CRISPR : Prediction of Off-Target Levels by Using Latent Class Analysis and Bayesian Latent Class Analysis

Ali Mertcan Köse[1,2]

[1] Department of Computer Programming, Istanbul Ticaret University
[2] Department of Statistics, Mimar Sinan Fine Arts University

Introduction

Recently, CRISPR (Clustered Regularly Interspaced Short Palindromic Repeats) is one of the most popular applications in the field of biology. CRISPR is a system that enhances the immune system against viral diseases and infections. In the target sequence (DNA), which is separated into specific clusters, off-target and on-target positions are determined by the matching between DNA and guide RNA. The spread and occurrence of viral diseases and infections are prevented by modifying the off-target bases (nucleotides). Off-target and on-target positions are evaluated based only on two categories with CFD/MIT scores. Instead of two classes, this study focuses on off-target positions.

Objectives

1. to identify the levels of off-target by using the latent class analysis.
2. According to the number of classes (Specification of Priors), Bayesian Latent Class Analysis can be a more robust approach to predict the off-target levels.
3. to specify the off-target position in the genome for CRISPR.

Methods

The main objective of latent class analysis (LCA) is to categorize individuals into groups within different populations exposed to a specific disease, based on their characteristics in public health and medicine. LCA is a method used to analyze the relationship between categorical variables measured at nominal and ordinal variables. In LCA, the maximum likelihood (ML) estimation method achieves the best solution by using the Expectation-Maximization (E-M) algorithm. Conversely, as an alternative to the ML method, the Bayesian estimation method is also utilized to estimate unknown parameters in latent class models through the Markov Chain Monte Carlo (MCMC) method.

The dataset of the study was obtained from the CRISPOR database (http://crispor.tefor.net/). The analysis was conducted using a sample size of 5132, which included target sequences and gRNAs matching positions based on 23 bases. First of all, the matching of DNA and gRNA was encoded using binary (0/1) coding and multi (4x4 = 16) coding, considering that the combination of two bases in the same positions constitutes a variable. Subsequently, latent class analysis was applied separately for each encoding system. Bayesian latent class analysis was conducted, with the number of classes determined by the latent class analysis application.

Results

In binary and multi-category datasets, Lo-Mendel-Rubin (LMR) test results were found statistically significant for each latent class ($p < 0.001$). According to the Bayesian Information Criteria and the Consistent Akaike Information Criteria, the latent models were well-fitted to four classes for binary category datasets and three classes for multi-category datasets. Because the entropy values were higher than 0.60, the three and four latent class models were distinctly and deterministically classified. After determining class numbers, latent class models were estimated using the Bayesian approach with binary and multi-category datasets.

Table 1: Latent class analysis results for observed variables with binary-categories.

	Log-						LMR	
Class	Likelihood	AIC	BIC	SSABIC	CAIC	AWE	Test	Entropy
2	-50887.6	101869.2	102176.8	102027.4	102223.8	102719.3	<0.001	1
3	-50737.1	101616.1	102080.7	101855.1	102151.7	102900.3	<0.001	0.998
4	-50606.2	101402.4	102024	101722.1	102119	103120.6	<0.001	0.998
5	-50520.1	101278.1	102056.8	101678.6	102175.8	103430.4	<0.001	0.969
6	-50452.2	101190.4	102126.1	101671.7	102269.1	103776.8	<0.7167	0.787

AIC: Akaike Information Criterion, BIC: Bayesian Information Criterion,SSABIC:Sample Size Adjusted Bayesian Information Criterion, CAIC: Consistent Akaike Information Criterion, AWE:Approximate Weight of evidence Criterion, LMR Test: Lo-Mendel-Rubin Test

AIC: Akaike Information Criterion, BIC: Bayesian Information Criterion,SSABIC:Sample Size Adjusted Bayesian Information Criterion, CAIC: Consistent Akaike Information Criterion, AWE:Approximate Weight of evidence Criterion, LMR Test: Lo-Mendel-Rubin Test

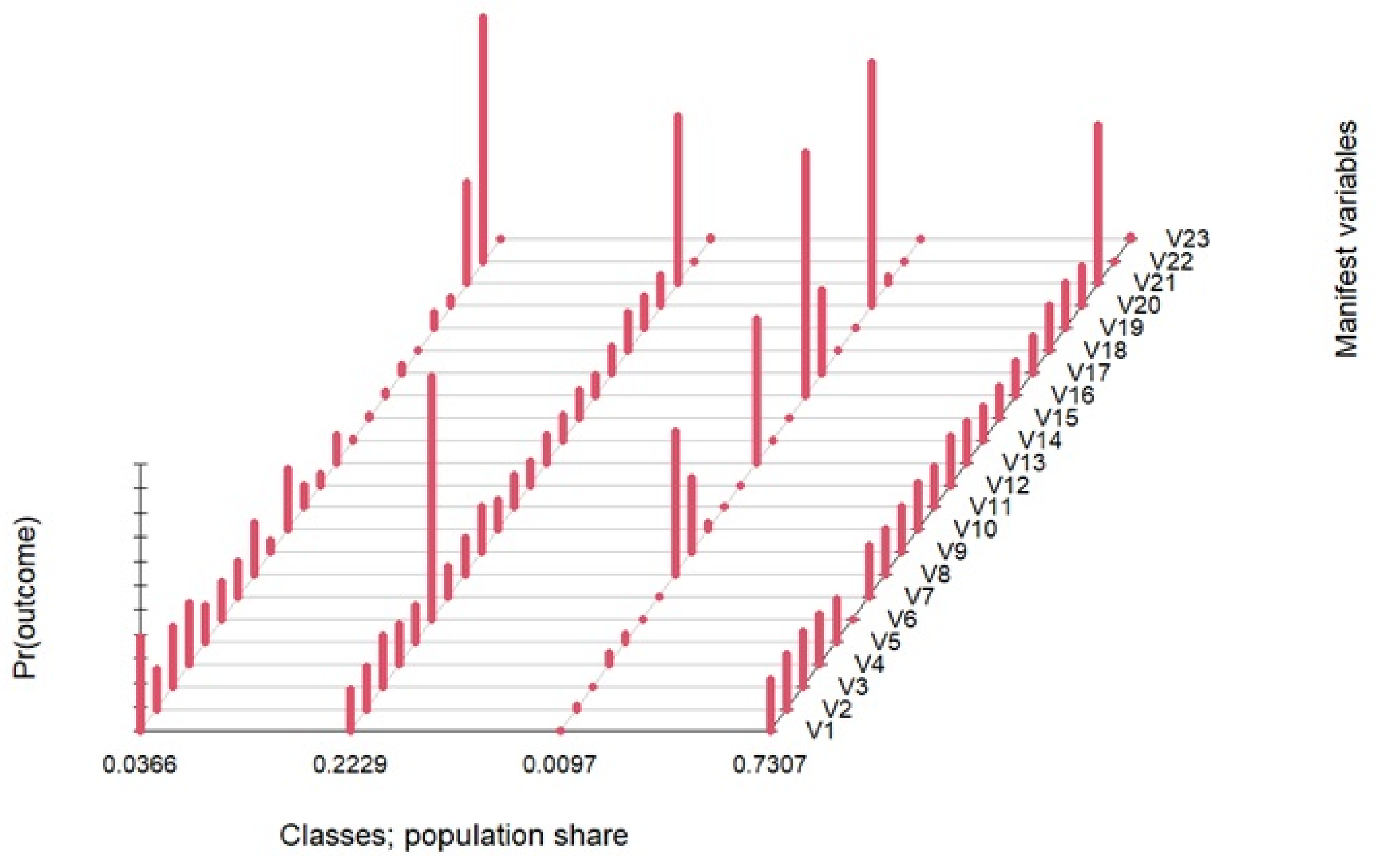

Figure 1: Latent Class Probabilities(Binary-Encoding)

Table 2: Latent class analysis results for observed variables with Multi-categories.

	Log-						LMR	
Class	Likelihood	AIC	BIC	SSABIC	CAIC	AWE	Test	Entropy
2	-102166.52	205323	208561.9	206989	209056.9	214275.8	<0.001	0.719
3	-101010.36	203506.7	208368.4	206007.4	209111.4	216945	<0.001	0.791
4	-100042.46	202066.9	208551.3	205402.2	209542.3	219990.6	<0.001	0.819
5	-99067.84	200613.7	208720.8	204783.6	209959.8	223022.9	<0.001	0.861
6	-98347.5	199669	209398.8	204673.6	210885.8	226563.6	<0.001	0.87

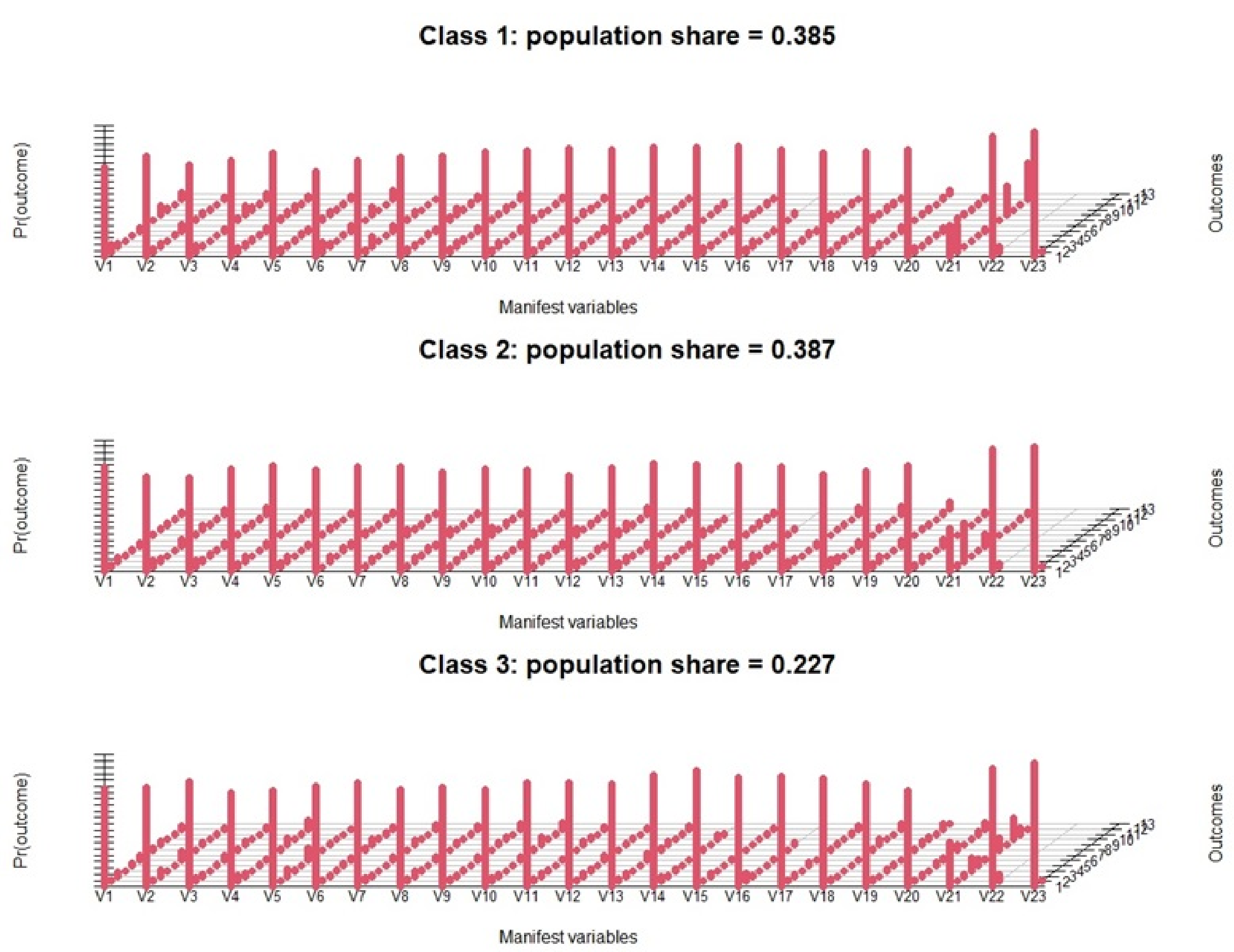

Figure 2: Latent Class Probabilities(Multi-Encoding)

Table 3: Bayesian Latent class analysis results for observed variables with Binary-categories.

Class	Median	Mean	SD	SSeff
1	0.741	0.727	0.0071	7614
2	0.044	0.037	0.0033	5133
3	0.236	0.223	0.0060	8125
4	0.015	0.011	0.0020	3699

Table 4: Bayesian Latent class analysis results for observed variables with Multi-categories.

Class	Median	Mean	SD	SSeff
1	0.346	0.346	0.0066	37485
2	0.046	0.046	0.0029	40000
3	0.607	0.607	0.0068	40000

Conclusion

Making too many changes to the bases in the DNA sequence can lead to a higher chance of developing another disease or weakening the immune system. This study systematically tested the impact of mismatches between the target sequence and guide RNA, laying the groundwork for minimizing base changes at each position. In the next stage, machine learning and deep learning algorithms may be used to predict the locations of multi-categorical off-target levels.In the next stage, machine learning and deep learning algorithms may be used to predict the locations of multi-categorical off-target levels.

Acknowledgement

✉ alimertcankose@gmail. com

http://alimertcankose.netlify.com/

https://orcid.org/0000-0002-5464-9441

https://github.com/mertcank1

References