}$ is the maximum likelihood function for the model.

AIC explicitly includes the number of parameters to be fitted in the model. Models with lower AIC are preferable. For smaller sample sizes, there is a higher risk that AIC will select models that are too complex with a large number of parameters, which leads to the issue of overfitting. It is suggested in this case to use a modified version of AIC [8]; AICc given by

$$AICc = AIC + \frac{2\lambda^2 + 2\lambda}{N - \lambda - 1}$$

3. Bayesian information criterion (BIC) was developed for model selection using the Bayesian approach. This statistic (divided by 2) is equal to the SC(Schwarz criterion), which was developed in 1978 by Gideon E. Schwartz [34]. A model with higher posterior distribution would be selected with this strategy. The standard formula for BIC is

$$BIC = \lambda \log(N) - 2 \log(\widehat{L})$$

where λ is the number of estimated parameters in the model, $\widehat{L}$ is the maximum likelihood function for the model and N is the number of data points used by the fitted model.

BIC imposes a heavier penalty on complexity than AIC.

As $N \to \infty$, BIC outperforms AIC because most likely BIC will select the correct model [14] for larger sample sizes while AIC tends to select models that are too complex [14].

When N is relatively small, BIC favors models that are too simple (with low numbers of parameters) [14].

4. The minimum description length (MDL)

Vapnik–Chervonenkis theory was developed by Vladimir Vapnik and Alexey Chervonenkis during 1960-1990, according to the Vapnik's structural risk minimization (SRM) principle, which stated that to achieve the smallest bound on the test error by controlling (minimizing) the number of training errors, the machine (the set or functions) with the smallest VC dimension should he used.[40]. This approach circumvents the issue of overfitting and strikes a balance between having a model that is too specific or too general.

1.2 Data Splitting a Sampling Problem

Subject matter experts and researchers are currently looking at ways to improve the data splitting mechanism. Data splitting has a significant impact on model performance [44] therefore appropriate methods of data splitting needs to be conducted.

The problem of appropriate data splitting can be handled as a statistical sampling problem [30].

Numerous studies have been done to investigate whether different data splitting methods can lead to an improvement in terms of bias and variance reduction. For example, Wu et al. in 2013 have evaluated the performance of three different data splitting methods when fitting an ANN model on three real datasets [44]. Earlier, Reitermanovà in 2010 presented a survey of existing splitting methods applicable to the data splitting problem [30]. Selecting a particular sampling design when performing data splitting can have major implications on the quality of the data subsets resulted from the 3-way partitioning of the data into training, testing, and validating during ANN model development [23].

May et al. in 2010 developed a novel approach to stratified sampling based on Neyman sampling of the self-organizing map (SOM) that can reliably generate high-quality data subsets used for training, testing, and validating ANN model [23].

This dissertation does not use any model assumption to develop an optimal validation dataset. It addresses the issue of data splitting in machine learning as an appropriate sampling problem to achieve efficient data splitting. The aim is to use sampling theory and

best practices to form the training data.

The focus in this dissertation is for the training data portion to conserve the multivariate probability distribution that exists between the features and the target variable over all instances after the data split.

The core of efficient data splitting is to enable the best practices of sampling designs. Treating the training data as the best possible sample representing the full available data is a fundamental step towards achieving the best model fit for machine learning.

This dissertation shows that fitting a given machine learning using training data created by this strategy will minimize the risk of overfitting and increase the fitted model's prediction accuracy. This is demonstrated through the simulation work in Chapter 3 and real data examples in Chapter 4.

In sampling, they generally want each subset to have three main fundamental properties:

- Randomness is the first requirement. For example, in survey sampling, the ideal goal is to produce the smallest possible sample that allows the investigator to make inferences about the population. This is done by taking a sample from the population of the study. Usually, the population is quite large, and conducting an entire census which consists of selecting every member of the population is costly and time-consuming. Survey statisticians will design a probability sample such that every unit has a known, non-zero probability of selection. The process of selection is made random. If the selection of a sample denoted by S from a given population denoted by Ω is not random, this is referred to as a non-probability sample. It can be shown that this could potentially lead to serious bias in any scientific research study; without randomness, results would not be reliable, and most likely, the conclusion inferred from the sample will be misleading.

- Representativeness of the actual population. Following basic principles of sampling theory to ensure that the sample units are diverse like the general population. Upon doing this, estimates of the target variable parameters such as the mean, median, or

quantiles using the sample will be similar and close enough to the entire population's target variable parameters.

- Replicability, reproducible randomization, refers to the idea that the selected random sample taken from the population can be reproduced again. This way, reproducible research can be achieved, and results can be repeated. For example, in the R software package, the set.seed() function can be used to perform this task by setting a non-random seed. The obtained random sample during simulation work can be identically replicated if the same seed is used prior to simulation. This will provide other researchers to independently replicate the randomization work to achieve similar results.

There are various ways of doing this and multiple sampling designs to choose from. Depending on the distribution of the population data, different designs should be selected.

1. Simple random sampling with replacement

Simple random sampling with replacement (SRSWR) is the simplest form of selecting a probability sample [4]. In an SRSWR of size n from a population of size N, each unit must have the same probability of being included in the sample. Units can be selected more than once. In theory, a unit can be selected anywhere between 0 to n times. There are N^n possibilities to select a sample of size n from a population of size N, this form of sampling is mathematically attractive because the observed values that get sampled will result in an iid (independent and identically distributed) sample; however, this is not practical in the context of surveys, and other disciplines since the information gained by repeating a unit more than once do not add any value or contribute any additional information about the population.

2. Simple random sampling without replacement.

Simple random sampling without replacement (SRSWOR) is the second most simple form of selecting a probability sample [4]. To select an SRSWOR sample of size n from a

population of size N, each unit in the population must have an equal inclusion probability; this means that all units in the population will have the same chance to be included in the sample. Furthermore, once a unit is selected, it can not be drawn again. Typically the selection is made "without replacement" to avoid choosing any member of the population more than once. The simple random sampling model without replacement can be represented with an urn model where N balls are numbered 1 through N are placed in an urn, and the process involves randomly shuffling the balls inside the urn and drawing n balls but one a time. If a ball is selected, it will not be placed back in the urn.

There are $\binom{N}{n}$ possible subsets that can be selected with this sampling procedure. Several algorithms can be used to design an SRSWOR sample. These randomization processes can be handled manually if the population is finite and relatively small, but it can be implemented through programming and using a discrete or continuous uniform distribution, which is a special probability distribution that plays an important role in simulation. The random events generated by the uniform distribution have an equal probability of occurrence. One can number the units in the population from 1 to N or equivalently label the elements in the population $\Omega = \{\omega_1, \ldots, \omega_n\}$ and associate a randomly generated uniform number between 0 and 1 to each element, then sort the generated numbers in ascending or descending order and keep the first n numbers in the sorted list, at the end select their associated (corresponding) units, because with the uniform distribution over the interval $[0, 1]$ all subintervals of [0,1] with equal lengths have the same probability measure. This algorithm will ensure that each unit has the same probability of inclusion.

With SRSWOR sampling design, every one of the $\binom{N}{n}$ distinct samples has an equal chance of being drawn.

3. Stratified sampling.

In Stratified sampling the population is partitioned into L subpopulations called **strata**. These strata form a partition of the population; means that they are disjoints; they do not overlap, and their union is equal to the full population. The population Ω of size N units

is divided into subpopulations $\Omega_1, \ldots, \Omega_L$ of $N_1, \ldots, N_L$ sizes respectively. so that

$$N_1 + \ldots + N_L = N$$

and

$$\Omega_1 \cup \ldots \cup \Omega_L = \Omega.$$

A probability sample is selected from each stratum. The selections of the L samples are independent of each other. Combining the L samples by taking their union creates a sample known as a stratified random sample.

Depending on the population, without stratification choosing simple random sampling procedures with and without replacement might lead to a bad sample as it might not be representative or does not reflect the diversity of the population if it misses an important group in the population. If stratification is done properly by dividing a heterogeneous population into layers that are internally homogeneous, then a stratified sample will have better precision in the estimates of characteristics of the whole population compared to a simple random sample.

4. Cluster sampling.

Oftentimes a reliable list of all elements in the population is not available, or constructing such a list is prohibitively expensive. In these situations, sampling plans such as simple random sampling and stratified sampling methods would not work because these methods of sampling require direct access to the sampling units. Cluster sampling is a sampling plan that can be used in this situation because even though a complete list that identifies each and every population unit does not exist, a natural grouping of the population units into groups or **clusters** often can easily be defined. Initially, cluster sampling was developed in survey methodology mainly to reduce the cost for demographic surveys where the population elements are scattered over a wide geographic area. For example, in order to select a sample of households in a given city for purposes to investigate the utilization of public transportation services among residents of the city and if there are administrative records

or directory that list every household in the city, then it can serve as a sampling frame from which the sample can be selected. However, if such a list does not exist, it would be extremely costly to develop such a sampling frame.

It might be feasible, however, to construct a list of city blocks. The list of city blocks is used as the sampling frame. Each city block is a cluster of households, and the desired sample of households can be randomly selected by first taking a probability sample of clusters (city blocks), and every household in the selected block is surveyed and made part of the sample. This is known as a one-stage cluster sampling; if instead, a subset of households is selected in the selected block, this is known as a two-stage cluster sample. Generally, cluster sampling can have multiple stages. Each element of the population belongs to one and only one cluster.

Cluster sampling works best when the clusters and mutually homogenous but internally heterogeneous with respect to the variable(s) of interest. In other words, an optimum cluster sample is obtained when the variance of the target variable y is small between the clusters but large within the clusters.

5. Multi-Stage sampling.

Cluster sampling often requires multiple stages of selection to reduce cost. The selection of the units is done in various stages. For example, The Centers for Disease Control and Prevention (CDC) conducted an epidemiologic research study in April of 2020 to estimate SARS-CoV-2 prevalence among frontline healthcare personnel who care for COVID-19 patients. [36] The sampling design involves three stages;

(a) Twelve states participated in the survey

(b) For each participating state, a specified number of enrolled hospitals were selected

(c) Finally, for each hospital, 250 medical workers were sampled

6. Systematic sampling.

Figure 1.1 describes the algorithm of random selection using systematic sampling. In systematic random sampling, a random start and a sampling interval is determined based

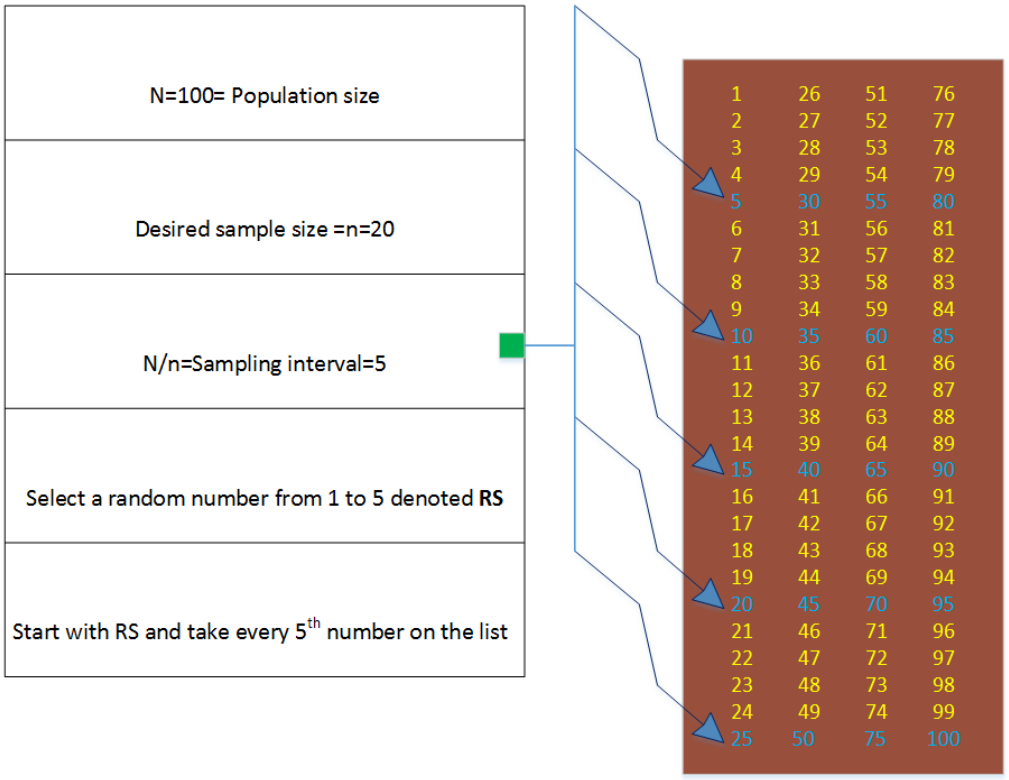

Figure 1.1: Diagram Describing Simple Systematic Sampling.

on the desired sample size. The choice of the starting point is the only random aspect of systematic random sampling. If the starting point is chosen uniformly from the first $\frac{N}{n}$ elements, then every unit in the population has the same probability of inclusion in the sample.

Suppose that N elements in the population are numbered 1 to N in some order. Suppose that the desired sample size in n. Without loss of generality we can assume that N is a multiple of n, that is N can be expressed as the product of n and some integer k

$$N = nk.$$

k is known as the sampling interval and basically calculated by dividing the population size N by the desired sample size n.

The steps needed to draw a systematic sample of size n are as follows:

(a) select a random number between 1 and k.

(b) if the selected number in step 1 is i then the first i^{th} element in the list is selected to be part of the sample.

(c) select every k^{th} unit after the i^{th} unit to be part of the sample.

In the end, the sample is the list with the following indices:

$$i, i+k, \ldots, i+(n-1)k$$

The current sampling methodology described above is known as linear systematic sampling (LSS); it was first introduced by Madow, W.G. and Madow, L. H. in 1944 [21]. This design of sampling leads to k systematic samples with each of sample size n assuming the population size N is a multiple of k, $N = nk$. Since N is not always a multiple of k, this methodology will lead to k but different sample sizes. Lahiri suggested circular systematic sampling (CSS) that will provide a constant sample size and unbiased sample mean [18].

Assume the N population units are arranged to form a circle, select a random number r , $1 \leq r \leq N$, then select the r^{th} element in the circle and every k^{th} unit going around the circle thereafter, once n elements are accumulated stop the selection.

Systematic sampling is equivalent to simple random sampling if the population is randomly ordered, that is; if the numbering in the population is random.

Systematic sampling is operationally convenient and leads to estimators with lower mean square errors than simple random sampling if the data file is "sorted appropriately". Sorting will offer implicit stratification. Stratification will guarantee that the sample do not misrepresent any groups in the population.

I recommend never to sort the data, but just the index vector for efficiency. Systematic sampling outperforms simple random sampling if the data file to sample from is preprocessed properly. The pre-process involves a sort by a variable that is highly correlated with

the key variable of interest. Systematic sampling is related to cluster sampling as it can be regarded as a special cluster sample, the population is divided into $k = N/n$ possible systematic samples and a simple random sample of one cluster is randomly chosen.

7. And much more complex sampling designs.

To properly design a sampling strategy that is efficient and at a manageable cost, both stratification and clustering are required elements. The population to sample from is organized hierarchically before the randomly selected sample is obtained [39]. Complex sampling designs involving multiple stages are necessary to ensure that the sample selected represents the population. Clustering is needed to reduce cost, and stratification is necessary to increase the precision of the sample. For example, for household surveys in the United States, such as the American Housing Survey (AHS), the first stage of the sample selection is accomplished by dividing the country into counties or groups of contiguous counties known as PSUs (Primary Sampling Units). The second stage of sampling design involves selecting the housing units within the selected PSUs in the previous step. The selection is made systematically after sorting the file geographically for implicit stratification.

Data miners employ random sampling to create only a subset of the entire set of data of interest. The goal is to reduce the size of the data to be processed when more expensive algorithms are executed. This strategy is successful as long as the sample is representative. [37].

Tan et al. [37] argued that this could be achieved if the sample has approximately the same property (of interest) as the original set of data. For example, if the mean of the data is the parameter of interest, then a sample is representative if its average is close to that of the original data [37]. This dissertation provides a novel sampling algorithm that can preserve the property (of interest) of the original data for the randomly selected sample.

One of the contributions in this dissertation is to use any known characteristic of the population to ensure that the sample randomly drawn "covers" the variation of the target variable in the population and also preserves the relationship between all variables. The

variance-covariance matrix of the sample should resemble the variance-covariance matrix of the population.

The **Frobenius** norm, also known as the Hilbert-Schmidt norm, can be used to assess the quality of a given data split; for example, computing the difference of the two correlation matrices and evaluating the Frobenius norm of the difference would provide us with useful information about the data split.

The Frobenius norm for a given matrix $A = (a_{ij})$ is defined as follows

$$\|A\|_F = (\sum_{i,j} a_{ij}^2)^2$$

Because the variance is not bounded, I prefer to use the correlations matrix to compute the Frobenius norm due to the fact that the correlations are bounded between -1 and +1. So computing the Frobenius norm of the difference of the two correlation matrices and if the difference is high, then this is an indication that the data split or the sampling scheme produced a sample that has distorted the relationship between the original features. It is important to design a sample or a data split that constraints to keep the norm of this difference to a minimum. Preserving the relationship between the variables is important for machine learning.

Using the correlation matrices instead of the variance covariance matrix renders comparing data splits unit free and independent of the data. Thus, even if the data is astronomically large or microscopically small, it would not differ.

1.3 Efficiency Comparison of Systematic Sampling Relative to Stratified Sampling and Simple Random Sampling

The efficiency of simple systematic sampling designs has been studied by Madow, W.G. and Madow, L. H. [21] and Cochran [10]. They compared its efficiency with both of stratified and simple random sampling designs.

The efficiency of systematic sampling is a function of the population ordering [32]. Sampling designs that produce unbiased samples are desirable. The Mean Squared Errors of the different sampling strategies were compared. Sampling designs associated with lower MSE is the goal. The MSE can be written as the sum of variance and the bias squared. If the bias is zero then comparing the MSE is equivalent to comparing the variances. If the sampling designs produce unbiased estimates of the population means then it is sufficient to evaluate their variances.

Each column in Table 1.1 represents one possible systematic sample. There are k possible samples; they all contain n units each. The y_{ij} notation denotes the j^{th} unit of the i^{th} systematic sample, which corresponds to the $i + (j-1)k$ $^{\text{th}}$ unit of the population, $i = 1, 2, \ldots, k$ and $j = 1, 2, \ldots, n$.

The variance of the systematic sample as derived by Cochran [10] can be expressed as follows:

$$\text{var}(\bar{y}_{sys}) = \left(\frac{N-1}{N}\right) S^2 - \frac{k(n-1)}{N} S^2_{wsy} \tag{1.1}$$

where S^2 is the population variance of y

$$S^2 = \frac{1}{N-1} \sum_{i=1}^{N} (y_i - \bar{y})^2 \tag{1.2}$$

and the mean of the population is

$$\bar{y} = \frac{1}{N} \sum_{i=1}^{N} y_i \tag{1.3}$$

and S^2_{wsy} is the variation that exist among the units that lie within the same systematic sample defined by

Table 1.1: Layout of the k Systematic Samples

	Sample number					
	1	2	...	i	...	k
	y_1	y_2	$\cdots$	y_i	$\cdots$	y_k
	y_{k+1}	y_{k+2}	$\cdots$	y_{k+i}	$\cdots$	y_{2k}
	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$
	$y_{(n-1)k+1}$	$y_{(n-1)k+2}$	$\cdots$	$y_{(n-1)k+i}$	$\cdots$	y_{nk}
Means	$\bar{y}_1$	$\bar{y}_2$	...	$\bar{y}_i$	...	$\bar{y}_k$

$$S^2_{wsy} = \frac{1}{k(n-1)} \sum_{i=1}^{k} \sum_{j=1}^{n} (y_{ij} - \bar{y}_{i.})^2,$$

where $y_{ij} = y_{i+(j-1)k}$ and $\bar{y}_{i.} = \frac{1}{n} \sum_{j=1}^{n} y_{ij}$.

Cochran provided a second expression for the variance of the systematic sample [10], as it can also be written as follows:

$$\text{var}(\bar{y}_{sys}) = \left(\frac{S^2}{n}\right) \left(\frac{N-1}{N}\right) [1 + (n-1)\rho_w] \tag{1.4}$$

where ρ_w is the correlation coefficient between pairs of units that are in the same systematic sample. ρ_w is also known as the interclass correlation between the pairs of units that are in the same systematic sample.

The efficiency of systematic sampling is more prevalent when the population is autocorrelated; that is when two observations y_i and y_j will be more nearly alike when their indices i and j are close together in the frame than when they are distant [5].

A third way derived by Cochran [10] for the variance of systematic sample using the

within strata variation is:

$$\text{var}(\bar{y}_{sys}) = \left(\frac{S^2_{wst}}{n}\right)\left(\frac{N-n}{N}\right)[1+(n-1)\rho_{wst}] \tag{1.5}$$

where S^2_{wst} is the variance among units that lie in the same stratum, it is the mean sum of squares within strata (or rows in this case) and can be expressed as follows

$$S^2_{wst} = \frac{1}{n(k-1)}\sum_{j=1}^{n}\sum_{i=1}^{k}(y_{ij}-\bar{y}_{.j})^2,$$

where $\bar{y}_{.j} = \frac{1}{k}\sum_{k=1}^{k} y_{ij}$.

ρ_{wst} is the correlation between the deviations from the stratum means of pairs of items that are in the same systematic sample.

The variance of simple random sample is

$$\text{var}(\bar{y}_{srs}) = \left(\frac{N-n}{N}\right)\left(\frac{S^2}{n}\right) \tag{1.6}$$

The variance of stratified random sample is

$$\text{var}(\bar{y}_{st}) = \left(\frac{N-n}{N}\right)\left(\frac{S^2_{wst}}{n}\right) \tag{1.7}$$

From equations: 1.1 and 1.6, it can be deduced that systematic sampling is more efficient than simple random sampling if and only if $S^2_{wsy} > S^2$ (see Appendix A).

This important result implies that if the population is sorted such that units within the same sample are heterogeneous, then the precision of systematic sampling increases, and if the units are homogenous, then the precision will decrease.

By comparing the formulas 1.4 and 1.6, it can be proved that systematic sampling will

have a larger variance than simple random sampling if $\rho_w > 0$ (see Appendix A), that is to say, the existence of a positive correlation between units in the same sample will inflate the variance of the sample mean if systematic sampling is used.

If the population is in random order, then systematic sampling is equivalent to simple random sampling [5]. One study found that "when the sampling frame is in increasing or decreasing order then systematic sampling is likely to be more precise than simple random sampling, adjacent elements tend to be more similar than elements that are farther apart: such a population is said to have positive autocorrelation." (Sharon Lohr, 2019, p. 160) [20].

1.4 Principal Component Analysis

Principal Component Analysis (PCA) was derived from the Principal axis theorem; it was invented initially by Karl Pearson in 1901 [25] and then in the 1930s, Harold Hotelling independently developed it by way of using covariance and correlation analysis [13]

PCA has mainly two objectives (1) data reduction and (2) interpretation [16].
Often, PCA is a technique used to simplify a large dataset; the idea is to reduce the number of variables but not the number of observations. Suppose that we have an $(n \times p)$ data matrix denoted by $X(n \times p)$
We may write $X = (X_1, \ldots, X_p)$, a row vector of p elements, where each element is itself a column vector of n elements.
PCA constructs new variables $Z_1, \ldots, Z_p$ from X where the first principal component Z_1 is a linear combination of the components of X that has the largest variance

$$\max_{(\alpha_1,\ldots,\alpha_p)} \text{var}(\alpha_1 X_1 + \ldots + \alpha_p X_p) = \text{var}(Z_1)$$

Z_2 is constructed linear combination of the components of X such that it has zero correlation

with Z_1 and has the maximum possible variance, each subsequent component Z_j have the largest possible variance and is mutually zero correlated with all Z_i's where $i < j$.

By definition, the variance-covariance structure requires the p set of components X to be used, but PCA is a technique that commingles the original p components to develop a new set of k components where k is smaller than p. PCA algorithm seeks to reduce the dimension of data without much loss of information; that is, the variance-covariance structure using the new set k components carries as much variability information as the original variance-covariance structure [12]. If all variables $Z_1, \ldots, Z_p$ are retained, i.e., if $k = p$, then no information is lost.

Technically, PCA analysis seeks to choose among all possible vectors that are linear combinations of the original $X = (X_1, \ldots, X_p)$ variables the linear compound that has a maximum possible variance. This process is iterative.

The basic idea is to find the linear combination of axes, that is, the direction in which there is the most variation in the data. That is the **principal component** or **first principal component**. Next, the algorithm finds an orthogonal direction to the first principal component in which there is the most remaining variation in the data.

Eigenvectors will be showing the direction of the spread of data, while eigenvalues will be indicating the magnitude of the spread. Covariance measures how much two variables change together.

PCA transformation takes the original p variables as an input to create p new variables. This transformation projects the data into a new space. This projection is made using a linear transformation. PCA attempts to engineer features that are linear combinations of the original p features, such as most of the information will be carried by the first k principal components.. The variation in the first component is greater than the variation in either of the original variables. The variation within the second principal component is less than the variation in the first principal component.

For any square symmetric matrix, It is a guarantee that the eigenvectors λ_i exist and

that they are real numbers [17].

By design, covariance matrices are square and symmetric. Also, any covariance matrix has the property of positive semi-definite. This property is a necessary condition to have all its eigenvectors non-negative, that is:

$$\lambda_i \geq 0 \text{ for all } i's.$$

1.5 Outline of Chapters to Follow in this Dissertation

The remainder of this dissertation is organized as follows.

In Chapter 2, the novel method of sampling of PCA-Systematic Sampling is developed and the theoretical evidence to support its benefits over the traditional systematic sampling is provided. The second contribution in this dissertation, which is the creation of an index quality for sampling is derived.

In Chapter 3, Monte Carlo simulation is used to provide evidence for the novel PCA-Systematic Sampling performance over state of the art sampling methods using simulated data.

In Chapter 4, real publicly available datasets are used to show how the novel PCA-Systematic can be used as an application on real surveys data using a demographic survey and another economic survey. Fisher Iris Flowers data in the context of classification is also used to show how novel PCA-Systematic outperforms current and best state of the art sampling methods.

Chapter 5 summarizes results and present some promising future research.

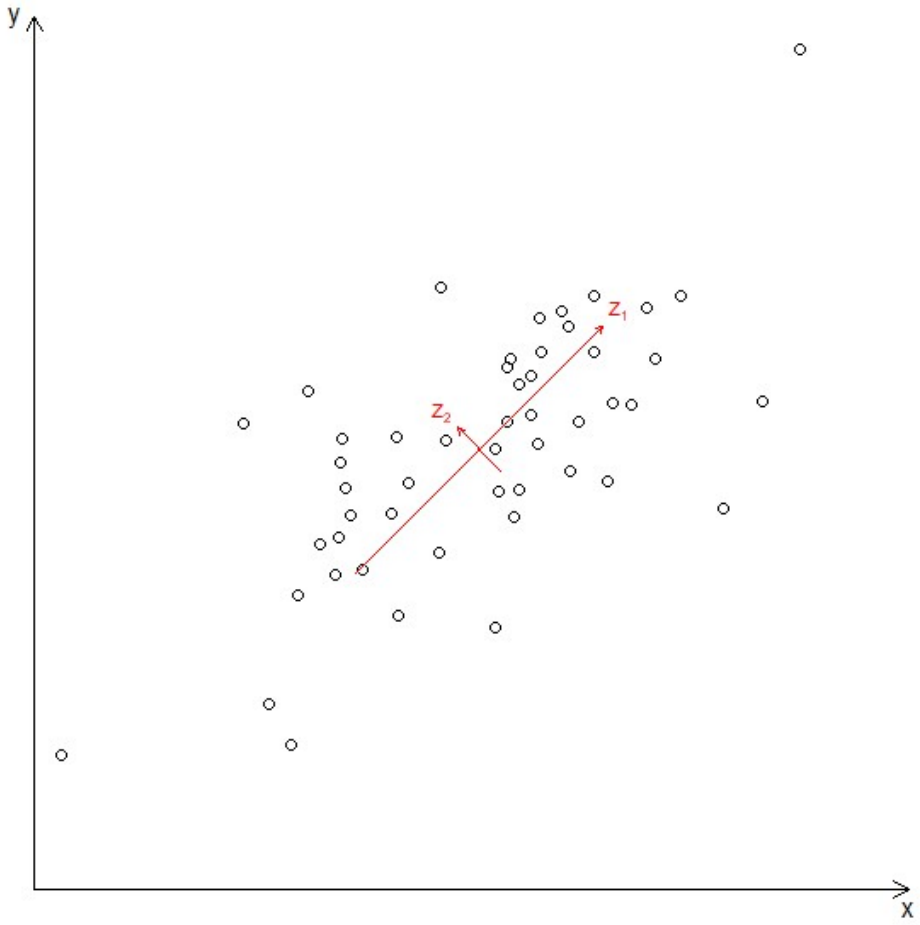

Figure 1.2: Diagram Describing PCA Change of Axis.
Length of principal component vector is proportional to the variance explained, plotting the new axis formed by the first two principal components Z_1 and Z_2 along the original X and Y-axis.

Chapter 2: Theoretical Proof

This chapter contains the following three elements

- 2.1 is the theoretical development to mathematically formulates the evidence to support the hypothesis that with a higher probability conducting principal component analysis (PCA) transformation will lead to an increase in Pearson correlation of the two-dimensional case.
- 2.2 is the extension of the two-dimensional case to the multivariate situation using the Bayesian approach.
- 2.3 is the development of the index of quality for data splitting.

The key contribution in this dissertation is to use linear algebra to engineer new features by combining the set of all available variables, this include the use of the target variable as well as all independent variables to create a set of new variables. In Linear Algebra, Spectral Decomposition Theorem states that every symmetric matrix can be factorized using its Eigenvectors [17]. This decomposition is sometimes referred to Eigen-Decomposition. This decomposition is useful in many practical applications where the goal is to reduce the dimensions of the data matrix.

In this dissertation, it is used for the decorrelation of the feature variables and to project the data matrix into a new space engendered by new axis where the variation is maximized along each axis [12].

To clarify the framework, I introduce the following standard definitions

Suppose that X is real valued continuous random variables with probability density function

$f(x)$ over a support R, the expected value of X is defined by

$$\mathrm{E}(X) = \int_R x f(x) dx.$$

The variance of X is defined by

$$\mathrm{var}(X) = \int_R (x - \mathrm{E}(X))^2 f(x) dx.$$

The standard deviation of X denoted by σ_X is the square root of the variance of X; that is $\sigma_X = \sqrt{\mathrm{var}(X)}$.

For any arbitrarily random variable Y, the covariance between X and Y is defined by

$$\mathrm{cov}(X, Y) = \mathrm{E}\left[(X - \mathrm{E}(X))(Y - \mathrm{E}(Y))\right],$$

and the Pearson's correlation coefficient between X and Y denoted by $\mathrm{cor}(X, Y)$ or $\sigma_{X,Y}$ is defined by

$$\mathrm{cor}(X, Y) = \frac{\mathrm{cov}(X, Y)}{\sigma_X \sigma_Y}.$$

Theorem 1 (Spectral decomposition theorem, or Jordan decomposition)**.** *Any symmetric matrix $A(p \times p)$ can be written as:*

$$A = \Gamma \Lambda \Gamma' = \sum_{i=1}^{p} \lambda_i \gamma_i \gamma_i' \tag{2.1}$$

where Λ is a diagonal matrix with eigenvalues of the matrix A, and Γ is an orthogonal matrix whose columns γ_i's are standardized eigenvectors.

In Theorem 1 the eigenvectors γ_i's are chosen to be orthonormal which means they satisfy the following two properties

1. For every i The L^2 norm of γ_i (which is defined as the square root of the sum of the squares of its coordinates) is equal to 1, that is

$$L_2(\gamma_i) = \|\gamma_i\|_2 = \sum_{j=1}^{p} \gamma_{j,i}^2 = 1.$$

2. The γ_i's are pairwise orthogonal, which means that distinct eigenvectors are orthogonal to one another, that is for each i and j if $i \neq j$ then $\gamma_i \perp \gamma_j$ which will be true if their inner product denoted by $\langle \gamma_i, \gamma_j \rangle$ or $\gamma_i \cdot \gamma_j$ is equal to 0, that is

$$\sum_{j=1}^{p} \gamma_{h,i}\gamma_{h,j} = 0$$

The two properties above are satisfied if and only if $\Gamma' = \Gamma^{-1}$.
Λ is a diagonal matrix; that is $\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_p)$.
The lambdas are configured such that $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p$, with this configuration, the first eigenvector is associated with the largest eigenvalue.

In this dissertation I demonstrate how Theorem 1 is applied for machine learning. In supervised machine learning we have a set of feature variables $(X_1, \ldots, X_p)$ and a target variable Y to be learned. One lets the multivariate vector composed by the response variable and all feature variables combined together and let Γ to be the variance covariance matrix of the multidimensional vector $(Y, X_1, \ldots, X_p)$.

2.1 Two Dimensional Case

In this section I provide the mathematical proof that with a higher probability principal component analysis transformation increases the correlation with the response variable Y.

The two dimensional is when there is one dependent variable and only one single feature

(independent) variable X. From survey sampling, as shown in Chapter 1, sorting the data file known as the frame by the most highly correlated variable with the study variable before conducting systematic sampling is beneficial and increases prediction when the goal is to estimate population parameters, such as the mean of the study variable. In machine learning, the goal is to determine the relationship between the study variable and how it relates to the feature variables in order to perform prediction. In this dissertation data splitting is treated as a sampling problem because when splitting data to create a testing data set for example, it is necessary to have a good data portion that represents the distribution of the target variable and how it relates to its predictors. From a practical point of view, the variable Y is used to denote the target variable (dependent variable) to be predicted and learned by a given machine learning algorithm and the X denotes a given feature variable (independent variable).

In the context of survey sampling the Y usually is the variable of interest or some key statistics to be estimated by the collected data sample, and X is another variable that could either obtained by the survey data that was collected by the sampled unit that responded to the survey, or it could very well be some auxiliary information existing in the sampling frame before the unit was sampled out and sent for data collection.

Theorem 2 (PCA Increases Correlation in 2-D Theorem)**.** *If X and Y are continuous variables and Z_1 , Z_2 are the principal components then:*

$$\max\{|\text{cor}(Y, Z_1)|, |\text{cor}(Y, Z_2)|\} \geq |\text{cor}(Y, X)|. \tag{2.2}$$

Proof. Suppose (X, Y) is a random vector with mean μ and covariance matrix Σ. Then the variance covariance matrix Σ in this case will be a 2×2 matrix and have the following form:

$$\Sigma = \begin{bmatrix} \text{var}(X) & \text{cov}(X, Y) \\ \text{cov}(Y, X) & \text{var}(Y) \end{bmatrix}$$

where $\text{cov}(X, Y)$ denote the covariance between X and Y defined as:

$$\text{cov}(X, Y) = \text{E}[(X - \text{E}(X))(Y - \text{E}(Y))]$$

and $E(\cdot)$ is the expectation of random variables or function of random variables. Σ is a square matrix that is symmetric because $\text{cov}(X, Y) = \text{cov}(Y, X)$, hence according to the spectral decomposition theorem 1 I can decompose Σ as follows:

$$\Sigma = \Gamma\Lambda\Gamma' \tag{2.3}$$

where Λ is a diagonal matrix of eigenvalues of Σ , and Γ is an orthogonal matrix whose columns are standardized eigenvectors of Σ.

The principal component transformation is defined to be the following linear transformation:

$$(X, Y)' \to (Z_1, Z_2)' = \Gamma'(X, Y)' \tag{2.4}$$

For simplicity, I assume for now that X and Y are standardized to have variance 1 so that

$$\Sigma = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}$$

The eigenvalues of Σ are the roots of $|\Sigma - \lambda I| = 0$.

$$\begin{vmatrix} 1-\lambda & \rho \\ \rho & 1-\lambda \end{vmatrix} = 0 \Rightarrow (1-\lambda)^2 - \rho^2 = 0 \Rightarrow (1-\lambda-\rho)(1-\lambda+\rho) = 0$$

This implies that $\lambda = 1 - \rho$ or $\lambda = 1 + \rho$.

It is simple to verify that the eigenvalues of Σ are $1 + \rho$ and $1 - \rho$ with corresponding standardized eigenvectors $\gamma_1 = (\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}})$ and $\gamma_2 = (\frac{1}{\sqrt{2}}, \frac{-1}{\sqrt{2}})$ and that $\Lambda = \begin{bmatrix} 1+\rho & 0 \\ 0 & 1-\rho \end{bmatrix}$

$$\text{because } \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{bmatrix} = (\frac{1}{\sqrt{2}}) \times \begin{bmatrix} 1+\rho \\ 1+\rho \end{bmatrix} = (1+\rho) \begin{bmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{bmatrix}$$

and

$$\begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{2}} \\ \frac{-1}{\sqrt{2}} \end{bmatrix} = (\frac{1}{\sqrt{2}}) \times \begin{bmatrix} 1-\rho \\ \rho-1 \end{bmatrix} = (1-\rho) \begin{bmatrix} \frac{1}{\sqrt{2}} \\ \frac{-1}{\sqrt{2}} \end{bmatrix}$$

γ_1 and γ_2 are the columns vectors of Γ.

From (2.4) I can write:

$$\begin{pmatrix} Z_1 \\ Z_2 \end{pmatrix} = \begin{bmatrix} \gamma_1' \\ \gamma_2' \end{bmatrix} \begin{pmatrix} X \\ Y \end{pmatrix} = \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{-1}{\sqrt{2}} \end{bmatrix} \begin{pmatrix} X \\ Y \end{pmatrix} = \frac{1}{\sqrt{2}} \begin{bmatrix} X+Y \\ X-Y \end{bmatrix}$$

$$Z_1 = \frac{1}{\sqrt{2}}[X + Y] \tag{2.5}$$

$$Z_2 = \frac{1}{\sqrt{2}}[X - Y] \tag{2.6}$$

From the variance covariance matrix Σ, the correlation between Y and X can be computed as:

$$\text{cor}(X,Y) = \frac{\text{cov}(X,Y)}{\sqrt{\text{var}(X)}\sqrt{\text{var}(Y)}} = \frac{\rho}{\sqrt{1}\sqrt{1}} = \rho.$$

The correlation between Y and the eigenvector Z_i can be computed as follows:

$$\text{cor}(Y,Z_i) = \frac{\text{cov}(Y,Z_i)}{\sqrt{\text{var}(Y)}\sqrt{\text{var}(Z_i)}}. \tag{2.7}$$

From equation (2.4) The variance covariance matrix of (Z_1, Z_2) is:

$$\text{var}(Z_1,Z_2) = \text{var}(\Gamma'(X,Y)). \tag{2.8}$$

The second right-hand side of equation (2.8) has the form

$$\text{var}(AV)$$

where A is constant and V is random which is equal to:

$$A\text{var}(V)A'.$$

So this leads to

$$\text{var}(Z_1,Z_2) = \Gamma'\text{var}(X,Y)\Gamma = \Gamma'\Sigma\Gamma = \Lambda$$

But we know that

$$\Lambda = \begin{bmatrix} 1+\rho & 0 \\ 0 & 1-\rho \end{bmatrix}.$$

This implies that we have the following

$$\text{var}(Z_1, Z_2) = \begin{bmatrix} 1+\rho & 0 \\ 0 & 1-\rho \end{bmatrix}.$$

In the meantime, by definition

$$\text{var}(Z_1, Z_2) = \begin{bmatrix} \text{var}(Z_1) & \text{cov}(Z_1, Z_2) \\ \text{cov}(Z_1, Z_2) & \text{var}(Z_2) \end{bmatrix}$$

That means that the variance-covariance matrix of the principal components is diagonal, because by matching the two equal matrices in the previous two equations element by element we can conclude that the variance of the eigenvector Z_1 is the first element of the diagonal matrix Λ which is $1 + \rho$.
The variance of the eigenvector Z_2 is the second diagonal element which is $1 - \rho$.
Furthermore Z_1 and Z_2 are uncorrelated because the off diagonal elements are zero.

So we have $\text{var}(Z_1) = 1 + \rho$. and $\text{var}(Z_2) = 1 - \rho$ and $\text{cov}(Z_1, Z_2) = 0$.

$$\text{cov}(Y, Z_1) = \text{cov}\left(Y, \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix} \begin{bmatrix} X \\ Y \end{bmatrix}\right)$$

of the form

$$\text{cov}(AV, BW)$$

where A and B are constant matrices and V and W are random matrices which is equal to

$$A\text{cov}(V, W)B'$$

In particular,

$$\operatorname{cov}(V, BW) = \operatorname{cov}(V, W)B'$$

so because $Z_1 = \frac{1}{\sqrt{2}}[X + Y]$ from (2.5)

$$\operatorname{cov}(Y, Z_1) = \operatorname{cov}\left(Y, \begin{bmatrix} X \\ Y \end{bmatrix}\right) \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix}',$$

$$\operatorname{cov}(Y, Z_1) = \begin{bmatrix} \operatorname{cov}(Y, X) & \operatorname{cov}(Y, Y) \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{bmatrix},$$

$$\operatorname{cov}(Y, Z_1) = \begin{bmatrix} \rho & 1 \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{bmatrix},$$

$$\operatorname{cov}(Y, Z_1) = \frac{1 + \rho}{\sqrt{2}},$$

$$\operatorname{cov}(Y, Z_2) = \operatorname{cov}\left(Y, \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{-1}{\sqrt{2}} \end{bmatrix} \begin{bmatrix} X \\ Y \end{bmatrix}\right),$$

$$\operatorname{cov}(Y, Z_2) = \operatorname{cov}\left(Y, \begin{bmatrix} X \\ Y \end{bmatrix}\right) \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{-1}{\sqrt{2}} \end{bmatrix}',$$

$$\operatorname{cov}(Y, Z_2) = \begin{bmatrix} \operatorname{cov}(Y, X) & \operatorname{cov}(Y, Y) \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{2}} \\ \frac{-1}{\sqrt{2}} \end{bmatrix},$$

$$\text{cov}(Y, Z_2) = \begin{bmatrix} \rho & 1 \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{2}} \\ \frac{-1}{\sqrt{2}} \end{bmatrix},$$

$$\text{cov}(Y, Z_2) = \frac{1-\rho}{\sqrt{2}}.$$

By substituting the numerator and denominator values obtained above, equation (2.7) becomes:

$$\text{cor}(Y, Z_1) = \frac{\frac{1}{\sqrt{2}}(1+\rho)}{(1+\rho)^{(\frac{1}{2})}} = \frac{\sqrt{1+\rho}}{\sqrt{2}},$$

and

$$\text{cor}(Y, Z_2) = \frac{\frac{1}{\sqrt{2}}(1-\rho)}{(1-\rho)^{(\frac{1}{2})}} = \frac{\sqrt{1-\rho}}{\sqrt{2}}.$$

$$\max\{\text{cor}(Y, Z_1), \text{cor}(Y, Z_2)\} = \begin{cases} \text{cor}(Y, Z_1) & \text{if } \rho > 0; \\ \text{cor}(Y, Z_2) & \text{if } \rho \leq 0. \end{cases}$$

$$\max\{\text{cor}(Y, Z_1), \text{cor}(Y, Z_2)\} = \begin{cases} \frac{\sqrt{1+\rho}}{\sqrt{2}} & \text{if } \rho > 0; \\ \frac{\sqrt{1-\rho}}{\sqrt{2}} & \text{if } \rho \leq 0. \end{cases}$$

$$\max\{\text{cor}(Y, Z_1), \text{cor}(Y, Z_2)\} = \frac{\sqrt{1+|\rho|}}{\sqrt{2}}.$$

The differences between the absolute value of the correlations of Y and the feature variable before and after the PCA transformation is:

$$\delta_i(\rho) = |\text{cor}(Y, Z_i)| - |\text{cor}(X, Y)|, i = 1, 2$$

because $|\text{cor}(Y, Z_i)|+|\text{cor}(X, Y)| \geq 0$ and multiplication by a positive number doesn't change the sign then the sign of $\delta_i(\rho)$ is the same sign of:

$$(|\text{cor}(Y, Z_i)| - |\text{cor}(X, Y)|) \times (|\text{cor}(Y, Z_i)| + |\text{cor}(X, Y)|)$$

which is the sign of:

$$(\text{cor}(Y, Z_i))^2 - (\text{cor}(X, Y))^2,$$

$$\text{cor}(Y, Z_i)^2 - (\text{cor}(X, Y))^2 = \begin{cases} \frac{1+\rho}{2} - \rho^2 & \text{if } i = 1; \\ \frac{1-\rho}{2} - \rho^2 & \text{if } i = 2. \end{cases}$$

The sign of $\delta_1(\rho)$ is the same sign of the polynomial $-2\rho^2 + \rho + 1$.

of the form

$$ax^2 + bx + c = 0.$$

The discriminant of the quadratic equation $-2\rho^2 - \rho + 1 = 0$ is $\Delta = b^2 - 4ac = 9 > 0$.

So the polynomial has two roots $\rho = \dfrac{-b \pm \sqrt{\Delta}}{2a} \Rightarrow \rho_1 = \dfrac{-1}{2}$ and $\rho_2 = 1$,

$$\text{sign}(\delta_1(\rho)) = \begin{cases} - & \text{if } \rho < \frac{-1}{2} \text{ or } \rho > 1; \\ 0 & \text{if } \rho = \frac{-1}{2} \text{ or } \rho = 1 \ ; \\ + & \text{if } \frac{-1}{2} < \rho < 1. \end{cases} \tag{2.9}$$

Because the correlation $\rho \in [-1, +1]$, 2.9 is reduced to:

$$\text{sign}(\delta_1(\rho)) = \begin{cases} - & \text{if } \rho \in]-1, \frac{-1}{2}[; \\ + & \text{if } \rho \in]\frac{-1}{2}, 1[\ ; \\ 0 & \text{if } \rho \in \{\frac{-1}{2}, 1\}. \end{cases} \tag{2.10}$$

The sign of $\delta_2(\rho)$ is the same sign of the polynomial $-2\rho^2 - \rho + 1$.

So the polynomial has two roots $\rho_1 = -1$ and $\rho_2 = \frac{1}{2}$

$$\text{sign}(\delta_2(\rho)) = \begin{cases} - & \text{if } \rho < -1 \text{ or } \rho > \frac{1}{2}; \\ 0 & \text{if } \rho = -1 \text{ or } \rho = \frac{1}{2}\ ; \\ + & \text{if } -1 < \rho < \frac{1}{2}. \end{cases} \tag{2.11}$$

Similarly because the Pearson correlation is bounded between -1 and $+1$ equation 2.11 can be reduced to:

$$\text{sign}(\delta_2(\rho)) = \begin{cases} - & \text{if } \rho \in]\frac{1}{2}, 1[; \\ + & \text{if } \rho \in [-1, \frac{1}{2}[\ ; \\ 0 & \text{if } \rho \in \{-1, \frac{1}{2}\}. \end{cases} \tag{2.12}$$

Combining 2.10 and 2.12 we have:

$$\forall \rho \in [-1, +1]\ \text{sign}(\delta_1(\rho)) = + \text{ or } \text{sign}(\delta_2(\rho)) = +$$

in other words:

$$\forall \rho \in [-1, +1]\ \ \max\{\delta_1(\rho), \delta_2(\rho)\} > 0.$$

Hence

$$\max\{|\text{cor}(Y, Z_1)|, |\text{cor}(Y, Z_2)|\} \geq |\text{cor}(Y, X)|\}. \tag{2.13}$$

Figures 2.1 and 2.2 show that PCA increases correlation in absolute value.

For a positive correlation between X and Y of value ρ, the first principal component Z_1 will increase the correlation from ρ to $\frac{\sqrt{1+\rho}}{\sqrt{2}}$ by $\delta_1(\rho) = \frac{\sqrt{1+\rho}}{\sqrt{2}} - \rho$. This increase reaches a maximum of $\frac{1}{\sqrt{2}} \approx 0.71$ when $\rho = 0$ and a minimum of 0 when $\rho = 1$ (Figure 2.1).

For a negative correlation between X and Y of value ρ, the second principal component Z_2 should be used as it will increase the correlation of Y and $-X$ from $-\rho$ to $\frac{\sqrt{1-\rho}}{\sqrt{2}}$ by $\delta_2(\rho) = \frac{\sqrt{1+\rho}}{\sqrt{2}} + \rho$. This increase reaches a maximum of $\frac{1}{\sqrt{2}} \approx 0.71$ when $\rho = 0$ and a minimum of 0 when $\rho = -1$ (Figure 2.2).

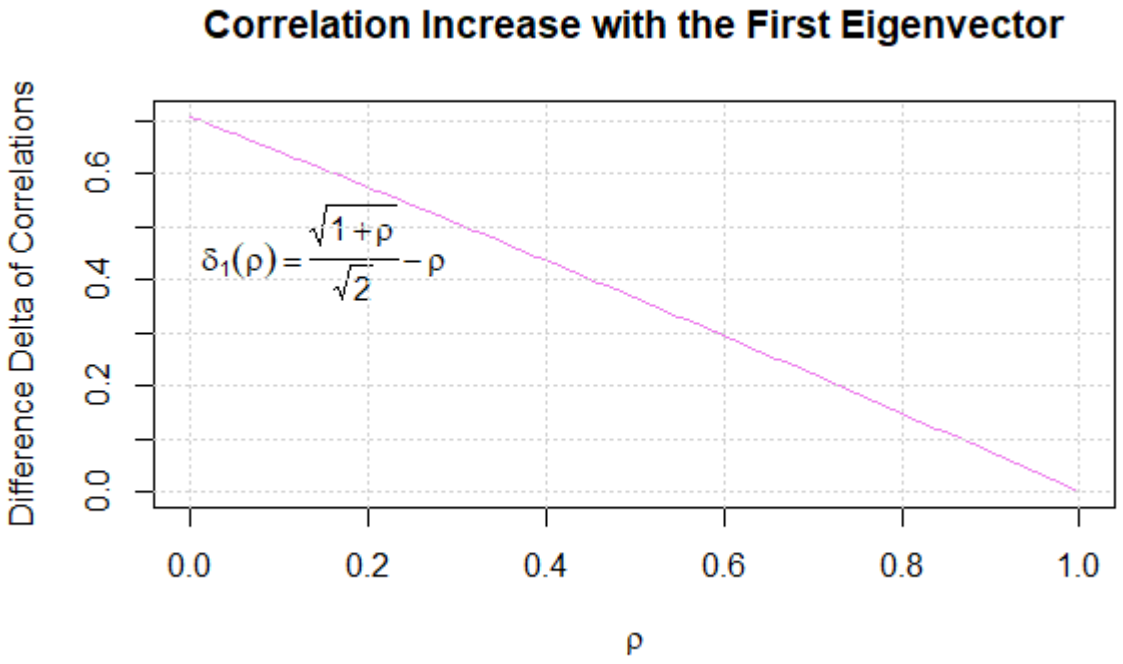

Figure 2.1: Correlation Increase with First Eigenvector.

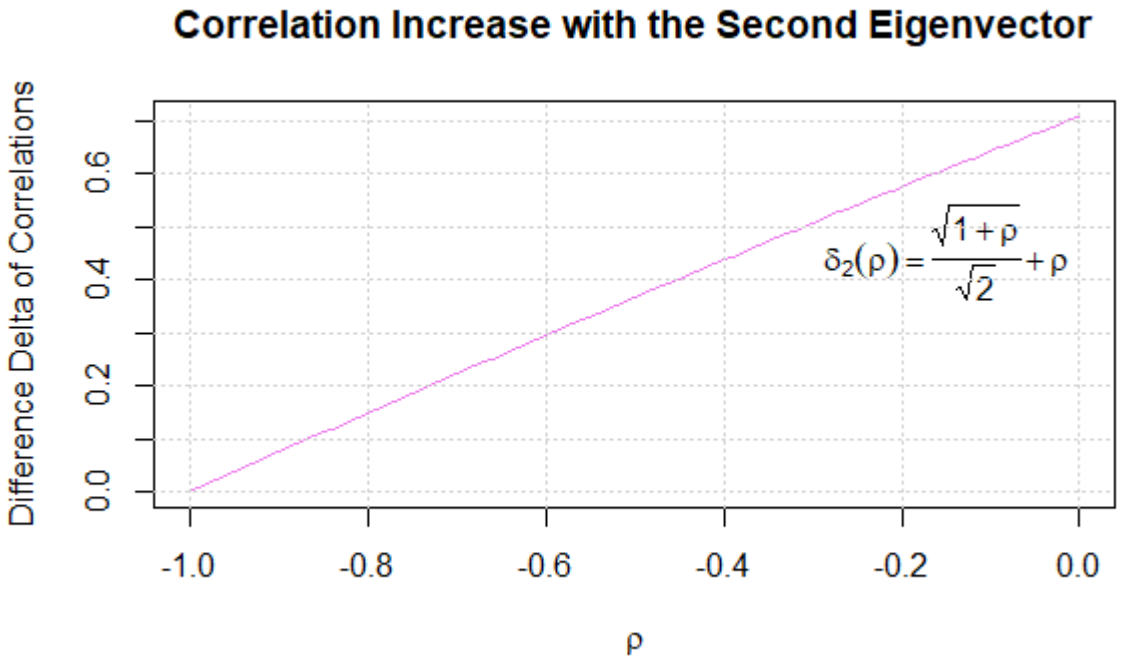

Figure 2.2: Correlation Increase with Second Eigenvector.

For simplicity and without loss of generality one can consider only feature variables with positive correlations with the target variable Y, If $\text{cor}(Y, X) = \rho$ is negative, then if I let $\tilde{X} = -X$, $\text{cor}(Y, \tilde{X}) = -\rho$ will be positive. In this case it is sufficient to use only the first principal component and equation 2.10 will be reduced and gives:

$$\text{sign}(\delta_1(\rho)) = + > 0 \text{ for all possible values of } \rho.$$

The result can be extended as this comes from the fact that correlation is invariant to scaling and shift; that is, if a_1, a_2, b_1 and b_2 are arbitrary real constants, then for any random variables V and W, the following result is true

$$\text{cor}(a_1 V + a_2, b_1 W + b_2) = \text{cor}(V, W) \tag{2.14}$$

Now, to continue the proof for the case when the variances of X and Y are not both equal to 1, that is when $(\sigma_X, \sigma_Y) \neq (1,1)$ I consider $X^* = \frac{1}{\sigma_X}X$ and $Y^* = \frac{1}{\sigma_Y}Y$ where σ_X and σ_Y are the standard deviations of X and Y respectively.
Using (2.14), I have in particular

$$\text{cor}(X, Y) = \text{cor}(X^*, Y^*). \tag{2.15}$$

if I let Z_1 and Z_2 be the principal components of (X^*, Y^*) then because it was proven for the unit variance case in the previous step, the following is inequality is true

$$\max\{|\text{cor}(Y^*, Z_1)|, |\text{cor}(Y^*, Z_2)|\} \geq |\text{cor}(Y^*, X^*)|\}. \tag{2.16}$$

From (2.14), It is also true that

$$\text{cor}(Y, Z_i) = \text{cor}(Y^*, Z_i) \text{ ; for } i = 1, 2 \tag{2.17}$$

combining (2.15), (2.15) and (2.17) I conclude that

$$\max\{|\text{cor}(Y, Z_1)|, |\text{cor}(Y, Z_2)|\} \geq |\text{cor}(Y, X)|. \tag{2.18}$$

■

The result of inequality (2.13) in Theorem 2 states that PCA increases the correlation of variables with Y; that is, Y is more correlated with at least one of the eigenvectors Z_1 or Z_2 than X. This gain in correlation is beneficial because the eigenvector associated with the highest correlation can serve as a sort variable to boost implicit Stratification with systematic sampling. This strategy of sampling will lead to a sample that better represents the distribution of Y compared to simple random sampling and systematic sampling with X being used as a sort variable.

Furthermore, in the inequality (2.13) X and Y are both arbitrarily continuous random

variables with no additional properties for Y over X, so by interchanging the role of X and Y, it is also true that

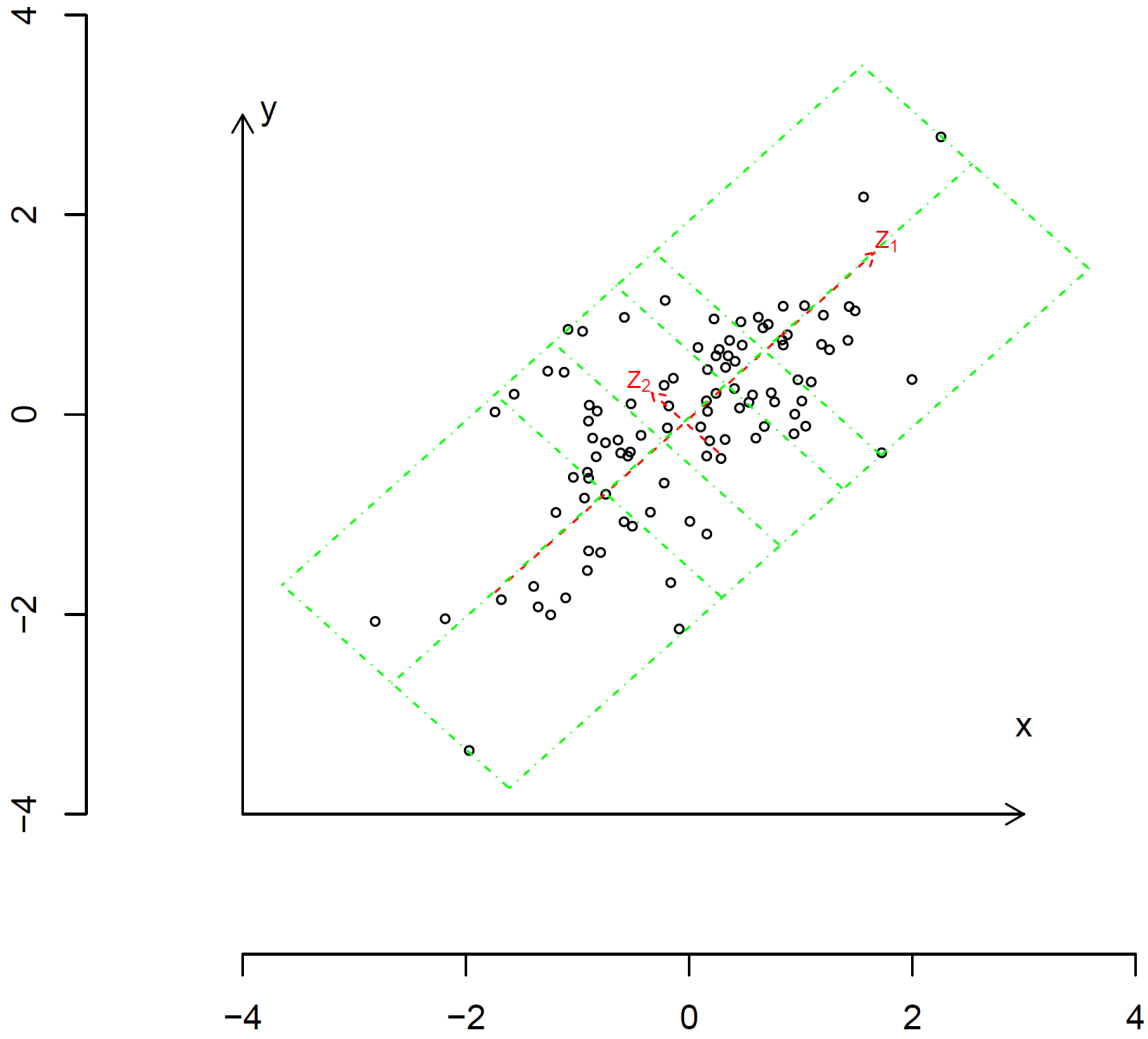

Figure 2.3: Design of Strata based on PCA Quantiles for Efficient Data Splitting.

$$\max\{|\text{cor}(X, Z_1)|, |\text{cor}(X, Z_2)|\} \geq |\text{cor}(X, Y)|. \tag{2.19}$$

The result in (2.19) also indicates a gain in correlation for the variable X is achieved by PCA.

Suppose that $\text{cor}(X, Y) = \rho$ and $\rho > 0$, the value of correlation coefficients between the first

eigenvector Z_1 and the two variables X and Y are

$$\text{cor}(X, Z_1) = \text{cor}(Y, Z_1) = \frac{\sqrt{1+\rho}}{\sqrt{2}} > \rho \text{ (See Proof of Theorem and Appendix A)}$$

Because of this important result, systematic sampling with Z_1 as a sort variable will lead to a sample that better represents both distributions of X and Y. Among all linear combinations between X and Y, the first eigenvector Z_1 has the largest variance; it accounts for as much variation in the data as possible, then the second eigenvector Z_2 captures the maximum possible remaining variation in the data. The novel sampling design in this dissertation uses the axis formed by the principal components as a stratification tool. For example, by segmenting the axis formed by Z_1 into regions using the quantiles, these regions will serve as strata for sampling. The number of strata is a parameter of the novel sampling design; it is proportional to the variance λ_1 of Z_1. This process of forming the strata can be done for the second component Z_2 and for each subsequent component.

The scatter plot in Figure 2.3 is an example of simulated 100 observations from a bivariate standard normal distribution with a correlation between X and Y equals to $\rho = 0.7$.

To draw an efficient random sample from this population, the novel sampling methodology in this dissertation introduces the step of conducting PCA on the full data as a first step; the resulting eigenvectors Z_1, Z_2 are shown in red arrows with lengths proportional to the variances they explained ($\lambda_1 = 1+\rho$ and $\lambda_2 = 1-\rho$); with this, the ratio of the two variances is approximately equal to 5.

Suppose the required sample size is $n = 10$, in that case, a five by two 2-D strata can be formed by segmenting the axis formed by Z_1 into five regions using the min, max and the percentiles of Z_1; the green dashed lines that are perpendicular to the axis formed by Z_1 passes through the min, max, and the following percentiles of Z_1: $p_{0.20}$, $p_{0.40}$, $p_{0.60}$ and $p_{0.80}$. The green dashed lines for the axis of Z_2 are perpendicular to Z_2 and passes through the mix, max and median of Z_2. In the end, to carry out the sampling, one data point is

randomly selected from within each of the stratum.

Chapter three, through simulation work, shows that fitting a linear regression model based on training data created by sampling from the strata formed by the PCA Quantiles as described in Figure 2.3 resulted in a better predictive model than a model if the training data was created using traditional data splitting methods such as simple random sampling with or without sampling. PCA creates eigenvectors that can potentially be used as newly engineered features as predictors in machine learning; however, this dissertation's focus was to only use these eigenvectors as implicit stratification variables tools at the stage of data splitting to create training, testing, or validation datasets.

2.2 Bayesian Method

Proving the increase of correlation resulting from PCA transformation gets complicated for higher dimensions; for example, in 3-D with a multivariate (Y, X_1, X_2), Σ becomes $\begin{bmatrix} 1 & \rho_1 & \rho_2 \\ \rho_1 & 1 & \rho_3 \\ \rho_2 & \rho_3 & 1 \end{bmatrix}$; where ρ_1, ρ_2 and ρ_3 are the correlation between the pairs (Y, X_1), (Y, X_2) and (X_1, X_2) respectively.

Solving the roots of the characteristic polynomial $|\Sigma - \lambda I| = 0$ requires solving the following equation

$$-\lambda^3 + 3\lambda^2 + (\rho_1^2 + \rho_2^2 + \rho_3^2 - 3)\lambda + 1 + 2\rho_1\rho_2\rho_3 - \rho_1^2 - \rho_2^2 - \rho_3^2 = 0$$

under the constraints $\det(\Sigma) \geq 0$ because it is a positive definite matrix. Fortunately, statistical inference using frequentist or Bayesian methodology is a tool than can be used to derive point estimates and interval estimates for the probability of increase in correlation. For a 3-D case, the goal is to determine the validity of the statement that with a higher

probability

$$\max\{|\text{cor}(Y, Z_1)|, |\text{cor}(Y, Z_2)|, |\text{cor}(Y, Z_3)|\} \geq \max\{|\text{cor}(Y, X_1)|, |\text{cor}(Y, X_2)|\}. \quad (2.20)$$

Let ρ_A denote the left-hand side of inequality 2.20; ρ_A is the maximum correlation with Y that can be achieved using PCA. Let ρ_B denote the right-hand side of inequality 2.20 ; ρ_B is the maximum correlation with Y without PCA.

The goal is to compare the distributions of the correlations coefficients before and after PCA. Obtaining an analytical expression for the posterior distribution for the difference of correlations $\rho_A - \rho_B$ can be tedious; however, Monte Carlo method can be used to derive the distribution of functions of other random variables.

The Bayesian framework to handle the statistical inference for $\rho_A - \rho_B$ is described in the steps below

First, assume a given data model:

$$(Y, X_1, X_2) \sim N\left(m = \begin{bmatrix} 1 \\ 0 \\ -1 \end{bmatrix}, \Sigma = \begin{bmatrix} 1 & \rho_1 & \rho_2 \\ \rho_1 & 1 & \rho_3 \\ \rho_2 & \rho_3 & 1 \end{bmatrix}\right).$$

The simulation;

1. For every replicate $b = 1, \ldots, B$

   - Draw $(\rho_2, \rho_2, \rho_3)_b$ from $U(-1, 1)$
   - Draw $\left(Y, X_1, X_2\right)_{ib} \sim N\left(m = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \Sigma = \begin{bmatrix} 1 & \rho_1 & \rho_2 \\ \rho_1 & 1 & \rho_3 \\ \rho_2 & \rho_3 & 1 \end{bmatrix}\right); i = 1, \ldots, n$
   - Conduct PCA on $\left(Y, X_1, X_2\right)_{ib}$; $i = 1, \ldots, n$

- Set a Boolean flag variable

  $T = \max\{|\text{cor}(Y, Z_1)|, |\text{cor}(Y, Z_2)|, |\text{cor}(Y, Z_3)|\} \geq \max\{|\text{cor}(Y, X_1)|, |\text{cor}(Y, X_2)|\}.$

  (T is computed using the n data points, and it is equal to 1 if the inequality 2.20 holds true and zero if false, and n is an arbitrarily data size.)

2. Estimate the Probability of statement 2.20 by $\frac{1}{B}\sum_{b=1}^{B} T_b$

The estimated probability of statement 2.20 can be written as an expected value in the three-dimensional probability measure of the random variable T created above. Using Central Limit Theorem, this expected value can be approximated using Monte Carlo. The probability above is only a point estimate. It is necessary to develop the uncertainty around this estimate by computing the standard error to derive a lower and upper bound. One way to achieve this is to repeat steps 1 and 2 above ten times and calculate the grand mean and the standard deviation of T; this will be used to determine the boundaries of the 90 percent approximate credible interval, which can be obtained using the following formula:

$$\bar{T} \pm Z_{\alpha/2} se(T)$$

where $se(T)$ is the standard error of T computed using its 10 data point estimates.

Another possible approach assumes that T follows a binomial distribution with B trials and a proportion parameter π. Based on the normal approximation of proportions for large sample sizes using the Central Limit Theorem, π can be estimated by $\hat{\pi}$ using

$$\hat{\pi} = \frac{1}{B}\sum_{b=1}^{B} T_b$$

and I can obtain an approximate $100(1-\alpha)$ confidence interval for π by using

$$\hat{\pi} \pm Z_{\alpha/2}\sqrt{\frac{(\hat{\pi}(1-\hat{\pi})}{B}}.$$

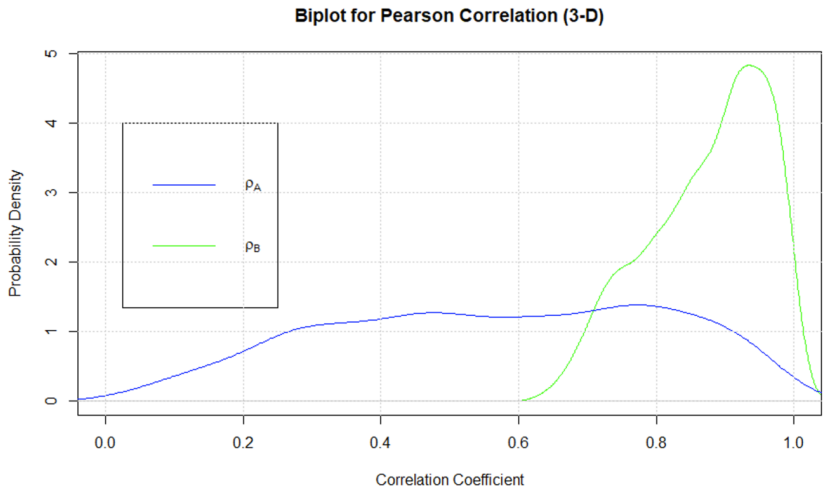

Figure 2.4: Bayesian Posterior and Prior Density plots (3-D).

A visual inspection of the superimposed distributions in Figure 2.4 shows that the values of the quantiles of ρ_A in green are higher than those of ρ_B in blue. However, the question that needs to be addressed is whether the difference is statistically significant. One way to answer this question is through hypothesis testing; set an alpha level of 0.1 and derive a confidence bound for the mean of the distribution of T or a 90% credible interval for the distribution of T. If 0 is below the lower bound of the interval, then one will conclude that the test statistic is significant at the alpha level of .10; that is, there is a higher probability that statement 2.20 is true. For the normal distribution data, similar results were obtained, the 90% approximate credible interval and the confidence interval were $(0.9712, 0.9747)$ and $(0.9700, 0.9754)$ respectively; I conclude that there is a 90% confidence and credibility that the estimated probability that PCA increases the correlation is greater than 0.97 with a small margin of error. It is essential to recognize that this credible interval is not for the difference of means of correlations between the after and before PCA is implemented; this

credible interval is for the proportion of times PCA increased the correlation.

The data model applies to any multivariate distribution and not only the multivariate normal distribution.

2.3 Utility Measure of Data Splitting

While data continues to grow and evolve, decision-makers need answers faster than ever, including the choice of using the right data to be analyzed. However, the quality of the analysis conducted by their statisticians and scientists is only as good as the quality of the data they are studying (GIGO garbage in garbage out). Ensuring good quality of data applies to the mechanic of data splitting when fitting machine learning methods. It is essential to have a novel index measure of data quality for sampling and data splitting. In this dissertation, I create a score (index) based on three different strategies:

1. I compute the five summary statistics for both S and S^c as a data quality control tool to assess data split and how different S is from S^c or even from $S \cup S^c$; a large deviation between the five summary statistics of the two subsets is not favorable.

2. I compute a distance measure between the quantiles of S and $S \cup S^c$; larger values for this distance would be an indication that the sample S does not fully represent $S \cup S^c$.

3. I compute a utility sample using the algorithm explained in the next section.

Woo, et al. in 2009, in the context of disclosure avoidance and utility measure of synthetic data, developed a framework for evaluating the utility of data masked to protect confidentiality [43]; in their work, they used the following algorithm:

1. Append original data to synthetic data.

2. Created an indicator variable $Z = 1$ if data is synthetic and $Z = 0$ otherwise.

3. Fit logistic regression and compute the propensity score mean square error as follows:

$$pMSE = \frac{1}{n_1 + n_2} \sum_{i=1}^{n_1+n_2} (\hat{p}_i - \frac{n_1}{n_1 + n_2})^2$$

Where n_1 and n_2 are respectively the size of original and synthetic datasets. Smaller values are favorable with this quantity and will lead to conclude that the synthetic data is similar to the original data which is the goal for data disclosure avoidance synthetic data creation.

The propensity score is a concept in observational studies developed in 1983 by Rosenbaum and Rubin where assignment to a particular treatment cannot be achieved using randomness [31]. The propensity score is the conditional probability of treatment assignment conditional on the observed covariates. The goal is for the distribution of covariates to be similar between the treated subjects and untreated subjects. In this dissertation, I use a similar idea, but I apply it in the framework of sampling. First, I change the above algorithm slightly as follows:

1. I append S^c to S and create a new indicator variable Z to account for group membership, it is equal to 0 for all the $N - n$ records that belong to S^c and equal to 1 for all the n records that belong to S. If the original data denoted by Ω before taking the sample can be represented by

$$\begin{bmatrix} X_{11} & X_{12} & \dots & X_{1p} \\ X_{21} & X_{22} & \dots & X_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ X_{n1} & X_{n2} & \dots & X_{np} \\ \vdots & \vdots & \ddots & \vdots \\ X_{N1} & X_{N2} & \dots & X_{Np} \end{bmatrix},$$

then this step involves adding a column Z to the data matrix as follows:

$$\begin{bmatrix} X_{11} & X_{12} & \dots & X_{1p} & Z_1 \\ X_{21} & X_{22} & \dots & X_{2p} & Z_2 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ X_{n1} & X_{n2} & \dots & X_{np} & Z_n \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ X_{N1} & X_{N2} & \dots & X_{Np} & Z_N \end{bmatrix}$$

where

$$Z_i = \begin{cases} 1 & \text{if } i \leq n, \\ 0 & \text{if } i > n. \end{cases}$$

2. Fit logistic regression and other binary prediction models such as: CART, RF (Random Forest), SVM (Support Vector Machine), kNN, ANN, Naive Bayes and then I use bagging ensemble method technique to build a more robust binary prediction model to predict group membership.

3. Compute the propensity score mean square and compute the sampling utility score using formula (2.21).

This dissertation introduces a novel index to assess sampling quality in definition 1, which I named the sampling utility score.

Definition 1. *If S is a sample from a population Ω, the Utility of the sample is defined as*

$$U = 1 - \frac{\sum_{i=1}^{N} (\hat{p}_i - f)^2}{n(1-f)^2 + (N-n)f^2} \tag{2.21}$$

where

$f = n/N$ *is the sampling fraction.*

and

$$p_i = probability(Z = 1|\underline{X} = \underline{x}i)$$

$\hat{p_i}$ is the estimated probability of p_i by a binary prediction ensemble model that includes logistic regression.

Theorem 3 (Sampling Utility Theorem)**.** *The Sampling Utility U defined in equation (2.21) has an upper bound of 1 ($U \leq 1$.)*

U is equal to 0 when the prediction of the propensity score is fully accurate over Ω

U reaches a maximum of 1 when the prediction of propensity score is significantly weak.

Proof. If I assume that the prediction model is fully accurate, then I have

$$\hat{p_i} = \begin{cases} 1 & \text{if } i \leq n, \\ 0 & \text{if } i > n. \end{cases}$$

This implies that the term in the numerator of(2.21) $(\hat{p_i} - f)^2$ will be equal to $1 - f^2$ whenever $i \leq n$ and equal to f^2 whenever $i \geq n + 1$ So we have

$$\begin{aligned} \sum_{i=1}^{N}(\hat{p_i} - f)^2 &= \sum_{i=1}^{n}(\hat{p_i} - f)^2 + \sum_{i=n+1}^{N}(\hat{p_i} - f)^2 \\ &= \sum_{i=1}^{n}(1 - f)^2 + \sum_{i=n+1}^{N}(0 - f)^2 \\ &= n(1 - f)^2 + (N - n)f^2 \end{aligned}$$

and therefore

$$U = 1 - \frac{n(1-f)^2 + (N-n)f^2}{n(1-f)^2 + (N-n)f^2}$$

$$= 0.$$

Both numerator and denominator of the fraction term in (2.21) are positive, that is

$$\sum_{i=1}^{N} (\hat{p}_i - f)^2 >= 0$$

and

$$n(1-f)^2 + (N-n)f^2 > 0.$$

This implies that

$$\frac{\sum_{i=1}^{N} (\hat{p}_i - f)^2}{n(1-f)^2 + (N-n)f^2} \geq 0$$

Hence

$$1 - \frac{\sum_{i=1}^{N} (\hat{p}_i - f)^2}{n(1-f)^2 + (N-n)f^2} \leq 1 - 0$$

We conclude that $U \leq 1$

The denominator term does not depend on the p_i's, so the maximum value of U occur when $\sum_{i=1}^{n} (\hat{p}_i - f)^2$ is minimized which will happen if and only if $\hat{p}_i = f$ for all i's so we have

$$\max(U) = 1 - \frac{\sum_{i=1}^{N} (\mathbf{f} - f)^2}{n(1-f)^2 + (N-n)f^2}$$

$$\max(U) = 1.$$

■

Upon fitting the binary prediction ensemble model, the propensity scores p_i are estimated for each of the N records, if the estimated probabilities for the sample S are very close to the true proportion (sampling fraction) $\frac{n}{N}$ then this is an indication that the split resulted in two parts of data that very similar, that is S behaves similar to S^c however larger deviation from this ratio is an indication that the data split resulted in a biased sample. In other words the prediction propensity scores of the binary prediction model "belonging to S and not S^c " is a way to assess the risk associated with the data split. I run Monte Carlo simulation to assess the efficiency of the sampling utility U under all traditional sampling designs and also for skewed distributions and showed how our sampling utility is able to detect nonrepresentative sample from representative samples, the Monte Carlo simulation results are presented in Chapter 3

Chapter 3: Monte Carlo Simulation

Monte Carlo (MC) methods are widely used in the research literature to evaluate properties of statistical methods, and more recently, computational inference using Monte Carlo methods is replacing asymptotic approximations [12]. In this chapter of this dissertation, Monte Carlo Simulation is used to evaluate the properties of my novel data splitting method and evaluate its performance with the different data splitting methods. The initial version of Monte Carlo was documented as early as the 1870s by Erastus Lyman while studying the properties of a statistical procedure [12]. The method has been widely used since, it has been applied as an alternative method to solve physics and mathematics problems that are deterministic in nature and difficult to solve. The method uses randomness to generate repeated random sampling to generate draws from a probability distribution. Monte Carlo uses random simulation to study the performance of different estimators for example, and evaluate different algorithms.

In this section of the dissertation controlling the seeds in Monte Carlo studies is important as it allows to reproduce the same results. This is done by using the same seed from one run of the program to another.

In this dissertation, Monte Carlo simulation tool is conducted to evaluate the performance of the different data splitting methods with comparison to the novel PCA-Systematic data splitting method. The next section introduces the metrics used to assess the method's performance.

This chapter describes the Monte Carlo simulations method to address three main domains; the context of canonical correlation, survey sampling, and machine learning:

1. MC method performed to demonstrate how PCA transformation will increase the correlation of the features with the target variable.

2. MC method performed to evaluate my proposed PCA-Systematic sampling design against traditional simple random sampling and best current systematic sampling design.

3. MC method performed to evaluate and demonstrate the benefits of my proposed PCA-Systematic data splitting method in the context of Machine learning, such as binary classification and ordinary linear regression.

3.1 Monte Carlo Sample Size

Monte Carlo sample size is set to 10,000. The required number of simulated random samples to test a given hypothesis h can be derived using power analysis, if I let p be the true proportion of errors of the hypothesis h, then p can be estimated by assuming the draws to be independent Bernouilli random variables. If I set the goal to have a two-sided 95% confidence interval for p with a margin of error(MoE) of 0.01, then the minimum required sample size can be determined.

$$MoE = z_{\frac{\alpha}{2}} se(\hat{p}) = z_{0.025} \times \sqrt{\frac{p(1-p)}{M}} \approx 2 \times \sqrt{\frac{p(1-p}{M}}.$$

Because $p \in [0, 1]$ the quantity $p(1-p)$ has its maximum when $p = \frac{1}{2}$.
This can be verified by setting the first derivative to 0, solving for p and verifying that it is a maximum because the sign of the second derivative is negative.

Therefore $MoE \leq \frac{1}{\sqrt{M}}$, and since the goal is to have

$$MoE \leq 0.01.$$

So it is sufficient to require $M \geq 10000$.

3.2 Measures of Accuracy

The properties of the methods are analyzed in terms of bias and variance. A method is preferred if it reaches the goal of minimizing both the variance and the bias simultaneously; however, there is a trade off between the two. The performance of data splitting methods can be evaluated using different measures of accuracy.
Depends on the context, different metrics can be used.

If T is an estimator of some parameter θ, then by definition the bias of the estimator is defined by:

$$\text{Bias}(T) = \text{E}(T) - \theta. \tag{3.1}$$

The bias is the difference between the expected mean (arithmetic mean) of the estimator T and the true value θ being estimated by the estimator T. In the context of this dissertation the estimator T is the data splitting method being employed, the parameter θ will depend on the context of the machine learning the data splitting is being used.

The variance of T is defined by:

$$\text{var}(T) = \text{E}(T - \text{E}(T))^2. \tag{3.2}$$

The variance measures the overall average of square deviations from the estimator mean. Estimators with small deviations are preferable therefore it is desirable to have estimators with low variance.

In order to minimize both the mean and variance the MSE (mean squared error) can be used, it is defined as follows:

$$\text{MSE}(T) = \text{E}(T - \theta)^2. \tag{3.3}$$

We have the following result:

$$\text{MSE}(T) = \text{var}(T) + (\text{Bias}(T))^2.$$

The relative root mean squared error (RRMSE) is an interesting measure, it is defined by:

$$\text{RRMSE}(T) = \frac{\sqrt{\text{MSE}(T)}}{\theta} = \frac{\sqrt{\text{MSE}(T)}}{\text{True Value}}. \tag{3.4}$$

The RRMSE is a unit free measure, hence can be used to evaluate different machine methods and algorithms on completely different datasets. The relative bias is another unit free measure of accuracy that can be used, it is denoted by RB and can be computed as follows:

$$\text{RB}(T) = \text{E}\Big(\frac{T}{\theta} - 1\Big). \tag{3.5}$$

Sometimes the direction of the bias is not important and the absolute relative bias (ARB) can be used instead of RB i.e.

$$\text{ARB} = |\text{RB}|.$$

3.3 Monte Carlo to Show Increase in Correlation

In this section, Monte Carlo Simulation investigates the hypothesis that states with a high probability, principal component analysis transformation will increase the correlation between the response variable Y and the feature variables. After that, MC simulation is used to evaluate the performance of the PCA-Systematic data splitting method for fitting Machine learning models such as Linear regression, k-NN, and CART. This evaluation involves comparing its prediction accuracy with a comparison with other existing data splitting methods.

My idea for machine learning in order to enhance the task of prediction problems is

to engineer additional features that increase correlation with the target variable Y. These would be suitable for situations where the target variable and the features are continuous variables. It turns out that linear combinations of the original feature variables using PCA will result in new variables for which their correlation with Y increases.

My hypothesis is the following for the 3-dimensional cases:
If Y is positively correlated with X_1 with correlation $\text{cor}(Y, X_1) = \rho_1 > 0$
and if Y positively correlated with X_2 with correlation $\text{cor}(Y, X_2) = \rho_2 > 0$,
and if however the is a conflict of direction between X_1 and X_2; with X_1 being negatively correlated with X_2 with correlation $\text{cor}(X_1, X_2) = \rho_3 < 0$,
If I let (Z_1, Z_2, Z_3) be the PCA of (Y, X_1, X_2) then with a higher probability the following inequality is true:

$$\max\{|\text{cor}(Y, Z_1)|, |\text{cor}(Y, Z_2)|, |\text{cor}(Y, Z_3)|\} \geq \max\{|\text{cor}(Y, X_1)|, |\text{cor}(Y, X_2)|\}. \tag{3.6}$$

I conducted Monte Carlo experiment to evaluate and validate the Inequality 3.6 for two scenarios. The first scenario is when $\rho_1 = \rho_2 = 0.3$ and a second scenario is when $\rho_1 = \rho_1 = 0.5$.
For both scenarios I generated 10000 independent random samples simulated from a multivariate vector (Y, X_1, X_2). The 3-dimensional random vector (Y, X_1, X_2) is assumed to follow the multivariate normal distribution with mean $m = (1, 0, -1)$ and variance covariance matrix $\Sigma = \begin{bmatrix} 3 & 0.5 & 0.5 \\ 0.5 & 1 & -0.5 \\ 0.5 & -0.5 & 1 \end{bmatrix}$

This can be denoted by

$$(Y, X_1, X_2) \sim N\left(m = \begin{bmatrix} 1 \\ 0 \\ -1 \end{bmatrix}, \Sigma = \begin{bmatrix} 3 & 0.5 & 0.5 \\ 0.5 & 1 & -0.5 \\ 0.5 & -0.5 & 1 \end{bmatrix}\right).$$

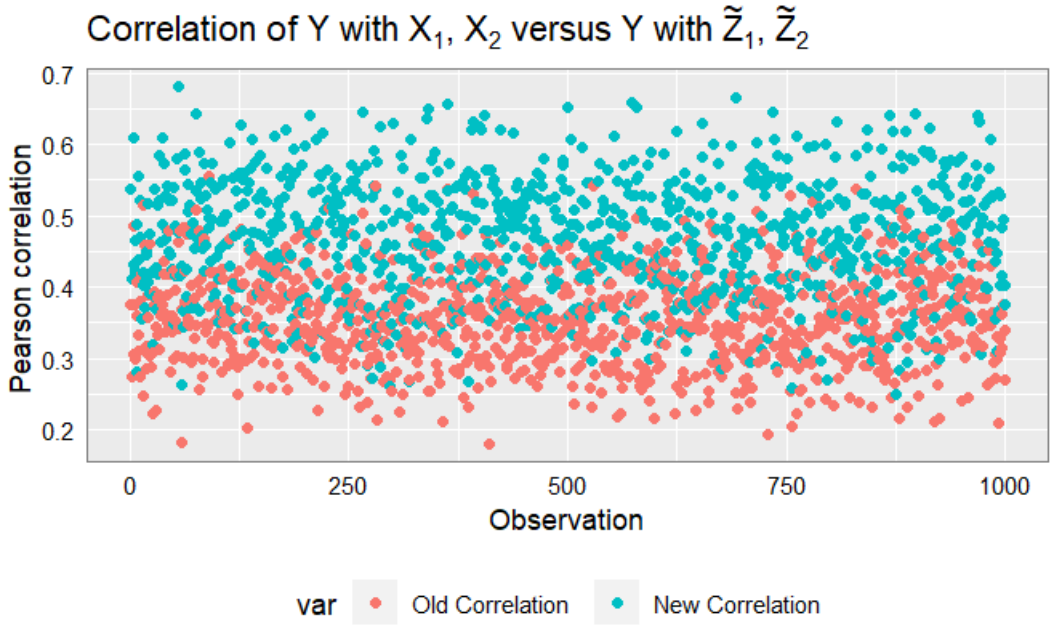

Figure 3.1: Correlation Increase for Mild Conflict between X_1 and X_2 Scenario.

With this setting the mutual correlations between the variables can be computed as follows:

$$\text{cor}(Y, X_1) = \rho(Y, X_1) = \rho_{Y,X_1} = \frac{\sigma^2_{Y,X_1}}{\rho_Y \times \rho_{X_1}} = \frac{0.5}{\sqrt{3} \times \sqrt{1}} \approx 0.289.$$

$$\text{cor}(Y, X_2) = \rho(Y, X_2) = \rho_{Y,X_2} = \frac{\sigma^2_{Y,X_2}}{\rho_Y \times \rho_{X_2}} = \frac{0.5}{\sqrt{3} \times \sqrt{1}} \approx 0.289.$$

$$\text{cor}(X_1, X_2) = \rho(X_1, X_2) = \rho_{X_1,X_2} = \frac{\sigma^2_{X_1,X_2}}{\rho_{X_1} \times \rho_{X_2}} = \frac{-0.5}{\sqrt{1} \times \sqrt{1}} = -0.5.$$

The estimated probability of correlation increase is 0.893.

For this scenario where the conflict between X_1 and X_2 scenario is relatively mild, the improvement in correlation between the target variable Y and the new feature variables engineered by the PCA transformation is shown in Table 3.1. The correlation improves from .3547 to .4675 in average.

Table 3.1: Correlation increase for mild conflict between X_1 and X_2 scenario

	Minimum	1st Quantile	Median	Mean	3rd Quantile	Maximum
$\max\{\rho_1, \rho_2\}$	0.1652	0.3094	0.3529	0.3547	0.33946	0.5677
$\max\{\tilde{\rho_1}, \tilde{\rho_2}\}$	0.2319	0.4124	0.4708	0.4675	0.5242	0.6846

In Figure 3.1, The green data points represent the correlation between the target variable and the new transformed variables after I implemented the PCA, and the red data points represent the correlation between the target variable y and the original feature variables.

For the second scenario,

$$(Y, X_1, X_2) \sim N\left(m = \begin{bmatrix} 1 \\ 0 \\ -1 \end{bmatrix}, \Sigma = \begin{bmatrix} 1 & 0.5 & 0.5 \\ 0.5 & 1 & -0.5 \\ 0.5 & -0.5 & 1 \end{bmatrix}\right).$$

The notation above means that the multivariate random vector (Y, X_1, X_2) is distributed as a multivariate normal with mean $m = (1, 0, -1)$ and variance covariance matrix $\Sigma = \begin{bmatrix} 1 & 0.5 & 0.5 \\ 0.5 & 1 & -0.5 \\ 0.5 & -0.5 & 1 \end{bmatrix}$ With this setting the correlation are as follows:

$$\text{cor}(Y, X_1) = \rho(Y, X_1) = \rho_{Y,X_1} = \frac{\sigma^2_{Y,X_1}}{\rho_Y \times \rho_{X_1}} = \frac{0.5}{\sqrt{1} \times \sqrt{1}} = +0.5.$$

$$\text{cor}(Y, X_2) = \rho(Y, X_2) = \rho_{Y,X_2} = \frac{\sigma^2_{Y,X_2}}{\rho_Y \times \rho_{X_2}} = \frac{0.5}{\sqrt{1} \times \sqrt{1}} = +0.5.$$

$$\text{cor}(X_1, X_2) = \rho(X_1, X_2) = \rho_{X_1,X_2} = \frac{\sigma^2_{X_1,X_2}}{\rho_{X_1} \times \rho_{X_2}} = \frac{-0.5}{\sqrt{1} \times \sqrt{1}} = -0.5.$$

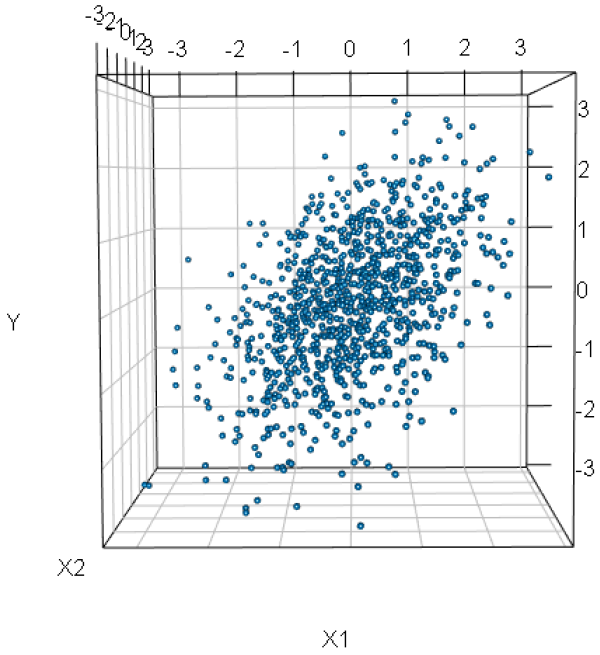

Figure 3.2: Correlation between Y and X_1 of + 0.5.

Table 3.2: Correlation Increase for Strong Conflict between X_1 and X_2 Scenario

	Minimum	1st Quantile	Median	Mean	3rd Quantile	Maximum
$\max\{\rho_1, \rho_2\}$	0.3984	0.5155	0.5482	0.5500	0.5795	0.7177
$\max\{\tilde{\rho_1}, \tilde{\rho_2}\}$	0.5328	0.7631	0.8101	0.8052	0.8540	0.9902

0.5, 0.5 and -0.5 is the maximum conflict that one can have with 3 dimensional dataset. This scenario is visualized in Figures 3-1 through 3-4.

The estimated probability of correlation increase is 0.983, the separation between the green data points and red data points is clear in Figure 3.2.

The improvement in correlation between Y and the feature variables after the PCA transformation is more prominent now. For this scenario where the conflict between X_1 and X_2 scenario is strong.

The improvement in correlation between the target variable Y and the new feature variables engineered by the PCA transformation is shown in Table 3.2. The correlation

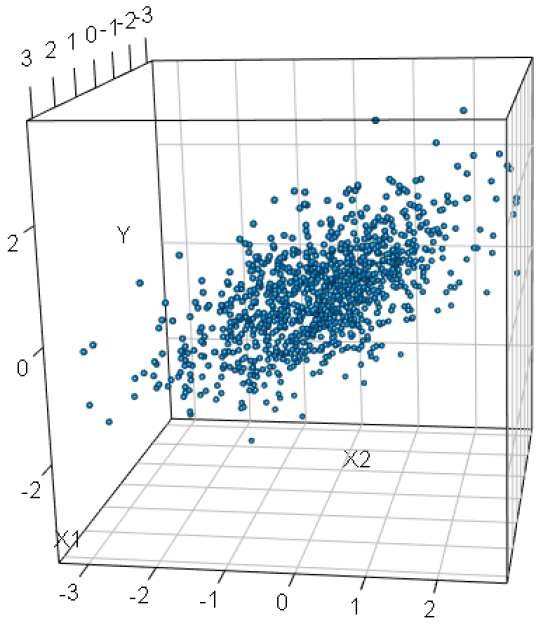

Figure 3.3: Correlation between Y and X_2 of + 0.5.

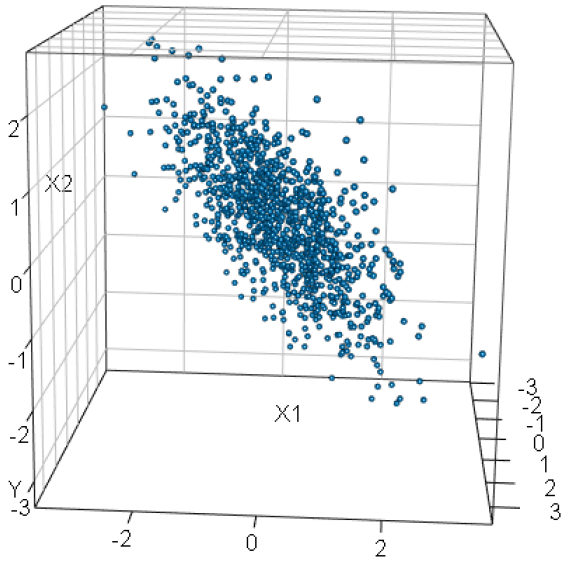

Figure 3.4: Correlation between X_1 and X_2 of - 0.5.

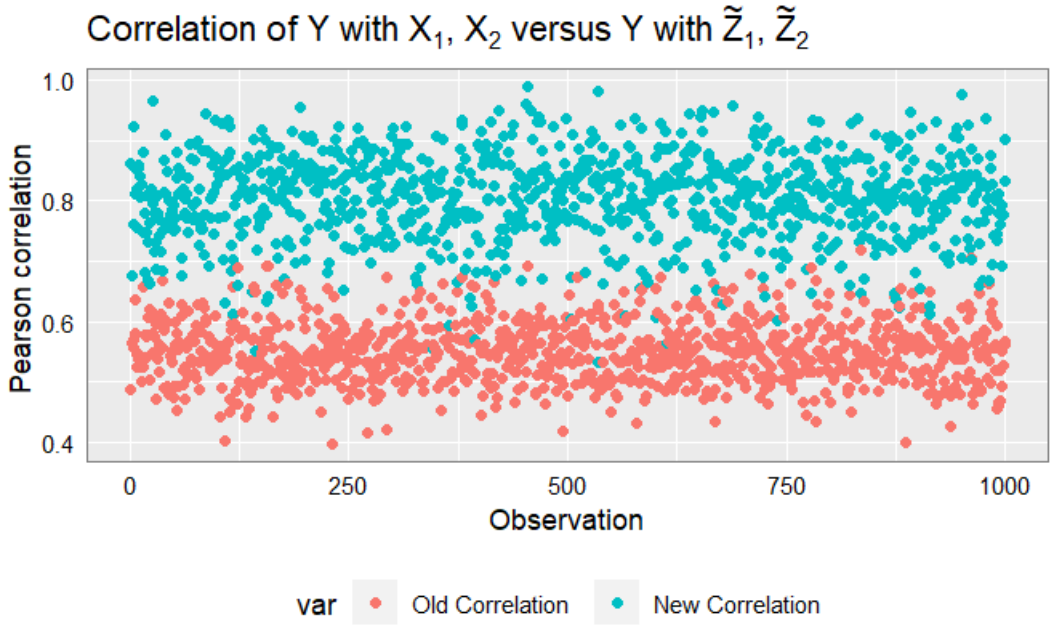

Figure 3.5: Correlation Increase for Strong Conflict between X_1 and X_2 Scenario.

improves from .55 to .81 in average.

3.4 PCA-Systematic Sampling for Surveys Estimates

My Monte Carlo simulation work for PCA in the context of surveys is as follows:

If Y is positively correlated with X_1 and Y is positively correlated with X_2 (but with less correlation than with X_1), however there is a conflict between X_1 and X_2 which means they are negatively correlated, I demonstrate that sorting on the PCA between Y, X_1, X_2 produced better performance while fitting the model $y = f(\underline{x})$ than just sorting by X_1 and X_2 for both survey sampling and *k-NN* classification.

In the context of surveys, the model $y = f(\underline{x})$ is equivalent to $\bar{Y} = \bar{y} + \epsilon$ where $\bar{Y}$ is the average of the variable Y when the entire populations members are used, i.e; $\bar{Y} = \frac{\sum_{i=1}^{N} Y_i}{N}$ and $\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$ is the average of the variable Y using the members selected by the sample of size n.

In my Monte Carlo simulation work example, I simulate the data vector (Y, X_1, X_2) from a multivariate normal distribution as follows:

$$(Y, X_1, X_2) \sim N\left(m = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \Sigma = \begin{bmatrix} 1 & 0.9 & 0.2 \\ 0.9 & 1 & 0.3 \\ 0.2 & 0.3 & 1 \end{bmatrix} \right). \tag{3.7}$$

With this structure $\text{cor}(Y, X_1) = 0.9$ and $\text{cor}(Y, X_2) = 0.2$.

Current sampling methodology suggests to sort the file by X_1 since it has the highest correlation of 0.9 and then do systematic sampling in order to construct a good sample to estimate the mean of Y which should be 0 in this case.

However in the novel PCA-systematic sampling, if I Let $(\tilde{Z}_1, \tilde{Z}_2, \tilde{Z}_3)$ be the PCA of (Y, X_1, X_2) I sort by $\tilde{Z}_1$ instead of X_1.

The MC experiment consisted of

- Simulate a population of $N = 1000$ data points randomly according to (3.7),
- Select a sample of size $n = 100$ according to each sampling method; that is all the possible different sampling strategies, SRSWR, SRSWOR, optimum stratified sampling with Neyman allocation of sample size, traditional systematic where the sort key is X_1, and finally using the novel PCA-Systematic using the sort key $\tilde{Z}_1$.
- I compute the sample mean of the variable Y using the n data points.

Next, this process is repeated $M = 10000$ times, that is, the MC sample size not to be confused with the population size N or the size of the sample n.

The goal is for sample mean to be close to the population's true mean, which is computed using all N data points. The RRMSE is computed for each data splitting method.

Table 3.3 shows how the mean squared error has the lowest value of 0.00362 for PCA-systematic sampling.

Table 3.3: Performance of PCA-Systematic in the Context of Survey Sampling

Data splitting Method	MSE for $\bar{Y}$	MSE for $\bar{X}_1$	MSE for $\bar{X}_2$	Overall MSE
SRSWR	0.01008	0.0099	0.01068	0.01022
SRSWOR	0.00895	0.0087	0.00967	0.00911
Optimum Stratified Sampling Using Neyman Allocation	0.00572	0.00441	0.00915	0.00643
Cluster Sampling	0.06097	0.0534	0.01159	0.04199
Systematic Sampling after sorting by X_1	0.00194	0.00039	0.01095	0.00443
Systematic Sampling stratified by $\tilde{Z}_1, \tilde{Z}_2$ and sort by $\tilde{Z}_1$	**0.00164**	0.00079	**0.00103**	**0.00115**

PCA-Systematic outperforms both SRSWOR and traditional (best existing systematic practice).

3.5 PCA-Systematic Sampling for Statistical Learning

If the task is to fit a model $Y = f(\underline{X})$, it is important to do some preliminary analysis to determine the features that are highly correlated with the response variable Y.
The vector $\underline{X}$ is the set of feature variables used to predict the target variable y.
f is the learner to learn the relationship between the input $\underline{x}$ and y.
Sorting the data file by the feature with the highest correlation with the target variable y prior systematic sampling "to randomly create the test data or training data" has great benefits. Relation among the variables will be preserved and not distorted in the sample training and test file.

Sorting on a Principal Component is my proposed PCA-modified systematic sampling. In a situation where there is a group of features that are important to the target (response) y variable, according to sampling theory if one wants to create systematic sample then the recommended methodology is to sort the file by the variable that is highly correlated with the target variable Y. In my method I do canonical correlation analysis by using principal component analysis on all the variables combined together then sort the data file by the

resulting PCA prior systematic sampling.

3.5.1 Simple Linear Regression

Simple linear regression is the most popular supervised machine learning tool to describe a quantitative variable Y and a single predictor variable X. This is a parametric model that assumes that the relationship between X and Y can be approximated using a linear equation of the form:

$$Y = \beta_0 + \beta_1 X + \varepsilon.$$

The two parameters β_0 and β_1 which are the Y intercept and the slope respectively are known as the coefficient of the regression model and can be estimated using the observed data. The term ε is an error term, it represents the noise not being explained by the model. Fitting the model leads to choosing β_0 and β_1 that minimize the error term and equivalently the distance between Y and $\beta_0 + \beta_1 X$. This distance is also the norm $\|.\|$ of the vector difference

$$Y - (\beta_0 + \beta_1 X)$$

There are different choices of norms that can be used. For example the L_1 norm is based on the least absolute deviations [6]. Let

$$(x_1, y_1), \ldots, (x_n, y_n)$$

represent the n training example pairs, given the observed training examples, L_1 can be regarded as a multivariate function with respect to β_0 and β_1 and the goal is to minimize this function;

$$L_1(\beta_0, \beta_1) = \|Y - (\beta_0 + \beta_1 X)\|_1 = \sum_{i=1}^{n} |y_i - (\beta_0 + \beta_1 x_i)|.$$

Solving the least absolute deviations minimization problem:

$$\min_{(\beta_0,\beta_1)} L_1(\beta_0,\beta_1) = L_1(\hat{\beta}_0,\hat{\beta}_1)$$

does not have an explicit analytic solution and often requires a numerical approximation. Algorithms such as gradient descent can be used to achieve this goal.

Let $y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ be the predicted value of Y when $X = x_i$. Then the difference between the true value y_i and the predicted value $\hat{y}_i$, $e_i = y_i - \hat{y}_i$, represents the i^{th} residual.

The L_2 norm is based on the residual sum of squares (RSS) and defined as

$$L_2(\beta_0,\beta_1) = \|Y - (\beta_0 + \beta_1 X)\|_2 = RSS = \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_i))^2.$$

Unlike L_1, Solving the least sum of squares minimization problem:

$$\min_{(\beta_0,\beta_1)} L_2(\beta_0,\beta_1) = L_2(\hat{\beta}_0,\hat{\beta}_1)$$

has an explicit analytic solution given by

$$\begin{aligned} \hat{\beta}_1 &= \frac{S_{xy}}{S_{xx}}, \\ \hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x}, \end{aligned} \tag{3.8}$$

where $S_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$ is the sum of cross products,

and $S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2$ is the sum of squares.

The regression slope β_1 is equal to the covariance between X and Y divided by the variance of X.

In order to explain how my new data splitting method works, I had to show the steps through an example in the context of simple linear regression. My method consists of an innovative method of sampling strategy, it uses PCA. The method is efficient in the sense that it requires only a small sample size to be effective. In this section, I compare it with current state-of-the-art one-time data splitting methods. I run it through simulation to show its effectiveness, but first, I had to show all steps of how and why it works through the example provided below. A simulated multivariate data (X, Y) consisting of 100 observations according to

$$(X, Y) \sim N\left(m = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \Sigma = \begin{bmatrix} 1 & 0.7 \\ 0.7 & 1 \end{bmatrix}\right) \tag{3.9}$$

Because the data is simulated from a known multivariate distribution that I specified in (3.9), I know how the data were generated, and I know the true parameters, for example, the correlation coefficient between the variables Y and X, ρ is equal to 0.7, it is the off-diagonal entry of the matrix Σ, that is true in this case because both Y and X have a variance of 1. For the simulated data in (3.9), the ordinary least squares fit for the regression of Y onto X is shown in Figure 3.7. The fit is derived by minimizing the residual sum of squares. The slope and the intercept of the black line are computed according to (3.8).

The equation of the fitted regression line to fit the simulated data according to (3.9) can be obtained using the *lm* function of the *stats* package in *R*.

$$y = -0.05 + 0.67x$$

after rounding to two digits.

The red arrows in Figure 3.7 represent the first and the second eigenvector Z_1 and Z_2 They are orthogonal and both having different lengths. The length of each vector is equal to its variance.

as it was shown in Chapter 2

$Z_1 = \frac{X+Y}{\sqrt{2}}$ and $\text{var}(Z_1) = 1 + \rho = 1.7$

$Z_2 = \frac{X-Y}{\sqrt{2}}$ and $\text{var}(Z_2) = 1 - \rho = 0.3$

PCA projection used the variance covariance of the data to develop a new set of orthogonal axis as shown in Figure 3.7.

The new x-axis is obtained using the direction of the vector Z_1 that has the largest variance of 1.7.

The new y-axis is defined using the direction of the vector Z_2

So, I set the goal to split the dataset into training set denoted by S with sample size equals to n, and testing dataset (which would be simply the complement of S denoted by S^c) with sample size equals to $N - n$.

This is a two-way one time data split. If the sample size n is relatively large then the effect of data splitting has less effect on the quality of the training dataset. Because as n increases it will be close to N (meaning if $n \approx N$) regardless of the method being used, it will produce a training dataset that has properties similar to the full dataset Ω. So for this reason in order to conduct a good assessment of data splitting methods with the novel data splitting methods developed in this dissertation reserving only a small size for the training is the approach that was used in this simulation work.

For the choice of the sample size of the training dataset denoted by n; it was set to 10. now, I like to show two novel data splitting for this simulated data based on PCA. First, I set the sample size to be 10. Given the obtained simulated data, I create a grid of values of Z_1 from

Table 3.4: RRMSE for Estimating the Intercept

Data Splitting Method	β_0 (True Intercept)	$\overline{\beta_0}$ (Estimated Intercept)	RRMSE
SRSWOR	-0.0533	-0.0493	4.3079
SRSWR	-0.0533	-0.0507	4.4451
Stratified by cubes formed by PCA Quantiles	-0.0533	-0.0726	**2.4442**

its minimum value to its maximum value. This can be done by partitioning the interval $[\min(Z_1), \max(Z_1)]$ into $M_1 + 1$ equally spaced grid points or using the quantiles of Z_1. Using the quantiles techniques will avoid having blank regions, especially if the distribution is skewed. The number of subintervals M_1 is an option but can I made it proportional to the variance of Z_1 and the requested sample size n. Similarly, I create a grid of values of Z_2 from its minimum value to its maximum value using $M_2 + 1$ equally spaced grid points. M_2 is proportional to the variance of Z_2 and the requested sample size n.

The method requires that $n \geq M_1 M_2$ M_2 is smaller than M_1. for this simulated data $[\min(Z_1), \max(Z_1)] = [-3.700, 3.647]$ and $[\min(Z_2), \max(Z_2)] = [-1.399, 1.474]$ so I made a 5×2 grid and my sampling methods is to use stratified random sample using the $M_1 M_2$ cubes as strata, ideally drawing a random sample of size 1 is the goal.

The coefficient of determination defined by

$$R^2 = \frac{\sum_{i=1}^{n} (\widehat{y}_i - \bar{y})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}$$

is a statistic that can be used to assess the goodness of fit and how well the fitted straight line describes the data [9].

The RRMSE results shown in Table 3.4 are in absolute values show how data splitting methods when stratification is added based on PCA outperform current splitting methods based on SRSWOR and SRSWR.

The second column in the table is the true intercept β_0 that resulted from fitting OLS simple linear regression model on the full data consisting of the simulated 100 data points

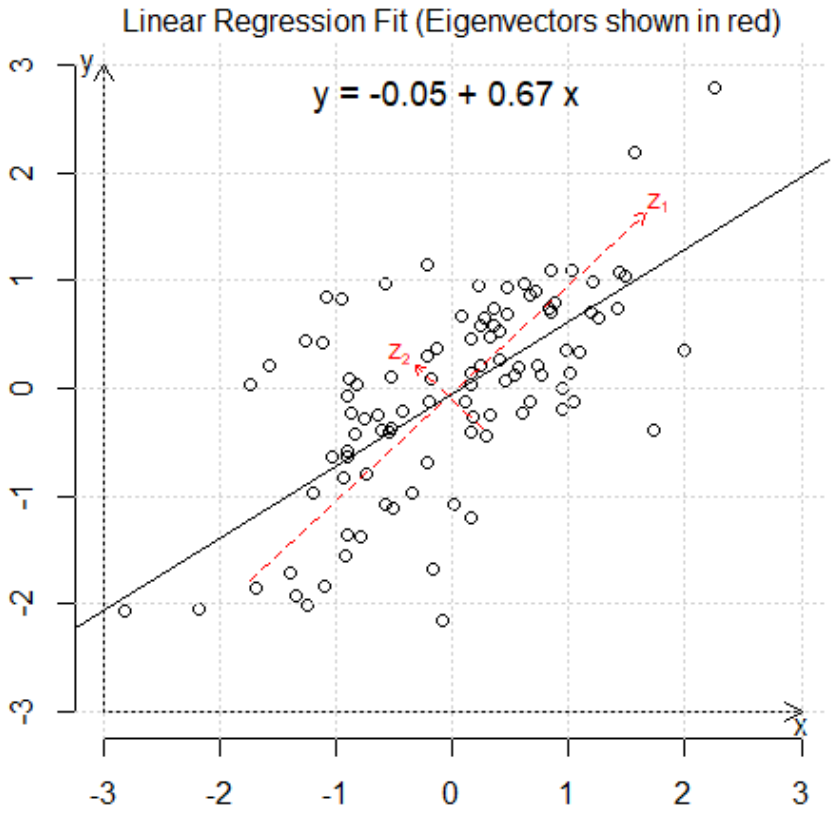

Figure 3.6: Simple Linear Regression Fit.

before a given data split is implemented.

For every replicate r of MC simulation, a slope using only the 10 data points that resulted from a given data split is computed; this is an estimate of the true slope of the linear regression. The process is repeated 10,000 times, and the third column is the average of all these estimates.

Central Limit Theorem justifies that the MSE can be estimated using MC simulation, so

Table 3.5: RRMSE for Estimating the Slope

Data Splitting Method	β_1 (True Slope)	$\overline{\beta_1}$ (Estimated Slope)	RRMSE
SRSWOR	0.6718	0.6524	0.4521
SRSWR	0.6718	0.6583	0.4612
Stratified by cubes formed by PCA Quantiles	0.6718	0.6803	**0.3260**

that is

$$\widehat{\text{MSE}} = \frac{1}{M}\sum_{r=1}^{M}(\hat{\beta_0}^{(r)} - \beta_0)^2$$

and therefore RRMSE defined in equation (3.4) can be computed for column four of the table as

$$\frac{\sqrt{\hat{\text{MSE}}}}{\beta_0}.$$

Table 3.5 show similar results for the slope estimation.

The data splitting methods were further evaluated in terms of their ability to predict the unfitted 90 data points. For each MC replicate, and after a split is conducted to produce a training data set S of size 10, OLS simple linear regression model is fitted, and the predicted $\hat{y}_i's$ are computed using $\hat{\beta}_0 + \hat{\beta}_1 x_i$ for each data point in S^c and compared with the true values y_i. The testing prediction errors for each splitting method are computed using the L^2 and shown in Table 3.6 demonstrate how stratification based on PCA improved the representativeness of the relationship between the variables Y and X using Linear regression; The training data set formed when data splitting using a sample stratified by cubes formed by PCA quantiles resulted is a better strategy than using a traditional simple random sample.

Table 3.6: Distribution of Test Prediction Errors for Simple Linear Regression

Data Splitting Method	Min	1st Quantile	Median	Mean	3rd Quantile	Max
SRSWR	0.5378	0.5763	0.6339	0.6843	0.7361	2.4231
SRSWOR	0.5378	0.5750	0.6304	0.6773	0.7277	3.7036
Stratified by cubes formed by PCA Quantiles	**0.5378**	**0.5549**	**0.5800**	**0.5996**	**0.6234**	**1.1190**

3.5.2 K-Nearest Neighbors

K-Nearest Neighbors (k-NN) is a nonparametric method that addresses both classification and regression prediction problems. When the target variable is categorical, it is known as a classification problem, and if the target variable is continuous, it is a regression problem. k-NN is conceptually a simple algorithm; the intuition behind it is that neighbors tend to be alike. It is a reasonable assumption to predict a member's class membership based on the proximity to its neighbors.

To fit k-NN, a distance d measure to compute distances between the different data points in the feature space is needed, and an arbitrarily positive integer k set by the user is required, this parameter k is the number of nearest neighbors. To classify a given new observation x^*, its distance to all training records is computed. The algorithm finds the set denoted by N^* of the k points in the training dataset that are closest to x^* according to the distance d. In classification, k-NN algorithm chooses the class with the highest conditional probability of $Y = j$ given x^*, so the x^* will be classified to the class j^* such that:

$$j^* = \underset{j}{\operatorname{argmax}}\, P(Y = j | X = x^*).$$

Under no information about these conditional probabilities that are harder to obtain unless some assumptions are made, but still need to be verified and validated, these conditional probabilities can be estimated using the proportion of points in the neighborhood N^*. This will be the same thing as using the majority vote to decide the class membership that x^* needs to belong to when the task is a classification problem.

Therefore we have the following:

$$j^* = \underset{j}{\operatorname{argmax}}\, P(Y = j | X = x^*) = \underset{j}{\operatorname{argmax}}\, \frac{\sum_{i \in N^*} I(y_i = j^*)}{K}.$$

The choice of k is critical for this algorithm and affects the results depending on the data at hand. For example, choosing the parameter k to be an odd integer will eliminate the possibility of ties in the case of binary classification, which is when the target variable Y is dichotomous and take only two possible values. If k is too small, then the algorithm is sensitive to outliers or noise points as the decision boundaries become complex, increasing the risk of overfitting [15].

If k is too large, then the algorithm might create a neighborhood that may include points from other classes.

Current methods of choosing the parameter k use data splitting as a tool to search for an optimal k, measuring the prediction error for different values of k using the testing data portion of the data split and picking the k associated with the lowest value.

This dissertation provides a method on how to optimally split the data without any bias, it shows how the data split can affect the overall fit of the model enormously and how the measure of the test error for the model can be affected as well.

Feature variables might need to be scaled to prevent that some variables might dominate one another.

k-NN is considered to be a type of instance-based-learning or lazy learning because classification is deferred until all computation is completed.

I have done an additional MC simulation work in the context of k nearest neighbors (k-NN) classification problem to evaluate my novel data splitting method and compare its performance against standard methods of data splitting. The PCA-systematic seeks to conserve the variance-covariance of data structure after it gets split, which will maintain coverage of the relationship between Y and the covariate vector $\underline{X}$.

Table 3.5 shows the result from a 10000 simulation study.

The 10,000 independent samples are generated from a multivariate normal distribution composed of the target variable Y along with the covariates X_1, X_2. The joint distribution is assumed to be multivariate normal distribution with mean $m = (1, 0, -1)$ and variance

covariance matrix $\Sigma = \begin{bmatrix} 3 & 0.5 & 0.5 \\ 0.5 & 1 & -0.5 \\ 0.5 & -0.5 & 1 \end{bmatrix}$

$$(Y, X_1, X_2) \sim N\left(m = \begin{bmatrix} 1 \\ 0 \\ -1 \end{bmatrix}, \Sigma = \begin{bmatrix} 3 & 0.5 & 0.5 \\ 0.5 & 1 & -0.5 \\ 0.5 & -0.5 & 1 \end{bmatrix}\right).$$

From a population size of $N = 1,000$, a training dataset of size 333 is formed, and the remaining 667 records are used for testing.

I create a new target variable L with two labels that would take the value "+" or "-" depending on the value of Y.

$$L = \begin{cases} \text{“+”} & \text{if } |Y - 1| \geq h; \\ \text{“-”} & \text{if } |Y - 1| < h. \end{cases} \tag{3.10}$$

where h is an arbitrarily real number, in this experiment h is set to 2.5.

Then I used k-nearest neighbor(k-NN) to predict the "+".

The green density curve $SRSWOR$ in Figure 3.8 is the distribution of test error resulting from fitting k-NN model when data splitting is done using the standard method, which consists of using the traditional SRSWOR data splitting design.

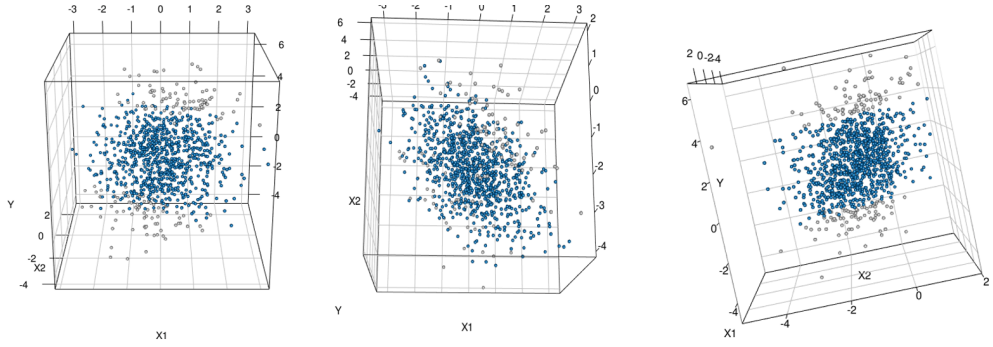

Figure 3.7: The data points colored in silver represent the prediction region, this corresponds with label ="+" according to equation 3.9.

Table 3.7: Confusion Matrix

	(+)	(-)
(+)	20	84
(-)	561	2

Systematic curve in light blue is the distribution error resulting from fitting k-NN when data splitting is done using systematic sampling.

PCASystematic represents the error resulting from fitting k-NN model using the PCA-systematic proposed data splitting method.

The errors are the mean prediction errors resulting from predicting the model and comparing it to the truth. Using the confusing matrix, the off diagonals counts are false positive and false positives; adding the two numbers and dividing by the overall count will give the *mpe* (mean prediction error).

Table 3.8: Comparing Test Errors of Data Splitting Methods for k-NN

Data splitting Method	Min	1^{st} Quantile	Median	Mean	3^{rd} Quantile	Max
SRSWOR	0.01349	0.03298	0.03898	0.03943	0.04498	0.07196
SRSWR	0.00987	0.03395	0.04076	0.04143	0.04809	0.07576
Systematic	0.00600	0.02099	0.02549	0.02532	0.02999	0.05105
PCA-Systematic	**0.00450**	**0.01349**	**0.01799**	**0.01798**	**0.02249**	**0.03748**

		Predicted labels: Positive	Predicted labels: Negative	Total
True labels	Positive	a	b	$a+b$
	Negative	c	d	$c+d$
	Total	$a+c$	$b+d$	N

Table 3.7 displays one confusion matrix that resulted from fitting a k-NN model on the simulated data. This is for one single iteration of 10,000 replicates. From this table the mean prediction error can be computed as follows

$$mpe = \frac{2+20}{2+84+20+561} \approx 0.033.$$

.

Table 3.8 illustrates how the PCA-systematic has better prediction accuracy compared to both SRSWR, SRSWOR and simple systematic, the distribution of the errors mean resulting from PCA-systematic has the lowest average of 0.01807.

Figure 3.8 shows the distribution errors superimposed, it is clear that the PCA-systematic accuracy stochastically dominate both SRSWOR and simple systematic.

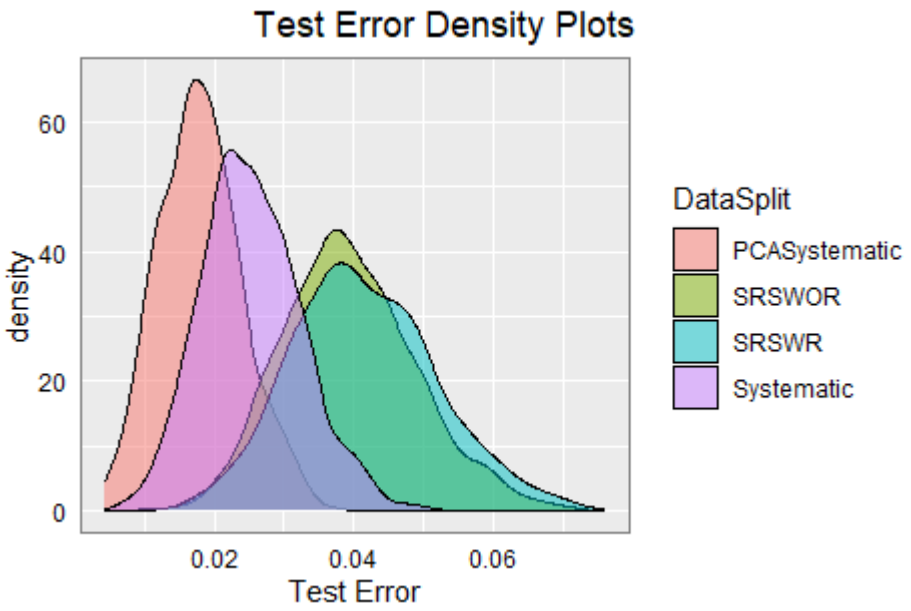

Figure 3.8: Performance of PCA-Systematic compared with SRSWR, SRSWOR and Traditional Systematic for k-NN.

3.5.3 Classification and Regression Tree

In this dissertation, PCA-Systematic was compared with traditional splitting methods such as simple random sampling without replacement to create training datasets with Classification and Regression Tree (CART) Methodology. CART methods for binary prediction model to predict the probability of correctly labeling the labels of L was considered. The same simulated data used in the previous section was used here.

CART partitions the predictor space, or the set of possible values of the covariates (X_1, X_2) into J distinct non overlapping cells or regions $\Lambda_1, \Lambda_2, \ldots, \Lambda_J$.
Minimizing the within cell variability is the primary criterion to construct the cells. In other words, the units inside the cells are as homogenous as possible with respect to the label variable L. The CART algorithm partitions the feature space into smaller subsets recursively. Step 0 of the algorithm starts with the root node, which is the entire dataset, and then subsequent child nodes are formed iteratively. At each step, node impurity can be measured using cross-entropy, Gini index, or residual sum of squares. I used entropy

defined by:

$$\text{Entropy}(N) = -\text{P}(L = \text{“}+\text{”}) \log_2 \text{P}(L = \text{“}+\text{”}) - \text{P}(L = \text{“}-\text{”}) \log_2 \text{P}(L = \text{“}-\text{”}). \quad (3.11)$$

where $p(L = \text{“}+''\text{)}$ and $p(L = \text{“}-''\text{)}$ are the proportion of positive labels and negative labels at node N respectively. The algorithm splits a given node N into subsequent child nodes N_{Right} and N_{Left} by seeking to maximize the information gain defined by

$$\text{Gain} = \text{Entropy}(N) - \frac{|N_{right}|}{|N|}\text{Entropy}(N_{right}) - \frac{|N_{left}|}{|N|}\text{Entropy}(N_{left}). \quad (3.12)$$

As shown in equation (3.12), the information gain is equivalent to the decrease in entropy after a split. The decision to split at node N is made based on the values of a best single predictor variable X_i among all predictors at one time. Thus the child nodes can be written as:

$$N_{left} = \{X_i \in \chi_{left}\}$$

and

$$N_{right} = \{X_i \in \chi_{right}\} = \{X_i \notin \chi_{left}\} = \overline{N_{left}}.$$

This will result in a recursive top-down binary splitting that is greedy because the algorithm always makes the choice that seems best at that moment to maximize (3.12). Once it makes its decision, the algorithm never goes back and reverses it [15].

In this dissertation, I used SAS/STAT software's PROCHPSPLIT procedure [33] to develop the output decision tree. This procedure implements the CART algorithm and has multiple functions, including tree pruning. The pruning approach developed by Breiman et al. in 1984 [7] is a technique that favors smaller trees rather than larger trees and includes a cost-complexity parameter (Cp) as a criterion of pruning. SAS PROCHPSPLIT selects a subtree instead of the full tree, starting from the bottom of the final three and then

Table 3.9: Comparing Test Error of PCA-Systematic Data Splitting with Existing Methods for Fitting CART Model

Data splitting Method	Min	1st Quantile	Median	Mean	3rd Quantile	Max
SRSWOR	0.00300	0.01502	0.02102	0.02318	0.03003	0.04805
SRSWR	0.00576	0.01769	0.02552	0.02593	0.03156	0.05780
Systematic	0.00387	0.00725	0.01354	0.01563	0.02369	0.03675
PCA-Systematic	**0.00000**	**0.00601**	**0.00901**	**0.00946**	**0.01201**	**0.02703**

going back to undo some of the splits as they are considered unnecessary according to the cost-complexity parameter. This approach avoids overfitting and strikes a balance between fitting training data and predicting unobserved data. This procedure allows the users to specify a parameter that controls the minimum number of observations in a terminal node.

3.6 Cost Analysis

This section addresses the computation cost and impact of using the novel PCA-Systematic data splitting method. The method has two main steps:

- Computing the covariance matrix of the data.
- Computing PCA on the covariance matrix from the previous step.

Computing the PCA involves deriving the eigenvalue decomposition as in Theorem 1. The PCA is computed for the $p \times p$ covariance matrix and not the actual $n \times p$ data. The computational complexity to derive the covariance matrix Σ is $\mathcal{O}(p^2 n)$. The eigenvalue decomposition of PCA computational complexity is $\mathcal{O}(p^3)$. So the overall complexity of PCA-Systematic splitting method is $\mathcal{O}(p^2 n + p^3)$.

The PCA algorithm in this dissertation performs spectral decomposition on the variance-covariance matrix Σ, which is a $p+1$ by $p+1$ matrix. The principal components are the results of matrix multiplication of the form $\Gamma' \times (Y, X_1, \ldots, X_p)'$ where Γ is an orthogonal matrix whose columns are the standardized eigenvectors of the Σ (See equation (2.4)).

The process of matrix multiplication can be optimized using parallel processing because

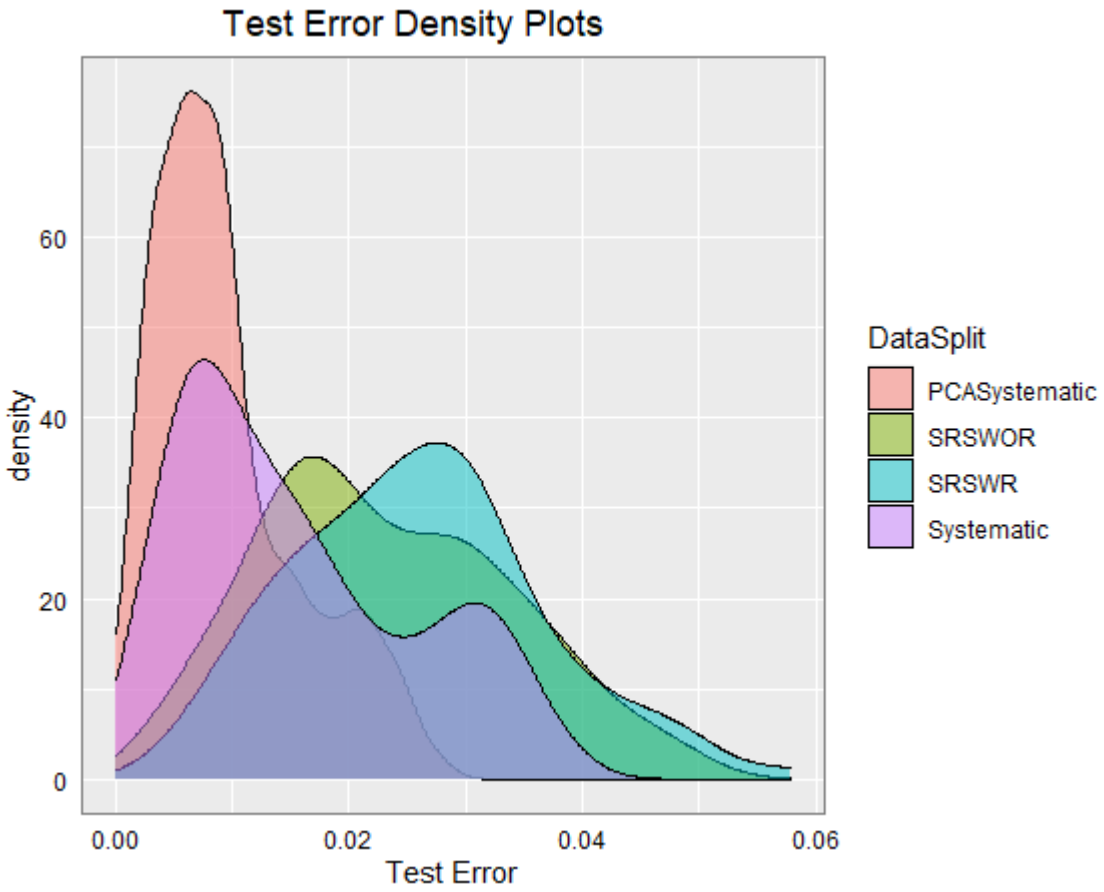

Figure 3.9: Performance of PCA-Systematic Compared with SRSWR, SRSWOR and Traditional Systematic for CART.

each principal component Z_i is the product between the i^{th} row of Γ' and the data matrix $(Y, X_1, \ldots, X_p)$ which can be calculated independently from one another if multiple copies of the data matrix are made available for each subprocess.

The execution time of running PCA on a simulated large-scale dataset composed of one billion records and four variables using SAS took less than 5 minutes, as shown in Table 3.9. However, the file size exceeded 39 gigabytes.

Table 3.10: CPU and Real Time of Running PCA on a one Billion Records and Four Variables Data File

Number of Records	**Real Time**	**CPU Time**	**File Size**
10^5	0.04 seconds	0.03 seconds	4,032 KB
10^8	17.23 seconds	19.87 seconds	3.92 GB
10^9	5:19.96	4:12.68	39.22 GB

Chapter 4: Real Data and Applications

In this dissertation, three real data sets have been used to demonstrate the effectiveness of my data-splitting approach. The next section provide details of the data sets.

4.1 Iris Flower Dataset

Edgar Anderson an American botanist collected the Iris flower data to analyze the morphology of three different species of Iris flowers [2]. This data became very popular when Ronald Fisher, a famous British statistician, and biologist published his research work when he introduced LDA (linear discriminant analysis) in 1936 paper in a an article titled "The use of multiple measurements in taxonomic problems" [11]. At the time of writing this dissertation, Ronald Fisher's paper was cited 17,021 times in Google Scholar . The Fisher Iris data is well known within the statistical community and machine learning experts and became a mainstream test case of many classification problems. For each type of the three species of Iris (Iris Setosa, Iris Virginica, and Iris Versicolor) 50 flower samples were randomly selected. The structure of each flower was measured using four features: the length and the width of the sepals and petals, in centimeters. The Iris data is a multivariate dataset that consists of 150 observations and 5 variables, the fifth variable is the class membership of the Iris flower.

The Iris data can be easily downloaded from the web, for example the University of California Machine Learning Repository has these data. Iris data is available at Kaggle, an online community acquired by Google LLC that maintains an online database with over 1 million users and more than 19,000 public datasets. Iris data are also available with almost every statistical software such as SAS, R and STATA. Crystal Vision software has it available for testing as well. In SAS software it can be accessed in the SAShelp library among

over 200 data sets for users to use for testing codes, in R it is available in the datasets package.

The method of Linear Discriminant Analysis was introduced by Fisher as a classification problem. He used the IRIS data as an application to predict the flowers membership, the task is to be able to predict the correct membership of the Iris flowers to the possible species Iris Setosa, Iris Virginica, or Iris Versicolor.

As part of any machine learning prediction modeling task, it is necessary to do data splitting, which consists in this case of partition the 150 data points into two groups, the first group is used for fitting the LDA model, I refer to this as the training dataset, the second group, which is in this case the complement group, I refer to it as the testing dataset.

The traditional method of sample size allocation during data splitting to form the training and testing datasets is to designate a large percentage of data for the training dataset, usually around 70 to 80 percent but sometimes it could be as large as 90 percent of the full data. The remaining complement is dedicated to form the testing dataset.

This dissertation introduces a new sampling designs to randomly select the training and testing data and compared it with the traditional data splitting using the state of the art methods of sampling schemes.

The goal of this dissertation is to show how data splitting affects the prediction accuracy of a given fitted model. The dissertation also introduces new data splitting methods for the IRIS data. These methods can be generalized to other datasets as well.

To effectively demonstrate the effects of data splitting I will allocate only 15 data points for the training data and the remaining 135 data points for testing, I used LDA as the classification method.

Data splitting is used as a way to validate machine learning predictions model, but current methodology is to allocate a large portion of data for fitting and the remainder data portion for testing the model. Accuracy of the model is measured by comparing the predicted values of the target variable against the true value of the response using the test dataset. In this dissertation, my approach in order to ensure that the data splitting is

optimal is to focus on randomly selecting the best test dataset possible. Because the test dataset is small compare to the training dataset the risk is higher for the test to misrepresent the full data set. Using the best sampling designs to form the test dataset is critical. Once the test dataset is randomly selected using the best sampling design; the training dataset is simply the complement of the test dataset. Choosing the best sampling designs is a function of the full data set to sample from and especially the probability distribution of the target variable and how it relates to the feature variables. There is no single sampling design that works all the time for all types of data. Depending on the nature of data and the target variable, different sampling designs should be selected accordingly.

Among numerous useful EDA (Exploratory data analysis) tools that can be used to visualize Iris data, the overlay density plot based on the three Species for each of the four features shown in figure 4.1 reveals, for example, how Petal Length and Petal Width are both good predictors when it comes to discerning Virginica from Setosa.
CrystalVision is a Windows application designated as a multivariate visualization and exploration tool. One of the key features of this software is its ability to produce parallel coordinates to analyze high dimensional data, Iris data has four continuous variables and a label variable.

Parallel coordinates is data visualization tool that can be used with multi-dimensional data [41]. Parallel coordinates is ideal to compare multiple variables and determine the relationship in a single plot.
Tables 4.1 and 4.2 show the correlation analysis for the Iris predictor variables before and after PCA is conducted.

The principal components can be used as engineered features to boost the classification task; however, in this dissertation, they are not used as predictors but rather tools for data splitting. In this dissertation, one of the novel data splitting tools involves conducting PCA to create principal components and then used the components as sort key variables before systematic sampling. Another novel splitting method involves using the components as a stratification tool which will be explained in methods 4, 5, and 6 with Iris data.

Table 4.1: Correlation Analysis for Iris Flowers Data

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width
Sepal.Length	1	-0.12	0.87	0.82
Sepal.Width	-0.12	1	-0.43	-0.37
Petal.Length	0.87	-0.43	1	0.96
Petal.Width	0.82	-0.37	0.96	1

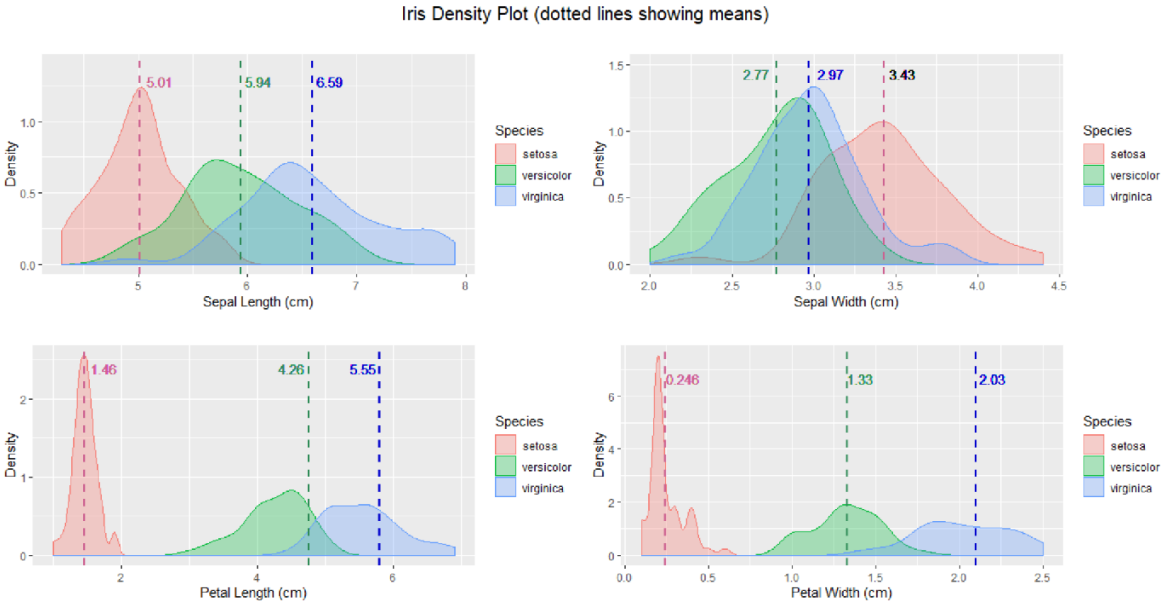

Figure 4.1: Iris Flowers Data Overlay Density Plots.

Table 4.2: Correlation Analysis of Iris Flowers Data with PCA

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	pc_1	pc_2
Sepal.Length	1	-0.12	0.87	0.82	0.90	0.39
Sepal.Width	-0.12	1	-0.43	-0.37	-0.40	0.83
Petal.Length	0.87	-0.43	1	0.96	1.00	-0.05
Petal.Width	0.82	-0.37	0.96	1	0.97	-0.05
pc_1	0.90	-0.40	1.00	0.97	1	0
pc_2	0.39	0.83	-0.05	-0.05	0	1

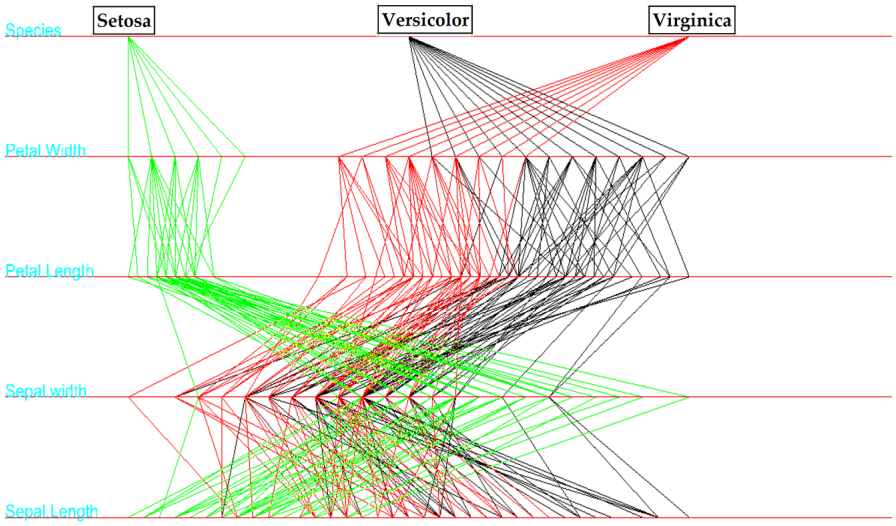

Figure 4.2: Parallel Coordinates to Visualize Iris Data.

The data splitting methods are explained below

1. SRSWOR

15 data points out of 150 available data points are randomly selected without replacement. With the SRSWOR design, the total number of all possible data splits is very large. Because it is done without replacement, it is equal to the number of all subsets of size 15 that can

be withdrawn from a dataset that has a size 150; it is 150 choose 15, and the formula is

$$\binom{150}{15} = \frac{150!}{15!(150-15)!} \approx 1.623922 \times 10^{20}.$$

2. SRSWR

15 data points out of 150 available data points are randomly selected with replacement. With this data splitting design, the total number of all possible data splits is very large and equals 150^{15}; this is larger than that of SRSWOR of $\binom{150}{15}$ because duplicates are acceptable.

3. Stratified by Species

15 data points out of 150 are selected using a multistage stratified random sample with an equal allocation of the sample size. Where the strata are the Species; this means from within each species, 5 data points are randomly selected using SRSWOR. With this design, for every stratum, each data point has the sample probability of inclusion of 1/10.

4. Stratified by regions created by the quadrants formed by the PCA.

The first step is to compute the variance-covariance Σ of the numerical part of the data matrix, so for the Iris data, all four variables, the length and the width of the sepals and petals are used. Then similar to how PCA was shown in Chapter 2, I perform PCA on Σ and derive the eigenvectors. As a result, with Iris data, much of the variation is captured by the first two principal components.

Then I consider the cartesian product created by the first two principal components for this data. Using the first component PC_1 as the X-axis and the second component PC_2 as the Y-axis four regions can be formed; they can be labeled North East (NE), North West (NW), South East (SE) and South West (SW) defined as follows:

$$Region = \begin{cases} NE & \text{if } PC_1 \geq x_0 \text{ and } PC_2 \geq y_0; \\ NW & \text{if } PC_1 \leq x_0 \text{ and } PC_2 \geq y_0; \\ SE & \text{if } PC_1 \geq x_0 \text{ and } PC_2 \leq y_0; \\ SW & \text{if } PC_1 \leq x_0 \text{ and } PC_2 \leq y_0. \end{cases} \tag{4.1}$$

where $x_0 = \frac{1}{n}\sum_{i=1}^{n} PC_{1_i}$ and $y_0 = \frac{1}{n}\sum_{i=1}^{n} PC_{2_i}$ are the sample means of PC_1 and PC_2 respectively

This Cartesian space can be expanded by including more axes, but for the Iris data the first two components are sufficient and explains much of the variation. Once the regions created by the quadrants formed by the PCA and created, the allocation of the sampled 15 data points is conducted using allocation proportion to the size of the regions. With these Iris data, the value of PC_1 ranges approximately from -3.23 to 3.80, and PC_2 ranges from -1.27 and 1.37. With this, the resulted population were partitioned to 32 data points to the NW region, 42 data points to the NE region, 49 data points to the SE region, and 27 data points to the SW region.

The sample size in each region is taken in proportion to the size of the region. With this strategy, each region h gets a sample size n_h according to the formula.

$$n_h = n\frac{N_h}{N} \tag{4.2}$$

where N_h is the size of the region (stratum) h;

$N = \sum_{h=1}^{H} N_h$; and $H = 4$ here.

Table 4.3 shows the details of the sample size allocation. The 15 data points are proportionally allocated to the size of the stratum. The stratum is defined based on the PCA quadrant defined in (4.1).

With this sampling scheme, depending on the region h a given unit u_k it belongs to; its

Table 4.3: Proportional Allocation of Sample Size

PCA Quadrant Region	Stratum Size (N_h)	Sample Size n_h
North East	42	4
North West	32	3
South East	49	5
South West	27	3

inclusion probability to be selected by the sample S denoted by p_k can be computed as follows

$$p_k = \frac{n_h}{N_h}.$$

However, because of the proportional allocation given by (4.2), this implies that $p_k = \frac{n}{N}$. With this design, the population elements have an equal probability of selection. This property is known as EPSEM (equal probability of selection method)

5. Stratified by the grid formed by PCA.

With this sampling scheme the variance explained by each principal component is taking into consideration. First I create a 2-D grid in the cartesian space formed by the two PCA components. For the first PCA component PC_1 partition the interval $[\min(PC_1), \max(PC_1)]$ into $M+1$ equally spaced grid points, M is proportional to the relative variance explained by PC_1. For the second PCA component PC_2 partition the interval $[\min(PC_2), \max(PC_2)]$ into $N+1$ equally spaced grid points, N is proportional to the variance explained by the PC_2. There are $M \times N$ squares created which then are used as strata for the data splitting. Another way to create the $M \times N$ 2-D grid is to use the quantiles, so instead of creating equally spaced grid points across the axis formed by PC_1 I can use $p_1, \ldots, p_{M+1}$ the $M+1$ quantiles of PC_1 this will be suitable when the data exhibits some volatility such as financial and economic data. for the Iris data this was not needed. with skewed data using equidistant points to subdivide an interval will lead most likely to sparse or empty regions that has no data points.

6. Stratified by Species PCA-Systematic.

An additional novel data splitting method that I developed in this dissertation is as follows First, I create a synthetic sort key defined as follows:

$$srtkey = \sigma_1 PC_1 + \sigma_2 PC_2, \tag{4.3}$$

where σ_i is the standard deviation of the i^{th} principal component. Second, Sort the data file by *srtkey* and then implement systematic sampling afterward. The two steps above are implemented within each Species stratum.

The first three sampling designs described above are known as equal probability of selection method (EPSEM) sampling because every unit that is actually included in the sample had the sample probability of being selected in the sample.

Another measure that can be used to evaluate the performance of a particular data splitting method is based on its variability, which is the dispersion of the validation errors that can result by repeated implementation of that method.

The coefficient of Variation (CV) of the distribution of the errors also known as relative standard deviation (RSD) measures the relative variability; it is defined as the ratio of the standard deviation of the distribution to the mean of the distribution:

$$CV = \frac{\sigma}{\mu} \times 100 \tag{4.4}$$

Coefficient of variation measures the spread of the distribution relative to its mean, it is a unit free measure, which means it does not depend on the unit of measurements and often expressed as percentages.

The variability of the test errors obtained using a particular data splitting method using the variance and the coefficient of variation is given in table 4.11

As it can be seen, the variability of the validation errors obtained using our method PCA-Systematic Stratified by Species is low with a coefficient of variation of 1.86 percent

Table 4.4: Comparing Accuracy of Novel Data Splitting Methods against Standard Methods in the Context of LDA on Iris Data

Data Splitting Method	Min	1^{st} Quantile	Median	Mean	3^{rd} Quantile	Max
SRSWOR	0.533	0.933	0.956	0.941	0.963	0.993
SRSWR	0.511	0.926	0.949	0.938	0.963	1.000
Stratified by Quadrants formed by PCA	0.630	0.933	0.956	0.943	0.963	0.993
Stratified by grid formed by PCA	0.689	0.933	0.956	0.947	0.970	1.000
Stratified by Species	0.667	0.933	0.956	0.947	0.963	0.993
Stratified by Species and **PCA-Systematic**	**0.933**	**0.950**	**0.956**	**0.960**	**0.974**	**0.985**

which is desirable. In contrast, the variability of the validation errors obtained using current state of the art method of data splitting which is stratified by species for these data resulted in a coefficient of variation of 3.09 percent.

This dissertation used the Iris Flowers data as a prototype to describe the algorithms of the different novel data splitting methods and how they can compare with traditional data splitting methods.

Table 4.4 presents the results of comparing these novel splitting methods against current methods of data splitting into testing and training. The quality of a given data splitting method can be assessed by quantifying how much distortion of information resulted from the split. Data splitting will sample from the full data Ω and creates a subset data set S. Regardless of the specific machine learning model to be fitted, a distortion of information can occur when the resulting sample S do not preserve the property of (interest); the target and its relationship with the other variables.

This can be measured by computing the variance-covariance matrix for both Ω and S, taking the difference of the two matrices and computing its Frobenius norm $\|\Sigma_\Omega - \Sigma_S\|_F$

This resulting difference matrix is a square symmetric holding has $p \times p$ element wise differences.

The element of the i^{th} row and and j^{th} column of Σ_Ω is the covariance between X_i and X_j

Table 4.5: Frobenius Distance between Covariance Matrix of the Full Data and the Training Data

Data Splitting Method	Min	1st Quantile	Median	Mean	3rd Quantile	Max
SRSWOR	0.008	0.261	0.573	1.018	1.291	14.137
SRSWR	0.004	0.292	0.651	1.122	1.410	15.098
Stratified by Quadrants formed by PCA	0.007	0.169	0.371	0.662	0.850	9.409
Stratified by grid formed by PCA	0.012	0.164	0.292	0.371	0.496	2.781
Stratified by Species	0.005	0.166	0.341	0.538	0.692	6.006
Stratified by Species and **PCA-Systematic**	**0.083**	**0.121**	**0.210**	**0.244**	**0.286**	**0.714**

using all data points. The element of the i^{th} row and and j^{th} column of Σ_S is the covariance between X_i and X_j when only the data points that belong to S are used. Table 4.5 presents shows this result.

The second columns in Tables 4.7 through 4.10 labeled "True Value" holds the value of the mean for the variables *Sepal Length*, *Sepal Width*, *Petal Length* and *Petal Width* respectively; these averages are computed using the full data set of Iris of 150 data points, these values are considered to be the truth. However the third columns are their corresponding estimated values; they are estimated because they are derived using the training examples that resulted from the various data splitting methods listed in the first columns. The $RRMSE$ column in each table is computed as follows

Table 4.6: RRMSE for estimating Sepal Length

Data Splitting Method	True Value	Mean of Yhat	RRMSE
SRSWOR	3.057	3.057	0.166
SRS	3.057	3.058	0.166
Stratified by Quadrants formed by PCA	3.057	3.043	0.164
Stratified by grid formed by PCA	3.057	2.988	0.339
Stratified by Species	3.057	3.057	0.166
Stratified by Species and **PCA-Systematic**	3.057	3.057	**0.044**

Table 4.7: RRMSE for estimating Sepal Width

Data Splitting Method	True Value	Mean of Yhat	RRMSE
SRSWOR	5.843	5.843	5.659
SRS	5.843	5.845	5.657
Stratified by Quadrants formed by PCA	5.843	5.819	5.625
Stratified by grid formed by PCA	5.843	5.929	5.592
Stratified by Species	5.843	5.844	5.659
Stratified by Species and **PCA-Systematic**	5.843	5.843	**4.760**

Table 4.8: RRMSE for estimating Petal Length

Data Splitting Method	True Value	Mean of Yhat	RRMSE
SRSWOR	3.758	3.758	0.086
SRS	3.758	3.762	0.087
Stratified by Quadrants formed by PCA	3.758	3.738	0.085
Stratified by grid formed by PCA	3.758	4.032	0.292
Stratified by Species	3.758	3.759	0.086
Stratified by Species and **PCA-Systematic**	3.758	3.758	**0.009**

Table 4.9: RRMSE for estimating Petal Width

Data Splitting Method	True Value	Mean of Yhat	RRMSE
SRSWOR	1.199	1.199	5.131
SRS	1.199	1.201	5.132
Stratified by Quadrants formed by PCA	1.199	1.190	5.097
Stratified by grid formed by PCA	1.199	1.320	5.080
Stratified by Species	1.199	1.199	5.131
Stratified by Species and **PCA-Systematic**	1.199	1.199	**4.277**

Table 4.10: RRMSE for each Data Splitting Method with regards to the True Mean of each Feature

Data Splitting Method	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width
SRSWOR	5.659	0.166	0.086	5.131
SRSWR	5.657	0.167	0.087	5.132
Stratified by Quadrants formed by PCA	5.625	0.164	0.085	5.097
Stratified by grid formed by PCA	5.592	0.339	0.292	5.080
Stratified by Species	5.659	0.166	0.086	5.131
Stratified by Species and **PCA-Systematic**	**4.760**	**0.044**	**0.009**	**4.277**

Table 4.11: Variability of the Validation Errors for the Data Splitting Methods for LDA on Iris Data

Method	Variance	CV
SRSWOR	0.0017	4.45 percent
SRSWR	0.0023	5.08 percent
Stratified by Quadrants formed by PCA	0.0014	4.00 percent
Stratified by the grid formed by PCA	0.0010	3.39 percent
Stratified by Species	0.0009	3.09 percent
Stratified by Species and PCA-Systematic	0.0003	1.86 percent

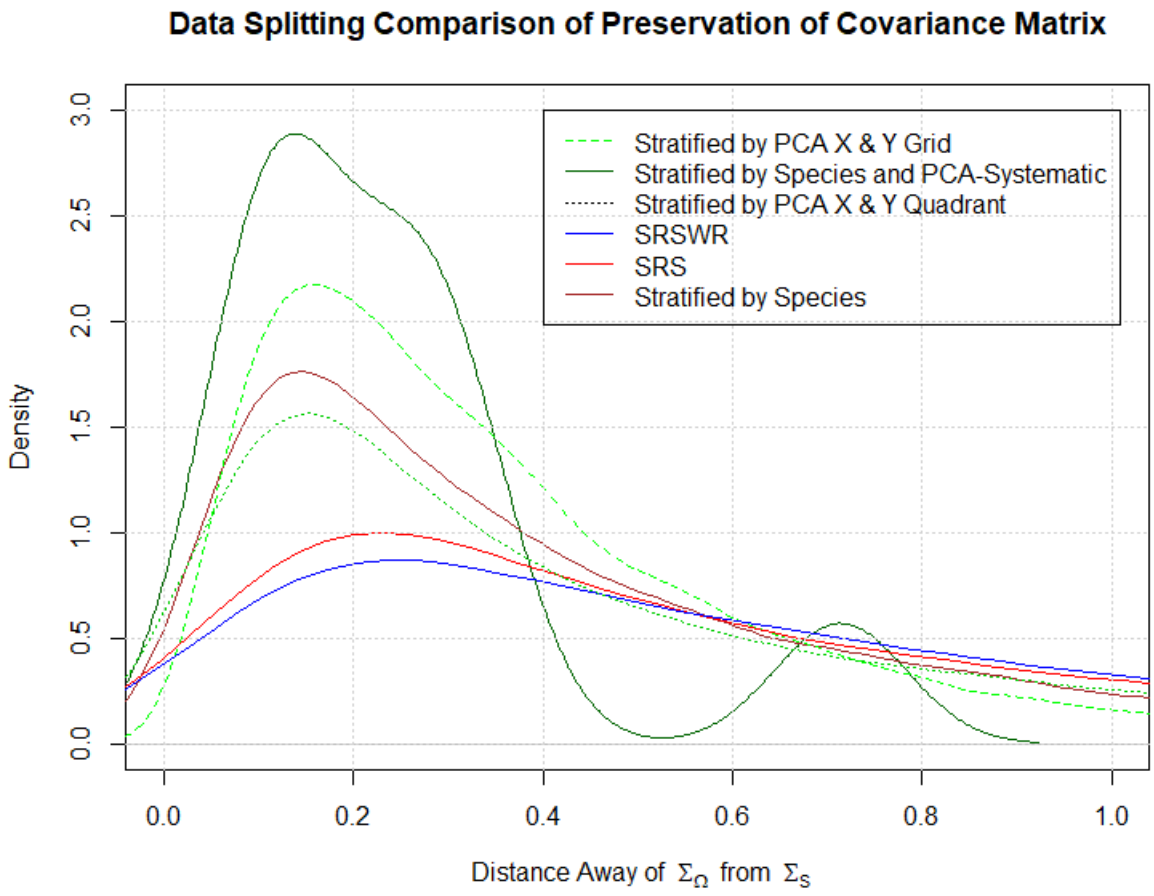

Figure 4.3: Preservation of Variance-Covariance Comparison for Iris Flowers Data.

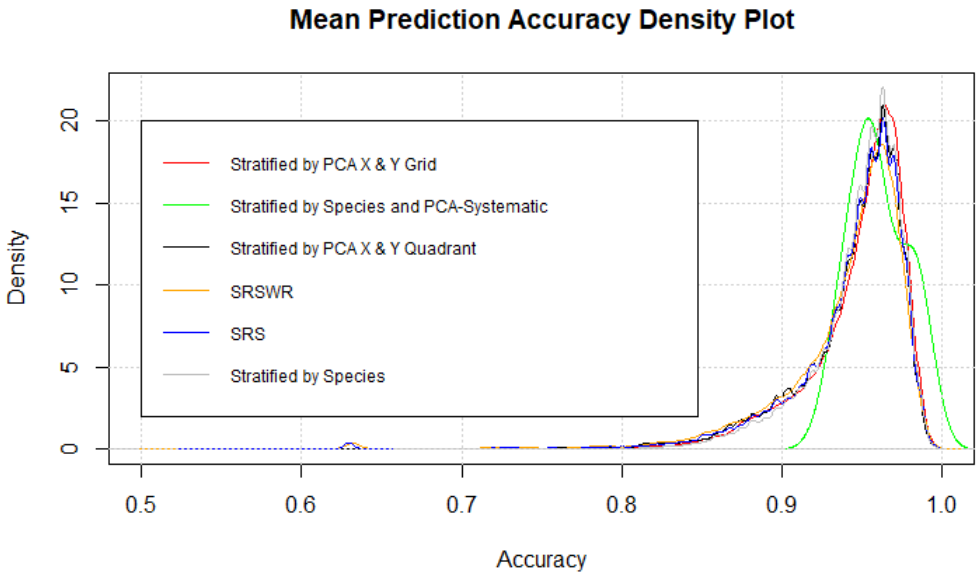

Figure 4.4: Performance of PCA-Systematic compared with all Data Splitting Methods.

4.2 The Annual Survey of Public Employment and Payroll (U.S. Census Bureau)

This section used data from the ASPEP (Annual Survey of Public Employment and Payroll) as a prototype of real survey data to illustrate how my novel PCA-Systematic sampling can be employed and show how it outperforms standard systematic sampling.

4.2.1 ASPEP Suvey Data File and Study Variables

Let us give some details concerning the ASPEP survey, the collected file and the study variables.

The U.S Census Bureau conducts economic Censuses of about 90,000 federal, state, and local government units every five years to collect data on the number of full-time and part-time federal, state, and local government employees and their payroll. Between two consecutive censuses (in the years ending with 2 and 7, e.g., (2012 and 2017), the U.S. Census

Bureau also conducts The Annual Survey of Public Employment and Payroll (ASPEP) to collect similar data on a nationally representative sample basis on federal, state, and local governments' civilian employees and their gross payrolls [42].

ASPEP survey is of significant importance because it is the only source of public employment data that provide state and local government data on full-time and part-time employment, part-time hours worked, full-time equivalent employment, and payroll statistics by governmental function. These governmental functions include school data, which covers elementary and secondary education, higher education, and other functions of the government such as libraries, police protection, judicial and legal, parks and recreation, fire protection, electric power, gas supply, financial administration, central staff services, highways, public welfare, solid waste management, sewerage, social insurance administration, health, hospitals, water supply, transit, natural resources, correction, air transportation, water transport and terminals, and housing and community development. The data is intended for public access and use and made available on the web online. This dissertation used the 2017 Census of Governments: Employment Component (CoG-E). A Census is an entire sample that covers the entire population. Census will include every unit in the population. In probability sampling, a Census is a probability sample where every element is taking by certainty.

The CoG-E collects government data classified into five types of governments: counties, cities, townships, special districts, and school districts. The different types of governments perform various governmental activities which are designated for the survey by governmental function codes.

The key statistics measured by this survey and variables of interest are:

- Total of Full-time employees
- Total of Full-time pay payroll
- Total of Part-time payroll
- Total of Part-time hours

- Total to Total Pay

ASPEP defines *Full-time employees* as any person whose employment during a pay period that averages at least 30 hours of service weekly or 130 hours of service monthly; with this definition, regardless of the job type, full-time temporary or seasonal employees are included.

Full-time pay for ASPEP survey is defined as Gross payroll amounts for the one-month period of March for full-time employees. This includes all salaries, fees, commissions, and overtime paid to employees before withholdings for taxes and insurance. It also includes regular incentive payments that are paid at regular time-period intervals. But these gross payroll amounts exclude employer share of fringe benefits like retirement, Social Security, health and life insurance, lump-sum payments, and so forth.

ASPEP defines *Part-time pay* as gross payroll amounts for the one month of March for part-time employees only. This gross payroll should include all salaries, fees, commissions, and overtime paid to employees before withholdings for taxes and insurance. This definition should also include incentive payments paid at regular pay intervals but exclude employer share of fringe benefits like retirement, Social Security, health and life insurance, lump-sum payments, and so forth.

Part-time hours can be defined as the number of hours worked by part-time employees during the pay period. Note that these data are not collected for publication but rather are used to calculate full-time equivalent employment statistics.

4.2.2 PCA Impact on Correlation for ASPEP Data

Pearson correlation analysis is carried out to measure the linear dependence between ASPEP variables before and after the PCA transformation is performed.

The correlation matrix results presented in Table 4.13 and Table 4.12 show how the first component PCA Z_1 improves the overall Pearson correlation for all of the study variables.

Table 4.12: Correlation Analysis before PCA for ASPEP Variables

	totpay17	ftemp17	ptemp17	pthours17
totpay17	1.00	0.96	0.55	0.53
ftemp17	0.96	1.00	0.46	0.41
ptemp17	0.55	0.46	1.00	0.96
pthours17	0.52	0.41	0.96	1.00

Table 4.13: Correlation Analysis after PCA for ASPEP Variables

	Z1	totpay17	ftemp17	ptemp17	pthours17
Z1	1.00	**0.86**	**0.83**	**0.87**	**0.85**
totpay17	**0.86**	1.00	0.96	0.55	0.52
ftemp17	**0.83**	0.96	1.00	0.46	0.41
ptemp17	**0.87**	0.55	0.46	1.00	0.96
pthours17	**0.85**	0.52	0.41	0.96	1.00

The novel PCA-systematic sampling design is compared with the five possible systematic sampling methods. For example, systematic sampling using *FTEMP* as a sort variable method consists of two steps;

1. Sort the ASPEP data file by the variable *FTEMP*

2. Draw a systematic sample afterward

There are five choices of variables to sort by in step1. So that is why there are five methods to be compared with PCA-systematic. The ASPEP data set is considered as the population. First, I compute the population truths: the average of each variable using the full (unsampled) ASPEP data file.

I denote The population average for a given study variable y by $\bar{Y}$. We have

$$\bar{Y} = \frac{1}{N}\sum_{i=1}^{N} y_i \tag{4.5}$$

N is the population size and in this case is the number of observations of the ASPEP file.

The second column in Table 4.15 labeled "True Value" holds the value of the population mean for the variable *FTEMP*

The same thing applies to Tables 4.16 through 4.19 but for the study variables *FTPAY*,*PTHOURS*,*PTPAY*, and *TOTPAY* respectively.

Each table presents the Monte Carlo simulation results of performance comparison between the different systematic sampling methods for one of the study variables. The Metric used for performance is the RRMSE. Lower values are preferable.
With this simulation work, for each sampling design method, a replicate of 10,000 random samples is produced and for each sample an estimated population parameter is derived using the weighted sample mean using the survey weights.
The weighted sample mean of the r^{th} MC replicate is an estimate of the population average $\tilde{Y}$ and computed as

$$\hat{Y}^{(r)} = \frac{1}{n}\sum_{i=1}^{n} y_i^{(r)} \tag{4.6}$$

The Tables 4.16 through 4.19 are sorted by ascending order of RRMSE. Table 4.16 shows how PCA-Systematic sampling which consists of sorting by the first principal component Z_1 has the lowest RRMSE of 13.3 and therefore outperforms all other systematic sampling for the study variable *FTPAY*.

Table 4.14: RRMSE for estimating FTEMP

Key Variable	Sort Variable	True Value	Mean of Yhat	RRMSE
FTEMP	FTEMP	41	41	10.5
FTEMP	Z1	41	41	10.7
FTEMP	FTPAY	41	41	11.6
FTEMP	PTPAY	41	41	13.0
FTEMP	PTHOURS	41	41	13.7
FTEMP	TOTPAY	41	41	16.3

Table 4.15: RRMSE for estimating FTPAY Full Time Pay

Key Variable	Sort Variable	True Value	Mean of Yhat	RRMSE
FTPAY	Z1	202.73	199.608	13.3
FTPAY	TOTPAY	202.73	199.225	13.4
FTPAY	PTHOURS	202.73	209.944	15.3
FTPAY	FTPAY	202.73	203.029	17.5
FTPAY	PTPAY	202.73	203.598	19.6
FTPAY	FTEMP	202.73	204.432	20.5

4.2.3 Results

I assess accuracy of PCA-Systematic data splitting method compared to traditional systematic data splitting methods by computing the Relative Root Mean Squared Error. Tables 4.14 through 4.18 present RRMSE for the ASPEP study variables *FTEMP, FTPAY, PTHOURS, PTPAY* and *TOTPAY*. The tables below show the results:

Table 4.16: RRMSE for estimating Part Time Hours

Key Variable	Sort Variable	True Value	Mean of Yhat	RRMSE
PTHOURS	Z1	906	900	16.1
PTHOURS	PTPAY	906	901	16.5
PTHOURS	PTHOURS	906	914	17.3
PTHOURS	TOTPAY	906	903	17.5
PTHOURS	FTPAY	906	915	17.5
PTHOURS	FTEMP	906	920	19.2

Table 4.17: RRMSE for estimating Part Time Pay variable

Key Variable	Sort Variable	True Value	Mean of Yhat	RRMSE
PTPAY	FTPAY	15,980	16,026	19.0
PTPAY	PTHOURS	15,980	16,010	19.1
PTPAY	FTEMP	15,980	16,008	19.8
PTPAY	PTPAY	15,980	15,941	20.1
PTPAY	TOTPAY	15,980	16,045	21.0
PTPAY	Z1	15,980	16,238	22.0

Table 4.18: RRMSE for estimating Total Pay variable

Key Variable	Sort Variable	True Value	Mean of Yhat	RRMSE
TOTPAY	TOTPAY	218,712	217,700	13.7
TOTPAY	PTHOURS	218,712	217,326	13.8
TOTPAY	Z1	218,712	219,055	14.9
TOTPAY	FTPAY	218,712	218,170	15.8
TOTPAY	PTPAY	218,712	219,831	16.7
TOTPAY	FTEMP	218,712	220,283	17.9

I further combine the results from Tables 4.15-4.19 to derive a global RRMSE as a measure of performance by computing for each sort-variable of the systematic data splitting an average RRMSE across all the tables. Table 4.20 is the resulting global RRMSE which shows how PCA-Systematic data splitting method outperforms all other possible systematic data splitting methods.

Table 4.19: Overall RRMSE for all the Possible Sort Key Variables

Sort Variable	RRMSE
Z_1	**15.4**
PTHOURS	15.8
FTPAY	16.3
TOTPAY	16.4
PTPAY	17.2
FTEMP	17.6

4.3 The American Housing Survey Public File (U.S. Census Bureau)

In this section, I use a data from the 2011 American Housing Survey (AHS) to illustrate the proposed data splitting technique.

4.3.1 AHS Data File

The American Housing Survey (AHS) is one of the oldest surveys in the U.S that started in 1997. The survey is authorized by an act of the U.S Congress with funding and oversight by the Department of Housing and Urban Development agency (HUD).
HUD sponsors the AHS and delegates to the U.S census bureau the role and responsibility of designing the sample, collecting the data, and producing the estimates at the national level and at the state level for the selected metropolitan areas.

The mission of the American Housing Survey is to inform the public and help economists and policymakers. The survey provides the public and data users from different backgrounds and expertise with detailed and timely information about housing quality, housing costs, and neighborhood assets to support effective housing policy, programs, and markets.

The AHS consists of two surveys; a national survey conducted every other odd-numbered year and multiple metro surveys in selected regions. The surveys are independent of one another. However, In 2011 both surveys were combined together to form one aggregated big file. The Census Bureau developed a new set of weights to accommodate this change. This combination resulted in a larger file compared to what a regular national file would be.

One of the advantages of this survey is the fact that HUD, with collaboration with the U.S Census bureau, release the entire microdata file and make it available online every year the survey is conducted. the file is so detailed to the point that individual responses to survey questions are available at the housing unit level and not just aggregated summary tables like most other surveys do in the U.S and elsewhere.

It is to be noted that the file is examined, and data disclosure avoidance techniques such as top coding, bottom coding, and noise infusion are employed and performed before the release of this microdata file to protect the identity of the respondents to the survey and maintain confidentiality. This file is known as the Public use file (PUF); just as its name indicates, it is a file to be used by everyone in public.
The AHS PUF file is offered in two different data sets:

- An integrated National Sample PUF which includes individual responses from a representative sample of the entire nation, from representative samples of the largest metropolitan areas, and from a representative sample of households receiving HUD rental assistance.

- An independent Metropolitan Area Samples PUF includes individual responses from representative samples of a number of metropolitan areas selected from among America's top 51 largest metropolitan areas.

The 2011 AHS survey file that combined the National survey data with the metropolitan areas contains a large number of variables. It has 186,448 records and 3,078 variables that measure a comprehensive inventory of all housing in the United States of America. This survey is considered to be demographic survey data. It covers data on a wide range of housing subjects, including single-family homes, apartments, manufactured housing, vacant units, family composition, income, housing, and neighborhood quality, housing costs, equipment, fuel type, and recent moves

4.3.2 Selected AHS Variables

AHS measures several cost variables that are good candidates for this dissertation as they are continuous; some of these variables are monthly such as Average monthly cost of Gas (AMTG) or the monthly average cost of electricity (AMTE); some are annual, such as Homeowners insurance(AMTI) or the annual cost of water consumption (AMTW).

Each cost variable can be treated as a dependent variable, and one possibility would be to fit a prediction model for each cost separately, assuming independence. Statistical learning methods can be employed after clustering is made on continuous variables after they get transformed to ordinal categorical variables by creating ranges. The methods include but not limited to KNN(K nearest neighbors), LDA(Linear discriminant analysis), QDA(Quadratic discriminant analysis). If the purpose is to make a prediction, then employing more flexible parametric methods would also be the right choice since the goal would be to increase the prediction's accuracy, not Statistical Inference. Depending on the purpose, there is a tradeoff between Interpretability and increasing the accuracy of prediction.

I seek to select the best covariates or features or "independent variables" that can predict each cost, since some explanatory variables are nominal, such as the type of the living unit, for example, a housing unit could be located in an attached building, detached building, or in a multi-Unit types building, etc.... It would have been a good idea to expand the model above to account for that. one can either do an Analysis of Variance or do separate regression for each type. This dissertation uses the AHS data to show how can traditional systematic sampling be improved by using the novel PCA- Systematic sampling.
Related to cost variables that are used are presented in Table 4.20:

These variables are mainly chosen because they are good candidates as they are numeric variables and, therefore, good candidates for this dissertation.

The variable *AMTE* is the response to the following survey question: In the past 12 months, what was the average MONTHLY cost for electricity?

For *AMTG*, the survey questionnaire is: In the past 12 months, what was the average MONTHLY cost for gas?

For *AMTI*, it is: In the last 12 months, what was the total cost? If the head of household (householder's) response exceeds $5,582 or more, then for confidentiality reason, it is kept at $5,582. This technique is known as top-coding and used for privacy protection as the risk of disclosure increases; it would become easy to identify the housing unit if values become

Table 4.20: American Housing Survey Cost Variables

Variable Name	Description
AMTE	Average monthly cost of electricity
AMTG	Average monthly cost of gas
AMTI	Annual cost of homeowners insurance
AMTO	Annual cost of fuel oil
AMTT	Annual cost of garbage and trash
AMTX	Annual Real Estate Tax Payments
AMTW	Annual cost of water and sewage
CPRICE	Cost of construction plus value of land
CSTMNT	Annual cost for routine maintenance
FMHOTF	Average regular cost of other required monthly fees
HKRAC	Cost of alteration/repair due to Hurricane Katrina
RAC	Cost of replacements/additions to unit
ZSMHC	Monthly housing costs

extreme or outliers, especially if they are combined with other information.

The variable AMTO is the answer to the following question: From 12 months ago to the current month and year, what was the total cost for fuel oil? for this variable, responses that are equal or greater than $ 6,057 are also kept at $ 6,057.

The variable AMTT is the answer to the following question: From 12 months ago to the current month and year, what was the total cost for garbage and trash collection? the minimum and maximum allowed values for this variable are $ 1 and $ 2,990.

To collect the variable AMTW, the surveyed housing unit head of the household is asked the question: "From 12 months ago to the current month and year, what was the total cost for water supply and sewage disposal? " Topcoding value for this variable is also $ 3,358.

4.3.3 AHS Exploratory Data Analysis

First, I start to explore the data for each of the selected variables. This process involves the following:

- Compute summary statistics to include both measures of central tendency such as

the mean, median, and mode and measures of spread, such as the variability and the standard deviation.

- Compute descriptive statistics based on moments such as the skewness and kurtosis to identify the shape of the probability distribution. Skewness measures asymmetry; values near zero indicate that the distribution is symmetric. If the value is positive, then it is positively skewed; otherwise, it is negatively skewed. Kurtosis provides a measure for the tails of the distribution; If The value is higher than three, then the distribution is heavy-tailed. Otherwise, if it is less than three, then the distribution is considered to be platykurtic.

- Visualize the data distribution using the frequency histogram, compute the kernel density estimate, and overlay the fitted density curve with the histogram; this helps with outliers detections. Kernel density estimate is a non-parametric method and does not make any assumptions about the data's underlying probability distribution. One can visually verify, for example, if normality assumptions about the distribution are reasonable by adding and superimpose a fitted normal curve to the graph. In addition to assessing normality, other parametric distributions can also be fitted and verified.

Results from analyzing the data structure of the selected variables are shown in figures 4.5 through 4.11. For each cost variable, the analysis shows that the distribution is positively skewed. It also reveals the presence of an outlier on the right tail of the distribution. The outliers are essentially caused by top coding. With top coding, among other techniques, a unique single estimate is allocated to a large group of housing units. Some other characteristics and information were omitted or modified to protect the privacy of respondents; for example, the state information is not made available for those units.
All selected variables except for $AMTO$, with a kurtosis value of 1.6, all variables have heavy-tailed distributions.

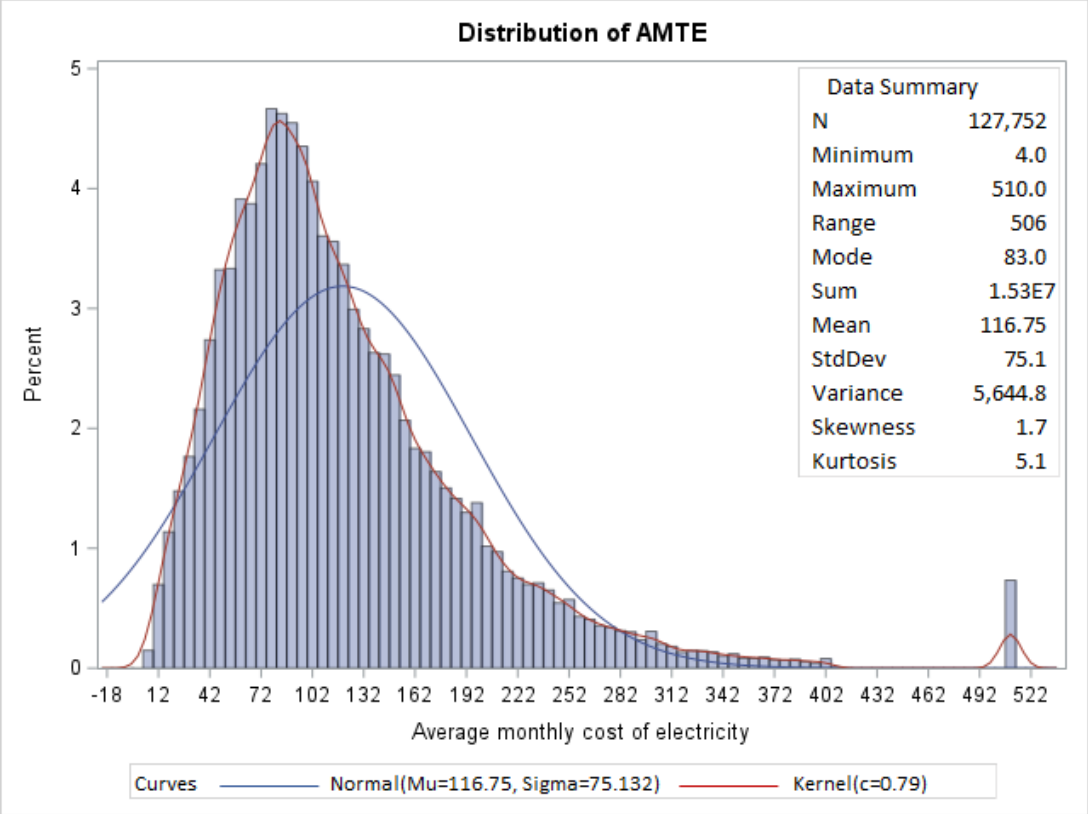

Figure 4.5: Distribution for Average Monthly Cost of Electricity.

4.3.4 PCA Impact on Correlation

This subsection analyzes the correlation among the study variables before and after PCA. The PCA is used to engineer new features that can be used only as a tool for sorting before Systematic method is implemented as data splitting and not as features variables or predictors per se to have a fair comparison between different data splitting.

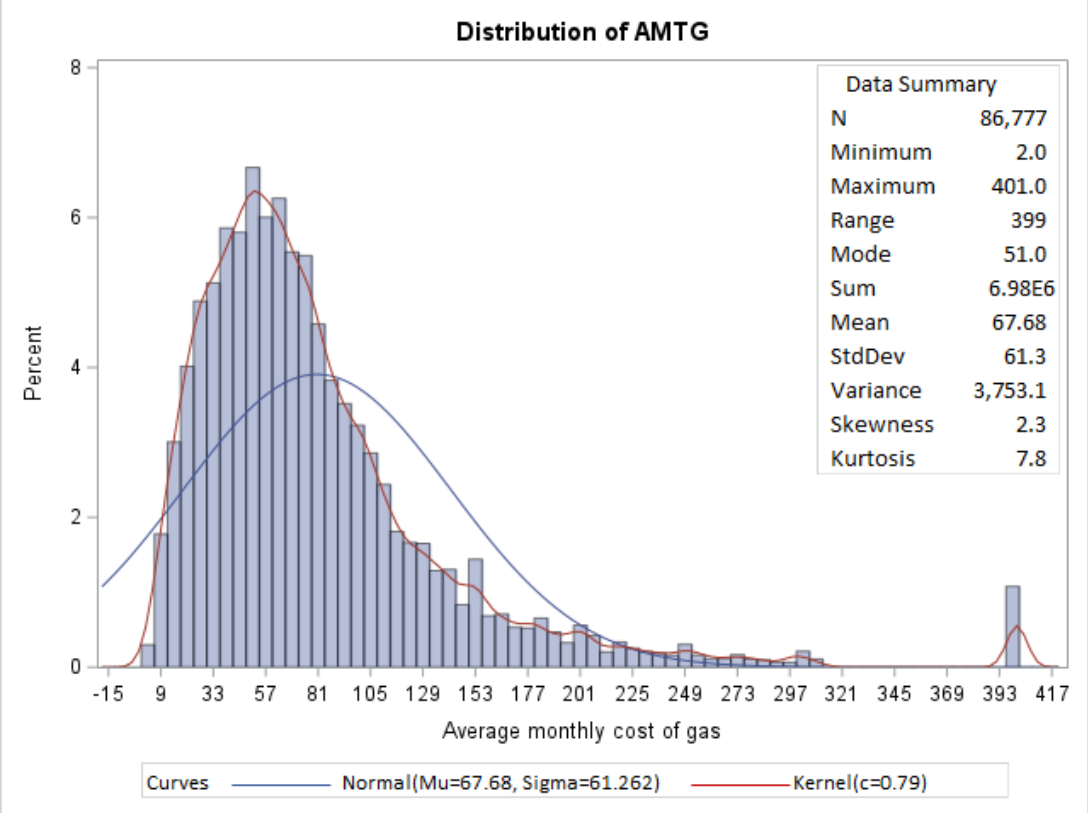

Figure 4.6: Distribution for Average Monthly Housing Cost of Gas.

Table 4.21: Correlation Analysis Before PCA for AHS Variables

	AMTE	AMTF	AMTG	AMTI	AMTO	AMTT	AMTW
AMTE	1.00	0.02	0.24	0.25	0.13	0.17	0.17
AMTF	0.02	1.00	0.01	-0.04	-0.11	-0.00	-0.03
AMTG	0.24	0.01	1.00	0.13	0.11	0.03	0.10
AMTI	0.25	-0.04	0.13	1.00	0.27	0.12	0.18
AMTO	0.13	-0.11	0.11	0.27	1.00	0.07	0.13
AMTT	0.17	-0.00	0.03	0.12	0.07	1.00	0.18
AMTW	0.17	-0.03	0.10	0.18	0.13	0.18	1.00

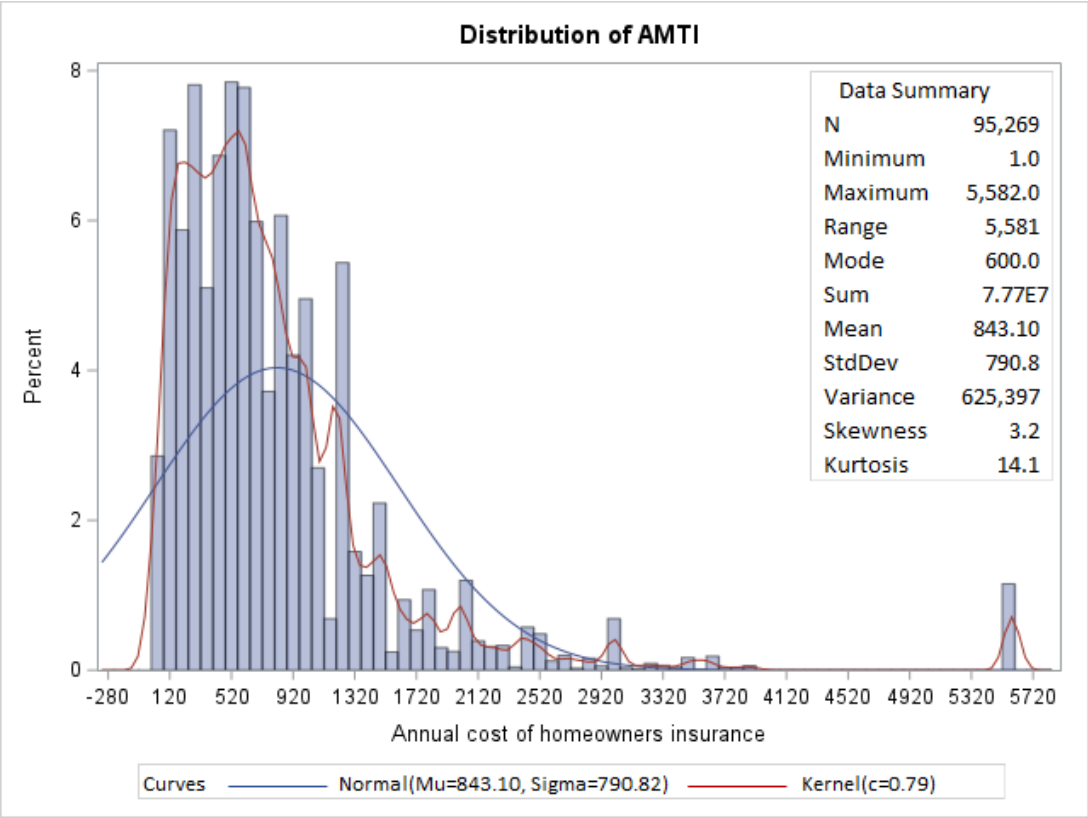

Figure 4.7: Distribution for Average Monthly Housing Cost for Homeowners Insurance. homeowners insurance

Table 4.22: Correlation Analysis After Adding The First PCA Component for AHS Variables

	Z_1	AMTE	AMTF	AMTG	AMTI	AMTO	AMTT	AMTW
Z_1	1.00	**0.92**	-0.22	-0.06	**0.90**	0.41	0.06	**0.73**
AMTE	**0.92**	1.00	0.02	0.24	0.25	0.13	0.17	0.17
AMTF	-0.22	0.02	1.00	0.01	-0.04	-0.11	-0.00	-0.03
AMTG	-0.06	0.24	0.01	1.00	0.13	0.11	0.03	0.10
AMTI	**0.90**	0.25	-0.04	0.13	1.00	0.27	0.12	0.18
AMTO	0.41	0.13	-0.11	0.11	0.27	1.00	0.07	0.13
AMTT	0.06	0.17	-0.00	0.03	0.12	0.07	1.00	0.18
AMTW	**0.73**	0.17	-0.03	0.10	0.18	0.13	0.18	1.00

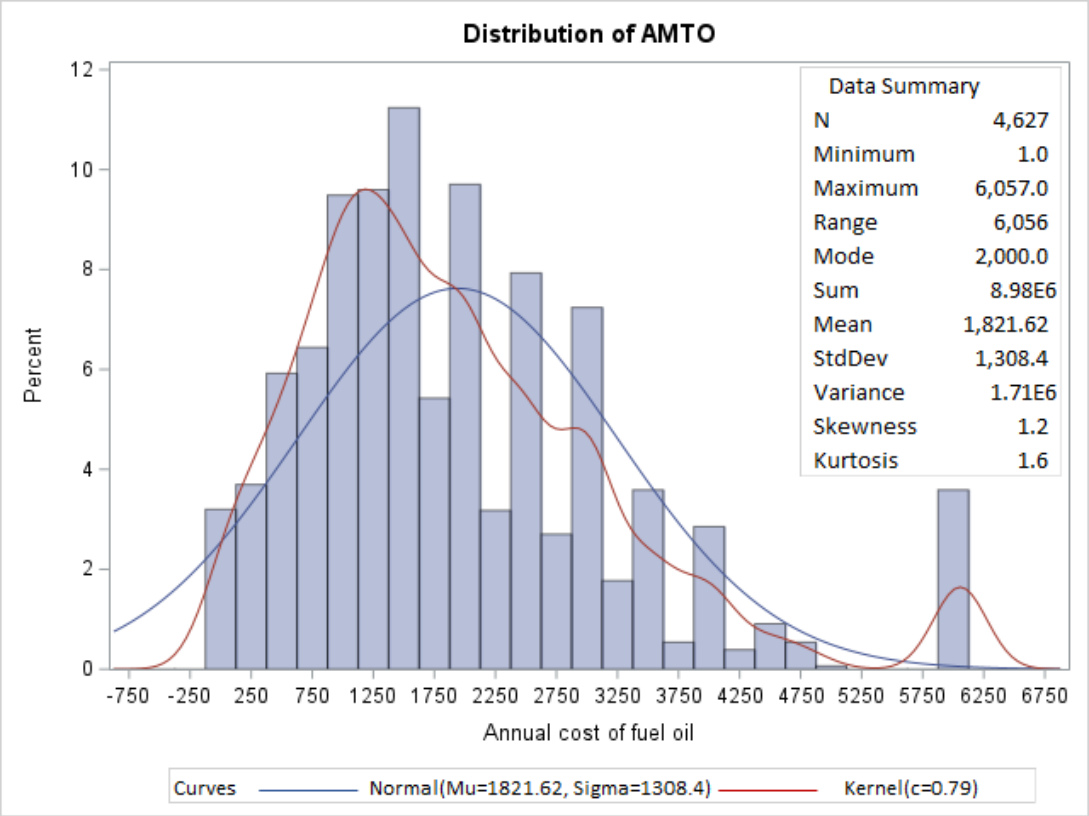

Figure 4.8: Distribution for Annual Cost of Fuel Oil.

Table 4.23: Correlation Analysis After Adding the First and Second PCA components for AHS variables

	Z_1	Z_2	AMTE	AMTF	AMTG	AMTI	AMTO	AMTT	AMTW
Z_1	1.00	0.00	**0.92**	-0.22	-0.06	**0.90**	0.41	0.06	**0.73**
Z_2	0.00	1.00	-0.14	-0.26	**0.80**	-0.17	**0.70**	0.31	-0.04
AMTE	**0.92**	-0.14	1.00	0.02	0.24	0.25	0.13	0.17	0.17
AMTF	-0.22	-0.26	0.02	1.00	0.01	-0.04	-0.11	-0.00	-0.03
AMTG	-0.06	**0.80**	0.24	0.01	1.00	0.13	0.11	0.03	0.10
AMTI	**0.90**	-0.17	0.25	-0.04	0.13	1.00	0.27	0.12	0.18
AMTO	0.41	**0.70**	0.13	-0.11	0.11	0.27	1.00	0.07	0.13
AMTT	0.06	0.31	0.17	-0.00	0.03	0.12	0.07	1.00	0.18
AMTW	**0.73**	-0.04	0.17	-0.03	0.10	0.18	0.13	0.18	1.00

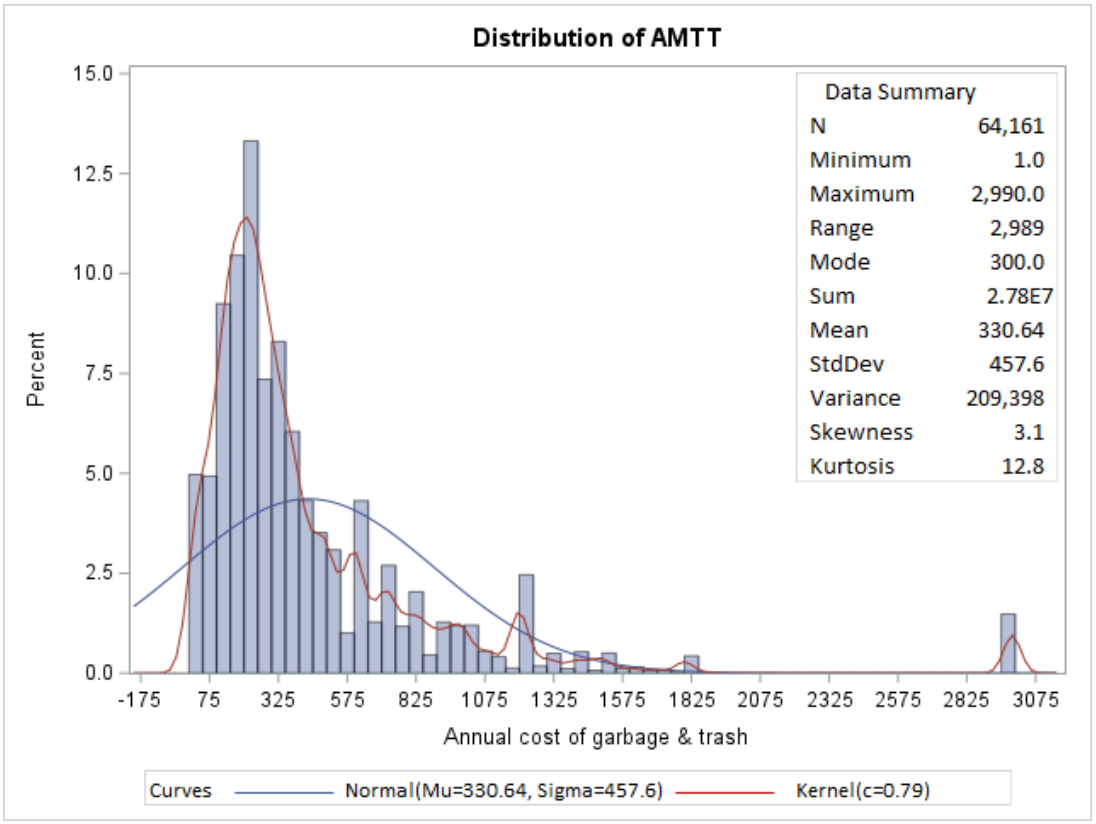

Figure 4.9: Distribution for Annual Cost of Garbage and Trash.

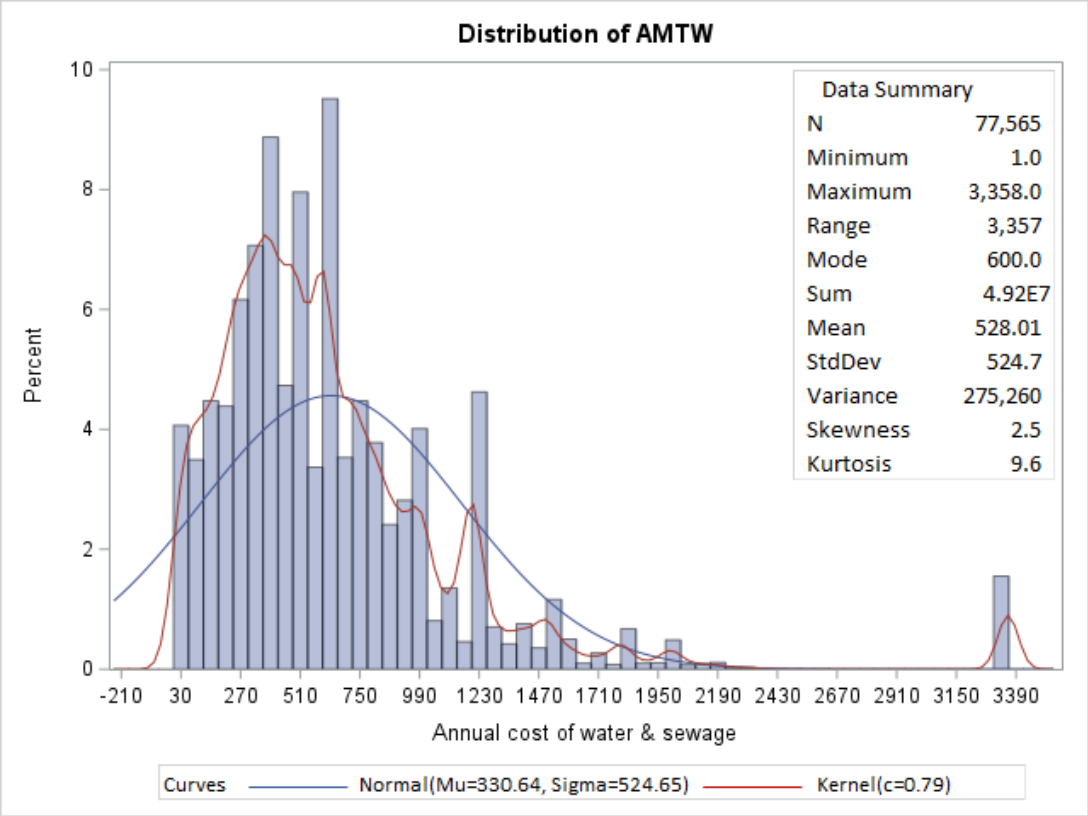

Figure 4.10: Distribution for Annual Cost of Water and Sewage.

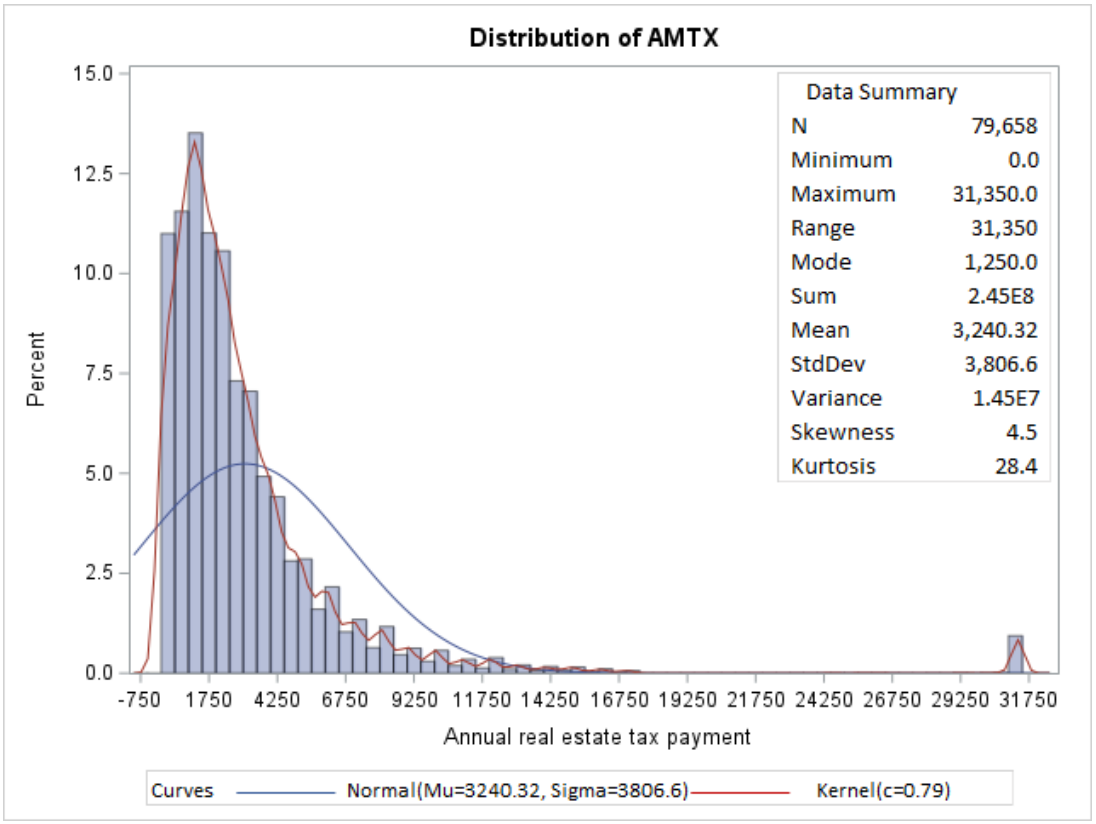

Figure 4.11: Distribution for Annual Cost of Real Estate Tax Payments.

4.3.5 Data Splitting Methodology for AHS

Canonical correlation analysis carried out for AHS data shows that the number of components to be kept is two as the first two eigenvectors capture the most of the variation. Adding more vectors will only add noise. The stratification made using two principal components is sufficient. Tables 4.22 and 4.23 show the correlation coefficients when the Z_1 and Z_2 are sequentially added.

4.3.6 Results

I assess accuracy of PCA-Systematic data splitting method compared to traditional systematic data splitting methods by computing the Relative Root Mean Squared Error. Tables 4.25 through 4.31 present RRMSE for the key AHS study variables *AMTE*, *AMTF*, *AMTG*, *AMTI*, *AMTO*, *AMTT* and *AMTW*. I further assessed the accuracy of PCA-Systematic data splitting method by computing a global performance measure RRMSE...

Table 4.24: RRMSE for estimating Average Monthly Cost of Electricity

Study Variable	Sort Variable	True Value	Mean of Yhat	RRMSE
AMTE	AMTE	116.75	116.58	0.0257
AMTE	Z_1	116.75	115.76	0.0527
AMTE	AMTX	116.75	116.57	0.0573
AMTE	AMTW	116.75	116.41	0.0826
AMTE	AMTI	116.75	116.56	0.0843
AMTE	AMTT	116.75	117.09	0.0873
AMTE	AMTG	116.75	116.31	0.0882
AMTE	AMTO	116.75	116.59	0.1490
AMTE	SRSWR	116.75	116.62	0.0835
AMTE	SRSWOR	116.75	116.62	0.0784

Table 4.25: RRMSE for estimating Average Monthly Cost of Gas

Key Variable	Sort Variable	True Value	Mean of Yhat	RRMSE
AMTG	AMTG	67.68	67.86	0.0359
AMTG	Z_1	67.68	67.70	0.0851
AMTG	AMTE	67.68	67.49	0.0873
AMTG	AMTT	67.68	67.70	0.1046
AMTG	AMTI	67.68	67.52	0.1070
AMTG	AMTX	67.68	67.86	0.1183
AMTG	AMTO	67.68	67.56	0.1611
AMTG	AMTW	67.68	67.16	0.1767
AMTG	SRSWR	67.68	67.77	0.1239
AMTG	SRSWOR	67.68	67.59	0.1164

Table 4.26: RRMSE for estimating Average Monthly Cost of Homeowners Insurance

Key Variable	Sort Variable	True Value	Mean of Yhat	RRMSE
AMTI	AMTI	843.10	842.35	0.0259
AMTI	Z_1	843.10	825.42	0.0633
AMTI	AMTG	843.10	844.87	0.0727
AMTI	AMTW	843.10	841.85	0.0773
AMTI	AMTE	843.10	843.42	0.0883
AMTI	AMTX	843.10	840.36	0.0986
AMTI	AMTT	843.10	840.79	0.1221
AMTI	AMTO	843.10	848.54	0.1587
AMTI	SRSWR	843.10	844.54	0.0961
AMTI	SRSWOR	843.10	842.36	0.0899

Table 4.27: RRMSE for estimating Average Annual Cost of Fuel Oil

Key Variable	Sort Variable	True Value	Mean of Yhat	RRMSE
AMTO	AMTO	1821.62	1819.92	0.0178
AMTO	AMTE	1821.62	1819.97	0.0497
AMTO	AMTW	1821.62	1820.46	0.0594
AMTO	Z_1	1821.62	1790.30	0.0617
AMTO	AMTI	1821.62	1817.98	0.0735
AMTO	AMTG	1821.62	1826.68	0.0759
AMTO	AMTT	1821.62	1830.96	0.0782
AMTO	AMTX	1821.62	1823.37	0.0870
AMTO	SRSWR	1821.62	1823.64	0.0909
AMTO	SRSWOR	1821.62	1821.59	0.0861

Table 4.28: RRMSE for estimating Average Annual cost of Garbage and Trash

Key Variable	Sort Variable	True Value	Mean of Yhat	RRMSE
AMTT	AMTT	330.64	331.62	0.0388
AMTT	AMTG	330.64	331.89	0.0686
AMTT	AMTI	330.64	330.55	0.0905
AMTT	Z_1	330.64	326.17	0.0945
AMTT	AMTO	330.64	330.60	0.0946
AMTT	AMTX	330.64	331.63	0.0995
AMTT	AMTW	330.64	331.11	0.1214
AMTT	AMTE	330.64	332.51	0.1303
AMTT	SRSWR	330.64	330.11	0.1259
AMTT	SRSWOR	330.64	330.64	0.1169

Table 4.29: RRMSE for estimating Average Annual Cost of Water and Sewage

Key Variable	Sort Variable	True Value	Mean of Yhat	RRMSE
AMTW	AMTW	528.01	527.10	0.0500
AMTW	AMTT	528.01	527.98	0.0715
AMTW	AMTO	528.01	526.12	0.0849
AMTW	AMTG	528.01	530.46	0.0934
AMTW	Z_1	528.01	540.59	0.0958
AMTW	AMTE	528.01	525.65	0.1199
AMTW	AMTI	528.01	528.00	0.1200
AMTW	AMTX	528.01	529.54	0.1343
AMTW	SRSWR	528.01	527.78	0.1154
AMTW	SRSWOR	528.01	528.02	0.1100

Table 4.30: RRMSE for estimating Average Real Estate Tax Payments

Key Variable	Sort Variable	True Value	Mean of Yhat	RRMSE
AMTX	AMTX	3240.32	3237.19	0.0310
AMTX	AMTO	3240.32	3239.85	0.0749
AMTX	Z_1	3240.32	3161.62	0.0773
AMTX	AMTT	3240.32	3250.79	0.0797
AMTX	AMTG	3240.32	3221.13	0.1038
AMTX	AMTW	3240.32	3235.97	0.1039
AMTX	AMTE	3240.32	3252.38	0.1050
AMTX	AMTI	3240.32	3231.33	0.1427
AMTX	SRSWR	3240.32	3248.42	0.1195
AMTX	SRSWOR	3240.32	3236.59	0.1118

Table 4.31: Global RRMSE

Sort Variable/Method	RRMSE%
Stratified by (Z1,Z2)	**7.58**
AMTG	7.69
AMTT	8.32
AMTE	8.66
AMTX	8.94
AMTI	9.20
AMTW	9.59
AMTO	10.59

Chapter 5: Conclusion and Summary of Results

The goal of this section is to summarize and highlight the contribution of my dissertation. This section explains how this work is novel and different from current methods and presents future work and possible improvements.

There are two main contributions in this dissertation:

1. The first one is the development of a novel method of sampling. This novel sampling design will benefit any application that requires sampling from a population. For example, for survey sampling, the PCA-Systematic sampling introduced in this dissertation is a methodology used as an alternative to existing sampling designs. This methodology can also be used as an application for machine learning when allocating parts of the main data into training and testing.

   This sampling design, when applicable, renders the sample more representative of the population to sample from in comparison to other sampling deigns, Chapter 2 explains the reason for this theoretically, Chapter 3 shows the simulation work using multinormal distributions, and Chapter 4 demonstrates how this works for real datasets such as the famous Fisher Iris flowers data.

2. The second contribution is creating an index quality for sampling, regardless of the sampling design used. For example, in the context of machine learning, when datasplit is conducted, this index can be computed and used to assess the quality of both training, testing, and validation datasets. If this index is below a certain threshold set by the user, then the split is rejected and conducted again until all three thresholds are above the threshold.

- The training data is treated as a random sample that gets carefully derived from the full dataset.

- Simple random sampling is not the only method in which the data can be split into training and testing. There have been several methods in survey sampling from a finite population in the literature that can serve machine learning. This dissertation's approach is interdisciplinary; it was to join survey sampling methods of sample selection with data splitting in machine learning. The process involves running PCA, then used its components to serve as sort variable(s) for systematic sampling. Chapters 2 and 3 of this dissertation demonstrated how to use the PCA components as a stratification tool. As shown in both simulated data and real data, The designed samples, when stratified based on PCA, produced better representative samples than traditional methods. the strata were formed based on two elements; the eigenvectors to determine the number of axes and the eigenvalues to determine the length for the grid in each dimension.

- Perhaps the most popular method that is currently used is SRSWOR. However, often, it does not properly represent the entire dataset and leads to a biased sample, especially the relationship between important variables that are essential for the machine learning model fitting is distorted. So the measure of the test error is not reliable.

- Splitting data can be handled as a proper randomized sampling procedure. For example, this dissertation introduced a new sampling scheme that is based on systematic random sampling procedure and PCA. In Chapter 3 of this dissertation, Monte Carlo simulation analysis provided evidence to demonstrate how this novel method outperformed current known data splitting methods.In chapter 4, real datasets were used to show how the novel sampling methodology, PCA systematic sampling, produced better training data than traditional sampling methods for an array of different machine learning methods; Models fitted using that training dataset had higher prediction accuracy.

My contribution in this dissertation consisted of the development of a novel sampling method; this method is used to efficiently split the data into training and testing for machine learning. The method is based on using PCA as a stratification tool when data splitting is conducted. My contribution consists of improving systematic sampling; the goal is to preserve the relationship between important variables for the training dataset after the data split is performed.

The best current existing method of systematic sampling in order to draw the best representative sample consists of two steps:

1. Sorting the file to be sampled by the feature(s) that are highly correlated with the variable of interest (the target variable).

2. Implement systematic sampling afterwards.

However, in my method, I added a step of canonical correlation analysis; so there are three main steps:

1. Perform Principal component analysis (PCA) on the dataset and compute the correlation between the target variable, all the features, and the principal components, then determine the component with the maximum absolute value correlation; this is the first principal component.

2. Sort the file by the variable chosen in step 1.

3. Implement systematic sampling afterward.

At the end of the three steps above, the method will create a systematic sample for a given desired sample size. The three steps above can be repeated by sorting by the second, third, or any other principal component instead of the first one.

The algorithm can also be used to create even a better training dataset that is fully representative of the main dataset; it is best to create multiple samples and then take the union of the subsamples.

This modified version of implementing systematic sampling resulted in an improved estimator for the population parameter such as the mean, this is of an important value for survey sampling, also it resulted in an improved prediction accuracy in the context of machine learning example are shown in section 3.5.
In my work(dissertation) I prove three hypotheses:

- Chapter 2 of this dissertation proved that with a higher probability, the correlation between the first principal component and the target variable is higher than the correlation between the target variable and the original feature variables.
- Using the PCA- Systematic sampling design produced a more efficient sample as it represented the true population better than using standard systematic sampling.
- Using the PCA-Systematic sampling method as a data splitting method resulted in better prediction accuracy for the testing data.

5.1 Limitations

- The novel sampling methods developed in this dissertation are based on the use of principal component analysis methodology; however, PCA requires to have numerical variables. At least two numerical and probably continuous variables are required in order for PCA to be implemented. The absence of numeric feature variables renders the PCA-Systematic splitting methodology not applicable; this is when all the variables are categorical. The remedy for this situation is to use stratified simple random sampling without replacement to split the data into training and testing. A sensible thing would be to use the proportional allocation methodology of the sample size when creating the smaller data set of the data split, which is most likely the test dataset for the two-way data split scheme. The target variable should be used as a stratum variable for this multistage sampling, so the test data set should be a stratified sample by y while preserving the test to train sample size ratio within each label of y.

- The proposed method of data splitting in this dissertation does not apply to time series data.

5.2 Future Work

1. In this dissertation, the focus was on using Pearson correlation. However, there are other measures of statistical dependence among variables worth exploring. For example, both Kendall and Spearman are nonparametric measures of correlation between the ranking of two variables. Because these correlations are defined using the rank of data and not the raw data, it gives them the property of being invariant under monotone transformation, a property that makes them robust and less sensitive to outliers than Pearson Correlation [45].

   Other transformations might include Exponential, Quadratic, Power, Spearman, inverse, square root.

   For example, Box-Cox is a popular transformation used to achieve linearity, Chapter one mentions how a higher value of Pearson correlation will make systematic sampling more efficient, and then Chapter two later shows how PCA transformation increases this correlation. In future work, I like to explore the idea of using Kendall's rank correlation and also Spearman's rank correlation instead of Pearson correlation. The goal would be to:

   (a) Initially investigate whether sorting by the variable that is highly Kendall-correlated or Spearman-correlated with the response variable has an effect on the efficiency of systematic sampling

   (b) The second step would be to investigate if the use of PCA transformation would still increase the correlation of Kendall and Spearman.

   The end goal would still be to seek a transformation to engineer a new feature that will increase these measures of correlation and the sort by the newly engineered variables.

2. I like to explore the possibility and investigate if using perhaps a nonlinear transformation to increase the relationship between variables for example using some form of a non linear PCA.

3. I like to explore the possibility of using MDS (Multidimensional scaling as an additional alternative method whenever PCA is not applicable. This would be the case when all the feature variables are not continuous. This would be the case when all the independent variables are categorical, for example, all nominal variables

Appendix A: Efficiency of Systematic Sampling Proofs

This Appendix provides the Proof that (i) $S^2_{wsy} > S^2$ is a necessary and sufficient condition for systematic sampling to be more efficient than simple random sampling and (ii) that systematic sampling will have a larger variance than simple random sampling if $\rho_w > 0$. The variance of simple random sample in 1.6 is:

$$\text{var}(\bar{y}_{srs}) = \left(\frac{N-n}{N}\right)\left(\frac{S^2}{n}\right) \tag{A.1}$$

The variance of systematic sample in 1.1 is:

$$\text{var}(\bar{y}_{sys}) = \left(\frac{N-1}{N}\right)S^2 - \frac{k(n-1)}{N}S^2_{wsy} \tag{A.2}$$

Suppose that the following inequality is true

$$S^2_{wsy} > S^2 \tag{A.3}$$

Then the inequality A.3 is true if and only if the following inequalities are true

$$-\frac{k(n-1)}{N}S^2_{wsy} < -\frac{k(n-1)}{N}S^2 \qquad \text{(Multiplication by a negative term)}$$

$$\left(\frac{N-1}{N}\right)S^2 - \frac{k(n-1)}{N}S^2_{wsy} < \left(\frac{N-1}{N}\right)S^2 - \frac{k(n-1)}{N}S^2 \qquad \text{(Adding a constant to each side)}$$

$$\text{var}(\bar{y}_{sys}) < \frac{n(N-1-kn+k)}{Nn}S^2 \qquad \text{(Left side of A.2)}$$

$$\text{var}(\bar{y}_{sys}) < \left(\frac{n(k-1)}{N}\right)\left(\frac{S^2}{n}\right) \qquad \text{(Because nk=N)}$$

$$\text{var}(\bar{y}_{sys}) < \left(\frac{N-n)}{N}\right)\left(\frac{S^2}{n}\right) \qquad \text{(Also because nk=N)}$$

$$\text{var}(\bar{y}_{sys}) < \text{var}(\bar{y}_{srs}) \qquad \text{(Left side of A.1)}$$

A second expression for the variance of systematic sample as derived by Cochran [10] in 1.4 is

$$\text{var}(\bar{y}_{sys}) = \left(\frac{S^2}{n}\right)\left(\frac{N-1}{N}\right)[1+(n-1)\rho_w]. \qquad \text{(A.4)}$$

Subtracting 1.6 from 1.4, denotes that by Δ, analyzing that expression as a function of ρ_w (i.e., $\Delta = \Delta(\rho_w)$)leads to

$$
\begin{aligned}
\Delta(\rho_w) &= \mathrm{var}(\bar{y}_{sys}) - \mathrm{var}(\bar{y}_{srs}) \\
&= \left(\frac{S^2}{n}\right)\left(\frac{N-1}{N}\right)[1+(n-1)\rho_w] - \left(\frac{N-n}{N}\right)\left(\frac{S^2}{n}\right) \\
&= \left(\frac{S^2}{Nn}\right)\Big[(N-1)(1+(n-1)\rho_w) - (N-n)\Big] \\
&= \left(\frac{S^2}{Nn}\right)\Big[n-1+(n-1)(N-1)\rho_w\Big] \\
&= \left(\frac{(n-1)S^2}{Nn}\right)\Big[(N-1)\rho_w + 1\Big].
\end{aligned}
$$

From the equation above it is clear that because $\frac{(n-1)S^2}{Nn} > 0$ and $N-1>0$, then $\rho_w > 0$ implies that $\Delta(\rho_w) > 0$.

Bibliography

Bibliography

[1] Hirotugu Akaike. A New Look at the Statistical Model Identification. *IEEE Transactions on Automatic Control*, 19(6):716–723, 1974.

[2] Edgar Anderson. The irises of the Gaspe Peninsula. *Bull. Am. Iris Soc.*, 59:2–5, 1935.

[3] Sylvain Arlot and Alain Celisse. A Survey of Cross-Validation Procedures for Model Selection. *Statistics Surveys*, 4:40–79, 2010.

[4] Raghunath Arnab. *Survey Sampling Theory and Applications*. Academic Press, 2017.

[5] DR Bellhouse. Systematic Sampling Methods. *Wiley StatsRef: Statistics Reference Online*, 2014.

[6] Peter Bloomfield and William L Steiger. *Least Absolute Deviation Regression: Theory, Applications, and Algorithms*, pages 299–302. Springer, New York, NY, 2008.

[7] Leo Breiman, Jerome Friedman, Charles J Stone, and Richard A Olshen. *Classification and Regression Trees*. CRC press, 1984.

[8] Kenneth P Burnham and David R Anderson. Multimodel Inference: Understanding AIC and BIC in Model Selection. *Sociological Methods and Research*, 33(2):261–304, 2004.

[9] George Casella and Roger L Berger. *Statistical Inference*, volume 2. Duxbury Pacific Grove, CA, 2002.

[10] William G Cochran. *Sampling Techniques*. John Wiley and Sons, New York, USA, 2007.

[11] Ronald A Fisher. The Use of Multiple Measurements in Taxonomic Problems. *Annals of eugenics*, 7(2):179–188, 1936.

[12] James E Gentle. *Elements of Computational Statistics*. Springer, New York, USA, 2011.

[13] Daniel Granato and Gaston Ares. *Mathematical and Statistical Methods in Food Science and Technology*. John Wiley and Sons, 2014.

[14] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Science and Business Media, 2017.

[15] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An Introduction to Statistical Learning*, volume 112. Springer, 2013.

[16] Richard Arnold Johnson and Wichern Dean W. *Applied Multivariate Statistical Analysis*, volume 6. Prentice Hall Upper Saddle River, NJ, 2014.

[17] JT Kent, JM Bibby, and KV Mardia. Multivariate Analysis (probability and mathematical statistics), 2006.

[18] Deba B Lahiri. A Method of Sample Selection Providing Unbiased Ratio Estimates. *Bulletin of the International Statistical Institute*, 33(2):133–140, 1951.

[19] Blake LeBaron and Andreas S Weigend. A Bootstrap Evaluation of the Effect of Data Splitting on Financial Time Series. *IEEE Transactions on Neural Networks*, 9(1):213–220, 1998.

[20] Sharon L Lohr. *Sampling: Design and Analysis*. Chapman and Hall/CRC Texts in Statistical Science, 2019.

[21] William G Madow and Lillian H Madow. On the Theory of Systematic Sampling, i. *The Annals of Mathematical Statistics*, 15(1):1–24, 1944.

[22] Colin L Mallows. Some Comments on Cp. *Technometrics*, 42(1):87–94, 2000.

[23] Robert J May, Holger R Maier, and Graeme C Dandy. Data Splitting for Artificial Neural Networks using SOM-based Stratified Sampling. *Neural Networks*, 23(2):283–294, 2010.

[24] Thomas M. Mitchell. *Machine Learning*. McGraw-Hill Science/Engineering/Math, New York, 1997.

[25] Karl Pearson. LIII. on Lines and Planes of Closest Fit to Systems of Points in Space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11):559–572, 1901.

[26] Richard R Picard and Kenneth N Berk. Data Splitting. *The American Statistician*, 44(2):140–147, 1990.

[27] Tomasz Puzyn, Aleksandra Mostrag-Szlichtyng, Agnieszka Gajewicz, Michał Skrzyński, and Andrew P Worth. Investigating the Influence of Data Splitting on the Predictive Ability of QSAR/QSPR Models. *Structural Chemistry*, 22(4):795–804, 2011.

[28] Maurice H Quenouille. Notes on Bias in Estimation. *Biometrika*, 43(3/4):353–360, 1956.

[29] Maurice H Quenouille et al. Problems in Plane Sampling. *The Annals of Mathematical Statistics*, 20(3):355–375, 1949.

[30] Z Reitermanovà. Data splitting. In *Proceedings of the 19th International Week of Doctoral Students Conference on Physics*, volume 10, pages 31–36, Prague, Czech Republic, June 2010.

[31] Paul R Rosenbaum and Donald B Rubin. The Central Role of the Propensity Score in Observational Studies for Causal Effects. *Biometrika*, 70(1):41–55, 1983.

[32] Carl-Erik Särndal, Bengt Swensson, and Jan Wretman. *Model Assisted Survey Sampling*. Springer Science & Business Media, 2003.

[33] SAS Institute Inc . *SAS/STAT 14.1 User's Guide*. SAS, Cary, NC: SAS Institute Inc., 2015.

[34] Gideon Schwarz. Estimating the Dimension of a Model. *The Annals of Statistics*, 6(2):461–464, 1978.

[35] Ronald D Snee. Validation of Regression Models: Methods and Examples. *Technometrics*, 19(4):415–428, 1977.

[36] William B Stubblefield, H Keipp Talbot, Leora Feldstein, Mark W Tenforde, Mohammed Ata Ur Rasheed, Lisa Mills, Sandra N Lester, Brandi Freeman, Natalie J Thornburg, Ian D Jones, et al. Seroprevalence of SARS-CoV-2 among frontline healthcare personnel during the first month of caring for COVID-19 patients—nashville, tennessee. *Clinical Infectious Diseases*, 2020.

[37] Pang-Ning Tan, Michael Steinbach, and Vipin Kumar. *Introduction to Data Mining*. Pearson Education India, 2016.

[38] John Tukey. Bias and Confidence in not Quite Large Samples. *Annals of Mathematical Statistics*, 29:614, 1958.

[39] Anthony G Turner. Sampling Frames and Master Samples. *United Nations Secretariat Statistics Division*, pages 1–26, 2003.

[40] Vladimir Vapnik. *The Nature of Statistical Learning Theory*. Springer science & business media, 2013.

[41] Edward J Wegman. Hyperdimensional Data Analysis Using Parallel Coordinates. *Journal of the American Statistical Association*, 85(411):664–675, 1990.

[42] Robert Jesse Willhide. Annual Survey of Public Employment and Payroll Summary report: 2013. Economy-Wide Statistics Briefs: Public Sector. *US Census Bureau*, 2014.

[43] Mi-Ja Woo, Jerome P Reiter, Anna Oganian, and Alan F Karr. Global Measures of Data Utility for Microdata Masked for Disclosure limitation. *Journal of Privacy and Confidentiality*, 1(1), 2009.

[44] Wenyan Wu, Robert J May, Holger R Maier, and Graeme C Dandy. A Benchmarking Approach for Comparing Data Splitting Methods for Modeling Water Resources Parameters using Artificial Neural Networks. *Water Resources Research*, 49(11):7598–7614, 2013.

[45] G. Udny Yule and M. G. Kendall. An Introduction to the Theory of Statistics. *An Introduction to the Theory of Statistics.*, (14th ed), 1950.

Curriculum Vitae

In 1996, I obtained a B.Sc. in Probability and Statistics from the Department of Mathematics, USTHB, Algiers, Algeria.
In 2009, I completed an M.S. in Statistical Science from George Mason University.
In the Fall semester of 2013, I enrolled in the Ph.D. program in Computational Science and Informatics (CSI) at George Mason University.
I held multiple teaching positions in my earlier career, including a graduate teaching assistant for George Mason University's statistics department. I started my carrer as a Statistician in the private sector before joining the U.S. Census Bureau in 2008 to work on many projects, including production and research related to record linkage of federal administrative records such as the IRS, Social Security Administration, and the Medicare enrollment database. From 2012 to 2015, I served on the demographic statistical division to support the sampling and design of educational and many demographic surveys, including the AHS and CPS. between 2015 and 2019, I served as a Mathematical Statistician technical lead for the sampling design and estimation of four Public Sector Government economic surveys, where I conducted production and research and published three JSM papers

1. Impact of Certified Mail on Nonresponse Rates
2. The Performance of the Empirical Best Linear Unbiased Predictor in Annual Survey of Local Government Finances
3. Outliers Research in the Annual Survey of Annual Survey of Local Government Finances

Since 2013, I have consulted for the International programs (I.P.) area of the U.S. Census Bureau to provide numerous technical assistance in survey sampling and related software programming (SPSS, Stata, and R) during multiple workshops in different countries: Jordan, Nepal, and Mozambique through collaborative efforts with the USAID. From late 2019 to the present, I joined the U.S. Census Bureau's I.P. area. I worked as detailed for the U.S. Global AIDS Coordinator (S/GAC) Office to support the U.S. President's Emergency Plan For AIDS Relief (PEPFAR) to provide technical guidance in HIV modeling. Since then, I have had the opportunity to work on multiple ad-hoc production and research on issues related to HIV and COVID-19. I facilitated a virtual UNAIDS workshop for Francophone countries to assist with HIV 2020 estimation. I presented at the 23rd International AIDS virtual conference, and I also provided technical assistance to the National Statistical Office in Malawi to develop a Master Sampling frame.