Hitting the Wall: Mixture Models of Long Distance Running Strategies

Joseph D. Pane

May 1, 2015

Abstract

The International Association of Ultrarunners 24 Hour World Championships holds a 24 hour race during which each entrant tries to run as many laps as they can. There are several different strategies to running a race, and these may be characteristic to the runner. Some runners may race at a consistent pace, dropout, take breaks and/or fluctuate their pace. Records containing how many laps a particular entrant runs over each hour in the race allow us the opportunity to model the different types of strategies and how successful they are. Building on previous work by White and Murphy (2013), we use mixture models, latent class analysis, and model based clustering for mixed data. We alter the model based clustering for mixed data framework, *clustMD*, that extends the estimation capability of the method in scenarios it was not able to prior. We use this method to determine running strategies in 24 hour races. In future work we plan to develop a longitudinal model-based clustering approach using Poisson processes. This approach will have a large scale impact on other applications outside the field of sports. The same type of model-based clustering can be used in various medical research questions, including clinical depression scores of patients over a period of time.

Acknowledgements

I would like to express my sincere gratitude to my advisor, Rebecca Nugent, for her continuous support throughout the completion of my thesis. She is a mentor who patiently guided me with her expertise. I hope that one day people will look up to me the way in which people look up to her.

I would like to thank Dirk Strumane and Hilary Walker for giving us access to the International Association of Ultrarunners (IAU) 2012 World Championship data.

I also would like to thank my parents and brother because they have been there for me every step of the way. I could not have completed this thesis without their love, support, and guidance.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Our research focuses on methodology for determining different strategies in running long distance ultra-races. For example, the International Association of Ultrarunners (IAU) 24 Hour World Championships were recently held in Katowice, Poland on September 8^{th} and 9^{th} in 2012. [1]They hold this prestigious event with over 200 runners representing their respective countries. There are several possible race strategies. For instance, runner A may start off very fast in the beginning of the race and then decrease pace until picking up the pace at the very end. However, runner B might run very consistently the entire time. In addition, runners might fluctuate their speed due to injury or rest stops. Runner C might completely drop out of the race. Given count data on the number of laps a runner completes each hour we define continuous, ordinal, and nominal variables to characterize the different count trajectories and thereby strategies we might see.

After defining running terminology, some common racing strategies and variables, we explore different clustering methodologies. To analyze the continuous variables, we use a Gaussian mixture model. We then cluster our categorical data (nominal, binary, and ordinal variables) through the use of latent class analysis. Although more difficult to interpret, we are still

[1]This information was gathered from Arthur White and Dr. Thomas Brendan Murphys', "Exponential Family Mixed Membership Models for Soft Clustering of Multivariate Data

able to use these results to plot the trajectories that were identified for each class, similar to the continuous setting.

We then turn toward combining these approaches into one mixture model estimation procedure for data of mixed type. We started with the framework developed by McParland and Gormley (2015), *clustMD*, for use with continuous, ordinal, and nominal variables. This approach uses Monte Carlo sample approximations when faced with intractable parameter estimation for the nominal variables in particular. We found that this approach had limitations in some scenarios, for example, sparse or missing categories within a cluster, that did not allow some models to be fit. We extended the estimation capability of their framework by adding some minor assumptions with respect to the Monte Carlo samples. The extended approach now allows for a higher number of feasible models and performs well in our analyses.

We first run simulations to help assess performance. We conducted one simulation based off distance runner strategies and another on the results we obtained through mixture models with continuous features on the 24 Hour World Championship data. We then cluster the original 24 Hour World Championship data to determine the final different race strategies.

Chapter 2

Using Different Strategies to Optimize Race Performance

A person may run to get into better shape, but running is also a competitive sport. There are different types of races, which mainly vary by their distance, which ranges from as short as 60 meter races to as long as 26.2 marathons and 24 hour long races. Factors such as weather, competition, age, injury history, training, and time of the year impact how a runner will perform on any given race day. While the goal of every race is to finish the race as quickly as possible or cover as much distance as possible, each runner does not run the same strategy in every race. Over time the runner may find what works best for them and can approach a race in a manner that gives the best chance of running their optimal race. Determining this optimal race strategy is a necessary part of race preparation for any competitive runner.

A runner's race strategy depends on a combination of the type of runners and the type of race. The length of the race impacts the strategy but ultimately how a racer trains, what the racer is used to doing, the body build of the racer, the natural talent of the racer, and how the runner feels that day all impact the choice of strategy. Runners may choose a strategy

before the race based on past performance, but they may also need to change that strategy during the middle of the race given several different factors. A change due to weather or how fast or slow the other runners are going are two instances where changing racing strategies is done on purpose. Sometimes a runner changes their race strategy based off how he/she feels during the race. These sudden changes can be unintentional and may happen naturally.

2.1 Running Terminology and Common Strategies

For ease of explanation, we first define a few common running terms/phrases.

- **Pace:** A rate of the number of minutes it takes to complete a mile.
- **Steady Pace:** Consistently running not too fast and not too slow relative to the runners ability.
- **Holding on:** Still running at a consistent pace but almost at the point where it is not physically possible for that runner to run that fast.
- **Dying:** Having a very slow pace, and getting slower, after running at a faster pace for a period of time.
- **Dropping Out:** Stop running completely.
- **Picking Up The Pace:** Increasing the pace.
- **Hitting the Wall:** The point in a race when you cannot keep up your pace anymore and start to slow down considerably. This is what we call the event after you cannot "hold on" for any longer.

A few common strategies or outcomes to running races are:

- Starting off the race at a steady pace and maintaining that same pace for the length of the race.
- Starting off the race fast, relative to the runner's ability, and "holding on" at the end of the race.
- Starting off the race fast, relative to the runner's ability, and "dying" in the middle and end of the race.
- Starting off the race fast, relative to the runner's ability, and dropping out of the race.
- Starting off the race slow, relative to the runner's ability, and picking up the pace soon after.
- Starting off the race slow, relative to the runner's ability, holding a consistent pace, and picking up to a fast pace by the end of the race.
- Starting off the race at a steady pace, slowing down, and picking it up to a fast pace at the end of the race.

Using information about paces during a race, we are interested in the different types of strategies runners use. Additional analyses could then perhaps tie the type of racing strategy to final performance. Within a race, we are interested in describing and characterizing the number of different strategies with specific interest in finding strategies that might be unexpected or unplanned.

2.2 Ultra-running/Ultra-marathons

An ultra-marathon is any race longer than a marathon (26.2 miles or 42.195 kilometers). There are several kinds of ultra-marathons (50km, 100km 250km, 6 hour, 12 hour, 48 hour),

but our application of interest is the 24 hour long race. A 24 hour ultra-marathon does not have a specified distance. Instead, the winner is determined by who can cover the longest distance.

A 24 hour long race requires different types of strategies compared to a race over a specified distance. The extreme distance and time involved requires careful planning and conservation of effort and energy. A consistent racing strategy as well as a slow, steady pace changing to a fast, end pace are two types of strategies that are common in the 24 hour long race (versus running as fast as possible and trying to hold on). However many racers are not able to run consistently the whole race because it is simply very hard to do. Unlike a shorter race like the mile or 5k, it is more acceptable to stop running for a brief period in a 24 hour long race. A runner that is having a "bad day" may be forced to drastically change strategies in order to complete the race. Given the length and difficulty of this race, we expect to see clear fluctuation in running strategies, based on the ability of the runner (among other factors).

2.3 International Association of Ultrarunners (IAU) 24 Hour World Championship

Our application dataset is from the 2012 IAU 24 Hour World Championship in Katowice, Poland. The International Association of Ultrarunners (IAU) is the organization that holds the 24 Hour World Championship. This race is held annually and it is comprised of the best ultra marathon runners in the world. There were 34 countries who participated in this race. See Table 2.1 for the countries who participated. The 2012 race started at noon on September 8^{th}, 2012 and ended at noon on September 9^{th}, 2012. There were 260 runners who were scheduled to participate in the event and 248 who actually started the race (i.e. 12 runners were DNS (did not start)).

Table 2.1: There were 34 countries who participated in the 2012 IAU 24 Hour World Championship in Katowice, Poland.

Country	#	Country	#	Country	#	Country	#	Country	#
ALG	1	ARG	1	AUS	7	AUT	5	BEL	4
BLR	5	CAN	9	CZE	7	DEN	13	ESP	13
EST	12	FIN	13	FRA	12	GBR	9	GER	11
GRE	2	HUN	12	IRL	4	ISL	1	ITA	17
JPN	9	LAT	7	LTU	5	MKD	2	NED	3
NOR	7	NZL	7	POL	13	RUS	13	SRB	3
SVK	4	SWE	4	UKR	9	USA	16		

2.3.1 Cumulative Laps

The course that the race is being organized on is a 1554m looped route in Park Slaski in Chorzow very close to Katowice. The course the runners ran on was a 1554 meter loop located in Park Slaski in a town very close to Katowice, Chorzow. For reference, a mile is 1600 meters. For every runner, the cumulative number of laps the runner ran is recorded in the data file. See Table 2.2 for an example of how the data is recorded. For instance, if Mike Morton, the eventual champion, ran nine laps in the first hour and seven laps in the second hour, the data would show a nine under H1 (Hour 1) and 16 under H2 (Hour 2). Along with the cumulative laps, the data file also contains the name of the runners, their age, the country they are from, and their gender. The original data file showed information on every runner who was registered for the race, regardless of whether or not they started the race. Notice in Table 2.2, the last row has an index of 264. The numbers 89-91 and 217 were not used as indices (no explanation was given).

2.3.2 Running Trajectories

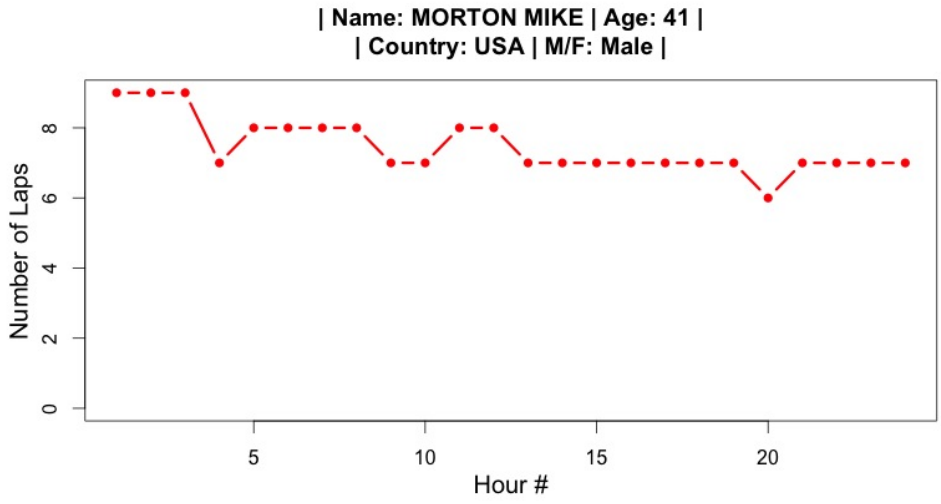
We could also view each runner as a trajectory of their lap counts for each hour. We might expect that characteristics of these trajectories would indicate the type of strategy

Table 2.2: Dataset Sample: The data was recorded from the race in the following format. Notice that Olsen has all 0s filled in his hour totals (DNS). We also can tell that Scholz beat Harvey-Jamieson because she completed more laps by Hour 24.

#	**Name**	**Age**	**Country**	**Gender**	**H1**	**H2**	$\cdots$	**H24**
1	SCHOLZ SHARON	36	AUS	Female	6	13	$\cdots$	37
2	HARVEY-JAMIESON SUSANNAH	26	AUS	Female	6	13	$\cdots$	50
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
264	OLSEN JON	38	USA	Male	0	0	$\cdots$	0

ultimately used. For example, the winner of this race, Mike Morton, had a very consistent race (Figure 2.1).

Figure 2.1: Mike Morton's Trajectory for the 2012 IAU 24 Hour World Championships in Katowice, Poland.

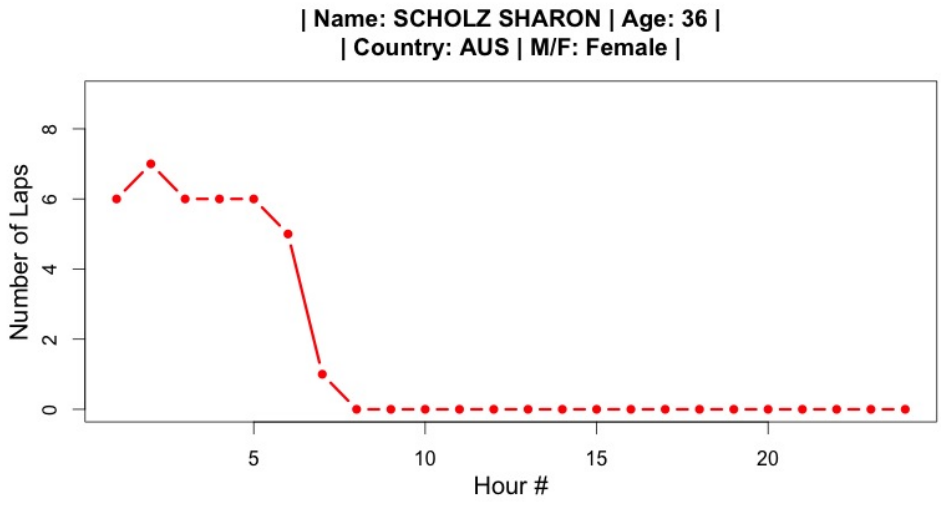


The trajectory shows that he ran the same number of laps plus or minus one lap from the fourth hour of the race until the end of the race. The fluctuation is most likely due to not being finished with a complete lap at the time the hour is up (e.g. running 7.5 laps two hours in a row would be recorded as seven for the first hour and eight for the second hour). This type of consistent racing strategy, like Mr. Morton's, comprises a large majority of the

top finishers in the World Championship.

Other trajectories that are characteristic to the racing strategies that we talked about in Section 2 can be seen in our runners. For instance, Figure 2.2 shows an example of a runner who dropped out of the race. We define a dropout as someone who started the race but at some point the hourly number of laps they ran was recorded as a zero for the remainder of the race. We also have runners who do not run any laps for one or more hours near the end of the race but then they begin running/walking again during the last hour. We decide to classify these people as dropouts as well. We will talk more about the definition of "dropout" in Section 3.4.

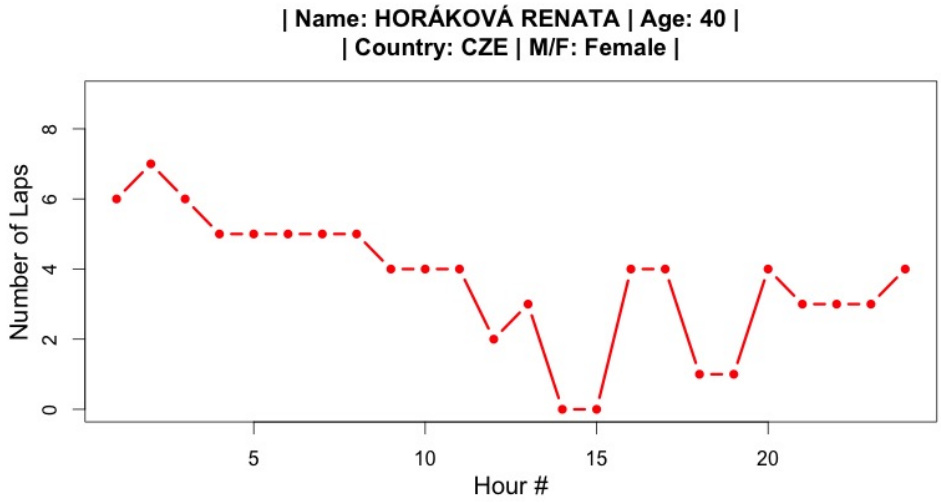
Figure 2.2: An example of a runner dropping out during the race.



During this type of race, it is common to stop running and then to return and continue running at a similar pace. This behavior is most likely characteristic of runners who are taking a break or an unexpected event occurred (e.g. injury or thunderstorm). Figure 2.3

gives an example of this kind of trajectory. Note there were no major unexpected events (e.g. bad thunderstorm or snowstorm) that happened in this race to our knowledge.

Figure 2.3: An example of a runner stopping completely for at least an hour(taking a break) and then running again after their break.

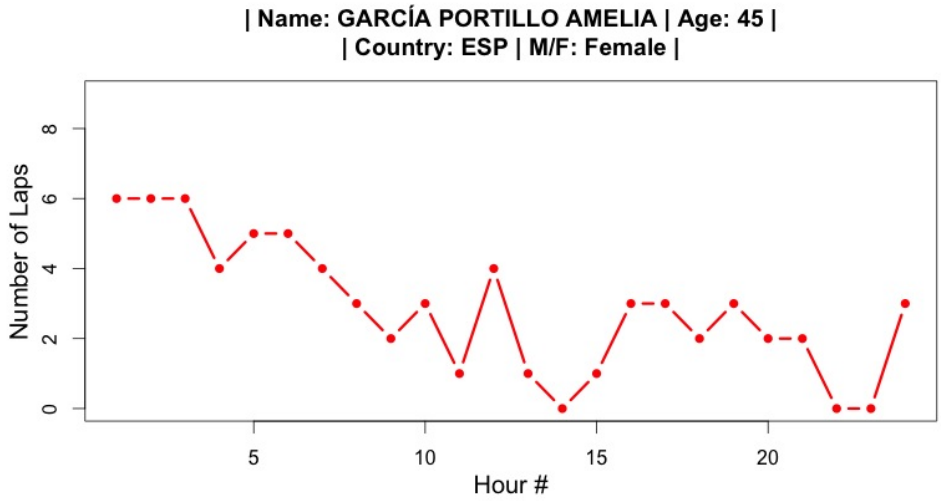


Another common characteristic for some runners is to finish a race strong. Figure 2.4 gives an example of what a trajectory might look like for someone who may have hit the wall but since it was the end of the race, they really pushed themselves. Runners have this type of characteristic because they may have saved too much energy for the end of their race or since they are almost finished, they put extra effort in at the end.

For these running trajectories, we brainstormed summary characteristics that might indicate a particular strategy. For example, a runner who drops out most likely has a big decrease in their race pace at some point followed by a large string of zeroes. A runner who is consistent may have the same number of laps ran each hour for an extended period of time. Since this race was run in the night, a runner's pace during the day may have been different than the night. The most laps gained from one hour to the next and the most laps dropped in an

hour from one to the next can also give indication of strategy. In Sections 3.4 and 4.2 we will look at all of the variables we defined to help explain these racing strategies.

Figure 2.4: An example of a runner finishing the race strong during the last hour of the race.



It is common for runners to have similar racing strategies. With the use of continuous and categorical variables, we can classify these runners into groups. The statistical approach of clustering will identify these strategies to describe the trajectories. Using continuous and categorical variables in combination is optimal, however first we fit the continuous data using model based clustering and the categorical data using latent class analysis. In Chapter 5 we will combine the estimations of continuous, ordinal, and nominal variables to identify racing strategies.

Chapter 3

Gaussian Mixture Models for Continuous Data

Given race information, we want to cluster runners into groups based off of their performance in an ultra-marathon race. We start with modeling continuous variables that describe our vector of lap counts using the commonly used Gaussian mixture model. After giving an overview of the mixture model and its estimation, we define and motivate the continuous variables used to describe the lap count trajectories. We then include some examples of the resulting clustered race strategies.

3.1 Mixture Model Notation

The Gaussian mixture model assumes that the population density $f(x)$ represents a weighted combination of the group densities, $f_k(x)$. We identify $f_k(x)$ as the normal distribution [6].

$$f(x) = \sum_{k=1}^{K} \pi_k \cdot f_k(x; \theta_k)$$

As shown in [6], the corresponding likelihood for a mixture model with distribution f_k (in our case Gaussian) is

$$\mathcal{L}_{MIX}(\theta_1, \cdots, \theta_K; \pi_1, \cdots, \pi_K|\mathbf{x}) = \prod_{i=1}^{n} \sum_{k=1}^{K} \pi_k f_k(\mathbf{x}_i|\theta_k),$$

The θ_k has a mean μ_k and variance matrix Σ_k that will identify the shape, volume, and orientation of the cluster for the k^{th} group/cluster. f_k can be any multivariate distribution but in our case we assume a multivariate Gaussian density, ϕ_k,

$$\phi_k(\mathbf{x}_i|\mu_k, \Sigma_k) = \frac{\exp\big(-\frac{1}{2}(\mathbf{x}_i - \mu_k)^T \Sigma_k^{-1}(\mathbf{x}_i - \mu_k)\big)}{\sqrt{\det(2\pi\Sigma_k)}},$$

where K is the number of groups or clusters. The π_k is the mixture weight or probability of the k^{th} component. In our mixture model context, each probability distribution will correspond to a race strategy cluster. The variables in $\mathbf{x}$ are assumed to be multivariate continuous. The total number of runners is n.

The resulting models are either ellipsoidal, diagonal, or spherical and correspond to different combinations of the features of the covariances, Σ_k. There are 10 different combinations of shape, volume, and orientation. Nugent and Meila outline the different combinations and share an illustration on page nine of "An Overview of Clustering Applied to Molecular Biology" (2010). We use either spherical, diagonal, or ellipsoidal covariance depending on the model. Each of the 10 models has some combination of equal or varying volume, round, equal, or varying shape, and axis parallel, equal, or varying orientation. For example, the "VVV" model is varying volume, varying shape, and varying orientation. This is assuming ellipsoidal variance. See Section 3.3 for our model selection procedure.

The Expectation-Maximization (EM) Algorithm is used to estimate our parameters. [6].

3.2 Estimation with the Expectation-Maximization (EM) Algorithm

Briefly described here, the Expectation-Maximization (EM) Algorithm is used to fit the Gaussian mixture model. See McLachlan [5] for more details. There are two iterated steps to the EM Algorithm. The first is the Estimation step (E-step) and the second is the Maximization step (M-step). In the E-step, we estimate the cluster assignments for all n runners across each cluster k. The algorithm assigns a probability, γ_{ki}, to represent the probability that each runner i belongs to each cluster k. These estimated probabilities can be written as

$$\gamma_{ki} = \frac{\pi_k f_k(x)}{f(x)} = \frac{\pi_k f_k(x)}{\sum_{k=1}^{K} f_k(x)}$$

The M-step estimates the mixture parameters, π_k, μ_k, and Σ_k. Here π_k is the mixture weight assigned to the k^{th} cluster (running strategy). μ_k is estimated by the following weighted average,

$$\hat{\mu}_k = \frac{\sum_{i=1}^{n} \gamma_{ki} \mathbf{x}_i}{\Gamma_k},$$

where Γ_k is the total number of runners that belong in a particular cluster k. The covariance matrix is estimated depending on the model we use for defining the shape, volume, and orientation. One example of the covariance matrix using the "VVV" model is

$$\Sigma_k = \frac{\sum_{i=1}^{n} \gamma_{ki} (\mathbf{x}_i - \mu_k)(\mathbf{x}_i - \mu_k)^T}{\Gamma_k}.$$

After initialization, the EM Algorithm alternates between the E-step and the M-step until it converges. Each choice of K and choice of covariance structure give a final set of parameter estimates and cluster assignments. After estimation, each runner is assigned a vector of probabilities of length K that return the probability of being in each group.

3.3 Model Selection

Since we are estimating ten different clustering models for each of one to the total K clusters, we need to choose a final model. Kass and Raftery (1995) [8] explain a common model selection process. Briefly, the number of clusters K, and the clustering model is optimally chosen by maximizing over the likelihood given the particular cluster model with the addition of a penalized term for complexity. This is called the Bayesian Information Criterion (BIC). It is given by [8],

$$\text{BIC} = -2 \times \text{logL}(\mathbf{x}|\theta) + \log(n) \times p$$

The estimation procedure returns BIC values for all of the possible models and numbers of clusters. We then determine which model returns the highest BIC and choose this model as the best fit for the data.

3.4 Continuous Variables

We now define a set of continuous variables that describe our lap count trajectories. Many of these are based on experience in competitive running and are not necessarily statistically derived. They are also variables that runners and their coaches naturally track and measure and so are hopefully more easily understood in the field. Splitting up some variables, for example, average pace into day and night pace will allow us to potentially see if runners took breaks at points in the night or simply did not run as well certain times of the day.

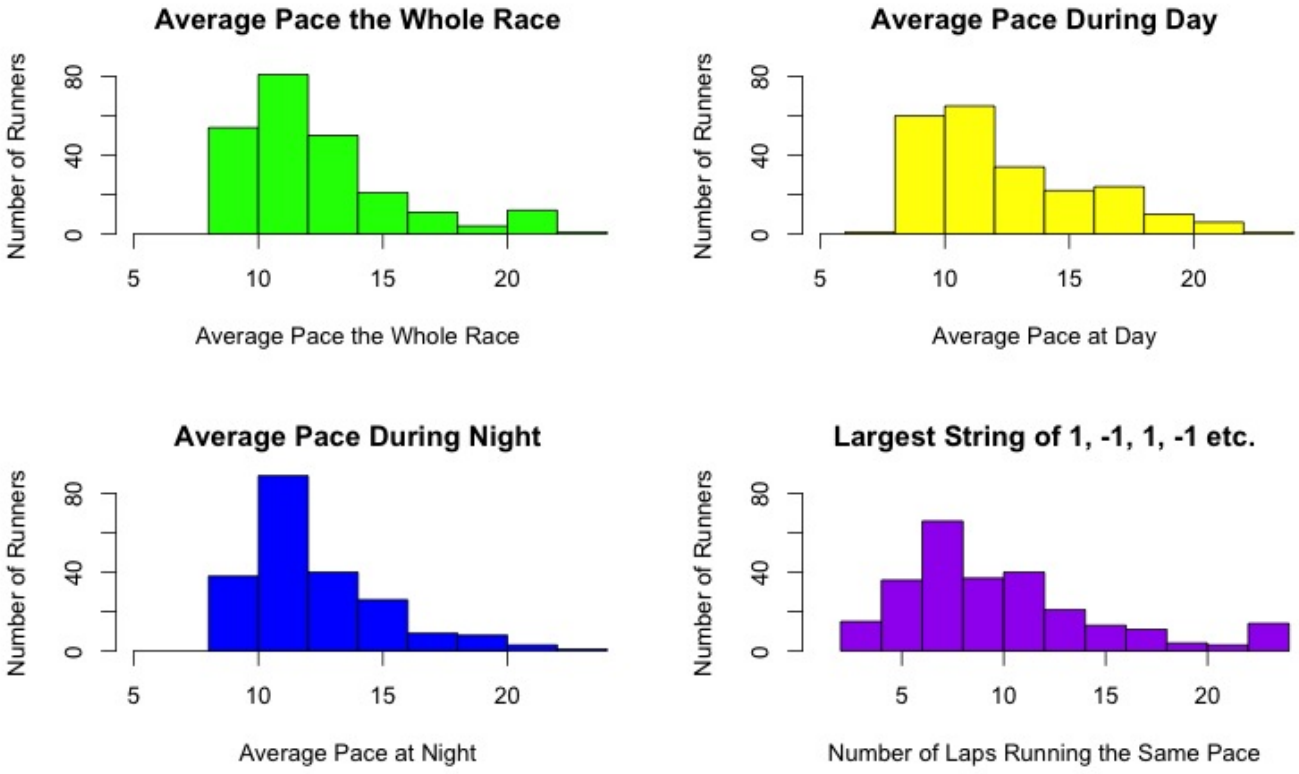
- **Average Pace For the Whole 24 Hours**

  We first define a variable that calculated the average pace for the entire 24 hours of the race. The original dataset only had the cumulative number of laps ran for each hour

over the course of the race. The average pace for the 24 hours of the race allows us to more easily compare runners. Runners, in particular, would rather track an overall pace compared to the number of laps they ran. For the purposes of quantifying pace, we calculated all pace variables with respect to per mile. Since each lap in the race was just under one mile (0.966 miles or 1.554 km), we converted the pace variables to minutes per mile. For example, the winner, Mike Morton's average pace for the length of the race was just over eight minutes (eight minutes and two seconds per mile). This was the fastest average pace. Figure 3.1 has the distribution of average pace for runners who averaged less than 23 minutes per mile. We chose 23 minutes per mile as our cutoff because Reynolds (2010), writer for the New York Times Wellness Blog, made the following conclusion from the National Walkers' Health Study, "The majority of the walkers in this group in fact required at least 20 minutes to complete a mile, and many had a pace of 25 minutes or more per mile." [10]. We decided to take 23 minutes per mile as the cutoff for a slow to medium paced walk. The distribution of average pace for the whole race was skewed right, unimodal, and the majority of the data falling between eight and 15 minutes per mile.

We also calculated the pace for selected, shorter time frames. We broke up the average pace into certain time frames as we expect paces to change for specific reasons during the 24 hours. See Figure 3.2 for an example of how the paces might be broken up into hour segments of the day and night. We eventually selected to look at the average pace during the night (6pm to 6am) and the average pace during the day (12pm to 6pm on the first day and 6am until 12pm on the second day). Since the race started at noon, the hours where the runners were running in the day were split up at the beginning and end of the race. As a result, we found the average pace of the runner at night (6pm-6am) and the average pace of the runner during the day (12pm-6pm) and

Figure 3.1: Exploratory data analysis for continuous variables. We removed the runners who did not start the race as well as runners who did not average less than 23 minutes per mile.

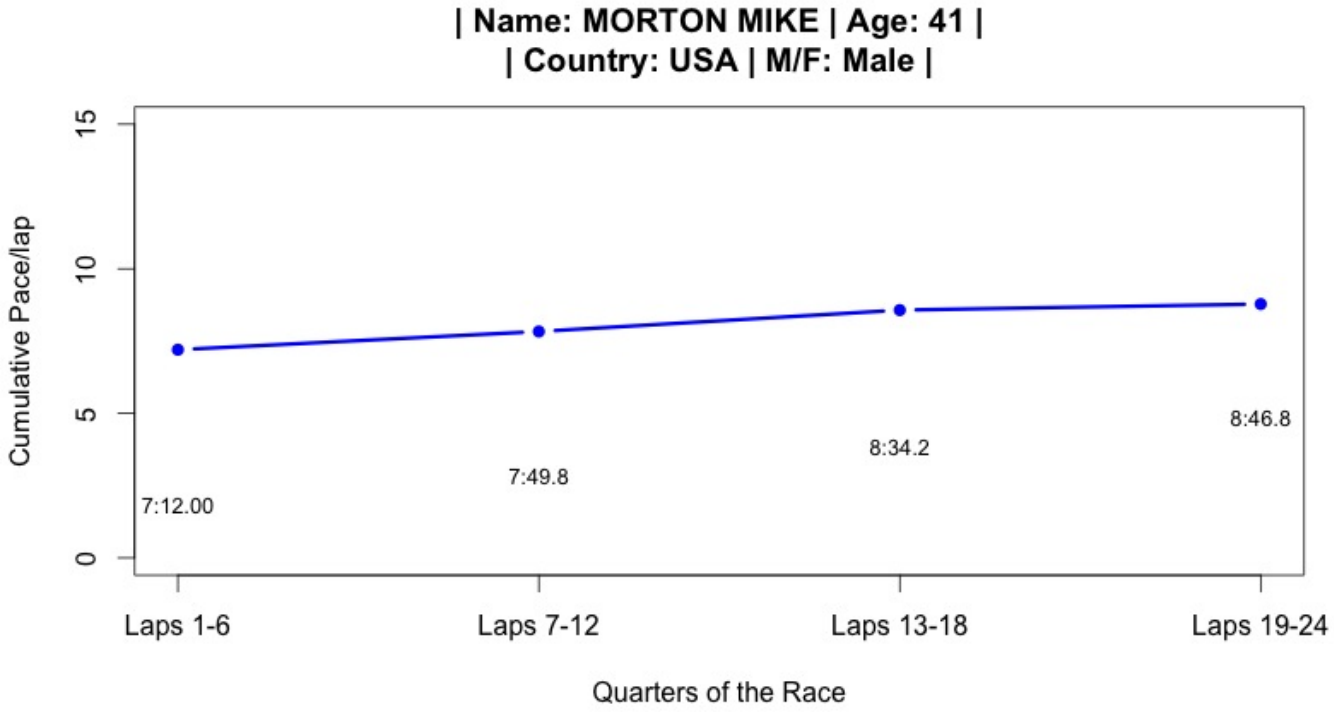


(6am-12pm). We believe breaking the pace up into day and night will identify patterns in the runners.

- **Average Pace at Night (6pm-6am)**

  Figure 3.1 shows the skewed right distribution of pace during the night for runners who ran less than 23 minutes per mile. The average pace during the night's distribution looks pretty similar to the average pace during the whole race (skewed right and unimodal), however there seems to be a higher spike of runners around 10 to 12 minutes per mile and less falling in the 8-10 and 12-14 minutes per mile range.

Figure 3.2: The Pace Variable is demonstrated in time frames. The average overall pace for the 24 hours would be the average of the four paces listed on the Figure.



- **Average Pace During Day (12pm-6pm) AND (6am-12pm)**

  Figure 3.1 shows the skewed right and unimodal distribution of the pace during the day for runners who ran less than 23 minutes per mile. The data look more uniform from paces of eight through 12.5 minutes per mile and in the range of 15 and 20 minutes per mile compared to the overall average pace and average pace during the night.

- **Average Pace for the Whole 24 Hours when hourly laps do not equal zero (Average Pace Nonzero)**

When we calculated the average pace for the whole 24 hours, we came across a potential feature of interest. Several runners did not run at all during one or more hours of the race. As a result, their overall average pace significantly increased. To address this, we also took the average pace only when hour laps did not equal zero. For instance, if a particular runner ran five laps in hour one, zero in hour two, and three in hour three, we would eliminate hour two when calculating the average pace. Therefore, the average pace when the total laps does not equal zero is the average pace when the runner has completed at least one lap per hour. This does not mean that the runner did not stop at some point during the hour; but we have no way of knowing this.

- **Average Pace for Day when hourly laps do not equal zero (Average Day Pace Nonzero)**

  Similarly, we calculated the average pace during the day for hours with non-zero hours when the hourly laps do not equal zero.

- **Average Pace for Night when hourly laps do not equal zero (Average Night Pace Nonzero)**

  We similarly calculated the average pace during the night when the hourly laps do not equal zero.

- **Longest String of Hours of Number Laps Being Same**

  We wanted to define a variable that represented runners being consistent because we suspected that the elite group of runners would run a very similar pace for the entire race. To do this, we calculated the longest string of consecutive hours when the runner ran the same number of laps.

- **Longest String of 1, -1, 1, -1 etc. (Number of Hours Running Same Pace)** One problem we identified with the previous variable is that a runner could be in the middle of a lap at the end of the hour. As a result, even though they may be running the same pace as the previous hour, they would be recorded as one fewer laps compared to the previous hour. For example, a runner may have ran five laps in hour one and four laps in hour two but was .01 miles away from completing five laps in the second hour. The second hour may have one more lap recorded than the first hour even though the pace has been consistent. We decided to define an additional variable that would account for this type of artifact in pace consistency. Specifically, we defined a variable that was the longest string where the difference from one lap to the next is 1, -1, 1, -1 or a similar pattern. For example, in Figure 2.1, we can see that in hour five Morton ran seven laps, the next four hours he ran eight laps, and the following hour he ran eight laps. The resulting string of differences between laps would be 1, 0, 0, 0, 0, -1. Note that when there is an increase or decrease from one lap to the next (meaning a difference of 1 or -1), it must be followed by a -1 or a 0 if a 1 is before and a 1 or 0 if a -1 was before. In addition, if the runner had an increase of one lap and then remained consistent for a few laps, the next time there is a change after the multiple zeroes, there must be a -1 or we know that he/she has changed the pace significantly since the previous increase. Simply said, the last 1 or -1 is followed by a -1 or 1, respectively after a string of zeroes. For example 1, -1, 1, 0, 0, 0, 1 would not be valid but 1, -1, 1, 0, 0, 0, -1 would be valid. In these two examples, the string of the same number of lap would be six and seven, respectively. The bottom right histogram of Figure 3.1 is skewed right but there is a spike between 22 and 24 hours, which we expect because of consistent runners and dropouts.

Table 3.1: Looking back on Figures 2.1 and 2.2, we can see what the trajectory variables look like for Morton and Scholz. We added Olsen's data in the table as well.

Name	Avg. Pace (Day)	Avg. Pace (Night)	Avg. Pace (Total)
MORTON	7.91	8.18	8.04
SCHOLZ	10.00	60.00	11.35
OLSEN	0.00	0.00	0.00

Name	Largest String Same	Hour Number Largest String Same
MORTON	15	19
SCHOLZ	16	24
OLSEN	23	24

- **Location of the End of the Largest String of 1, -1, 1, 0, -1 etc.**

  This variable is defined using the previous longest string of consistent laps. This variable is simply the hour in which the largest string of 1, -1, 1, 0, -1 or variation of the type of pattern we described above ends. The location of this string can give information as well. For instance, it may end one or two hours before the end of the race because the racer is finishing strong. There also may be other instances where a runner had an injury or something unexpected occurred that caused inconsistent pacing. If a runner was running very consistently for several hours and then all of a sudden dropped out, we might think that this runner was one in the elite group but either had an injury or an unexpected event come up that stopped him or her from completing the race at a consistent pace.

Table 3.1 is an example of what the data frame looks like for some of the continuous variables we mentioned. Notice that Olsen has the largest string of same laps (23) but an average pace of zero; this is due to the fact that he did not start the race (DNS). These variables are good representations of the running strategies/trajectories because they represent the speed and consistency of a runner (two of the most important factors that determine how well a runner is performing). These variables might be useful in clustering because the distribu-

tions of each of the variables show separation across their value ranges. This characteristic of these continuous variables might indicate different running strategies. While some of these variables are integer only, they were all treated as continuous. As we notice in Figure 3.1, the bottom right histogram is integer only.

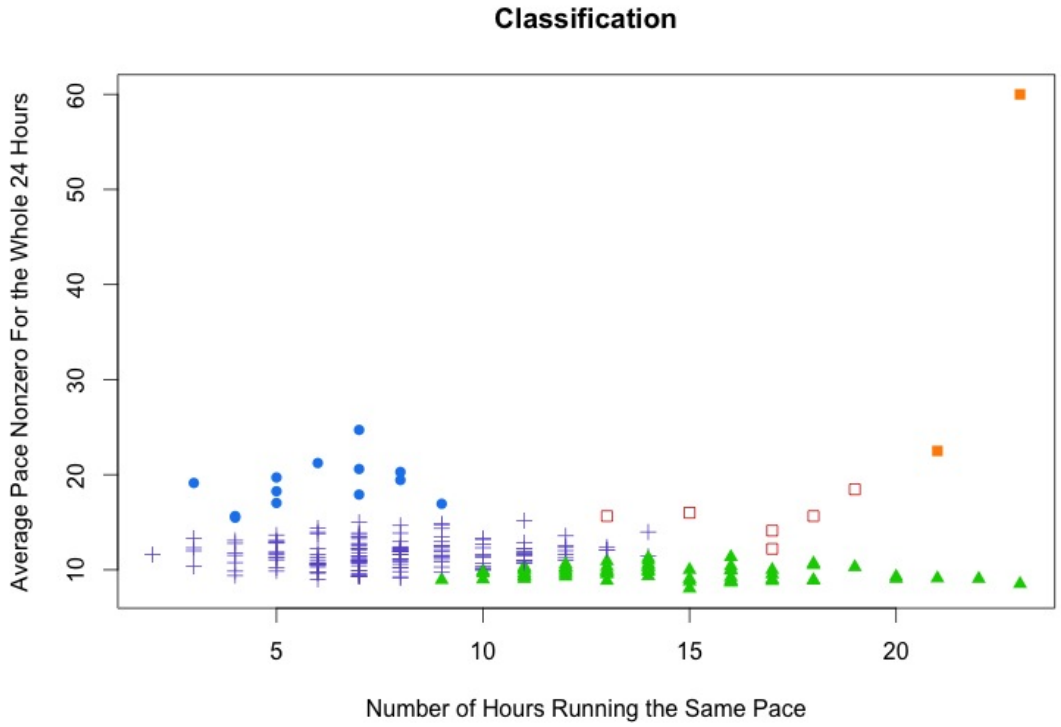
Before we look at the mixture model results, we set some standard definitions about the population of runners we were dealing with. First of all, as mentioned above, the population of runners are only those that started the 2012 World Championship race. Did not start (DNS) are those runners that were registered for the race but for some reason or another did not race. We briefly talked about dropouts in Section 2.3.2, but this is a tricky definition because a runner could not be running but could still be in the race because walking counts toward their lap total. However, a runner could be running for a very brief portion of the hour and then stop running. We consider a runner dropping out if they drop down to running zero laps and continue that for the rest of the race.

3.5 Mixture Model Results

We explored different combinations of continuous variables using the *mclust* package in R [3]. We did not cluster using all of the continuous variables at once because we wanted to visualize our results and some of the variables are highly correlated together. Our first example clusters the runners who started the race using the overall average pace while running and the number of hours at the same pace (Figure 3.3). Note that we adjusted the y-axis to have a minimum overall average pace of 60 minutes.

For these two variables, the chosen model was diagonal, equal volume, varying shape (EVI) with five components or racing strategies. We notice that there are five clear and defined

Figure 3.3: Normal Mixture Model Results when looking at Average Pace Nonzero and the Number of Hours Running the Same Pace. We see that five groups is the optimal number of groups given these two variables.



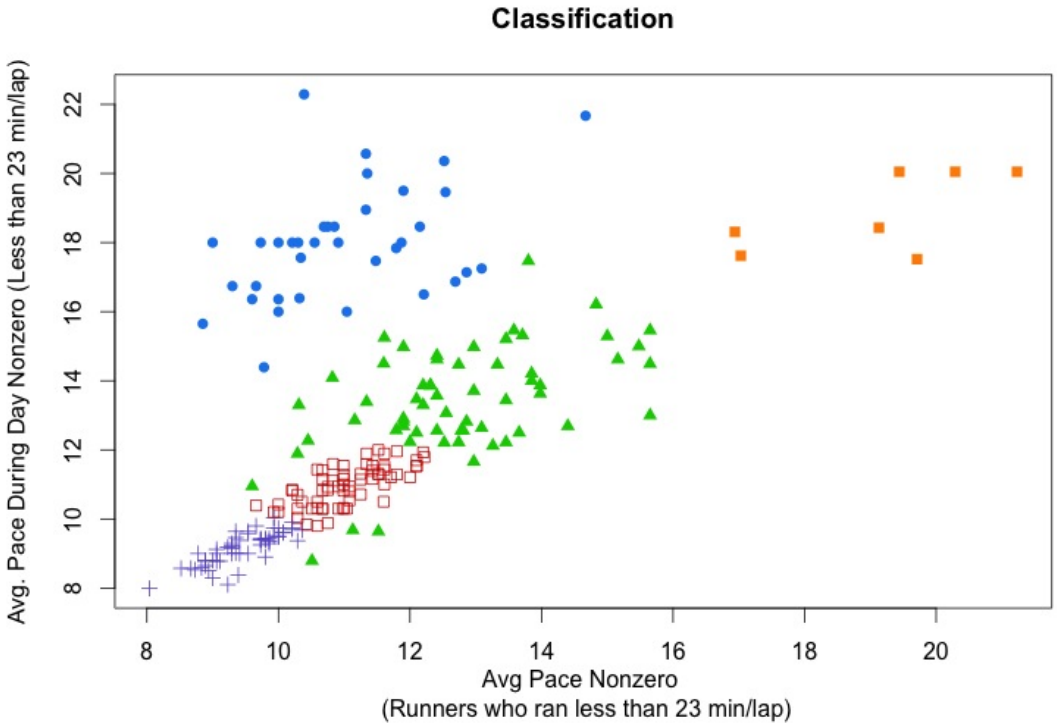
clusters in the plot (Figure 3.3). For example, looking at the blue cluster, their average pace, not including zeroes, is very slow to the point where they are almost walking (given our definition of a walk is 23 minutes per lap). Therefore, these people most likely did not run very long and did not run fast when they were running. In addition, the orange cluster contains two people who most likely dropped out due to their low number of hours running the same pace and their average pace nonzero around 23 and 60 minutes per lap, respectively. The red cluster could possibly be runners who were running at a fast pace when they were running but may have dropped out or started running again after dropping out for some time. The purple and green clusters are runners who most likely were running the majority of the whole race. The green cluster consists of the elite runners and the purple cluster has

a wider range of finishers.

Overall, we continued to explore cluster performance for continuous variables. Here we present three additional results that gave us useful information about the possible strategies/results of runners. Clustering the average pace nonzero and the average pace during the day nonzero (6am-6pm) selected five clusters/groups of runners. However, we chose to make a subset in the clustering of the overall pace to less than 23 minutes per lap, the casual walk pace (Figure 3.4). The purple cluster represents the elite runners we saw previously. However, we now have two clusters that were not as visible before (the blue and orange clusters). The orange cluster separated from the green cluster because their average pace nonzero and average pace during the day was in the 15-20 minutes per mile range. The blue cluster consists of runners who are running pretty fast overall throughout the entire race (9-12 minutes per mile) but not as fast during the day (16-22 minutes per mile). The purple cluster consists of the elite runners, both fast during the day and the race (8-10 minutes per mile). The red cluster runs fast during the day and the entire race as well, but the red cluster (10-12 minutes per mile) consists of runners who were slightly slower than the purple cluster. The green cluster has more variability in how fast they are running relative to the day time. We associate this with the runners who ran zero laps in hour three and then continued running in hour four.

We also cluster the average pace nonzero and the average pace during the day nonzero. We included all runners (aside from DNS), even casual walkers. In Figure 3.5, we see three clusters with the majority of the runners and then one cluster (green) that consists of the four runners who either ran a very slow pace during the day, a very slow pace throughout the entire race, or both. We see in Figure 3.4 that the red and blue clusters average pace during the day spread from eight to 13 minutes per mile and 8.5 to 17 minutes per mile,

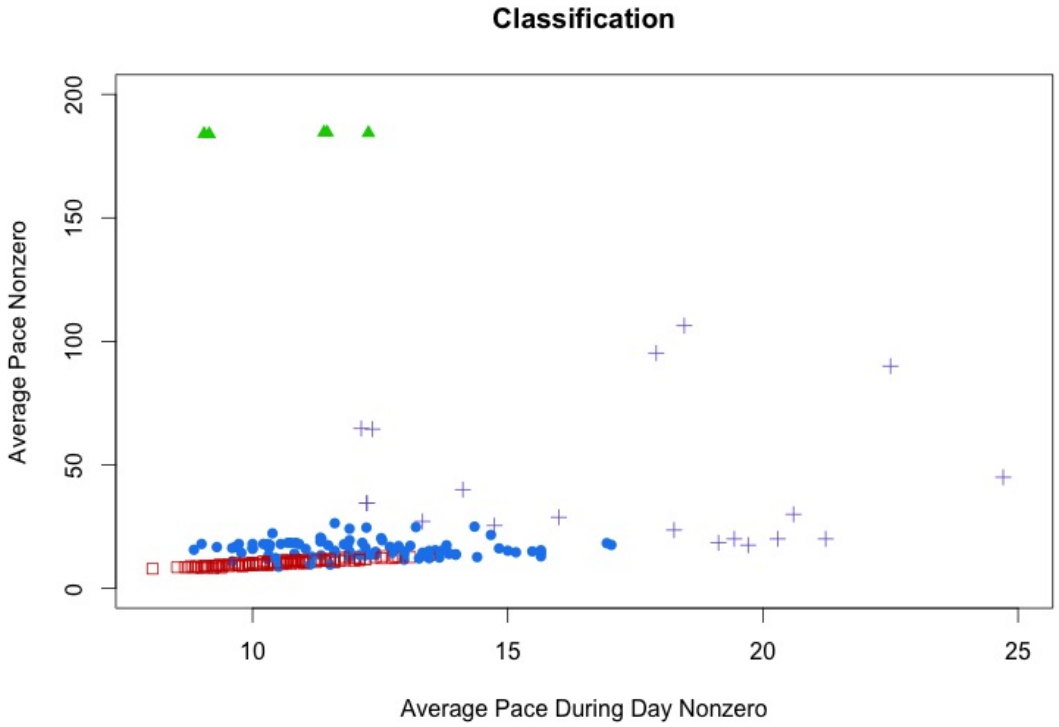
Figure 3.4: Normal Mixture Model Results when looking at overall pace and pace during the day nonzero for runners who kept an average pace nonzero less than 23 minutes per lap. We notice that five groups of runners seem to be the optimal number again.



respectively. The average pace nonzero ranges from eight to 12 minutes per mile and the blue cluster ranges from 12 to 25 minutes per mile.

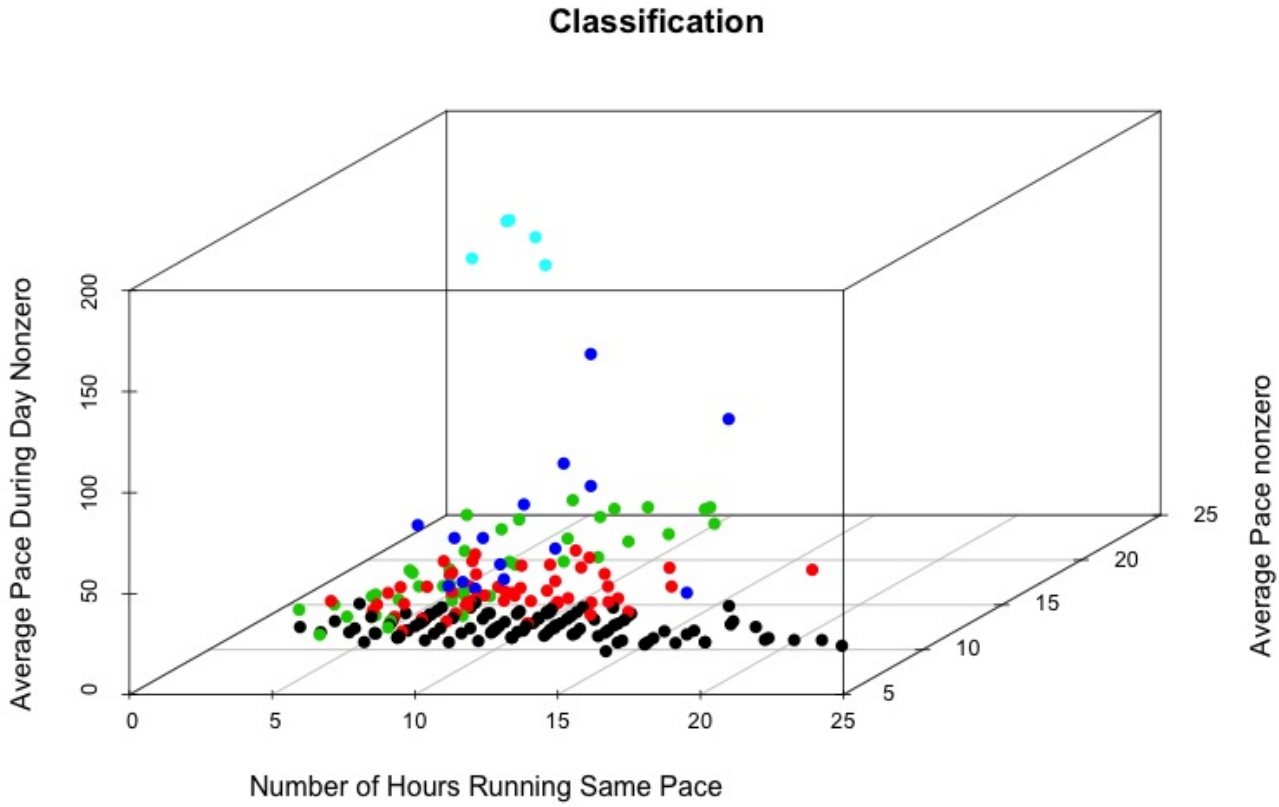
Finally, we chose to also cluster three variables: the number of hours running the same pace, the average pace during the day nonzero, and the average pace nonzero for runners who ran less than or equal to 23 minutes per lap. Looking at the black cluster in Figure 3.6 we see that it consists of the top (elite) runners. The cluster is not characterized by the runners having the same pace, but more by their overall pace and very slightly by their pace during the day. However, there is some overlap with respect to what cluster contains the fastest runners because the average pace nonzero and average pace during the day between the

Figure 3.5: Normal Mixture Model Results when looking at average pace nonzero and average pace during the day for all runners who started the race. We notice there being four groups; three of which consist of the five we have seen in the previous mixture models.



red cluster, blue, and green clusters have value ranges that overlap. The reason why they are not in the black cluster is that their average pace during the day nonzero is somewhat higher. This is either due to the sudden drop in the three to four hour range or these may be the runners that dropped out of the race. The 'red' runners, which overlap with the black cluster, are those who did not run as well during the day as the black cluster. We see that the blue cluster contains runners that have varied overall paces that are greater than 12 minutes per mile but their pace during the day is higher than all runners other than the light blue cluster, which contains the dropouts. The runners in the blue cluster may have stopped running during the race at some point but show more consistency when running compared to the light blue cluster.

Figure 3.6: Normal Mixture Model Results when looking at the number of hours running the same pace, average pace during the day nonzero (6am-6pm), and average pace nonzero for runners who ran less than or equal to 23 minutes per lap. A three dimensional plot gives five clusters.



In summary, so far we conclude that there is an elite cluster consisting of runners who are running roughly the same pace throughout the entire race. There is another group of runners with a very similar strategy but the variance of their overall pace is higher/consists of a larger range of runners. The difference between the elite runners and the "second group of runners" is consistency; specifically we partially attribute the race results for the second group to their lack of consistency across the full 24 hour race. We also conclude that there is a third group of runners who are similar to the second group but even more variable with

respect to their average pace nonzero and their pace during specific times of the day, which makes them slower. We also have reason to believe that there is a large group of runners who are very slow during the day relative to their overall pace, possibly due to the unexpected "third to fourth hour" time period. The "third to fourth hour" time period was the point in the race when the majority of runners had a large drop in their pace during hour three and immediately ran faster in hour four. We were unable to find information to explain these unexpected results.

The mixture model results were useful as an explanation but limited in the use of a small set of continuous variables. We next turn to latent class analysis, which allows clustering of categorical variables with an eye toward learning what other information might be useful in determining different running strategies.

Chapter 4

Latent Class Analysis (LCA)

Gaussian mixture models allowed us to look at clusters of runners but only based on the continuous measures. These variables were also dependent since all of them were related to the runners' pace. Our results seemed reasonable but there are several overlapping strategy clusters that might be more well-separated if other information was incorporated. There are many different categorical variables that might be able to explain the type of runner, including variables created by categorizing our integer-based variables. For example, the absolute value of the most laps dropped could be treated as a categorical variable where the values are one, two, and three, where one is the category "Greater than Four Laps", two is "One or Two Laps", and three is the category,"Three or Four Laps". In this chapter, we explore the use of Latent Class Analysis (LCA) as a tool to determine clusters based on categorical data.

4.1 LCA Estimation

Latent Class Analysis estimates groups with multivariate categorical variables. Although it is more difficult to visualize latent class analysis, we still are able to estimate and interpret

probabilities of observations being in certain groups/classes. LCA has a similar mixture model form [6],

$$\mathbf{x} \sim \sum_{g=1}^{G} \pi_g f_g(\mathbf{x}_n|\theta_g),$$

where G is the number of groups/classes/clusters and $\mathbf{x}_n$ are the multivariate observations. The function f is the density of an observation being in that gth group and θ_g are the parameters for each cluster g.

4.1.1 Notation

All of the categories in the groups are modeled with a multinomial density. For instance, looking at any variable, x, with d categories and given that it belongs in group g, the model is:

$$x|g \sim \prod_{j=1}^{d} p_{jg}^{1[x=j]}$$

where $x = j$ is the indicator function [6]. When equal to 1, the observation (runner) belongs to category j. The probability, p_{jg}, is the probability of being in group g for the variable that has value j. Assuming conditional independence, when we have k variables, Dean and Raftery describe In this case, p_{ijg} is the probability that some variable takes on the value j in group g. [6]

The overall density, is then a weighted sum of all of the individual product densities. The parameters in the model are estimated by maximum likelihood and using an EM algorithm and/or the Newton-Raphson algorithm. We use the EM and Newton-Raphson algorithms to maximize the latent class model log-likelihood function. The overall density follows:

$$x \sim \sum_{g=1}^{G} \left(\pi_g \prod_{i=1}^{k} \prod_{j=1}^{d_i} p_{ijg}^{1[x_i=j]} \right)$$

Determining whether we are able to fit a latent class model can be verified given the number of parameters in the model. Essentially, there needs to be enough cell counts in the model given by the following equation,

$$\prod_{i=1}^{k} d_i > \left(\sum_{i=1}^{k} d_i - k + 1\right) \times G.$$

The above formulas are taken from "Latent Class Analysis Variable Selection" by Nema Dean and Adrian E. Rafery [6].

4.1.2 Expectation-Maximization (EM) Algorithm

Similar to our Gaussian mixture models, the EM Algorithm is used to estimate the latent classes optimally by maximizing the log-likelihood. The general EM algorithm is described in Section 3.2. However, we do not only use the EM algorithm to construct the latent classes, we also use the EM algorithm with a step that contains the Newton-Raphson algorithm. See [11] for details.

4.1.3 Model Selection

Similar to Section 3.3, we choose the model with the maximum BIC as our best model. The BIC [8] is calculated as,

$$\text{BIC} = -2 \times \log(\text{maximum likelihood}) + (p) \times \log(n),$$

where we define n as the number of runners and p as the number of parameters. We compare the models that we fit for varying the number of groups G by taking the highest BIC and declaring that as the best model.

4.2 Categorical Variables

- **Most Number of Laps Dropped in 1 Hour**

  The most number of laps dropped in one hour was calculated by finding the largest decrease from one hour to the next (Figure 4.1).

- **Most Number of Laps Gained in 1 Hour**

  Similarly, we found the most number of laps gained in one hour.

- **Hour Number Having Largest Drop**

  The hour number having the largest drop is simply the hour number where their largest drop occurred. Note that we return the hour with the smaller number of laps. For example, if a runner has a largest drop from hour three to four, the recorded hour number would be four. Some runners have the same largest drop more than one time. For example, a runner may have dropped five laps from hours eight to nine and may have also dropped five laps from hours 23 to 24. As a result, we defined two more variables.

  - **Hour Number Having Largest Drop (Earliest)**

    The hour number having the largest drop (earliest) is simply the earliest point in the race where the runner had their largest drop.

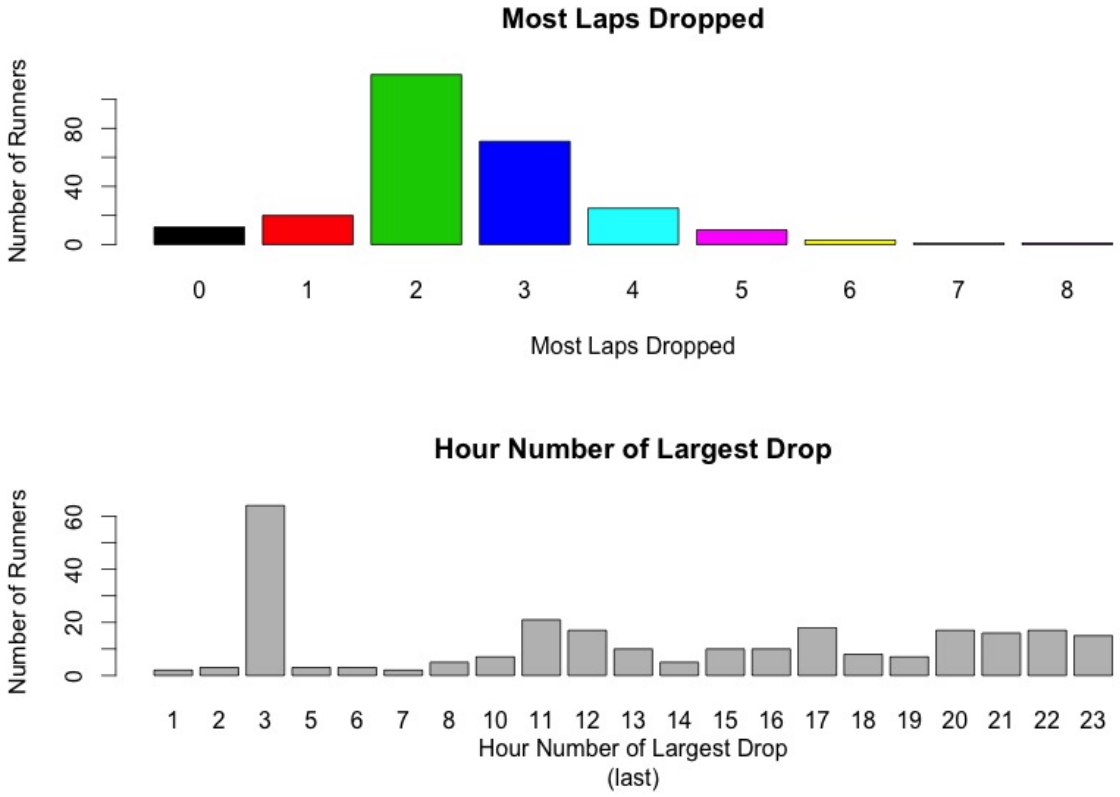
  - **Hour Number Having Largest Drop (Latest)**

    Similarly, we created this variable, as the latest point in the race where they had their largest drop.

- **Bounce Back From Largest Drop**

  After coming up with the largest drop variable, we also thought it would be interesting to look at how the runners might react to their largest drop. They may have dropped

Figure 4.1: Absolute Value of most laps dropped from one hour to the next and the hour number at which this largest drop occurred. If the runner had two hours where they dropped the same amount, we took the latest hour in the race as the value.



more laps after their largest decrease or rested for a small period of time and then continued at their normal pace.

- **Earliest Bounce Back From Largest Drop**

  The earliest bounce back from the largest drop is simply the hour after the earliest hour number having the largest drop occurs. For example, a runner may have dropped five laps from hours eight to nine but we are interested in what he did from hours nine to ten. If he dropped two more laps from hours nine to ten, then his "bounce back" variable would be negative two (-2).

- **Latest Bounce Back From Largest Drop**

The latest bounce back from the largest drop is the hour after the latest hour number having the largest drop occurs. In our previous example, since the runner's latest maximum drop was from hours 23 to 24 (the end of the race) his bounce back from his latest largest drop would be missing (NA).

- **Hour Number Having Largest Increase**

  We also similarly found the hour number of the largest increase for the earliest and latest point during the race.

  Similarly,

  - **Earliest Hour Number Having Largest Increase**
  - **Latest Hour Number Having Largest Increase**

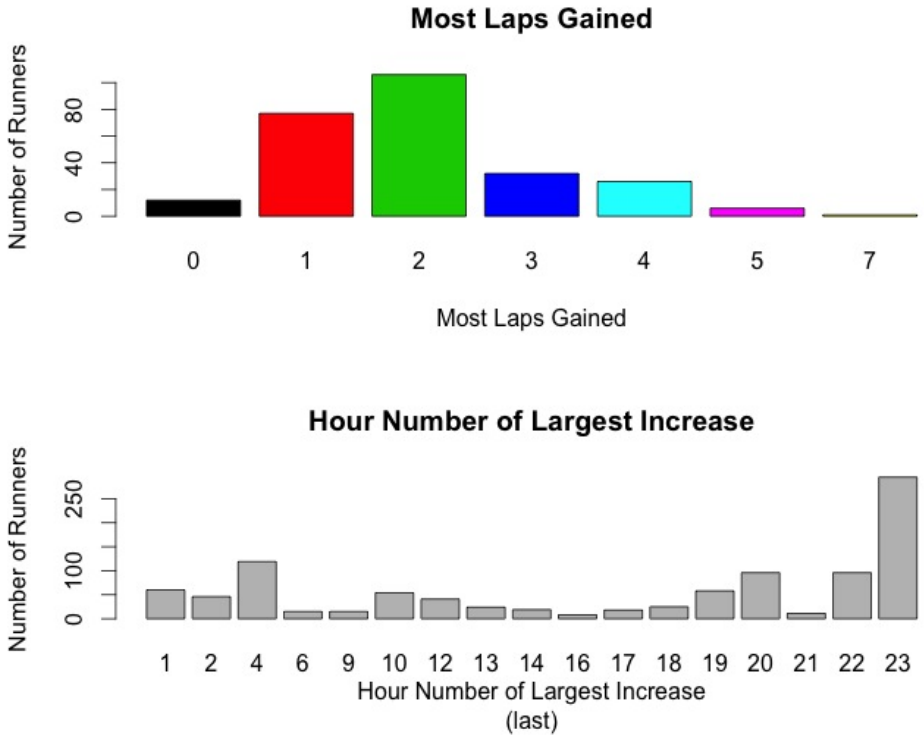
- **Next Lap Count After Largest Increase**

  As a counter part to our idea of bounce back after a drop, we look at the hour number directly after the runner's largest increase. We think that it is useful to understand the difference in the number of laps ran from their largest increase to their next hour. We expect that either the runner will increase a little more the next lap if it is the end of the race or we would expect that the runner would decrease because they went too hard the previous hour (Figure 4.3).

  Similarly,

  - **Earliest Result After Largest Increase**
  - **Latest Result After Largest Increase (If there were multiple increases of the same amount)**

Figure 4.2: Most laps gained from one hour to the next and the hour number at which this largest gained occurred. If the runner had two hours where they gained the same amount, we took the latest hour in the race as the value.

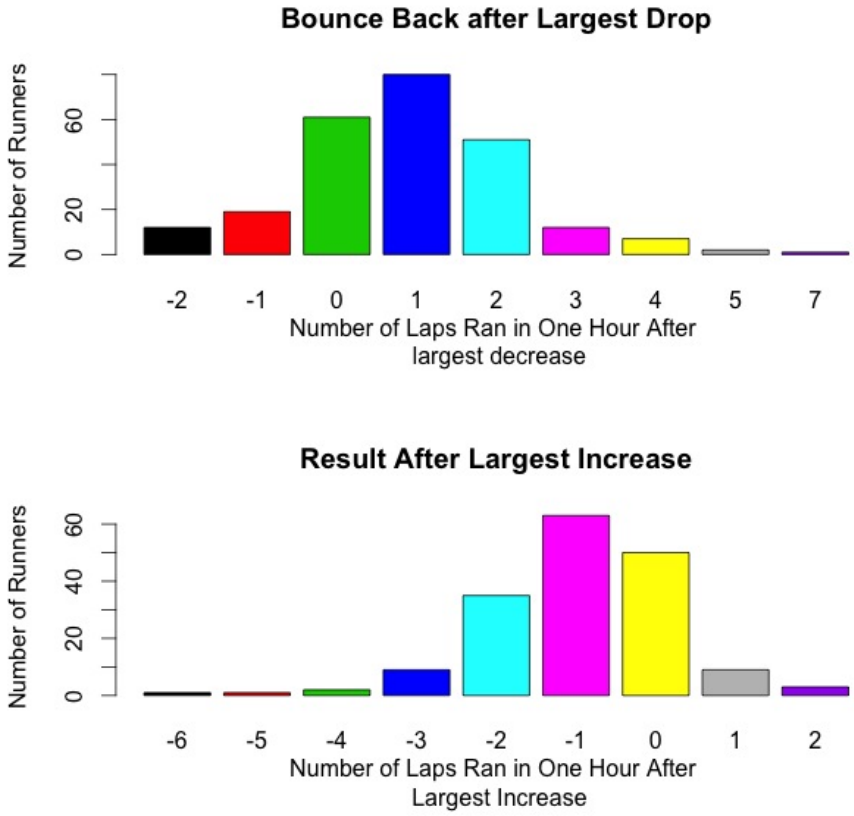


- **Largest String of Hours Consecutively Increasing**

  The largest string of hours consecutively increasing calculates the number of hours in a row when the runner ran more laps than the previous hour. This is a more common strategy in a shorter race because it is very difficult to make a drastic change in pace for multiple hours in a row. However, in some cases, we may find that when a runner took a break from the race completely, they may have come back into the race halfway through the hour so their pace for that hour was not actually representative of the actual pace they were running.

- **Largest String of Hours Consecutively Decreasing**

Figure 4.3: Bounce back after largest decrease and the result after the largest increase in the race. These results were taken the hour after their largest increase or decrease in pace.



Similarly, the largest string of hours consecutively decreasing calculates the number of hours in a row when the runner ran fewer laps than the previous hour. For instance, we tend to see this behavior when a runner is in the process of dropping out of the race. This tendency may also happen near the end of the race when a racer is "dying" or running out of energy to run at a fast pace. This could be a very useful variable in determining key characteristics of runners about the 24 hour race.

- **Hour Number of Largest String of Hours Consecutively Increasing (Shows three to four hour issue)**

  In relation to the largest string of hours consecutively increasing, we define a variable

Table 4.1: Table of LCA probabilities. The following table shows an example of one of the categorical variables (Absolute value of most laps dropped) that was included in the LCA model.

Class	OneorTwo	ThreeorFour	Greater4
Class 1	2.90e-01	6.18e-01	9.17e-02
Class 2	2.41e-01	7.25e-01	3.44e-02
Class 3	7.99e-01	2.01e-01	5.33e-227
Class 4	7.59e-64	7.77e-01	2.23e-01
Class 5	1.00	2.16e-08	3.79e-65

for the hour number in which the largest string ended. We found a very interesting pattern in the data approximately around the three to four hour mark in the race. Many people stopped running and then continued running the next hour.

- **Hour Number of Largest String of Hours Consecutively Decreasing**

  Similarly, we define a variable for the hour number of the largest string of hours consecutively decreasing. In particular, this allows us to make distinctions between someone who may have been dropping out as opposed to someone who may have been running out of energy to run at a consistent pace for the remainder of the race.

4.3 LCA Results

Using all of our categorical variables, we were able to fit a latent class model to estimate the probabilities that a particular category is in each group or class. We ran several models where the number of clusters ranged from one to five groups; a five group model was chosen as the best fit. The results gave us a table of probabilities of being in each class for each categorical variable (see, e.g. Table 4.1). Notice that there are three groups for the absolute value of the most laps dropped. These groups are one or two laps, three or four laps, and greater than four laps.

In addition to the absolute value of most laps dropped, we also looked at the following variables:

1. Most laps gained

   Three categories: One or two laps, three or four laps, greater than four laps

2. Largest number of consecutive hours decreasing in laps

   Three categories: One laps, two laps, greater than two laps

3. Largest number of consecutive hours increasing in laps

   Three categories: One Lap, Two Laps, Three Laps

4. Bounce back after largest number of consecutive hours decreasing in laps

   Three categories: Good (The number of laps increased from the previous hour), None (The number of laps stayed the same from the previous hour), Worse (The number of laps decreased from the previous hour).

5. (Last) Hour number with largest drop in number of laps ran from one hour to the next

   Four categories: Less than or equal to Hour 6, Hours 6-12, Hours 13-18, Hours 19-24

6. (Last) Hour number with largest increase in number of laps ran from one hour to the next

   Four categories: Less than or equal to Hour 6, Hours 6-12, Hours 13-18, Hours 19-24

7. Lap count after the last hour number with largest increase in number of laps ran from one hour to the next

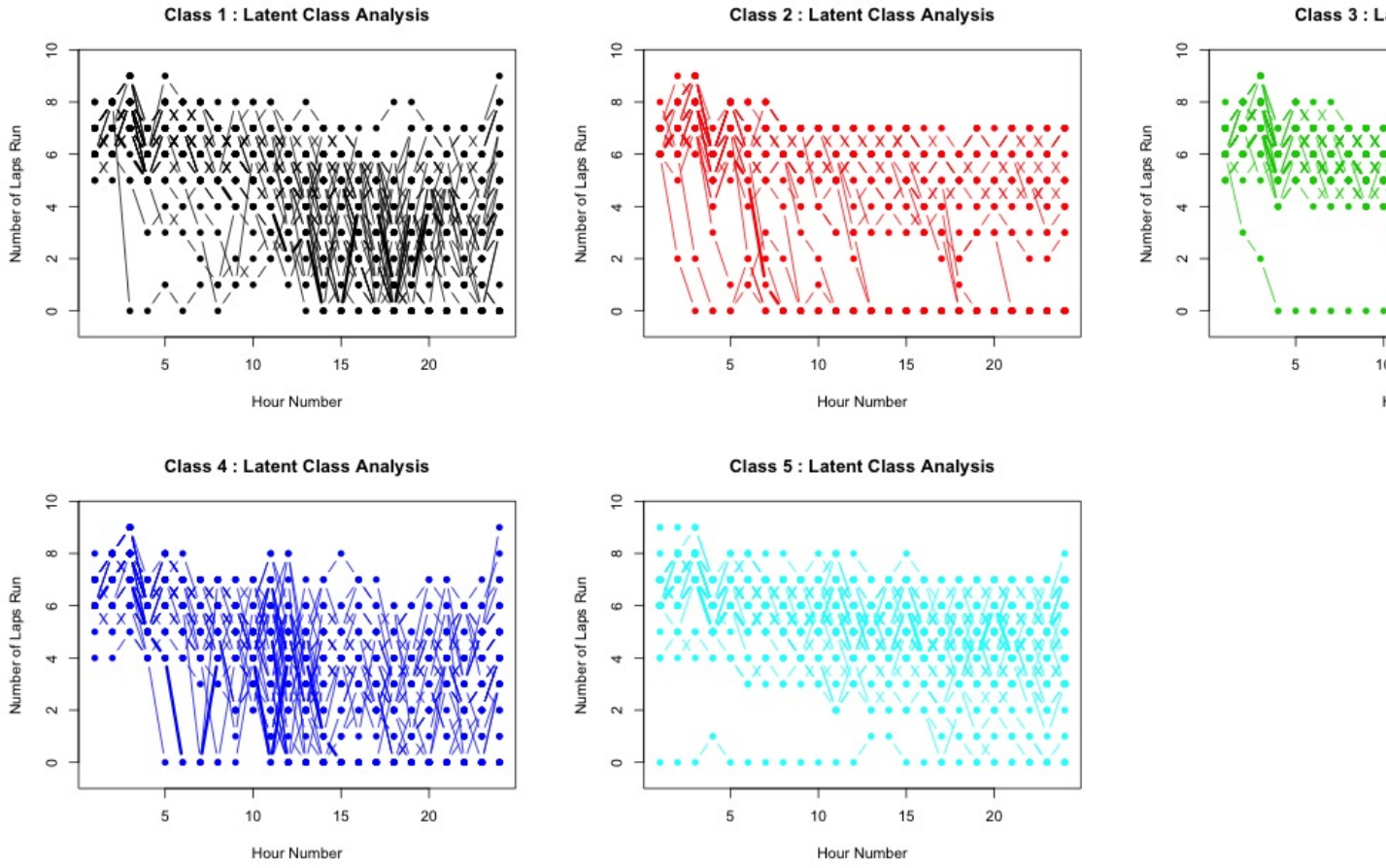
   Three categories: Zero Laps, Increased One or Two, Dropped Three or More, NA(meaning the largest increase was in Hour 24, thus there is no hour after that.

Figure 4.4 is a plot that shows the running trajectories for all of the runners in each final estimated cluster. There are several conclusions we can draw from this. First of all, although the plots are crowded, given the thickness of the lines in some areas, we can see where there are consistencies in terms of strategies. For instance, it looks like Class three (the green cluster) is the elite group of runners because their trajectories are the most consistent with their biggest gains at the end of the race. We also see that there is only one person who dropped out in this cluster (at hour four). Another interesting characteristic we see is the three to four hour drop and increase. This is present in the elite cluster but it is not as relevant when compared to classes one, two, and five. Looking at class five (the light blue cluster) we notice that there are not many dropouts here but the number of laps run is more variable than the green cluster. There also is large variation between laps 15 to 24, which shows signs of inconsistency. We would define class five as our second tier of runners behind the elite group.

The remaining three clusters are very interesting in that we can see several different characteristics that separate the classes. In particular, it seems that in class one there are several largest drops and largest increases around hours 13 to 21, however in class four it seems that the largest drops and largest increases are around hours 10 to 14. Therefore, it seems class four and one being separated by the time during the race when runners are being very volatile. Class two represents our class that contains the most dropouts and many of these seem to happen before the 13th hour of the race.

While we were able to successfully fit the latent class analysis to our set of categorical variables, this approach is still limited. Our next step is to build a mixture model that simultaneously estimates clusters for both continuous and categorical variables, or data of mixed type.

Figure 4.4: Latent Class Analysis Results: Running trajectories for each class, where the classes were made from latent class analysis



Chapter 5

Model Based Clustering with Mixed Data

Many problems arise when we have variables of mixed data types. In an all continuous setting we would use mixture modeling for model based clustering, as explained in Chapter 3, and in the case where we have categorical variables we would use latent class analysis, as shown in Chapter 4. Many times, like in our application of running, we have both continuous and categorical variables. Thus, ideally we want to use all of the information we have to make our conclusions about the number of clusters in the data and the explanation for each. The idea of mixture models with mixed data allow us to combine estimates using continuous, ordinal, and nominal variables into one estimation. This particular approach is relatively new to handling mixed data in model based clustering. The proposed model will allow us to put all of the variable information we have in our dataset and analyze the continuous, ordinal, and nominal variables separately but combine the estimation into one process. We plan to look at how the fit of the mixed data model changes the clusters in the simple continuous and categorical settings. We use the same method developed by McParland and Gormley [4], where they discuss the latent variable models use of Gaussian distribution mixtures to fit

the model and then extend it for broader use. The same Expectation-Maximization (EM) algorithm is used to complete the estimation, except in the case when there are nominal variables present. In these scenarios, a Monte Carlo EM algorithm is used. In our analysis of running, we use this method because of the presence of continuous, ordinal, and nominal variables. The goal with our analysis of mixed data types is to get all of the information we have about our data to identify the best groups and explanations of racing strategies used in an ultra-marathon. Although the latent variable model is hard to visualize, we see how the clusters change in small dimension space by comparing them to our results from the continuous setting of mixture models using only continuous variables in Chapter 3 and analyze the algorithm through simulations.

5.1 Mixed Data Type Likelihood

There are three separate models before coming up with a joint model. McParland and Gormley [4] briefly summarize developing a model for continuous, ordinal, and nominal variables. For detailed information about the mixed data type likelihood, refer to their paper.

5.1.1 Continuous Variables

The likelihood according to the *clustMD* model simply fits the continuous models based off of a multivariate normal distribution. That is for all observations, i, and continuous variable, j [4]:

$$y_{ij} = z_{ij} \sim N(\mu_j, \sigma_j^2)$$

We use this model for all of the continuous variables entered into the data frame for *clustMD*. That is if there are five continuous variables, the first five columns in the entering data frame

will correspond to the variables where the model is fit in this way.

5.1.2 Ordinal Variables

Ordinal variable j has K_j numbered categories at threshold parameter, γ_j. These threshold parameters give the value of the latent continuous variable, z_{ij}. z_{ij} is bounded from negative infinity to positive infinity while following a normal distribution $z_{ij} \sim \mathrm{N}(\mu_j, \sigma_j^2)$. The model can be estimated by the difference of two cumulative distribution functions. We can look at all of the ordinal variables and see the relationship one level has compared to the other level. Thus the probability is found for each level/category for the ordinal variables by taking the difference of two cumulative distribution functions. That is,

$$P(y_{ij} = k) = \Phi\left(\frac{\gamma_{j,k} - \mu_j}{\sigma_j}\right) - \Phi\left(\frac{\gamma_{j,k-1} - \mu_j}{\sigma_j}\right)$$

Φ represents the cumulative distribution function, i is the number of observations, j is the nominal variable we are looking at, k is the level/category in the j^{th} ordinal variable, and $\gamma_{j,k}$ represents the threshold level that comes from looking at the quantiles of a normal distribution [4].

5.1.3 Nominal Variables

The method for handling nominal variables is difficult and there are multiple ways to do it. We will use the method that McParland [4] used, where they state that a multivariate latent vector represents the nominal variable that is observed. The number of dimensions is one less than the number of categories in the nominal variable. In the nominal variable context $z_{ij} = (z_{ij}^1, \cdots, z_{ij}^{K_j - 1}) \sim \mathrm{MVN}_{K_j-1}(\mu_j, \Sigma_j)$, the scaled continuous vector we are looking at is coming from a threshold of 0 and is computed in respect to the reference level/category. That is, we see that the response y_{ij} is set at zero and then is either equal to 1 or k depending

on the calculated threshold. The threshold is defined as:

$$y_{ij} = \begin{cases} 1 & \text{if} \max_k \{z_{ij}^k\} < 0; \\ k & \text{if } z_{ij}^{k-1} = \max_k \{z_{ij}^k\} \text{ and } z_{ij}^{k-1} > 0 \text{ for } k = 2, \cdots, K_j \end{cases}$$

In the case of having binary data, we can treat the variable as having two responses instead of some k defined.

5.2 Estimation

We adopted the McParland and Gormley framework but we found that the use of the Monte Carlo samples (approximation required when there are nominal variables in the data) ran ended up limiting the estimation in some cases; we address that here. Within fixing the estimation in some of these cases, we found that we were able to fit more of the proposed models than we originally could.

5.2.1 Basic Setup of Variable Thresholds

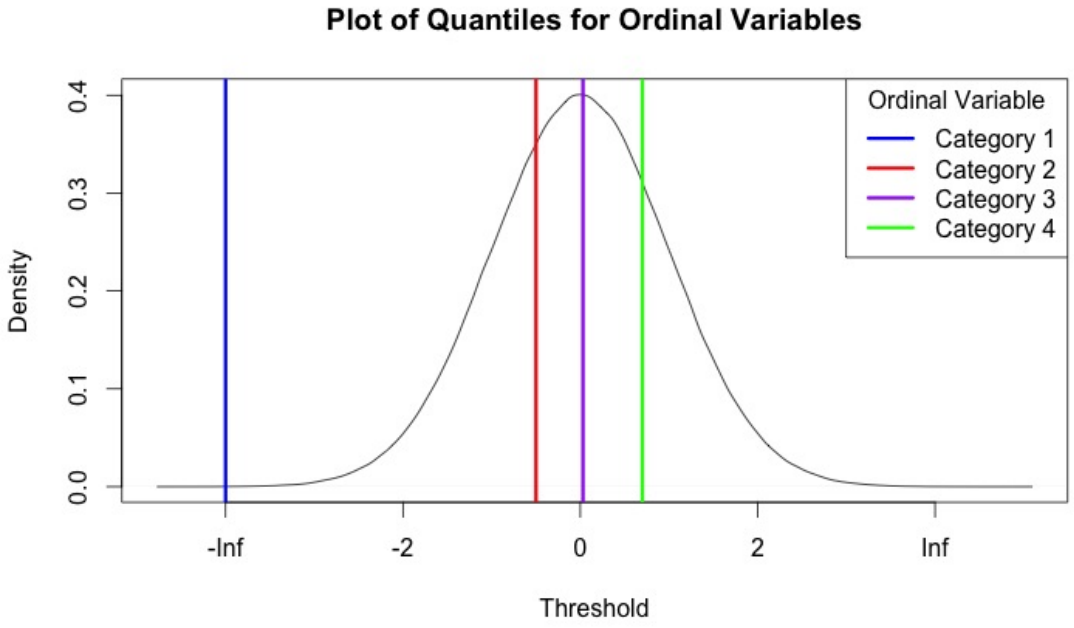
The function, *clustMD*, fits the mixture model for mixed data. The basic setup takes in an N by J matrix, where N is the number of observations and J is the number of variables in the dataset. This matrix, Y, must have the J variables in order where the continuous variables are first, the binary are second, coded as (1 or 2), ordinal third coded as (1,2, ...), and finally nominal variables coded as (1,2, ...).

We then check to see if there are any continuous variables and if so, scale the continuous variables by centering and dividing by the standard deviation. We then find the maximum value of these scaled continuous variables across all columns of Y, which are continuous, and

store them in a vector K. We then set the continuous variable max values to be NA in vector K and leave the number of categories for each categorical variable.

We calculate the threshold parameter for the ordinal variables by taking each ordinal variable and calculating the cumulative sum of the table of categories. We then proceed to take the quantiles from the normal distribution. See Figure 5.1 for an example.

Figure 5.1: Plot of the Quantiles assigned to an ordinal variable. For instance, the category that is coded as one for the ordinal variable would receive a threshold of negative infinity, category two would hold the value at the red line, three as the value at the purple line, and four as the value at the green line.



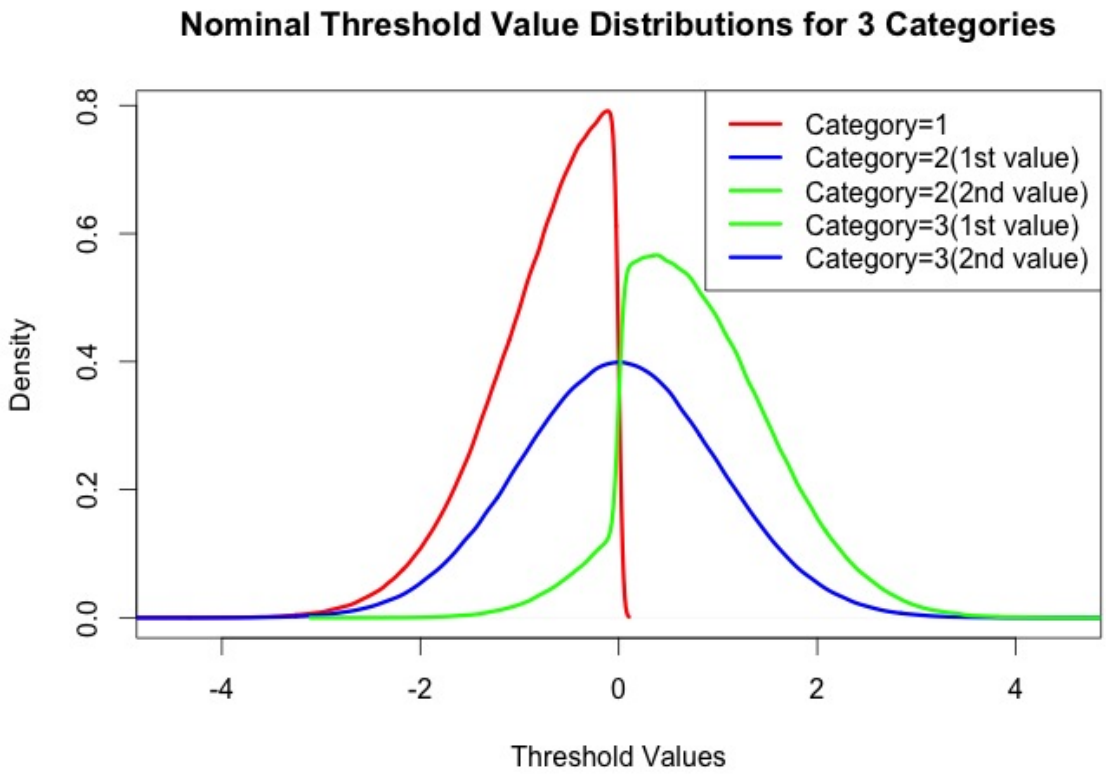
Notice in the figure that the blue line represents negative infinity and it is assigned to the category coded as a one in the nominal variable. Similarly, the red line holds a value that is assigned to the category coded as a two in the nominal variable. Similarly, we can say the same for categories three and four for the purple and green threshold values, respectively.

Before finding the threshold parameters for the nominal variables, we must check to see how many there are and take one less than the number of categories there are in each nominal variable. That is, we take the maximum number of categories for each nominal variable and subtract one, and then sum the nominal categories. We need this number so we can initialize a matrix for the threshold. In the nominal setting, we said that the threshold is centered at zero, therefore, the other categories are in reference to one group category. This reference group is the maximum number of categories for each nominal variable. That is if variable x had three categories, the reference group would be category three.

To calculate the threshold parameters for the nominal variables, the model takes truncated random deviates for all of the categories minus the reference category in each nominal variable. That is, the model takes a random deviate for all of the categories that are not the reference group. Therefore, if there are three categories in the nominal variable, the third group is the reference group. For the category that is one less than the reference category, the lower truncation point is the maximum of the random deviate for the previous categories that are not the reference category. Figure 5.2 shows an example in the case when the nominal variable has three categories. Notice that when the nominal variable, y, is equal to category one, the two threshold values (in our case the maximum value is three and we subtract one to get the two threshold values) are taken from the distribution that represents the red density. Notice that these values will be negative.

Second, when the nominal variable, y, is equal to category two, the first threshold value is taken randomly from the distribution that represents the blue density and the second threshold value is drawn randomly from the green density.

Figure 5.2: Nominal Threshold Value Distributions for three categories



The opposite holds true when the nominal variable, y, is equal to category three. That is the first threshold value is drawn from the green distribution and the second threshold is drawn from the blue threshold value. See McParland and Gormley for specific theoretical details [4].

5.2.2 Estimation Overview

The EM algorithm is used to fit the model and when there are nominal variables, an approximation using Monte Carlo samples is used in the expectation step of the EM algorithm. Refer to Section 3.2 for a summary of the EM algorithm. It differs in that now we calculate the complete log likelihood to latent data, which combines continuous, categorical,

and nominal variables. We must complete the expectation of the log likelihood three times. The formulas for the three expectations can be found on pages seven and eight of McParland and Gormley [4]. We will briefly discuss the expectation and maximization steps below.

The covariance matrix Σ_g^β is diagonal for several reasons. It is diagonal because this limits the number of parameters that we need to estimate. It limits the number of parameters we need to estimate because the covariance matrix can be shown as the product of probabilities. It is simple to calculate the probabilities for ordinal variables because we can use the threshold parameters we calculated in Section 5.1.2. In addition, we can find the mean and sigma components of these ordinal variables given these threshold parameters because they are the first and second components of a truncated Normal, respectively.

The maximization step of the algorithm depends on the model. There are six possible models in this setting, ("VVI", "EVI", "EEI", "VII", "VEI", and "EII"). The expected value of the log likelihood of the model parameters is maximized in this process. See McParland and Gormley [4] and McLachlan and Krishnan (referenced in McParland) for further detail.

There are constraints for identifiability with respect to nominal variables. We identity the nominal variable categories in Section 5.1.3. In the presence of nominal variables, we simulate from the multivariate normal distribution the number of times we initiate the Monte Carlo samples. We also fit the diagonal matrix based off of simulated values we get from the multivariate normal distribution. We then need to calculate the first and second moment of the latent data where the dimension of the data is p plus the number of categories we have in the nominal variables minus one for each nominal variable we have. This will return the expected value of the truncated normal after we calculate the empirical moments corresponding to the reference category for nominal variables. We are able to get the standard

deviation and probabilities of belonging to each category throughout this process.

5.2.3 Estimating Parameters with Nominal Variables

In the presence of nominal variables, we must use a Monte Carlo approximation that finds the actual probability of the response category based off of many simulated vectors from a normal distribution with mean μ_g^j and covariance Σ_g^j, where these are the means and covariances for each cluster, g. The mean and sigma components for nominal variables are also related to the first and second components of a truncated Gaussian distribution but the thresholds, as we mentioned in Section 5.1.3, are very difficult to compute so we use a Monte Carlo in this situation as well. The first component is the sample mean of all Monte Carlo samples that return a response k where k is the number of categories in the nominal variable. The second component is the inner product of all of the simulated vectors from the Monte Carlo samples that give response k and then finding the sample mean of the inner products.

We encountered many difficulties when running simulations in the nominal variable setting, specifically in the process of the algorithm when we check to see if the number of categories for the nominal variables are equal to the number of probabilities in the function. We were unable to fit the *clustMD* from the data we simulated and figure out what the problem was because our strategies were well separated. We found that the data we simulated in this simulation may have been too well separated because when the nominal variables broke down the categories in combination with the other variables, we had sets that did not contain any one of the three categories. This should be seen as a good thing because we believe that there is close to a zero percent chance that this subset of the observations belongs in category x for that nominal variable. As a result of this error, we were unable to fit the data using

clustMD, resulting in not getting misclassification rates and the ability to interpret our *clustMD* results. After trying to adjust the distributions of our strategies and not having much luck, we decided to alter the *clustMD* package by fixing the particular error we kept seeing. We will describe what we did and how adjusting the Monte Carlo samples led to extending our estimation capability.

5.2.4 Extending Estimation Capability

When fitting the *clustMD* model, we came across issues when iterating through the Monte Carlo samples. Specifically, we checked to determine if the number of categories for the nominal variables is equal to the number of probabilities. We changed the algorithm to cover the case if they are not equal. If they were not equal, we substituted a probability of $1e^{-10}$ in for the category that does not have its respective probability. We then subtracted ($1e^{-10}$ divided by the number of remaining categories) to get the new respective probabilities for the categories that did have category classifications. For instance, there is a vector with a length of the select number of Monte Carlo samples estimated from the nominal variables in the E-step. If there are three categories in the nominal variable, the vector will consist of ones, twos, and threes. We get the probability of each being a one, two, and three and store it to be passed into the Monte Carlo simulation. However, if there are no twos in the vector, the *clustMD* model simulation breaks. Therefore, we give the category (in our case two) that is missing from the Monte Carlo samples a very small probability to allow the simulation to continue. See Table 5.1 for the probability vector before and after substituting zero for the missing nominal category probability.

We then must update μ_g and Σ_g for the Monte Carlo sample to avoid encountering another error in the estimation process because of "NA" values for the subsequent iterations. If we suppose that z_i^{β} is the expectation of the complete log likelihood with respect to latent

Table 5.1: A table of the probability vector before and after substituting zero for the missing nominal category probability

Probability Vector	Category 1	Category 2	Category 3
Before	.4%		.6%
After	0.39999999995%	$1e^{-10}$%	0.59999999995%

data and l_{ig} is the latent cluster labels then we can define calculating $E(z_i^J|y_{ij} = k, l_{ij} = 1, \mu_g, \Sigma_g, \pi_g)$ and $E(z_i^{j^T} z_i^j|y_{ij} = k, l_{ig} = 1, \mu_g, \Sigma_g, \pi_g)$ for the Monte Carlo simulation where $y_{ij} = k$ is for nominal variable j [4]. We generate these Monte Carlo samples so that we can calculate the probabilities τ_{ig} where i represents the nominal variable and g represents the category in the nominal variable.

This is why we need to have the probability for each category. The μ_g and Σ_g need to be updated but cannot be updated when there are no simulations for one category in the previous iteration. Therefore, we stored the old μ_g and Σ_g so that if we run into the case where one category is missing their probability, we can use the old μ_g and Σ_g to plug in for that missing category and continue with the E-step estimation step for the following iteration.

5.3 Model Selection

Similar to our mixture model estimation in Section 3.3, we will use the Bayesian Information Criterion [8] to conduct our model selection. The BIC cannot be calculated but we can estimate the observed likelihood so we can calculate our BIC:

$$\text{BIC} = -2 \times \text{logL}(x|\theta) + \log(n) \times p$$

In order to choose the number of groups K/model, we iterate through all of the possible models and look to see what BIC is the highest. The majority of the time will take the number of groups and model with the highest BIC, but in some cases it may be appropriate to take a different model and group.

5.4 Simulated Distance Runner Strategies

We decided to run a few simulations to help interpret the results of the *clustMD* model. The issue we came across involved interpreting the clusters because there are reference groups when looking at the nominal variables and we also have no way to visualize what is going on like in the mixture models scenario. Once we simulated the data, we were able to fit *clustMD* on this data and see how the misclassification rate performs. We then hope this allows us to make some conclusions based off how well we believe the *clustMD* algorithm is working, as well as interpreting results based off our underlying assumptions from the simulated data.

We ran a simulation that was completely independent of the continuous and categorical results we obtained from the model based clustering and latent class analysis. Therefore, we decided to simulate three different strategies that we feel explain the majority of the groups in our data based off of guessing how the distributions of the variables would look. We wanted to make sure that the variables were separated among strategies to distinguish between the groups. We quickly noticed problems once we started fitting the *clustMD* model. These problems will be outlined as well. However, the three strategies were simulated as follows:

1. **Strategy 1** (Elite, Consistent, Low Variability)

   We simulated the dataset by looking at each variable for each strategy separately. For

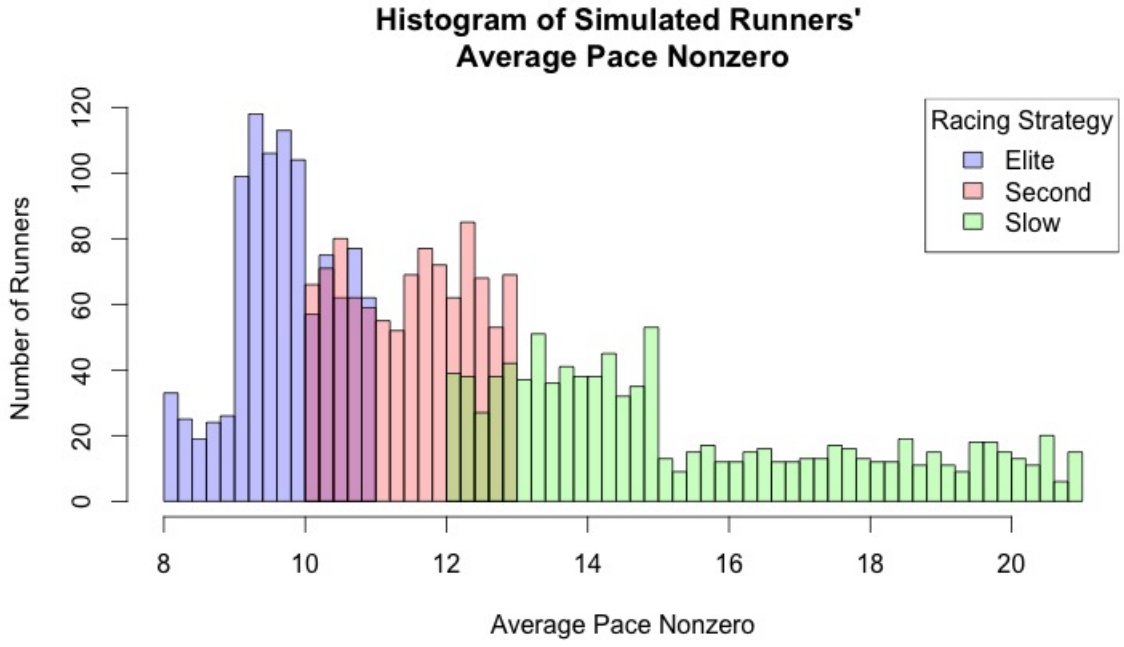
instance, we guessed that the average pace nonzero would range from eight minutes/lap (min/lap) to 10.99 min/lap for the elite group. See Figure 5.3 for the distribution of the elite group's average pace nonzero compared to the other two groups. Each of the distributions for all of the variables can be graphically visualized in this way. Continuing on with the average pace during the day nonzero and the average pace during the night nonzero, the elite group's range of pace is eight min/lap to 10.5 min/lap and 8-11.5 min/lap, for day pace and night pace respectively. All of the pace variables were constructed under a uniform distribution. We adjusted the distribution accordingly to account for a skewed left, skewed right, or normal distribution of paces. Finally, the same pace variable was distributed across 7 to 23 hours and skewed to the right. However, we made sure that the distribution picked up the most consistent runners since we think that the most consistent runners in the race are the ones who are going to run the fastest/be in the elite group.

The two ordinal variables are the hour number where the largest drop and largest gain occurred. For the largest drop, we decided that the range of laps would be from hour 5 to hour 23 with an equal probability of each hour. However, for the largest gain, the distribution would appear to be from all hours of the race with varying probabilities that would make the distribution skewed left.

Finally, the nominal variables, we had most laps dropped from one hour to the next ranging from one to three laps with probability of each of those categories as .2, .65, and .15, respectively. Similarly, the most laps gained from one hour to the next will range from one to two laps for the elite group with an equal probability for each of those two categories.

We simulated these variables for 1000 observations using the appropriate distributions

Figure 5.3: Continuous Variable: Distribution of runners' average pace nonzero by racing strategy.



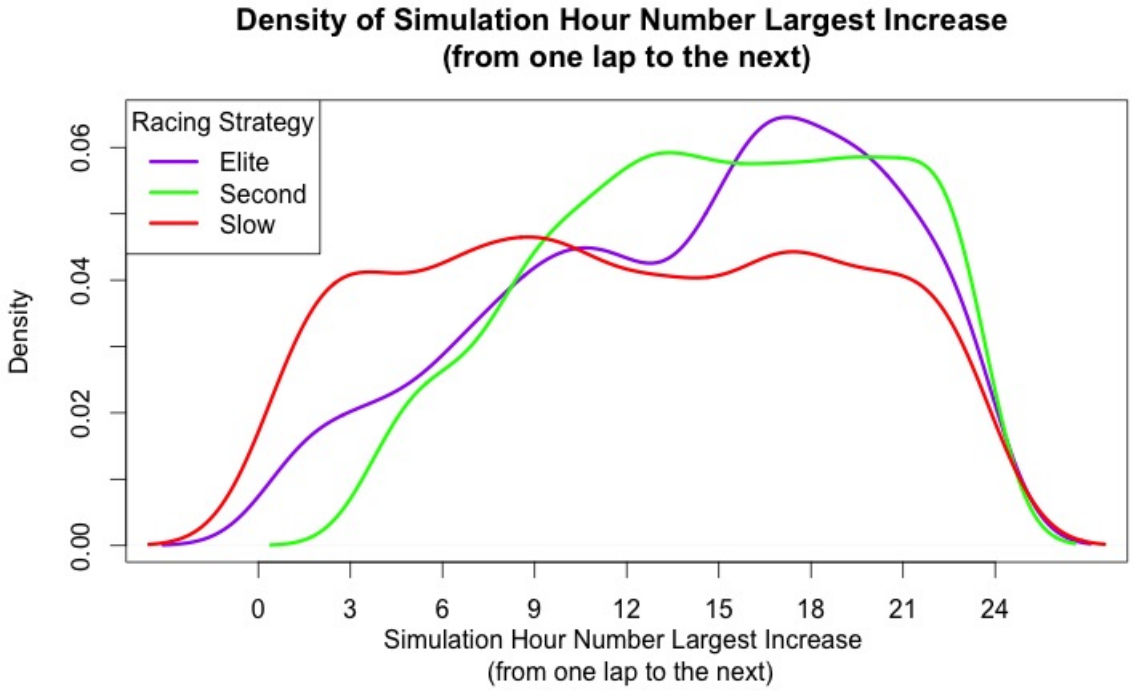
we described above. We then put this strategy into a data frame as the first 1000 observations with eight variables. The second and slow strategies will follow in the data frame as the next 2000 observations. We experimented with a different number of observations and different number of variables but we wanted to make our estimation consistent with how we fit the original *clustMD* model as described in the beginning of this chapter.

2. **Strategy 2** (Second, Somewhat consistent, High Variability)

   For strategy 2, we guessed that the average pace nonzero would range from 10 minutes/lap (min/lap) to 13 min/lap. See Figure 5.3 for the distribution of the second group's average pace nonzero compared to the other two groups. The second group's range of pace is 10 to 12.5 min/lap, for both the day and night pace. The same pace variable was distributed across 5 to 17 hours and a small skew to the right.

The largest drop range of laps would spread across all hours of the race with an equal probability of each hour. However, for the largest gain, the distribution would appear from hours 4 to 23 of the race with varying probabilities to make the distribution skewed left.

Figure 5.4: Ordinal Variable: Density of the simulated hour number largest increase from one lap to the next



For the second strategy the nominal variables we had was most laps dropped from one hour to the next ranging from one to four laps with probability of each of those categories as .25, .45, .25, and .05, respectively. Similarly, the most laps gained from one hour to the next will range from two to four laps for the second group with a probability vector of .6, .2, and .2 for the three categories, respectively. Figure 5.4 shows the density of the simulated hour number largest increase from one lap to the next. Notice that the distributions is uniform for the slow group but skewed left for

the elite and second groups.

3. **Strategy 3** (Slow, Inconsistent, High Variability)

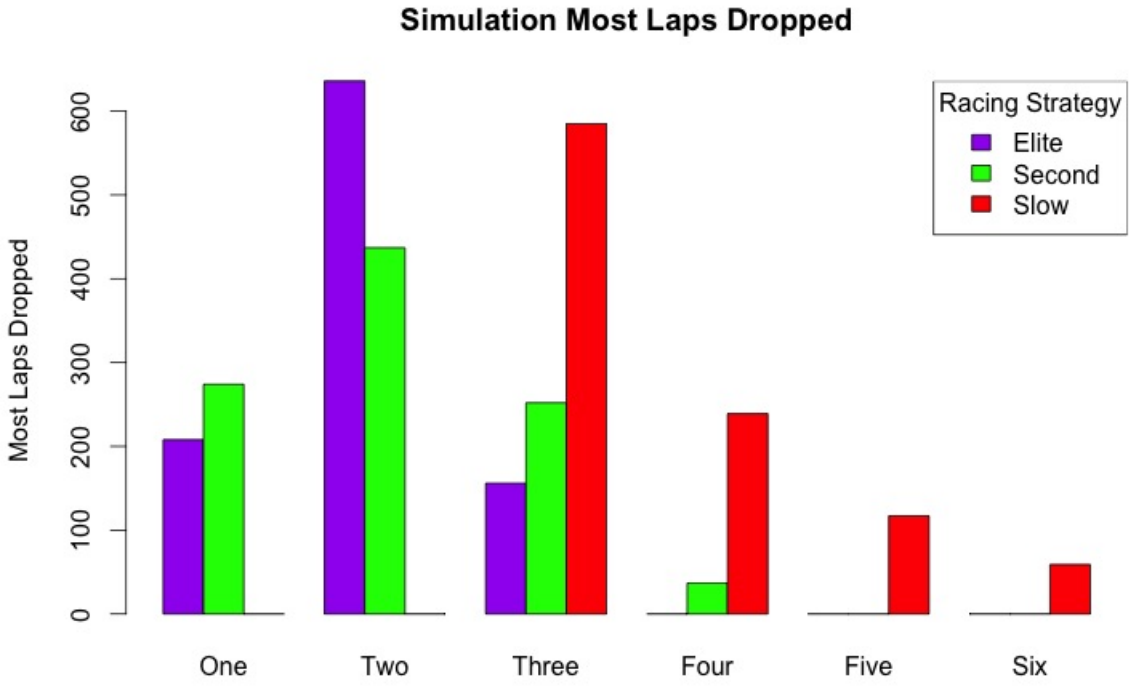
   We guessed that the average pace nonzero for the slow group would range from 12 minutes/lap (min/lap) to 21 min/lap for the slow group. See Figure 5.3 for the distribution of the slow group's average pace nonzero compared to the other two groups. The slow group's range of pace is 12 min/lap to 18 min/lap and 12 min/lap to 24 min/lap, for the day and night, respectively. The same pace variable was distributed across one to eleven consecutive hours.

   For the largest drop, we decided that the range of laps would be from hours 5 to 23 of the race with an equal probability of each hour. However, for the largest gain, the distribution would appear from hours 1 to 23 of the race with the same probabilities for each hour.

   The most laps dropped from one hour to the next ranging from three to six laps with probability of each of those categories as .6, .25, .1, .05, respectively. Similarly, the most laps gained from one hour to the next will range from three to six laps for the slow group with a probability vector of .5, .35, .1, .05 for the three categories, respectively. Figure 5.5 is a side-by-side bar plot showing the distribution of the simulated most laps dropped from one hour to the next hour.

Under the new model, the simulation worked and actually seems to be working as well if not better than before. We now are able to estimate using all of the models given in the *clustMD* model using our adjusted algorithm. We were able to calculate BICs for many different models and of many different group sizes for these models. The adjusted algorithm fit well according to the BIC values shown in Figure 5.6 and we got low misclassification rates again

Figure 5.5: Nominal Variable: Side-by-side barplot showing the distribution of the simulated most laps dropped from one hour to the next hour



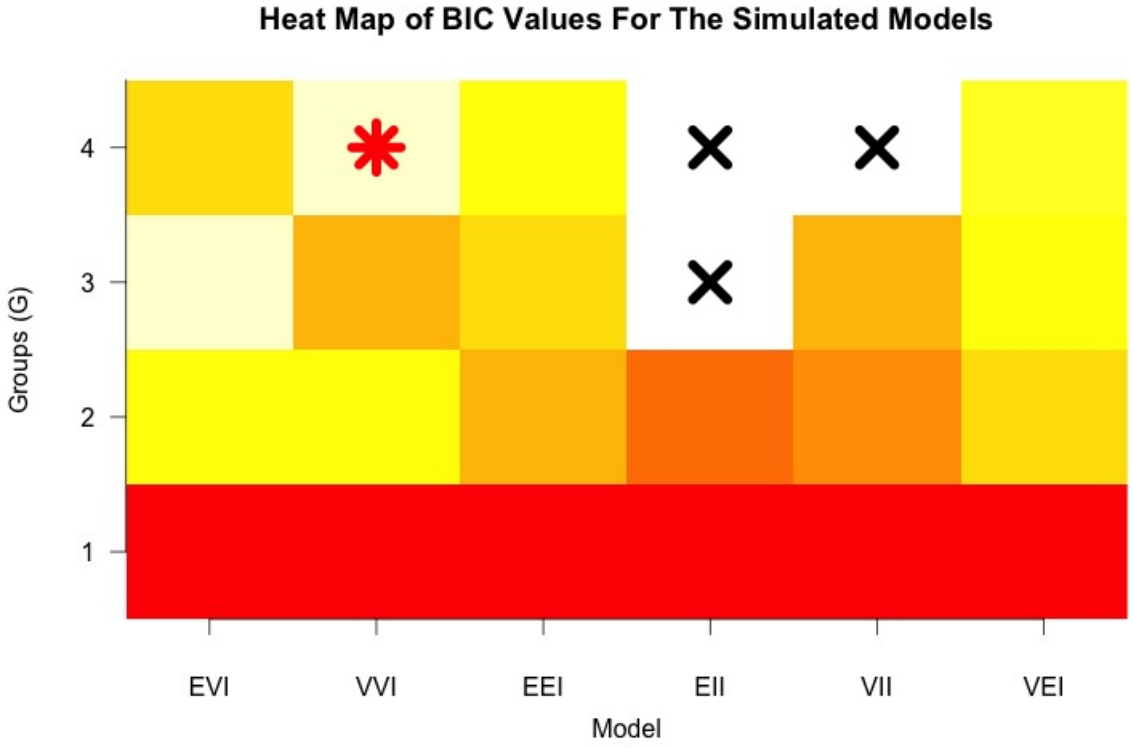
for the "EVI' model with three groups as shown in Table 5.2. Based off of this simulation, we find that the best model is using the "VVI" model with four groups. Figure 5.6 shows the BIC values for all models fit when using the adjusted *clustMD* algorithm with the distance runner strategy simulation data. In this case, we would most likely choose the "VVI" model with four groups, but the "EVI" model with three groups may be better. Since we simulated data corresponding to three groups, there is a reason why the "VVI" model is performing so well with four groups.

We analyzed to see why the "VVI" model with four groups was fitting better than the "EVI" model with three groups. Figure 5.3 shows the distribution of simulated runners' for the average pace nonzero and similarly the other figures show some of the other variables we

Table 5.2: Table of misclassification rates from the simulation of distance runner strategies. It was independent of any of the results we obtained thus far and was also fit using the altered *clustMD* algorithm.

Observations	1:1000	1001:2000	2001:3000	Total
Misclass. Rate	0.007%	3.3%	6.4%	3.47%

Figure 5.6: This shows the BIC values for all models fit when using the adjusted *clustMD* algorithm with the simulation of distance running strategies data.



simulated.

In particular, the histogram of the average pace nonzero has a very large slow group because the pace spreads from 12 min/lap to 21 min/lap. This is interesting because the "VVI" model at $g = 4$ actually splits up this group into separate groups (Figure 5.7). The "EVI" model for three groups almost has the identical plot of the distribution of simulated runners'

Figure 5.7: Using the VVI model with four groups, we compose the distribution of runners' average pace nonzero by racing strategy.

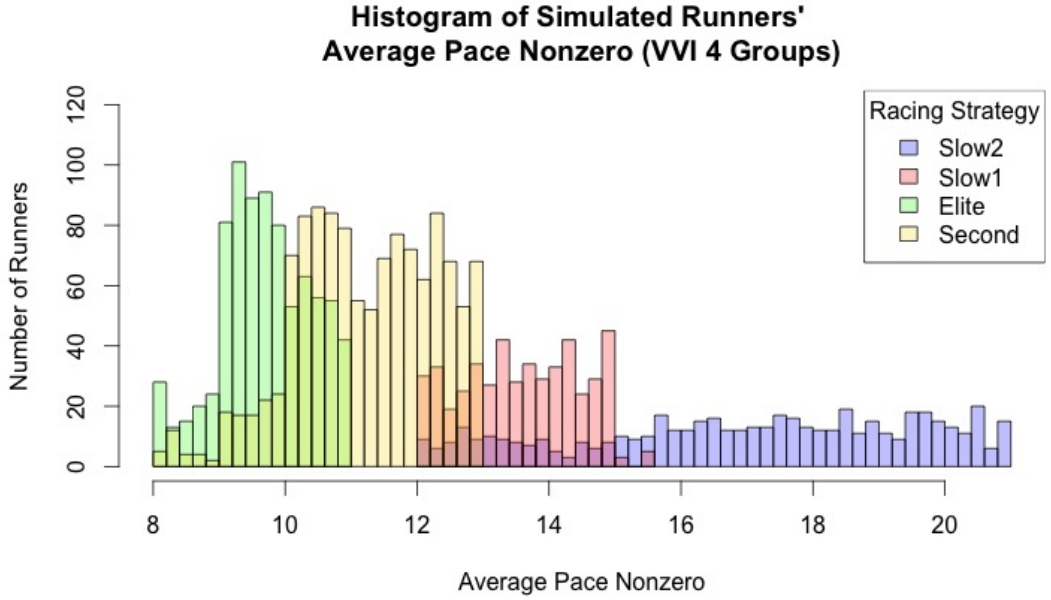
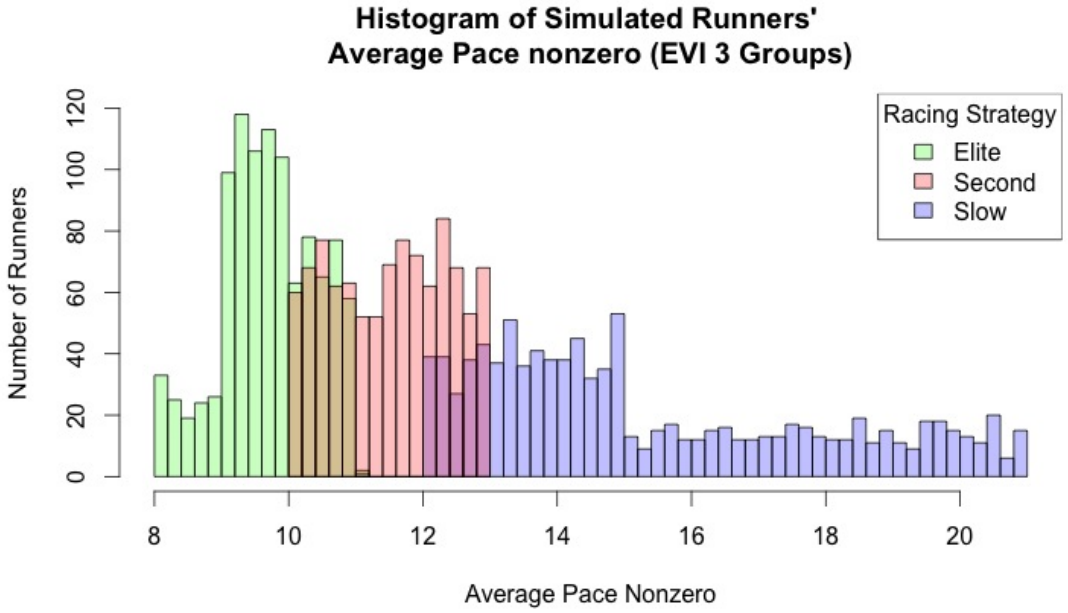


Figure 5.8: Using the EVI model with three groups, we compose the distribution of runners' average pace nonzero by racing strategy.



(Figure 5.8). Obviously we know how many groups are in this data because we simulated it in a way that composed three groups, but it is interesting to see the "VVI" model with four groups estimate better (in terms of BIC).

We believe that the adjusted version of the *clustMD* model is appropriate for fitting models that combine continuous, ordinal, and nominal variables. In any occurrence that we have more information about how the clusters should be separated, it should be beneficial for the estimation process of the model, not an issue. In the initial *clustMD* model, we are seeing this particular characteristic of the data becoming a problem for our estimation purposes, when it clearly should be benefiting our our model as a whole.

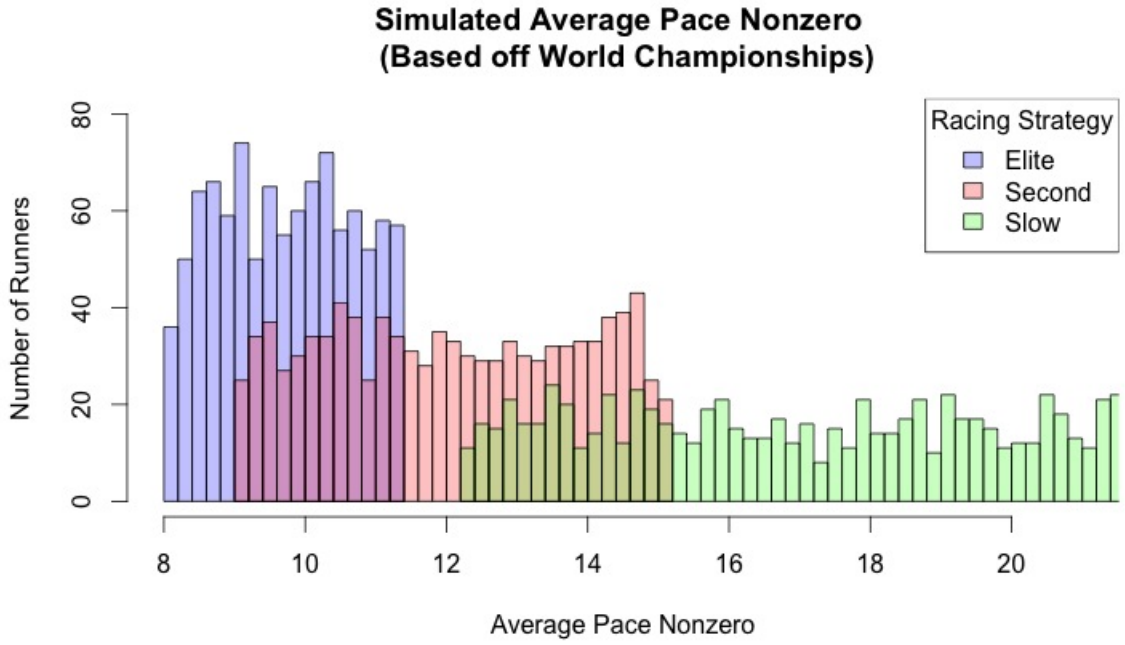
5.5 Simulation of World Championship Running Strategies

This simulation we ran was outlined by each strategy according to output from the *mclust* model. Throughout our analysis of the continuous and categorical variables using *mclust* and latent class analysis, we consistently were identifying an elite group, a group that was slower than the elite and less consistent, and then a group that was very inconsistent in terms of their characteristics. We based our simulation off of the normal mixture model result from Figure 3.3 but split the classifications into three categories. Therefore, this world championship running strategies simulation has dependence on our previous results and thus we would expect it to perform relatively well. In the simulation of distance runner strategies, we will construct our groups and data completely independent of the work we have done thus far. We have outlined the simulation of the world championship running strategies as follow:

1. **Strategy 1** (Elite, Consistent, Low Variability)

   Referencing Figure 3.3, we took the green classified cluster and claimed that this was the "elite" group and thus the basis for our Strategy 1. We claimed that these runners are the fastest runners (meaning their average pace nonzero is the fastest relative to the majority of the other runners). In addition, this group is also very consistent relative to the other runners as well. Going along with these two characteristics, the runners in this group have low variability from lap to lap.

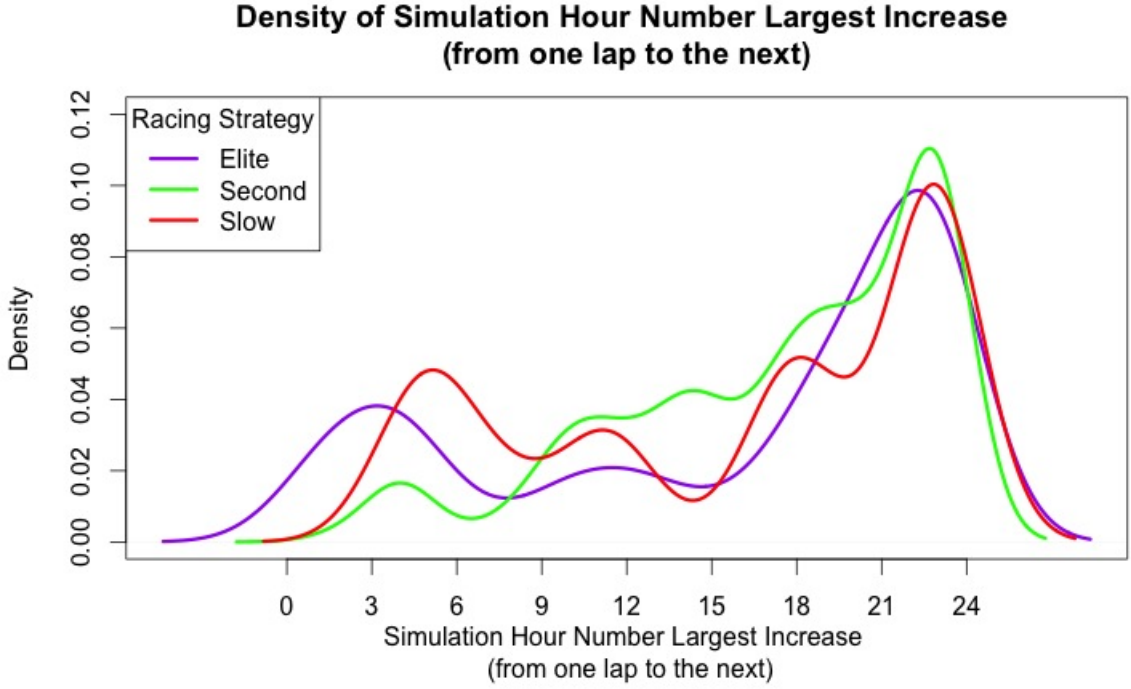
Figure 5.9: Distribution of runners' average pace nonzero simulated based off a uniform distribution where the minimum and maximum values are the minimum and maximum average pace nonzero of the elite group specified from the *mclust* model referring to Figure 3.3.



Our next step was to simulate the variables that we will be using in our *clustMD* model. We simply took the eight variables that we have been using throughout all simulations with the *clustMD* model. We simulated the continuous variables (average

pace nonzero, average pace during the night nonzero, and average pace during the day nonzero) from uniform distributions. We set our minimum and maximum bounds based off of the minimum and maximum paces ran from the "elite" cluster classifications from the *mclust* model. We removed outliers because we think that this will be accounted for in the variability of the uniform distribution model. For instance, Figure 5.9 shows the distribution of average pace nonzero from the simulation of all strategies.

Figure 5.10: Density of the simulated hour number largest increase from one lap to the next using the mixture models results for the basis of the simulation.



The rest of our variables were categorical and we simulated them by taking a sample of the category numbers with probabilities associated to the *mclust* category observation labels from the elite cluster. We looked at how many runners were in each of the categories and then found the probability of all of the runners to see the proportion

of runners who were in those respective categories. We used these probabilities as the probability vector when we were simulating our sample for the five categorical variables. This was done for the absolute value of most laps dropped, the most laps gained from one lap to the next, the hour number where the largest drop occurred, the hour number where the largest increase occurred, and the number of laps where the runner is running approximately the same pace. Figure 5.10 shows the distribution of the simulated hour numbers for largest increase from one lap to the next.

2. **Strategy 2** (Second, Somewhat consistent, High Variability)

   Similarly to strategy one, we looked at the mixture model results. This group has a slower average pace nonzero than the elite cluster and also a slower average pace during the day nonzero and night nonzero as well. In addition, they are less consistent runners and have higher variability in terms of their largest drops in the number of laps completed from one hour to the next as wells the largest increases.

   We came up with our simulated dataset by taking 1000 observations and simulating the eight variables in the same manner that we simulated strategy one.

3. **Strategy 3** (Slow, Inconsistent, High Variability)

   The third strategy takes on the three clusters that were not the elite or slow group from the normal mixture model results. The characteristics of these runners are that they have a much slower average pace nonzero and are highly variable in consistency and the other variables explaining the data. For instance, these runners have more categories for biggest increase and decrease in the number of laps from one lap to the next.

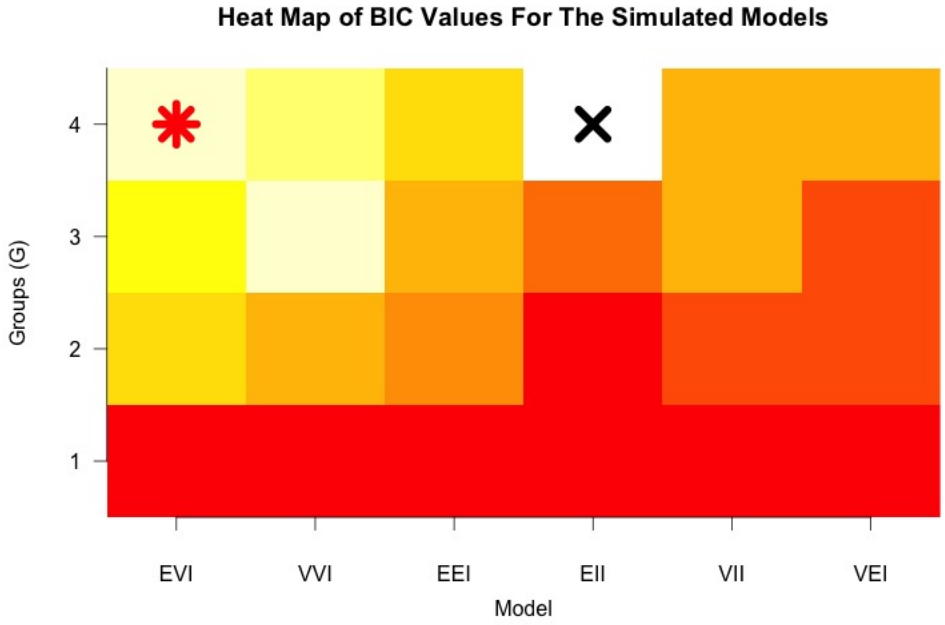
   We also simulated 1000 observations and the eight variables in the same manner that we simulated strategy one and two.

Table 5.3: Table of misclassification rates from the simulation of World Championship running strategies. This simulation was dependent on the results we obtained from our *mclust* models.

Observations	1:1000	1001:2000	2001:3000	Total
Misclass. Rate	1.00%	1.2%	3.7%	1.97%

The *clustMD* model fit appropriately as expected given that it returned a low misclassification rate as shown in Table 5.3. We are no longer having difficulty fitting the *clustMD* model for some of the models, but we still are for larger groups when looking into the World Championship Running simulation. From one to four groups, the "EII" model one model was the only model that could not fit four groups. Therefore, the new version of *clustMD* is fitting very well with the same or lower misclassification rate for all of the models. Based off of this simulation, we find that the best model is using the "EVI" model with four groups (Figure 5.11). We have shown that the new algorithm of *clustMD* seems to be estimating correctly given data that is based off of the continuous results. This aids us in our interpretation of the types of running strategies in the 24 Hour Ultra-Races.

Figure 5.11: This shows the BIC values for all models fit when using *clustMD* with the simulation data on the World Championship Data.

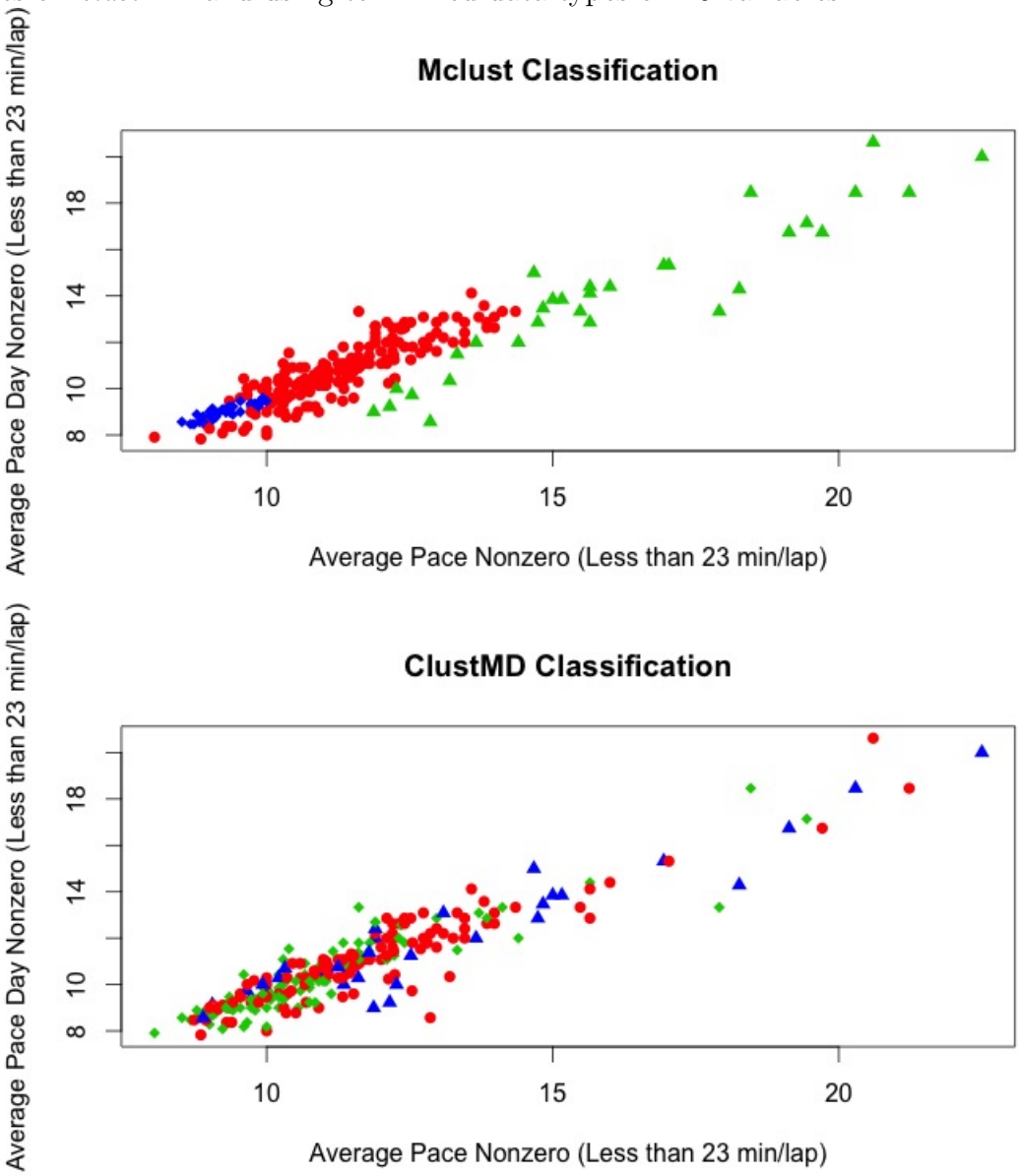


5.6 Mixture Models with Mixed Data Results

Notice the difference between the classifications of the plot on top and the plot on the bottom in Figure 5.12. The plot on the top was made by fitting model based clustering (*mclust*) on two variables. These two variables were average pace nonzero running less than 23 minutes per laps and the average pace during the nonzero running less than 23 minutes per lap. The result shows well distinguished clusters in two dimensions, however there is more information that we have in our dataset that has not been accounted for. Now, taking a look at the bottom plot, we notice that there are the same two dimensions but the classifications are different. This plot came from classifications made using mixed data types with *clustMD*. We would expect that the classifications would not be the same because we used eight different variables to fit the mixed data effect model. If we were to plot all eight variables in a ten

dimensional space, we would hope to find distinct clusters like the top plot. In summary, this plot is showing the difference of the two methods and how more explanation can lead to very different results looking in a reduced dimension space.

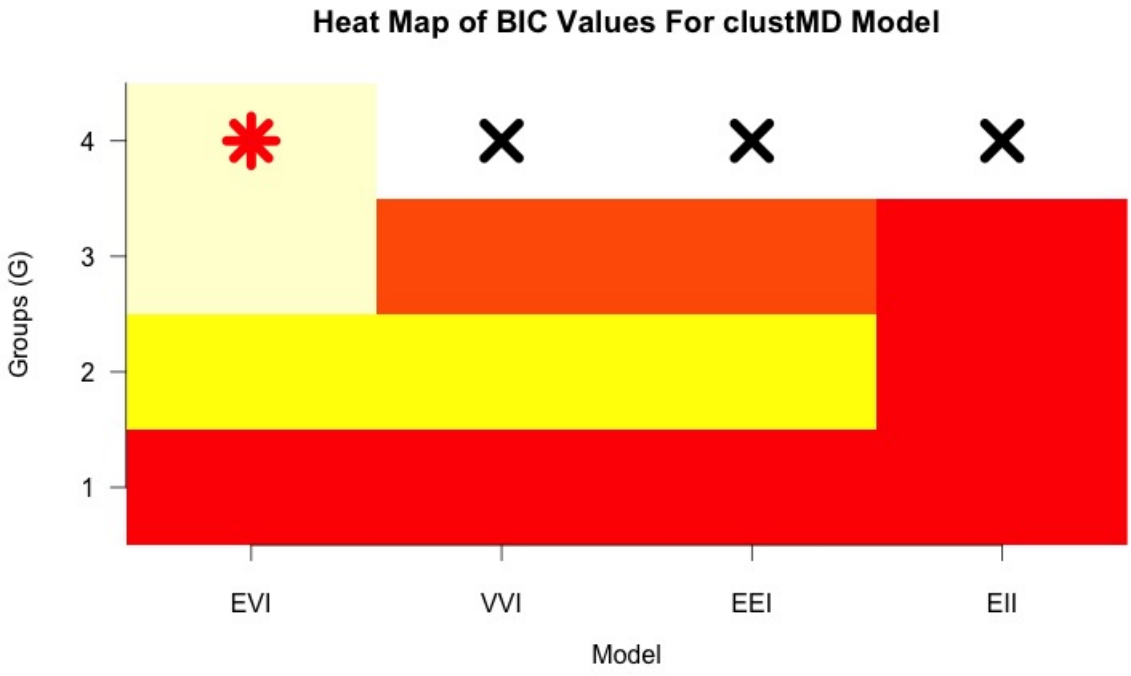
Figure 5.12: This is showing the comparison of the cluster classifications from using two variables on *clustMD* and using ten mixed data types on 10 variables.



We can see from Figure 5.13 that all of the models estimate better as the number of groups used goes up. However, once we reach five groups or in some cases four, the estimation

procedure stops working. In the simulated data we were able to estimate better because we had more observations. Part of future work will be to change the estimation so that we can estimate when the number of groups is greater than four. In addition, all of the models do not work when using the actual dataset of 248 observations. There are several reasons why the estimation procedure most likely stopped working but it most likely is because there are too many parameters to fit in the model given the small amounts of data we have. Figure 5.13 easily allows us to compare the models and their respective fitted BIC values and based off of the visualization, we can easily tell what model is the best. The model with the red star, EVI with four groups, had the highest BIC in absolute value. It is important to note that in this figure, the three squares that have an x on them could not be fit.

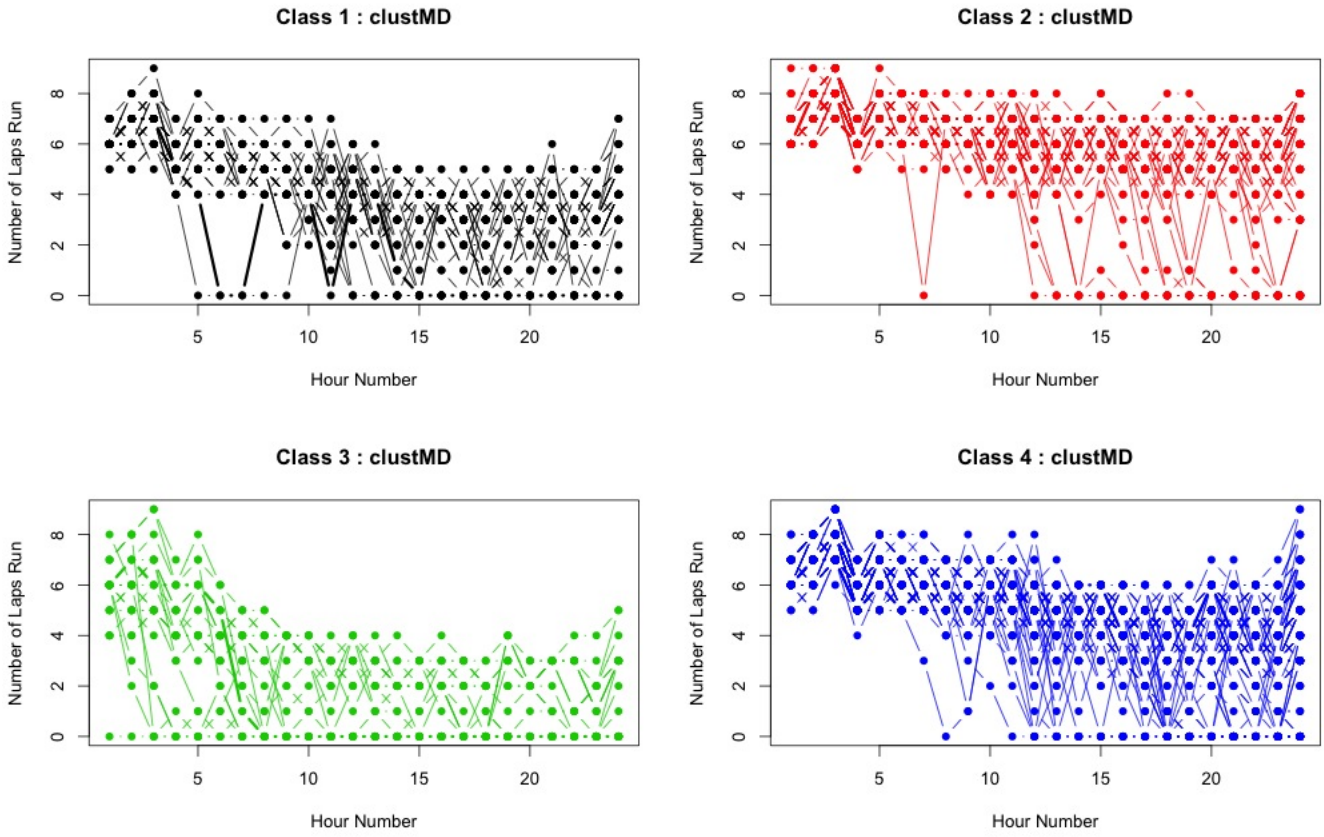
Figure 5.13: The heat map shows what model with group number G is the best fit based off of the adjusted *clustMD* algorithm. The square with the red star is the best model and the three 'X' models were not fit at all.



Looking at the results of the heat map for the BIC values for the fitted models, we chose

to fit the "VVI" model with four groups and then visualized the results by plotting the trajectories. Figure 5.14 breaks the runners down into four different groups, a number which we were seeing when we were running our *mclust* and latent class analysis models.

Figure 5.14: *clustMD* (VVI Four Groups) Results: Running trajectories for each class, where the classes were made from the EVI model with four groups using *clustMD*



The green group is definitely the slowest group and contains the runners who drop out. The red group (class two) looks like the elite group. The black group and blue group are very similar and definitely are the middle two groups in terms of pacing. It seems that the blue group has more variability in the middle and late hour numbers. It seems that the largest drops and largest increases and the time of when these occur distinguish these two groups. It seems that the blue cluster or class four has their largest increase in the race during the 23rd hour, whereas the black group seems to have their largest increase all throughout the

race. Also, it seems that the black cluster has largest drops around the 12 to 14 hour mark and the 17-18 hour mark whereas the blue cluster varies much more in terms of their largest drops.

5.7 Interpretability Comments

Interpreting the means returned from the *clustMD* model remain a challenge. The way in which the package chooses a reference variable and the threshold parameters that distinguish which group/category a runner is in makes it very difficult to see the difference between the runners. A plot like Figure 5.14 give us one the best visual pictures of how the *clustMD* model is breaking up the runners into different groups. Based off of that plot, it seems that we are getting both continuous and nominal/ordinal conclusions based off of how the runners are being clustered. We can tell the difference in the pace of runners from the slowest group to the fastest group, but in order to distinguish the runners who are in the middle two groups, we need to look at some of the nominal and ordinal variables.

Using the definitions of the nominal and ordinal variables we calculated, we can scan the trajectories in Figure 5.14 to get a general sense of how that class was constructed. We can perform exploratory data analysis on the runners in each of their respective clusters to get a better understanding as well. We constructed our data simulations to see how the *clustMD* model would react when we had more data and as we saw, the misclassification rate was very good and we were able to estimate using all six of the possible models. Interpreting the nominal variables remain difficult because the mean values returned depend on which category is the reference group. Our discussion indicated that the maximum category in the reference group (that is if there are three categories in a nominal variable, three would be the reference category) in a nominal variable.

Chapter 6

Conclusions and Future Work

6.1 Conclusions

Throughout our analysis of running data from the 24 Hour World Championships, we were able to make several conclusions. We were able to understand how the use of different strategies can be used to optimize race performance through analyzing continuous, ordinal, and nominal data. The use of a Gaussian mixture model allowed us to analyze the continuous data by clustering. Our mixture model results showed three or four main strategies/groups of runners. The cluster results showed there was an elite group who ran very consistently throughout the entire race, one or two "middle" groups, which were less consistent than the elite group and paces were more variable than the elite group. Finally, there was a slow group or group of runners who stopped running at some point in the race. However, we could only analyze continuous data in the mixture model framework so we used latent class analysis to analyze categorical variables. We can conclude that the latent class analysis did not identify the three or four clusters we found from our Gaussian mixture model, but broke down the classes where the time frame in which a runner picked up their pace, dropped their pace, fluctuated their pace, dropped out, and stopped running. We were able to conclude

about a runner's strategy through mixture models and latent class analysis, but we needed to find a way to combine the estimation of continuous and categorical variables.

McParland and Gormley recently released the idea of using mixture models with mixed data to allow the combination of analysis on continuous, ordinal, and nominal variables. We used their framework but extended the estimation capability. The estimation of nominal variables was constructed in a way where we could not have a nominal variable that had an empty category once the Monte Carlo simulations started. We claimed that if the probability of not being in a class is zero, then this should give us more information on that combination of estimation. As a result, we adjusted the *clustMD* framework to set the probability to approximately zero and adjusted the remaining probabilities for each nominal category not equal to the empty probability. Changing the way the Monte Carlo samples were simulated allowed the new version of *clustMD* to fit appropriately. Our results with this new extension of *clustMD* allowed us to fit more types of models as well as have very good misclassification errors.

We concluded that the continuous pace variables explain the majority of the variability in the models that we are fitting compared to the nominal and ordinal variables through our newly extended mixture models with mixed data framework. However, we noticed the nominal and ordinal variables break down groups that are nearly indistinguishable when only looking at the different pace variables.

6.2 Future Work

We plan to fit a mixture of Poisson processes to cluster runners based on the number of laps completed over time. The particular combination with the Poisson process is not well-studied

or implemented often. This method could turn large scale if found successful. We can further extend this proposed approach for any application that centers on a count data trajectory over a certain period of time. We would cluster the actual trajectories of counts. In our 24 hour race application dataset, the number of laps ran is the unit of trajectory. For an example of another application, we could apply our approach using Poisson process mixtures to clinical depression scores over time. For instance, the unit in this case could potentially be the number of episodes the patient has in each time unit. A better understanding of predicting the different recovery (or non-recovery) trajectories will be very beneficial to both patients and clinicians.

In addition, we would want to try and see how we could approximate *clustMD* for higher than $k = 4$ or 5. This would involve looking further into the estimation issues and the problems in the methodologies of defining the nominal variables. We could also look into how the estimation changes if we change the reference category of the nominal variables. In addition, it would be useful to do further analysis in mixed memberships models.

Bibliography

[1] White, A. and Murphy, T. Exponential Family Mixed Membership Models for Soft Clustering of Multivariate Data.

[2] Fraley, C. and Raftery, A. Model-Based Clustering, Discriminant Analysis, and Density Estimation. *American Statistical Association*, June 2002.

[3] Fraley, C. and Raftery, A. Mclust Version 3 in R: Normal Mixture Modeling and Model-Based Clustering. *CRAN*, July 2010.

[4] McParland, D. and Gormley, C. Model Based Clustering for Mixed Data: clustMD, 2015.

[5] McLachlan, G. and Krishnan, T. The EM Algorithm and Extensions. Wiley, 2008.

[6] Dean, N. and Raftery, A.. Latent Class Analysis Variable Selection. *Ann Inst Stat Math*, February 2010.

[7] Haughton, D., Legrand, P., and Woolford, S. Review of Three Latent Class Cluster Analysis Packages: Latent Gold, poLCA, and MCLUST. *American Statistical Association Vol. 63, No. 1*, February 2009.

[8] Kass, R. and Raftery, A. Bayes factors. Journal of the American Statistical Association. *American Statistical Association 90(430)*, 1995. 773-795.

[9] Nugent, R. and Meila, M. An Overview of Clustering Applied to Molecular Biology. Springer/Humana Press, 2010.

[10] Reynolds, G. Why a Brisk Walk Is Better. http://well.blogs.nytimes.com/2013/12/04/why-a-brisk-walk-is-better/?_r=0. Accessed: 2015-04-30.

[11] Bandeen-Roche, K., Miglioretti, D., Zeger, S., and Rathouz, P. Latent Variable Regression for Multiple Discrete Outcomes. *Journal of the American Statistical Association*, 1997.