Performance of LDA and QDA on non-normally distributed predictors

Abstract

This paper explores whether LDA and QDA can extract useful information even from non-normally distributed variables in a binary classification setting by simulating a dataset, fitting LDA and QDA models, and evaluating the variable importance of each predictor included in the simulated dataset. As an example of a non-normal distribution, a Gamma distribution is used. In evaluating the variable importance, a model is fitted to the dataset after removing one predictor at each time and the model performance is evaluated with AUC values. The damage caused by the removal of the predictor is measured in terms of the drop in the AUC values, and the greater damage is interpreted as a sign of the greater importance of the variable. Through simulating whether a Gamma-distributed variable can be as important as a normally distributed variable in comparable situations, this paper attempts to examine the performance of LDA and QDA on non-normally distributed variables. LDA and QDA both proved its robustness against the predictor that follows a Gamma distribution through showing that they can extract meaningful information even from such a non-normal predictor. Some possible interpretations of the results and suggestions for future research are also provided.

1 Introduction

Classification problems in statistics are a branch of topics in supervised learning and encompass the prediction and analysis of categorical response variables. In other words, whether a person has a certain type of cancer, whether an athlete is going to be injured after a game, and whether a student will be admitted in a college are all classification problems statistics can help answer based upon the value of some explanatory variables (e.g. the number of certain cells or the amount of certain substances in a blood sample, the athlete's frequency of injury in the past few years, or the student's GPA and SAT scores, etc.).

Linear Discriminant Analysis (LDA) and Quadratic Discriminant Analysis (QDA) are two classic but still popular approaches to these problems. The models built with LDA and QDA take the value of predictors as inputs and predict which class an observation is likely to be coming from. After Fisher (1936) proposed the original version of discriminant analysis, these techniques have been extensively applied in a wide variety of areas such as operations research, climatology, medicine, and economics. Fisher himself developed discriminant analysis as a non-parametric method to most effectively discriminate between multiple classes of observations using the value of some predictors; thus, his original paper did not include any assumptions about the distribution of predictors used in the model. However, as described in Section 2, statisticians later discovered a way to interpret discriminant analysis as an implementation of a Bayes' classifier and concluded that the efficacy of LDA and QDA somewhat relies on the normality of the distribution of predictors.

LDA and QDA are great alternatives for solving classification problems when the logistic regression models do not perform well for some reasons. For example, there might be too clear-cut a separation among different classes or the dataset might have more than 2 classes to be discriminated (James et al. 2021). Hence, it is worthwhile to study how these methods perform when the dataset is not satisfying the assumption of normality. Specifically, can LDA and QDA extract important information from non-normally distributed explanatory variables? Put differently, in what kind of situations does removing a non-normally distributed predictor damage the model performance of LDA or QDA more than removing a normally distributed predictor?

This paper answers these questions by simulating datasets with non-normally distributed variables with different relationships to the response and evaluating the performance of models built with LDA and QDA on those simulated datasets. First, in Section 2, a brief overview of the theory behind LDA and QDA is provided along with the description of several actual application of the method to other disciplines. Then, Section 3 describes the process of simulation and reports the results. In Section 4, this paper discusses those results while suggesting possible avenues for future inquiries, and the final section offers some concluding remarks.

2 Literature Review

As mentioned in the introduction, LDA and QDA are currently theorized as an implementation of Bayes' Classifier, an approach to classification problems based upon the Bayes' theorem. Thus, in this section, this paper starts with a brief exposition of the fundamental idea of Bayes' classifier. Next, it illustrates how LDA builds a model based upon the concept of Bayes' classifier in the case of $k = 2$, where k denotes the number of classes in the dataset to which an observation could potentially belong. Afterwards, it discusses the theoretical difference of LDA from logistic regression models, another common approach to a binary classification problem, and the expansion

of LDA into problems where $k > 2$. Finally, it moves onto the review of QDA as a brief modification of LDA, followed by the discussion of some possible application of LDA and QDA to the real-world contexts.

2.1 Bayes' Classifier

A Bayes' Classifier is an approach to the classification problem which assigns a case into a class i where the posterior probability $P(Y = i|X = x)$ takes the greatest value (Izenman 2008). Note that in this notation Y denotes the class which the observation belongs to. Because of the Bayes' theorem, it can easily be proven that

$$P(Y = i|X = x) = \frac{P(Y = i)P(X = x|Y = i)}{P(X = x)} \quad (2.1)$$

Further, since each of the k classes are assumed to be non-overlapping,

$$\begin{aligned} P(X = x) &= \sum_{l=1}^{k} P(X = x \ \cap \ Y = l) \\ &= \sum_{l=1}^{k} P(Y = l)P(X = x|Y = l) \quad (2.2) \end{aligned}$$

Applying (2.2) to (2.1) and defining $P(Y = i) = \pi_i$ and $P(X = x|Y = i) = f_i(x)$, as explained by James et al. (2021), the Bayes' theorem for a classification problem can be written as

$$P(Y = i|X = x) = \frac{\pi_i f_i(x)}{\sum_{l=1}^{k} \pi_l f_l(x)} \quad (2.3)$$

Equation (2.3) suggests one concrete procedure to utilize the Bayes' classifier in the real-world data analysis. If there is a way to estimate π_i and $f_i(x)$ from the sample data, it is no longer necessary to directly try to estimate the posterior probability $P(Y = i|X = x)$; using the Bayes' theorem, the estimated values of each component of the right-hand side can be plugged in and thereby used to generate the probabilities for the categorization based upon the Bayes' classifier. Usually, π_i can be estimated from the sample with the formula $\widehat{\pi}_i = \frac{n_i}{n}$, where n denotes the sample size and n_i the number of observations in the sample which belongs to class i (James et al. 2021). Therefore, the central question in applying the Bayes' classifier based upon the equation (2.3) is how to estimate $f_i(x)$ with the sample data at hand.

2.2 Linear Discriminant Analysis for $k = 2$

Linear Discriminant Analysis (LDA) offers a simple but useful way to estimate $f_i(x)$, enabling the implementation and application of the Bayes' classifier. Concretely, LDA starts with the assumption that the conditional distribution of the vector of predictors $X|Y = i$, where $i \in \{1, 2, \ldots, k\}$ follows a multivariate Gaussian Normal distribution with its means different for each class and the covariance matrix shared by all the classes (James et al. 2021).

For the sake of simplicity, this section focuses on the case of $k = 2$ (binary classification problem). Thus, suppose there are two categories, category 1 and category 2, and it is needed to classify a case with $X = x$ into one of these two classes. The Bayes' classifier, as reviewed above, classifies this case into category 1 if the value of $P(Y = 1|X = x)$ is greater and into category 2 otherwise.

Equivalently, however, it is also possible to write this rule as following (Izenman 2008):

Classify the observation into 1 if $L(x) > 0$, and into 2 if $L(x) < 0$, where $L(x) = log\left(\frac{f_1(x)\pi_1}{f_2(x)\pi_2}\right)$.

Recall that it is assumed that $X|Y = i \sim N(\mu_i, \Sigma_X)$ for $i = 1, 2$, where Σ_X stands for the covariance matrix common between the two categories. Plugging the probability density function of multivariate Gaussian distribution into the aforementioned $L(x)$, the function can be rewritten as follows (Izenman 2008).

$$\begin{aligned}
L(x) &= log\left(\frac{f_1(x)\pi_1}{f_2(x)\pi_2}\right) \\
&= log\left(\frac{f_1(x)}{f_2(x)}\right) + log\left(\frac{\pi_1}{\pi_2}\right) \\
&= log\left(\frac{\exp(-\frac{1}{2}(x-\mu_1)^T\Sigma_X^{-1}(x-\mu_1))}{\exp(-\frac{1}{2}(x-\mu_2)^T\Sigma_X^{-1}(x-\mu_2))}\right) + log\left(\frac{\pi_1}{\pi_2}\right) \\
&= -\frac{1}{2}(x-\mu_1)^T\Sigma_X^{-1}(x-\mu_1) + \frac{1}{2}(x-\mu_2)^T\Sigma_X^{-1}(x-\mu_2) + log\left(\frac{\pi_1}{\pi_2}\right)
\end{aligned}$$

Here, some algebra allows the reexpression of the first two terms in the following form (Izenman 2008).

$$\begin{aligned}
-\frac{1}{2}(x-\mu_1)^T\Sigma_X^{-1}(x-\mu_1) + \frac{1}{2}(x-\mu_2)^T\Sigma_X^{-1}(x-\mu_2) &= (\mu_1-\mu_2)^T\Sigma_X^{-1}x - \frac{1}{2}(\mu_1-\mu_2)^T\Sigma_X^{-1}(\mu_1+\mu_2) \\
&= (\mu_1-\mu_2)^T\Sigma_X^{-1}(x-\bar{\mu})
\end{aligned}$$

In other words, by assuming the normality of predictors and the commonality of covariance matrices among between the two categories, LDA simplifies the Bayes' classifier into the following rule written as a linear function of predictors (Izenman 2008):

Classify the observation into category 1 if $L(x) > 0$, and into 2 if $L(x) < 0$, where

$$\begin{aligned}
L(x) &= (\mu_1-\mu_2)^T\Sigma_X^{-1}x - (\mu_1-\mu_2)^T\Sigma_X^{-1}\bar{\mu} + log\left(\frac{\pi_1}{\pi_2}\right) \\
&= b_0 + b^T x
\end{aligned}$$

$$b_0 = -(\mu_1 - \mu_2)^T \Sigma_X^{-1} \bar{\mu} + log\left(\frac{\pi_1}{\pi_2}\right)$$
$$b = \Sigma_X^{-1}(\mu_1 - \mu_2)$$

This classification rule is called Linear Discriminant Analysis (LDA) or Gaussian Linear Discriminant Analysis, with the term $(\mu_1 - \mu_2)^T \Sigma_X^{-1} x$ called Fisher's linear discriminant function (Izenman 2008). When Fisher (1936) originally developed this classification method in a non-parametric manner, he attempted to separate the two or more classes of observations in such a way as to maximize the difference of the explanatory variables' group means in proportion to their standard deviations within the groups. However, today's statistics usually interprets LDA as a method built on the idea of the Bayes' classifier, which means that LDA is currently used as a supervised learning method predicated on the assumption of normality of the explanatory variables.

In order for LDA to be actually used for building a model which tackles the classification problem, it is necessary to estimate the following parameters of the above function $L(x)$: μ_1, μ_2, Σ_X, π_1, and π_2. For π_i's, it is permissible to estimate these values based upon the prior beliefs and subject knowledge, but if no available information is available besides the sample data, they are estimated as $\widehat{\pi}_i = \frac{n_i}{n}$ (Izenman 2008). For the other parameters, Izenman (2008) provides the following estimates as one of the most plausible options:

$$\begin{aligned}
\widehat{\mu_i} &= \bar{x_i} \\
&= \frac{1}{n_i} \sum_{l=1}^{n_i} x_{il} \text{ for } i = 1, 2 \\
\Sigma_x &= \frac{1}{n-2} S_x, \text{ where} \\
S_x &= S_x^{(1)} + S_x^{(2)} \\
S_x^{(i)} &= \sum_{l=1}^{n_i} (x_{il} - \bar{x_i})(x_{il} - \bar{x_i})^T
\end{aligned}$$

For Σ_x, the maximum likelihood estimate is actually $\frac{1}{n} S_x$, but reducing the denominator by the number of categories can adjust the estimator so that it is unbiased (Izenman 2008).

To sum up, by assuming that all the predictors follow the normal distributions and that each class has its own mean values of X while sharing the common covariance matrix of predictors, LDA defines a specific linear function of X which in turn helps estimate the decision boundary of the Bayes' classifier. By doing so, it develops a model to classify observations based upon the explanatory variables and creates a solution to the binary classification problem.

2.3 Comparison to Logistic Regression when $k = 2$

From the definition of $L(x)$ above, it is not difficult to identify the similarity between LDA and logistic regression models, another popular approach to a classification problem. Logistic regression models estimate the coefficients β_0 and β in the following equation (Izenman 2008):

$$
\begin{aligned}
logit(P(Y=1|X=x)) &= log\left(\frac{P(Y=1|X=x)}{1-P(Y=1|X=x)}\right) \\
&= \beta_0 + \beta^T X
\end{aligned}
$$

Here, it is assumed that $k = 2$ and, therefore equivalently, $1 - P(Y = 1|X = x) = P(Y = 2|X = x)$ (remember that the problem requires to assign each case to either one of the category 1 and 2). Hence, the following mathematics proves that the left-hand side is equal to $L(x)$ for the implementation of LDA (Izenman 2008).

$$
\begin{aligned}
logit(P(Y=1|X=x)) &= log\left(\frac{P(Y=1|X=x)}{P(Y=2|X=x)}\right) \\
&= log\left(\frac{\frac{\pi_1 f_1(x)}{\pi_1 f_1(x)+\pi_2 f_2(x)}}{\frac{\pi_2 f_2(x)}{\pi_1 f_1(x)+\pi_2 f_2(x)}}\right) \\
&= log\left(\frac{f_1(x)\pi_1}{f_2(x)\pi_2}\right) \\
&= L(x)
\end{aligned}
$$

Thus, the above algebra shows that, ultimately, logistic regression models and LDA models are estimating the same function when $k = 2$ (Izenman 2008). The important difference between them, however, lies in their theoretical assumptions. Logistic regression models assume that $L(x)$ can be modeled as a linear function and estimate the coefficients of the function with maximum likelihood estimators, whereas LDA assumes the normality of predictors and the commonality of the covariance matrix between the two classes, which consequently makes it possible to write $L(x)$ in terms of the linear combination of all the explanatory variables (Izenman 2008).

Due to this difference in the assumptions, logistic regression and LDA has different strengths and weaknesses. In general, logistic regression models perform much better when the predictors are not normally distributed because their model fit does not rely on the parametric assumption of explanatory variables (Izenman 2008). In contrast, LDA requires a smaller sample size when the classes can be clearly separated and it is also asymptotically more efficient than logistic regression if the conditions of normality and common covariance matrix are satisfied (Izenman 2008). In short, though LDA and logistic regression has certain mathematical similarity, these two methods are theoretically justified in quite different ways, which makes their shortcomings also different from each other.

2.4 Linear Discriminant Analysis for $k > 2$

LDA can also be applied to a context with multiple classes to which an observation could be categorized. The underlying assumption of the model is exactly the same as the case where $k = 2$; in other words, it is presumed that all the classes share the common covariance matrix and all the predictors follow the normal distributions (James et al. 2021). According to the Bayes' classifier, each case with $X = x$ is assigned to the class k' such that

$$P(Y = k'|X = x) = \frac{\pi_{k'} f_{k'}(x)}{\sum_{l=1}^{k} \pi_l f_l(x)}$$
$$= \frac{\pi_{k'} \exp(-\frac{1}{2}(x - \mu_{k'})^T \Sigma_X^{-1}(x - \mu_{k'}))}{\sum_{l=1}^{k} \pi_l f_l(x) \exp(-\frac{1}{2}(x - \mu_l)^T \Sigma_X^{-1}(x - \mu_l))}$$

takes the maximum value, where π_l, μ_l, and Σ_X $(l = 1, 2, \ldots k)$ are all estimated in an analogous manner to the method for $k = 2$ (James et al. 2021). Though further clarification of the algebra behind the construction of decision boundaries is out of the scope of this paper, it is clear that the idea of Bayes' classifier helps easily apply LDA into contexts where $k > 2$, which logistic regression models are not good at dealing with.

2.5 Quadratic Discriminant Analysis

QDA is only a slight modification of the LDA in a sense that it classifies observations in the same way but with only one assumption removed from the process: QDA does not assume that all the classes share an identical covariance matrix (James et al. 2021). Thus, instead of applying a single covariance matrix Σ_X to all the calculations of the conditional probability $P(Y = l|X = x)$, QDA classifies each case to the class k' such that the value of

$$P(Y = k'|X = x) = \frac{\pi_{k'} f_{k'}(x)}{\sum_{l=1}^{k} \pi_l f_l(x)}$$
$$= \frac{\pi_{k'} \exp(-\frac{1}{2}(x - \mu_{k'})^T \Sigma_{k'}^{-1}(x - \mu_{k'}))}{\sum_{l=1}^{k} \pi_l f_l(x) \exp(-\frac{1}{2}(x - \mu_l)^T \Sigma_l^{-1}(x - \mu_l))}$$

will be maximized (James et al. 2021). Here, note that the covariance matrix is no longer denoted as Σ_X because it is not common among all the classes. This difference makes QDA more prone to the variability between samples because of its possibility of over-fitting (James et al. 2021). Since LDA only allows one covariance matrix to be applied to all the classes, it is much less flexible as a model estimator, although that also means that there is less variance between LDA models constructed with different sample data. In contrast, QDA adjusts its covariance matrices to each class, making the estimator much more flexible but more vulnerable to the problem of over-adjusting to the samples. Thus, a basic rule of thumb is to prefer QDA to LDA either when the sample size is quite large and the variance between different classifiers built with different samples is not a major concern or when the commonality of the covariance matrices is a clearly unreasonable assumption (James et al. 2021).

2.6 Applications of Discriminant Analysis

These methods of discriminant analysis have been extensively applied in multiple academic disciplines. For example, one of the most classic examples of the application of discriminant analysis is an article written by Altman (1968). In this article, Altman explores the way to predict the future bankruptcy of private enterprises using financial ratios. Back in 1960's, discriminant analysis offered a great solution to the problem of needing to examine each financial ratio one by one

by enabling the construction of a classification model which takes all the variables into account simultaneously. Through using discriminant analysis and constructing a Z score as a linear combination of various financial ratios often important in the realm of corporate finance, Altman (1968) discovers a way to accurately predict the future bankruptcy of companies with those ratios.

However, the application of discriminant analysis is not limited to the field of corporate finance; climatology research also extensively applies discriminant analysis to distinguish the severity of certain weather events or building a variable which may give a more nuanced representation of certain climate phenomena (Li and Colle 2016). For instance, Li and Colle (2014) define Convective Environment days through applying LDA to predict the occurrence of convective precipitation and classifying each day with the value of the discriminant function defined above. By doing so, Li and Colle invents a way to categorize weather events of each day based upon the likelihood of convective precipitations rather than its actual occurrence.

In addition, LDA is sometimes used in a variety of fields such as sports medicine, archaeology, and operations research. Schilaty et al. (2020) utilizes LDA in order to classify the outcomes of knee injuries based upon demographic and biomedical variables. By doing so, these researchers exhibit a potential for LDA to build a clinically valid model which produces insights about injury prevention for athletes in the future. Also, Buchanan, Mraz, and Eren (2016) applies LDA to successfully classify past archaeological stone tools into those created through pressure flaking versus those through soft hammer percussion. Ünsal and Nazman (2020) rank the cities in Turkey with LDA through first developing a binary categorical variable about the efficiency of those cities and then apply LDA to predict that newly developed binary variable and thereby generate numerical scores which in turn give a more nuanced information about the efficiency of each city. As shown by these examples, because of its ability to both produce numerical outcome variables with the discriminant function and classify observations based on the value of that function, discriminant analysis is quite rigorously used as an approach to solve problems in these various disciplines.

To sum up, its creation as a method dating back to 1930's, LDA and QDA are still extensively applied to a wide range of disciplines not only to predict categorical response variables but also to construct variables based upon the value of the linear discriminant function above. They are classic but still very important and useful approaches for the binary classification problems.

3 Simulation and Results

As shown by the previous section, LDA and QDA are approaches to classification problems predicated on the assumption of normality of the explanatory variables. Thus, it is intriguing to explore whether these methods can extract useful information even from non-normally distributed predictors, and if so, to what extent they can do that. This chapter offers a detailed view on how this paper approached this question with simulation as well as the results of that simulation. First, it explains the basic procedures of the simulation and why those procedures are reasonable or important. Then, it moves onto the description of the concrete conditions used for the simulation with LDA models and the report of its results, followed by the same set of information for the simulation with QDA models. At the end, it summarizes the overall results of the simulation so that this paper can develop further insights on those results in the next section.

3.1 Basic Procedures and Structures of the Simulation

In order to answer its central question, this paper simulates a dataset which contains the following five variables: a response variable Y which represents the class to which the observation belongs to, three control explanatory variables X_1, X_2, and X_3 that are always normally distributed, and the other explanatory variable X^*, which this simulation study manipulates to change the condition of the simulations. Since this paper plans to explore whether discriminant analysis can extract meaningful information even from the non-normally distributed variables, X^* is set to follow Gaussian distributions in some scenarios and Gamma distributions in others. By comparing the importance of X^* between these cases with different distributions, this report tries to examine to what extent models built with discriminant analysis can extract useful information even from non-normally distributed variables. The size of the dataset is fixed at $n = 360$ throughout all the simulations and the dataset is made to contain two classes (in other words, $k = 2$ for the rest of this paper). Since there is no reason to do otherwise, these two classes are set to be balanced. The observations are randomly assigned to one of these two classes in each iteration of the simulation.

This paper's simulation code determines the value of each observation's explanatory variables by executing the following two steps: first, it decides what probability distribution each predictor should follow for each class. Since the simulation process addresses four predictors and two classes in total, it means that this simulation code needs 8 probability distributions in this first step. For each variable, the probability distribution is set to come from the same distribution family (either Gaussian or Gamma), but the parameters of the distribution are changed depending on the class. Then, it makes exactly 180 draws from each probability distribution to assign specific values to each cell of the dataset.

After generating the dataset through the above process, the simulation code fits discriminant analysis models to its dataset and produces predictions which are in turn used to evaluate the model performance. Here, the data are split into training and test set, and the training set is made to contain 240 observations and the test set 120. In executing this split, it is guaranteed that both the training and test sets will have exactly the same number of observations of the two classes. Also, in each iteration of the simulation, the code fits 5 models to the dataset: one with all the predictors and the other four which do not have one of the four predictors (X_1, X_2, X_3, and X^*). By measuring how much the model performance deteriorates when it loses each of the predictor from the full set, this simulation produces a value that represents the importance of those four predictors.

In assessing the model performance, this paper utilizes AUC values generated from the ROC curve. The AUC value represents the area under the ROC curve and is always included in $[0, 1]$. Hence, basically, this value denotes how much compromise the model needs to make about the increase in the false positive rate to improve the true positive rate of the prediction, and the higher value means the greater predictive performance by the model. For each iteration of the simulation, since there are five models constructed and evaluated, there are five AUC values produced. The simulation code records these values every time after completing each iteration of the process of data generation, model fitting, and model evaluation. For each scenario with different parameters used for the distribution of predictors, this simulation research runs 5000 times of this process, generating a dataset of model performance to be examined later.

The basic idea behind this procedure is to represent and manipulate the usefulness of the four predictors, X_1, X_2, X_3, and X^* through its probability distribution's means and variances. Intuitively, it makes sense to assume that a predictor with a larger distance between means for different classes

is going to be more useful if its variance is always the same. Relying on this assumption, this paper examines whether a variable which should be useful given the distance between the means of different classes can lose its power when it is following a Gamma distribution instead of a normal distribution. By doing so, it attempts to evaluate the ability for discriminant analysis to utilize information contained in the non-normally distributed variables.

3.2 Simulation: LDA

3.2.1 Background and Settings of the Simulation

As explained above, the central purpose of this simulation is to compare the importance of X^* between when it is following a Gaussian distribution (called control scenarios) and when it is following a Gamma distribution (called treatment scenarios), examining how much taking X^* out of the model on average damages the model performance in each case. Thus, it is important to make those two scenarios comparable by at least using the identical set of values for the means and variances of those distributions. For that sake, in this part, this simulation only uses 10 probability distributions shown in Table 1 to generate predictors in the dataset.

Since predictors should follow different distributions depending on the class of the observation, this paper's simulation code needs to select 2 distributions for one predictor in each iteration of the simulation. For the control scenarios (X^* follows a Gaussian distribution), it always selects Control 1 distribution for one class, and uses one of the other four distributions for the control cases (Control 2, 3, 4, and 5) for the other class. In other words, there are four pairs of distributions for X^* which this paper needs to simulate for 5000 iterations for the control scenarios. For the treatment scenarios (X^* follows a Gamma distribution), similarly, it selects Treatment 1 distribution for one class, and uses one of the other four Gamma distributions (Treatment 2, 3, 4, and 5) for the other class, which again means there are four pairs of distributions needed to be simulated for 5000 times.

Table 1: Distributions used for the Simulation (LDA)

Distribution Name	Distribution Type	Shape	Rate	Mean	Variance
Control 1	Gaussian	N/A	N/A	$\frac{1}{\sqrt{3}}$	$\frac{1}{3}$
Control 2	Gaussian	N/A	N/A	$\frac{3}{2\sqrt{3}}$	$\frac{1}{3}$
Control 3	Gaussian	N/A	N/A	$\frac{2}{\sqrt{3}}$	$\frac{1}{3}$
Control 4	Gaussian	N/A	N/A	$\frac{5}{2\sqrt{3}}$	$\frac{1}{3}$
Control 5	Gaussian	N/A	N/A	$\frac{3}{\sqrt{3}}$	$\frac{1}{3}$
Treatment 1	Gamma	1	$\sqrt{3}$	$\frac{1}{\sqrt{3}}$	$\frac{1}{3}$
Treatment 2	Gamma	$\frac{9}{4}$	$\frac{3\sqrt{3}}{2}$	$\frac{3}{2\sqrt{3}}$	$\frac{1}{3}$
Treatment 3	Gamma	4	$2\sqrt{3}$	$\frac{2}{\sqrt{3}}$	$\frac{1}{3}$
Treatment 4	Gamma	$\frac{25}{4}$	$\frac{5\sqrt{3}}{2}$	$\frac{5}{2\sqrt{3}}$	$\frac{1}{3}$
Treatment 5	Gamma	9	$3\sqrt{3}$	$\frac{3}{\sqrt{3}}$	$\frac{1}{3}$

As can be seen, the distance between the mean of the Control 1 distribution and that of the Control 2 distribution is intentionally set the same as that between the mean of the Treatment 1

and Treatment 2. The same principle applies to all the other pairs (Control 1 and Control 3 versus Treatment 1 and Treatment 3, and so forth). Thus, by comparing how important X^* is between the scenario where the Control 1 and Control 2 are used and that where Treatment 1 and Treatment 2 are used, and conducting the same examination for all the other pairs, it is possible to explore how difficult it becomes for LDA to extract information from a predictor when it comes to follow a Gamma distribution instead of a Gaussian distribution.

For the other three predictors X_1, X_2, and X_3, they are made to follow the following distributions depending on the class to which the observation belongs to.

$$X_1 \sim N(-0.2, \frac{1}{3}) \text{ for one class and } X_1 \sim N(0.2, \frac{1}{3}) \text{ for the other}$$
$$X_2 \sim N(-0.36, \frac{1}{3}) \text{ for one class and } X_2 \sim N(0.36, \frac{1}{3}) \text{ for the other}$$
$$X_3 \sim N(-0.5, \frac{1}{3}) \text{ for one class and } X_3 \sim N(0.5, \frac{1}{3}) \text{ for the other}$$

As can be seen, the distance between the means of the two different distributions is 0.4 for X_1, 0.72 for X_2, and 1 for X_3. These values are chosen intentionally. That between the means of Control 1 and Control 2 or Treatment 1 and Treatment 2 is about 0.289, that between those of Control 1 and Control 3 or Treatment 1 and Treatment 3 is about 0.577, that between those of Control 1 and Control 4 or Treatment 1 and Treatment 4 is about 0.866, and that between those of Control 1 and Control 5 or Treatment 1 and Treatment 5 is about 1.15. Thus, by design, the usefulness of X^* as a predictor measured by the distance between means is made to surpass one of the three predictors X_1, X_2, and X_3 each time it is improved by enlarging the distance between the means of the two distributions for X^*. Note that, since the variance of the distribution is set to the same value throughout the simulation, there is no way X^* can lose its usefulness as a predictor because its probability distribution is more spread out.

Also, for all the four predictors, X_1, X_2, X_3, and X^*, it is randomly decided which of the two classes should follow each of the two distributions respectively for each iteration of the simulation. This randomization is conducted in every single simulation executed for the inquiry of this paper. For further inquiries about the code and replication of the results, it is recommended to contact the author and acquire the `rmd` file of this paper.

3.2.2 Results: Control Scenarios

The four density plots in Figure 3 and 4 show the distributions of the AUC values for the five LDA models constructed under the control scenarios. The graph title represents the two distributions used for X^*, and as shown by the legend, the navy blue line is the distribution of AUC values of the model without X^*.

In all the graphs, the green line corresponds to the model with all the four predictors, which means that the distance between this green density plot and the other plots represents the extent to which AUC values on average deteriorated as a result of the removal of the corresponding predictor. Thus, a greater degree to which the graph shifted from the location of the green line to the left is equivalent to the greater damage done by taking the predictor out of the model, which in turn means a greater importance of the predictor.

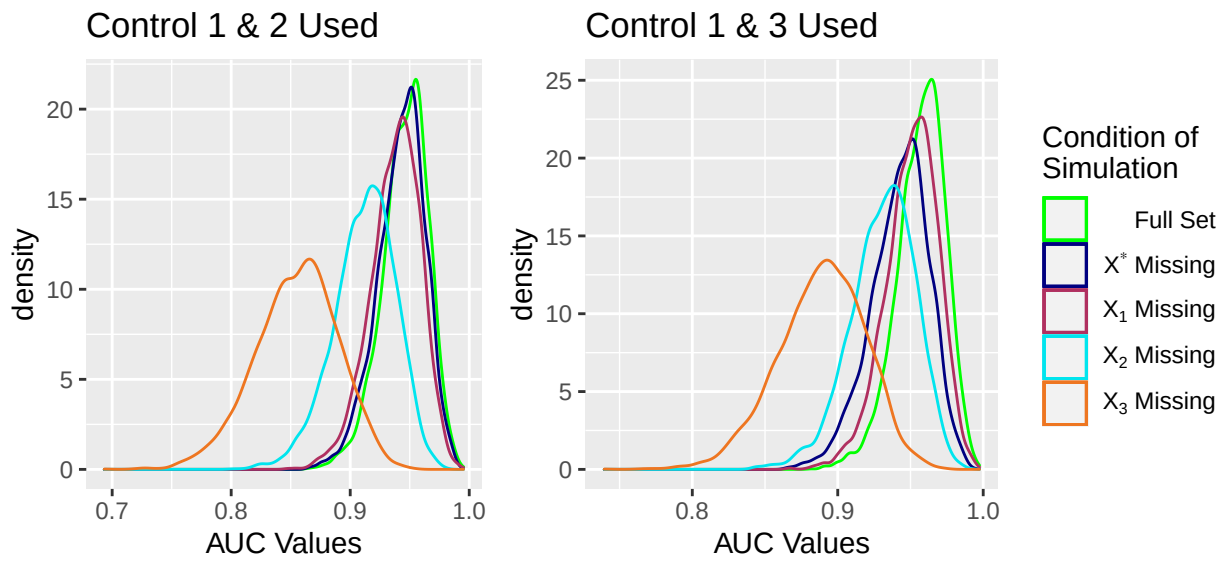

Figure 1: Distribution of AUC Curves of LDA Models of Control Scenarios 1

As can be seen from the graph title and Table 1, from the left plot in Figure 3 to the right plot in Figure 4, the distance between the means of the two distributions used for X^* is increased gradually. Therefore, from the first plot in Figure 3 to the last one in Figure 4, if the navy line is shifting to the left of the other density plots except for the green one (the ones corresponding to the models which do not have one of X_1, X_2, and X_3) which it is obviously, it means that the importance of X^* is surpassing that of those other three predictors as its two distributions for the two classes are separated from each other to the greater extent.

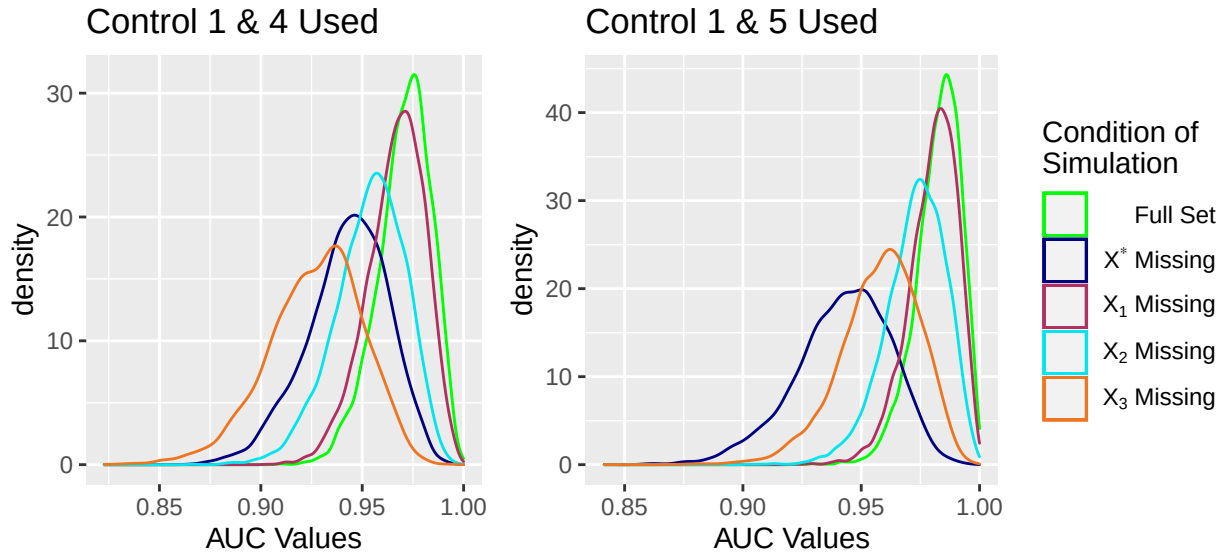

Figure 2: Distribution of AUC Curves of LDA Models of Control Scenarios 2

For example, from the left plot to the right plot in Figure 3, it can be observed that the navy

density line is shifting from the right of the purple density line to its left side. This means that the average magnitude of the deterioration of AUC values caused by the removal of X^* is smaller than that of the deterioration caused by the removal of X_1 when Control 1 and 2 distributions are used to simulate X^*, but this size relationship of the average deterioration of AUC values is reversed when Control 1 and 3 distributions are used. This means that X^* is, as expected from the way the entire simulation is structured, surpassing X_1 in its importance as a predictor as the two distributions change from Control 1 and 2 to Control 1 and 3.

Similarly, when the Control 1 and 4 are distributions are used to simulate X^*, the importance of X^* seems to surpass that of X_2 for the first time because, from the right plot of Figure 3 to the left plot of Figure 4, the navy density line is shifting from the right of the light blue line to its lest side. The same argument can be made about the relationship between the importance of X^* and that of X_3 because, in the right plot of Figure 4, the navy density line shifts to the left of the orange density line. In short, the above four plots show that X^* is surpassing X_1 in terms of variable importance first, and then, it is surpassing X_2, and finally, it is surpassing X_3, as somewhat expected given the way this simulation has chosen the mean values of the distributions of X_1, X_2, and X_3.

In short, the results shown above mean that an increased distance between the means of two different distributions used to simulate X^* indeed result in an increased importance of X^* as a predictor if the dataset is following an ideal distribution for LDA to analyze. This is an important result to be observed in these control scenarios because these results serve as the evidence to argue that the size relationship between the distance between the means used for the distribution of X^* and that between the means used for the other predictor signifies the size relationship between the importance of those two predictors, at least when all the predictors are following a Gaussian distribution. For instance, if the distance between the two values used as a mean of the distributions of X^* is larger than that between the two values used for X_1, it means that X^* should exceed X_1 in terms of variable importance.

Put differently, the results so far prove that, at least when all the explanatory variables are satisfying the assumption of normality, LDA extracts a greater amount of information from the variable which has a larger difference between the means of its two distributions. The question is whether this shift of the distribution of AUC values of the models missing X^* happens even when X^* is following a Gamma distribution. If it does, that clearly indicates that LDA can extract useful information from X^* to the same extent even if it is following a Gamma distribution. This is the question addressed in the next part with simulation results of the treatment scenarios.

3.2.3 Results: Treatment Scenarios

In Figure 5 and 6, the simulation results are shown in the exactly same way as it was for the control scenarios in Figure 3 & 4. These results are acquired through simulating the LDA models under the treatment scenarios, where X^* is made to follow a Gamma distribution. Again, the navy blue line is the density line of interest which corresponds to the model without X^*. Also, to iterate, when a density line of a model missing a predictor is more distantly left to that of the full model, it means that taking that predictor out of the model causes a greater damage to the predictive ability of the model.

The plots in those two figures show the exact same phenomenon as what is observed in the control scenarios. In the left plot of Figure 5, the navy density line has three density lines (purple, light blue, and orange, which correspond to the models missing one of X_1, X_2, and X_3) to its left and it only has the green line corresponding to the full model to its right. However, as the distance

between the means is increased from the left plot in Figure 5 to the right one in Figure 6, the navy line shifts to the left of all the other density lines. For example, from the left plot to the right plot in Figure 5, the density line corresponding to the distribution of the AUC values of the model missing X^* shifts to the left of the peak of the purple density line from its right side. Again, this means that, when Treatment 1 and 2 distributions are used, X^* is less important than X_1, but when Treatment 1 and 3 distributions are used, the order of the importance is reversed.

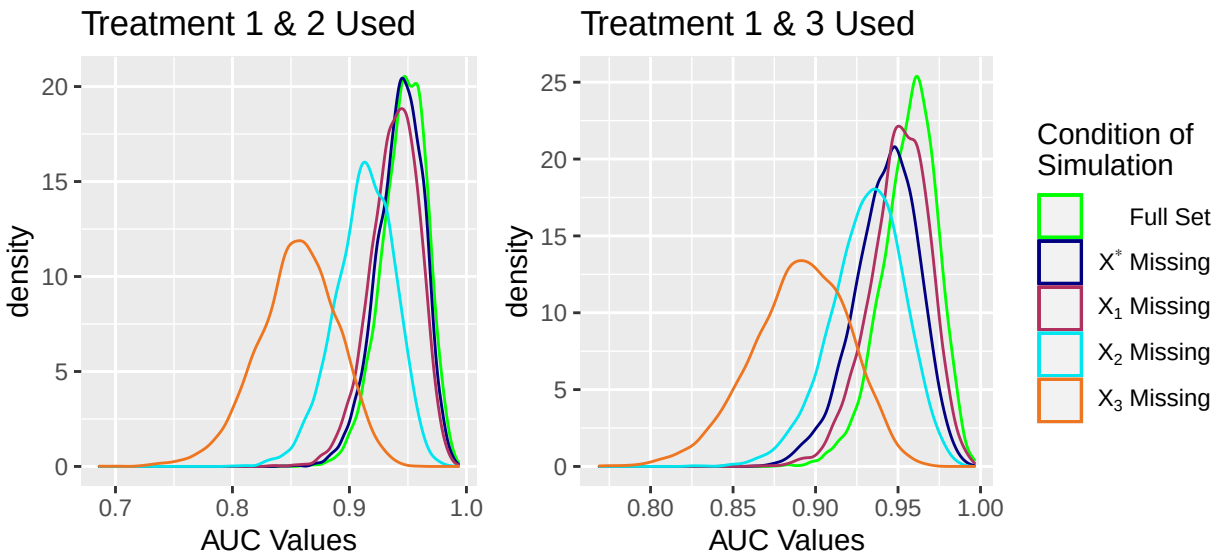

Figure 3: Distribution of AUC Curves of LDA Models of Treatment Scenarios 1

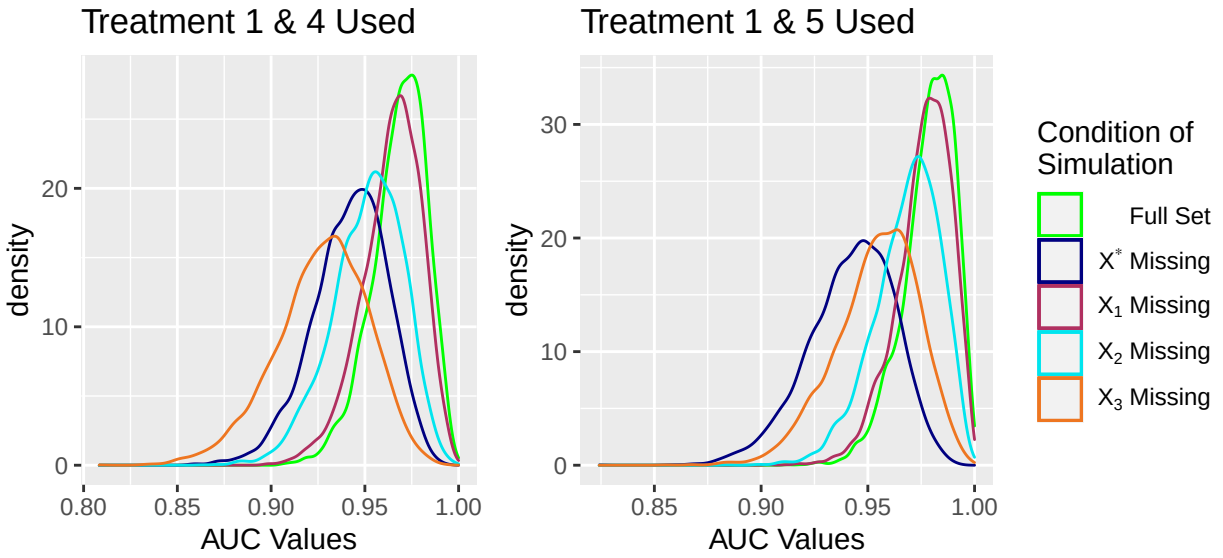

Figure 4: Distribution of AUC Curves of LDA Models of Treatment Scenarios 2

In other words, the above four plots show the exact same shift in the order of the importance of the predictors as what is detected in the control scenarios. X^* surpasses X_1 in terms of the variable

importance first, and then, it surpasses X_2 from the right plot in Figure 5 to the left plot in Figure 6, and finally it surpasses X_3 from the left plot to the right plot in Figure 6.

This result indicates that LDA models in general are capable of extracting meaningful information even from variables following Gamma distributions. If the LDA models simulated here were completely indifferent to the information contained by X^* which now follows a Gamma distribution, enlarging the distance between the means of the two distributions of X^* should not matter at all to the relative variable importance of X^*, X_1, X_2, and X_3 because whatever information contained in X^* would not be used for the calculation of the model. In contrast, the results above show that the amount of information contained in X^* is actually quite important in determining the relative variable importance of X^* in comparison to the importance of the other three predictors. This is a clear sign which indicates that the LDA models can extract important information even from a variable following a Gamma distribution.

3.3 Simulation: QDA

3.3.1 Background and Settings of the Simulation

As pointed out in the literature review section, QDA is different from LDA in a sense that it allows the presence of different covariance matrices depending on the classes to which the observation belongs to. In other words, without using a different value for the variance of the predictors between different classes, using QDA is meaningless in a sense that this simulation is obviously going to simply replicate the entire findings produced in the aforementioned simulation with LDA.

Therefore, for the investigation with QDA models, each time it simulates a dataset, the simulation code sets the variance at $\frac{1}{3}$ for the distributions with lower mean values and at $\frac{2}{3}$ for the distributions with higher mean values. In other words, X_1, X_2, and X_3 in this part will be following the distributions as described below. Note that the mean of the distributions is not changed at all from what was used to simulate LDA models.

$$X_1 \sim N(-0.2, \frac{1}{3}) \text{ for one class and } X_1 \sim N(0.2, \frac{2}{3}) \text{ for the other}$$
$$X_2 \sim N(-0.36, \frac{1}{3}) \text{ for one class and } X_2 \sim N(0.36, \frac{2}{3}) \text{ for the other}$$
$$X_3 \sim N(-0.5, \frac{1}{3}) \text{ for one class and } X_3 \sim N(0.5, \frac{2}{3}) \text{ for the other}$$

In a similar manner, the exact same set of values is used for the mean of X^*, but the variance of X^* will take the value $\frac{1}{3}$ only when its mean is equal to $\frac{1}{\sqrt{3}}$. In the other cases, it will take the value $\frac{2}{3}$. In order to achieve these conditions of simulation, the rate and shape parameters are also appropriately manipulated. The renewed set of distributions used for this QDA simulation research is provided below.

Table 2: Distributions used for the Simulation (QDA)

Distribution Name	Distribution Type	Shape	Rate	Mean	Variance
Control 1	Gaussian	N/A	N/A	$\frac{1}{\sqrt{3}}$	$\frac{1}{3}$
Control 6	Gaussian	N/A	N/A	$\frac{3}{2\sqrt{3}}$	$\frac{2}{3}$
Control 7	Gaussian	N/A	N/A	$\frac{2}{\sqrt{3}}$	$\frac{2}{3}$
Control 8	Gaussian	N/A	N/A	$\frac{5}{2\sqrt{3}}$	$\frac{2}{3}$
Control 9	Gaussian	N/A	N/A	$\frac{3}{\sqrt{3}}$	$\frac{2}{3}$
Treatment 1	Gamma	1	$\sqrt{3}$	$\frac{1}{\sqrt{3}}$	$\frac{1}{3}$
Treatment 6	Gamma	$\frac{9}{8}$	$\frac{3\sqrt{3}}{4}$	$\frac{3}{2\sqrt{3}}$	$\frac{2}{3}$
Treatment 7	Gamma	2	$\sqrt{3}$	$\frac{2}{\sqrt{3}}$	$\frac{2}{3}$
Treatment 8	Gamma	$\frac{25}{8}$	$\frac{5\sqrt{3}}{4}$	$\frac{5}{2\sqrt{3}}$	$\frac{2}{3}$
Treatment 9	Gamma	$\frac{9}{2}$	$\frac{3\sqrt{3}}{2}$	$\frac{3}{\sqrt{3}}$	$\frac{2}{3}$

Here, it should be noted that the shape parameter of the Gamma distribution used for simulation is always larger than 1 for Treatment 6 through Treatment 9 distributions. This is important because the Gamma distribution is unimodal only when the shape parameter is taking some value larger than 1. Otherwise, the distribution's probability density function becomes a monotonically decreasing function, which makes the appearance of the distribution quite different. In other words, setting a value so that it makes the shape parameters for the Treatment 6 to Treatment 9 distributions smaller or equal to 1 undermines the comparability of the results between the simulation of LDA and that of QDA.

Aside from this change in the values of the variance and the resulting changes in the shape and rate parameters, the procedure of the simulation is exactly the same as what it was for the LDA models. Basically, the simulation code will generate a dataset by making a draw from these distributions, fit five QDA models, and use the AUC value to examine the model performance. All the other properties of the distributions used to generate data are inherited from the settings for the LDA simulation above. Thus, the control scenarios will utilize normal distributions for simulating the values of X^* and the treatment scenarios will utilize Gamma distributions.

3.3.2 Results: Control Scenarios

The below plots in Figure 7 and 8 visualize the results of the simulation with QDA models under the control scenario in the same way as the plots used to analyze the simulation results with LDA. Each density line corresponds to the distributions of the AUC values of one of the following models: the one fit with the entire set of the predictors (green), the one with all the predictors but X^* (navy blue), the one with all the predictors but X_1 (purple), the one with all the predictors but X_2 (light blue), and the one with all the predictors but X_3 (orange). As for the graph titles, they again signify which distributions are used to simulate the values of X^*.

The results look quite similar to those yielded by the simulation of the LDA models and they together indicate the validity of the assumption of this research that a greater distance between the means of X^* should mean a greater importance of the variable in the context of QDA. As pointed

out in the previous discussion on simulation results about LDA, the navy blue density line shifting to the left of the other density lines means the reversal of the variable importance.

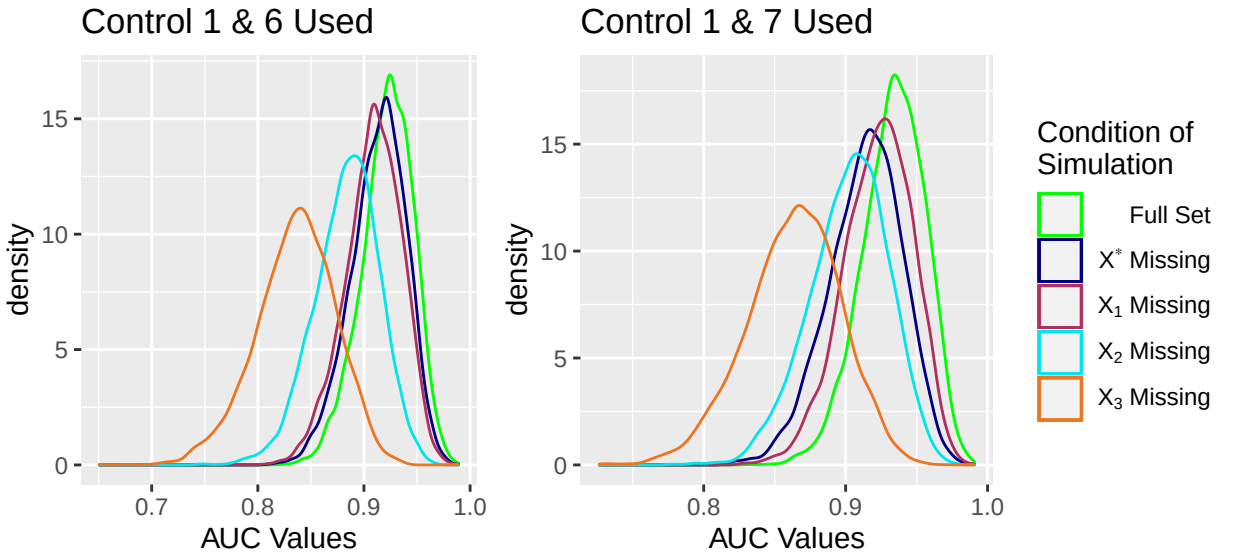

Figure 5: Distribution of AUC Curves of QDA Models of Control Scenarios 1

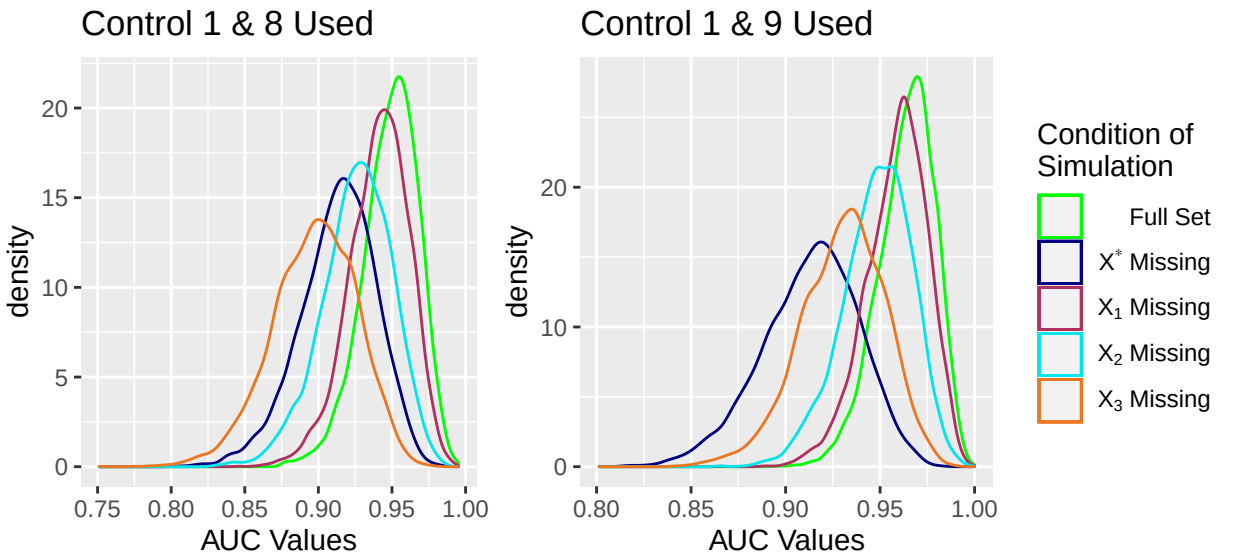

Figure 6: Distribution of AUC Curves of QDA Models of Control Scenarios 2

For example, from the left plot to the right plot in Figure 7, it can be seen that the navy density line is shifting from right to the left of the purple density line, which means that the average deterioration of the AUC values caused by the removal of X^* becomes greater than that of the AUC values caused by the removal of X_1 because of the enlargement of the distance between the two means from Control 1 and 6 to Control 1 and 7. Similarly, the navy line shifts to the left of the light blue line in the left graph in Figure 8 and to the left of the orange line in the right

graph in Figure 8. This means that X^* surpasses X_1, X_2, and X_3 one by one in terms of its contribution to the full model as the distance between the means of its two distributions is made larger throughout the course of the simulation. As can be seen in Table 2, these results in turn mean that the variable with the largest distance between the means of its two distributions is the most important variable in the model. For example, when X^* has the largest distance between the means of its two distributions, its removal on average causes the largest damage on model performance measured by AUC values.

Again, this is not a surprising result given that all the distributions are set to be normal by the design of this simulation. QDA should operate well in these kinds of datasets because the variables are all following its theoretical assumptions. In other words, similarly to the results of the simulation of the control scenarios for LDA, it is demonstrated that when QDA has no difficulty extracting information from the four predictors X^*, X_1, X_2, and X_3, the relative variable importance is identical to the size relation of the distance between the means of the different distributions of those four predictors. For example, if the distance between the means is greater for X^* than X_1, it means that X^* should be more important to the model than X_1 is. The important question is again whether this principle holds true even when X^* is not following a normal distribution, which is explored in the next part.

3.3.3 Results: Treatment Scenarios

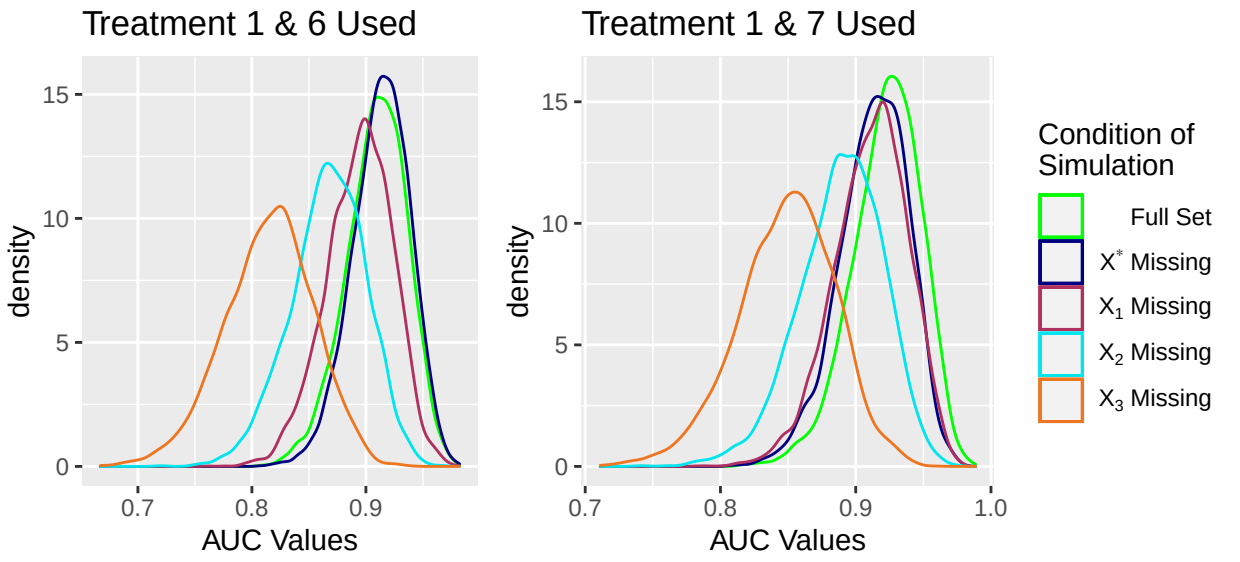

Figure 7: Distribution of AUC Curves of QDA Models of Treatment Scenarios 1

The plots in Figure 9 and 10 can again be seen in the same way as those in Figure 7 and 8 were examined. The density line of interest is the navy blue line, and when that density line is left to the purple, light blue, or orange line, it means that X^* is more important than the variable whose loss is corresponding to that line.

For example, the navy density line is located right to all the purple, light blue, and orange lines in the left plot of Figure 9, which means that the average deterioration of the AUC values caused by the removal of X^* is smaller than that of the AUC values caused by the removal of X_1, X_2, and

X_3. In contrast, similarly to all the other results so far, in the right plot of Figure 10, the navy density line is left to all the other density lines, which suggests that taking X^* causes the greatest damage to the model when the distance between the means of its two distributions is larger than all the counterpart values for X_1, X_2, and X_3.

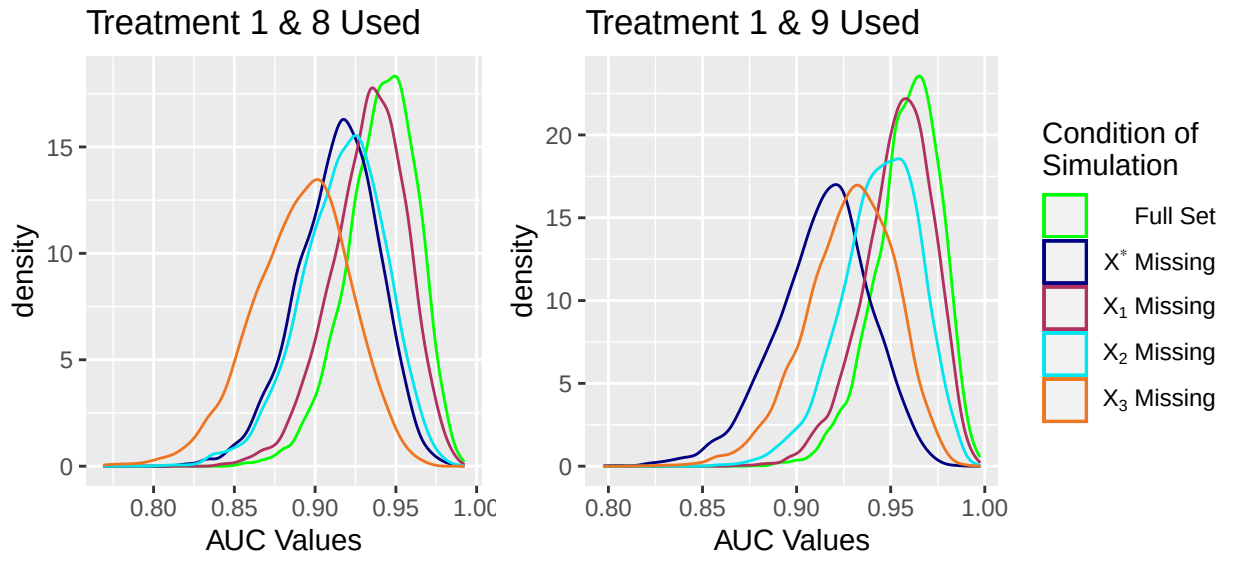

Figure 8: Distribution of AUC Curves of QDA Models of Treatment Scenarios 2

These results indeed resonate a lot with the simulation results and findings generated for LDA. Similarly to those results, what is shown in Figure 9 and 10 indicates that, even when X^* is following a Gamma distribution, it surpasses the other three predictors in its contribution to the model once the distance between the means of its two distributions gets larger than the equivalent values for X_1, X_2, and X_3. This is a phenomenon again that could not be observed if QDA were indifferent to the information contained X^* when it is following a Gamma distribution. In sum, it seems like that QDA is also capable of extracting meaningful information from a predictor which follows a Gamma distribution in the same way as LDA proved itself to be.

3.4 Summary of Results

In short, the simulation results so far consistently indicate that LDA and QDA can successfully extract information even from a predictor which follows a Gamma-distribution. This claim is proved by showing that the relative variable importance of X^*, X_1, X_2, and X_3 always follows the size order of the distance between the means of the distributions of those predictors regardless of the distribution family of X^*. The similarity between LDA and QDA is nothing surprising given that their theoretical backbones are so similar that they are almost identical. The difference is again that LDA does not allow for a different covariance matrix to be used for different classes, which means that it would not have worked well on the control and treatment scenarios prepared for QDA.

4 Discussion

Although the results and conclusions shown in the previous section seem quite intriguing, they have quite large a room of development given the nature of the Gamma distribution. As briefly mentioned at the beginning of the QDA simulation, Gamma distributions are all unimodal probability distributions as long as its shape parameter stays larger than 1. In other words, at least in terms of the appearance, Treatment 2 to 9 distributions should actually look quite alike normal distributions, or should look more similar to normal distributions compared to some other distributions. Also, Gamma distributions have the entire set of positive numbers for their support, which further renders the distribution similar to a normal distribution. These facts considered, the ability for LDA and QDA to extract meaningful information shown above seem less surprising.

Thus, in order to further investigate how LDA and QDA work with non-normally distributed variables, further simulation is required with different distributions used in lieu of the treatment scenario distributions in this paper. For example, a future study may implement LDA and QDA with predictors whose distribution follows an exponential distribution for all the classes, or may apply them to predictors which follow distributions radically different from Gaussian or Gamma distributions.

In short, this paper by itself is not enough to make a decisive conclusion about the functionality or usefulness of LDA and QDA on datasets whose predictors follow some distributions other than a Gaussian distribution. However, by showing that those methods can extract information even from predictors which follow a Gamma distribution, the observation above implies that discriminant analysis can indeed be quite useful in real-world contexts awash with non-normally distributed variables.

5 Conclusion

This paper examined whether LDA and QDA can extract important information in predicting the class of the observation even from non-normally distributed variables. In order to do so, it used a Gamma-distributed explanatory variable and measured its importance in a comparable situation to the importance of normally distributed variables. In order to examine the variable importance, this paper visualized how much damage the removal of each predictor could cause to the model's predictive ability through the average deterioration in the AUC values. According to the analysis of the simulation result, both LDA and QDA showed that, when the Gamma-distributed predictor acquired more importance in terms of the distance between the means of the two distributions, they attached a greater importance to the predictor in the same way as they did when that predictor was made to follow a Gaussian distribution in a comparable situation. In short, this paper's simulation indicates that, in a binary classification setting, as long as the predictor follows a Gamma distribution for the two classes and the shape parameter is larger than 1 for one of those two categories, LDA and QDA are robust to the non-normality of the Gamma-distributed predictor. This finding indicates that LDA and QDA may indeed be able to be utilized effectively to separate the classes even when some predictors are not following a Gaussian distribution.

References

Altman, Edward I. 1968. "Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy." *The Journal of Finance* 23 (4): 589–609. https://doi.org/10.2307/2978933.

Buchanan, Briggs, Veronica Mraz, and Metin I. Eren. 2016. "On Identifying Stone Tool Production Techniques: An Experimental and Statistical Assessment of Pressure Versus Soft Hammer Percussion Flake Form." *American Antiquity* 81 (4): 737–51.

Fisher, R. A. 1936. "The Use of Multiple Measurements in Taxonomic Problems." *Annals of Eugenics* 7 (2): 179–88. https://doi.org/10.1111/j.1469-1809.1936.tb02137.x.

Izenman, Alan J. 2008. *Modern Multivariate Statistical Techniques. [Electronic Resource] : Regression, Classification, and Manifold Learning*. Springer Texts in Statistics. Springer New York.

James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani. 2021. *An Introduction to Statistical Learning : With Applications in R*. 2nd ed. 2021. Springer Texts in Statistics. Springer US.

Li, Harrison, and Brian A. Colle. 2014. "Multidecadal Changes in the Frequency and Ambient Conditions of Warm Season Convective Storms over the Northeastern United States." *Journal of Climate* 27 (19): 7285–7300.

———. 2016. "Future Changes in Convective Storm Days over the Northeastern United States Using Linear Discriminant Analysis Applied to CMIP5 Predictions." *Journal of Climate* 29 (12): 4327–45.

Schilaty, Nathan D., Nathaniel A. Bates, Sydney Kruisselbrink, Aaron J. Krych, and Timothy E. Hewett. 2020. "Linear Discriminant Analysis Successfully Predicts Knee Injury Outcome From Biomechanical Variables." *American Journal of Sports Medicine* 48 (10): 2447–55.

Ünsal, Mehmet Güray, and Ezgi Nazman. 2020. "Investigating Socio-Economic Ranking of Cities in Turkey Using Data Envelopment Analysis (DEA) and Linear Discriminant Analysis (LDA)." *Annals of Operations Research* 294 (1/2): 281–95. https://doi.org/10.1007/s10479-017-2748-0.