Hamilton internship 2022: High-dimensional asymptotic forms of statistical estimators

Sebastian Chejniak
Trinity College Dublin
Supervised by Jason Wyse

July 6, 2022

Contents

1 Introduction

Statistical modelling is the methodology by which assumptions about the data-generating processes of potentially complex phenomena are quantified and verified, often with the goal of making future predictions. Parametric statistical models are those which are fully described by a finite number of parameters. Logistic regression is a widely popular example of such a model - it aims to quantify the effect of a set of explanatory variables on the outcome of a binary response.

Estimating the (in principle, unknown) parameters of such models is vital to verifying their assumptions and quantifying the process they describe (e.g. measuring the effect/lack thereof of explanatory variables in logistic regression). The theoretical properties of statistical estimators have been studied in detail since the dawn of statistics, and these are commonly used to verify hypotheses in all areas of scientific research. But many of the theoretical results used (such as those for logistic regression) only provide asymptotic results for the limit $n \to \infty$, i.e. as the sample size n becomes very large.

With the advent of big data, we are more often faced with circumstances where n is not necessarily very large compared to p (the no. of explanatory variables), and even circumstances with so-called "wide" data or "big p, small n" data, where $p \gg n$ (e.g. genomics, where $p =$ no. of genes), which makes the aforementioned classical results unsuitable. For example, in the case of logistic regression, when p is $o(n)$, asymptotic results relying on the limit $n \to \infty$ have empirically and theoretically been found to yield biased estimates of model parameters, therefore leading to incorrect conclusions [1]. Similarly, in the case of linear discriminant analysis (LDA), asymptotic ($n \to \infty$) estimators for the misclassification error rate are not reliable when p is $o(n)$, despite the fact that LDA is commonly used in a medical setting where obtaining many observations (of patients) is costly [2]. The aim of this report is to explore and give a brief overview of cases where such classical estimators are not accurate for high-dimensional data by summarising and replicating the results of [1] and [2].

The report is structured as follows: Section 2 summarises and discusses the results of [1] and [2] in clear language, without delving deep into the mathematics. In section 3.1 I give a brief background of the prerequisites to understand [1], namely of logistic regression. Following this, in section 3.2 I discuss the results of some of the simulations from [1], which I have replicated. In section 3.3 I explain the main results of [1], and then in 3.4 I discuss the limitations of the theory proposed by this paper. Section 4 is aimed towards [2]. I begin this section by providing a background of linear discriminant analysis in 4.1, after which I introduce the various error estimators for which [2] derives asymptotic results in 4.2. In 4.3 I state the results derived in [2] and then in 4.4 I replicate the simulations performed in [2] to test the accuracy of the asymptotic results.

2 Summary

2.1 A high-dimensional logistic regression theory

Traditionally, logistic regression estimates the regression coefficients $\boldsymbol{\beta}$ by Maximum-Likelihood estimation (MLE). It relies on the asymptotic normality of MLEs and Wilks' Theorem, in order to make inferences about the MLE estimates of its regression coefficients. Of course, both of these results require the number of observations n to be quite large in order for inference to be approximate. The Sur-Candés paper [1] argues that, in the case where the number of parameters p (i.e. the dimensionality of $\boldsymbol{\beta}$) is not negligible compared to n, these classical asymptotic results do not hold. More explicitly, these results do not hold in the limit $n \to \infty, p \to \infty, \frac{p}{n} \to \kappa \in \left(0, \frac{1}{2}\right)$. [1] shows empirically that, in contrast to what one would expect by taking $n \to \infty$ and using asymptotic normality of MLEs as well as Wilks' theorem:

1. The magnitude of the MLE estimates are positively biased
2. The standard errors of the MLE estimates are underestimated

3. The log-likelihood ratio (LLR) test statistic does not follow a χ^2 distribution

Moreover, [1] presents a theory of logistic regression that incorporates this fixed ratio $\kappa = \frac{p}{n}$, and provides an algorithm for finding the asymptotically (where by asymptotically I mean in the limit $p \to \infty$[1]) unbiased estimators of $\boldsymbol{\beta}$. It also provides, under the null hypothesis $\beta_{j_1} = \beta_{j_2} = \cdots = \beta_{j_m}$, the asymptotic distributions of the MLE estimates of the null coefficients, as well as the form of the distribution of the corresponding LLR. That being said, this theory relies on the assumptions that not made by the classic theory of logistic regression, namely that the components of the covariates are i.i.d. gaussian. This theory also requires the estimation of a "signal strength" parameter γ, which is defined by $\gamma^2 = \lim_{p\to\infty} \text{Var}\{\mathbf{x}_i^T\boldsymbol{\beta}\}$. In general, the covariates may be far from independent, and they may follow a more non-trivial distribution than a normal distribution. Furthermore, estimation of the signal strength parameter γ can affect the accuracy of the adjusted MLE estimates as well as the corresponding inferential statistics. The theory makes no comments on what would occur in these scenarios. As such, one should take caution when applying this algorithm. That being said, simulations using non-Gaussian covariates show that the theory can still hold for non-Gaussian covariates [1].

2.2 High dimensional drror dstimators for dinear discriminant analysis

[2] considers a two class linear discriminant analysis (LDA) problem, where the covariance matrix is known[2]. The focus of the paper is on the high-dimensional asymptotic behaviour of 3 various estimators of the error rate using a classifier based on Andersen's W-statistic[3]. In the limit $n \to \infty$, it follows from the continuous mapping theorem and the central limit theorem that the error rate associated with this classifier is the Bayes' error. Of course, for the case of a finite sample this will only hold approximately due to the estimation of the mean vectors. [2] improves this approximation by considering the following limit (which the paper refers to as the kolmogorov asymptotic conditions (k.a.c.))

$$\begin{aligned} n_0 \to \infty \quad & \frac{p}{n_0} \to J_0 \\ n_1 \to \infty \quad & \frac{p}{n_1} \to J_1 \\ p \to \infty \quad & \delta_p \to \delta, \end{aligned} \tag{1}$$

and considers the form of three different estimators[4] of the error rate in this limit, finding analytic solutions for the expected values of these estimators. These solutions agree very well with Monte Carlo estimates of these expected values in high-dimensional finite-sample circumstances, and are improvements upon the classical error estimator that is simply the Bayes' error. That being said, a limitation of these results is that they consider a known covariance matrix, and the asymptotic forms of the expectations of the error estimators are functions of the covariance matrix and means, which must be estimated. However it is still useful to look at the theoretical properties of these estimators so that we can choose one that fits our needs. For example, the plug-in and resubstitution error estimators turn out to be optimistic (negatively biased), while the smoothed resubstitution error estimator can be tuned such that it is a conservative estimator of the error rate.

[1] In what follows in this subsection, I let $p \to \infty$ denote the limit $n \to \infty, p \to \infty$, $\frac{p}{n} \to \kappa$.

[2] estimation of the covariance matrix is outside the scope of the paper, although in practice it is an important consideration

[3] $W(\bar{X}_0, \bar{X}_1, X) = \left(X^T - \frac{\bar{X}_0+\bar{X}_1}{2}\right)^T \Sigma^{-1} \left(\bar{X}_0 - \bar{X}_1\right)$, see more details in section 4.1

[4] The plug-in, resubstitution, and smoothed resubstitution error estimators. More details can be found in 4.2

3 Logistic Regression

3.1 Logistic regression

Logistic regression models the relationship between a p-dimensional covariate vector $\mathbf{x} \in \mathbb{R}^p$ and a binary response $y \in \{0, 1\}$ as being specified by a p-dimensional vector of regression coefficients $\boldsymbol{\beta}$ through the formula[5] $\mathbb{P}(y = 1 \mid X) = \rho'(\mathbf{x}^T\boldsymbol{\beta})$. Given n i.i.d. samples arising from this model $\{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_n, y_n)\}$, the likelihood function is given by

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} \{\rho'(\mathbf{x}_i^T\boldsymbol{\beta})\}^{y_i} \{1 - \rho'(\mathbf{x}_i^T\boldsymbol{\beta})\}^{1-y_i} \tag{2}$$

so the log-likelihood is

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left(y_i \log \rho'(\mathbf{x}_i^T\boldsymbol{\beta}) + (1 - y_i) \log \rho'(\mathbf{x}_i^T\boldsymbol{\beta})\right), \tag{3}$$

its gradient, *i.e.* the score function, is

$$\boldsymbol{S}(\boldsymbol{\beta}) = \frac{\partial}{\partial \boldsymbol{\beta}} \ell(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left(y_i - \rho'(\mathbf{x}_i^T\boldsymbol{\beta})\right) \mathbf{x}_i \tag{4}$$

and the negative Hessian of the log-likelihood which, by definition, is the observed Fisher information matrix, can be found have (i, j)-th entry

$$\begin{aligned} I(\boldsymbol{\beta})_{i,j} = -\frac{\partial^2}{\partial \boldsymbol{\beta}_j \partial \boldsymbol{\beta}_i} \ell(\boldsymbol{\beta}) = \sum_{k=1}^{n} \rho''(\mathbf{x}_k^T\boldsymbol{\beta}) x_{kj} x_{ki} &= \sum_{k=1}^{n} \frac{e^{\mathbf{x}_k^T\boldsymbol{\beta}}}{(1 + e^{\mathbf{x}_k^T\boldsymbol{\beta}})^2} x_{kj} x_{ki} \\ &= (\mathbf{X}^T D \mathbf{X})_{ij} \end{aligned} \tag{5}$$

Where $\mathbf{X}$ is the design or covariate matrix with rows $\mathbf{x}_i^T$, and $D = \operatorname{diag}(e^{\mathbf{x}_1^T\boldsymbol{\beta}}/(1+e^{\mathbf{x}_1^T\boldsymbol{\beta}})^2, \ldots, e^{\mathbf{x}_n^T\boldsymbol{\beta}}/(1+e^{-\mathbf{x}_n^T\boldsymbol{\beta}}))$ is a matrix which effectively assigns weights to the individual observations. Note that since $I(\boldsymbol{\beta})$ is independent of $y_1, \ldots, y_n$, it actually coincides with the expected Fisher information matrix $\boldsymbol{\mathcal{I}}(\boldsymbol{\beta})$. Maximum likelihood estimation is used to estimate the regression coefficients. Since the roots of (4) are not available in closed-form, one has to compute the maximum likelihood estimates(MLEs) $\hat{\boldsymbol{\beta}}$ numerically. This is typically done by Newton's method:

$$\hat{\boldsymbol{\beta}}^{(n+1)} = \hat{\boldsymbol{\beta}}^{(n)} + \boldsymbol{\mathcal{I}}\left(\hat{\boldsymbol{\beta}}^{(n)}\right)^{-1} \boldsymbol{S}\left(\hat{\boldsymbol{\beta}}^{(n)}\right) \tag{6}$$

Approximate inference on the MLEs $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ can be made[6] using asymptotic normality of MLEs:

$$\hat{\boldsymbol{\beta}} \dot{\sim} \mathrm{N}(\boldsymbol{\beta}, \boldsymbol{\mathcal{I}}(\hat{\boldsymbol{\beta}})^{-1}) \tag{7}$$

while null hypotheses of the form $\beta_{j_1} = \beta_{j_2} = \cdots = \beta_{j_k} = 0$ can be tested[7] by applying Wilks' theorem to the following log-likelihood ratio (LLR):

$$2\Lambda = -2 \log \frac{\ell(\hat{\boldsymbol{\beta}})}{\ell(\hat{\boldsymbol{\beta}}_0)} \dot{\sim} \chi_k^2 \tag{8}$$

where $\hat{\boldsymbol{\beta}}_0$ is the MLE estimate of the log-likelihood but with the constraint $\beta_{j_1} = \beta_{j_2} = \cdots = \beta_{j_k} = 0$.

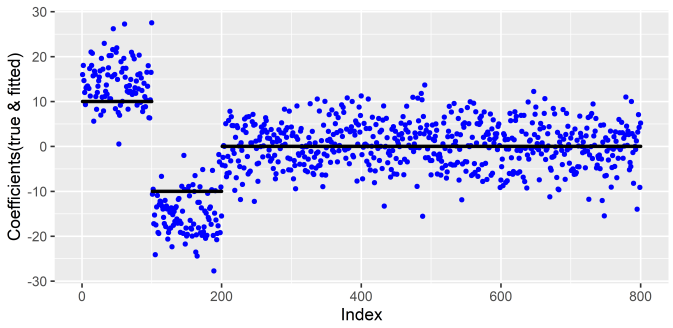


Figure 1: True values of the regression coefficients β_j vs j (black) and the MLE estimates $\hat{\beta}_j$ vs j (blue) when 600 of the β_j are zero, 100 are 10, and 100 are -10.

3.2 Motivating the need for a high-dimensional theory

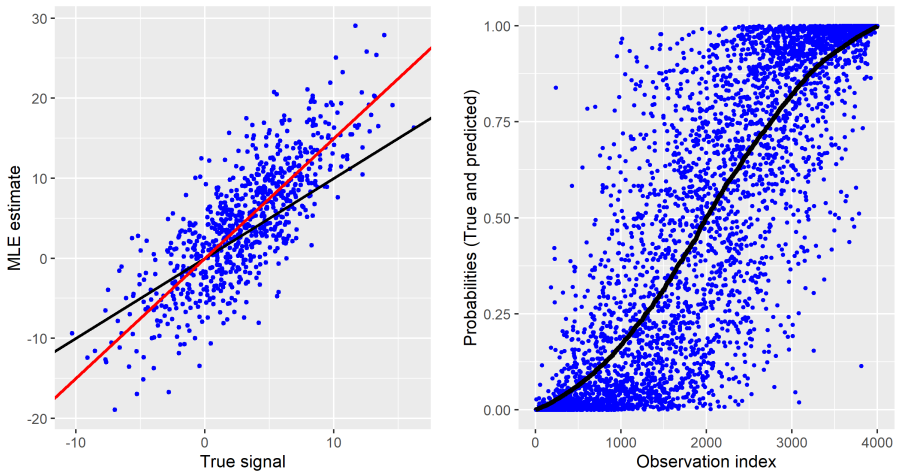
The inadequacy of classical maximum likelihood estimation for high-dimensional data is displayed in [1] by a number of simulation studies. I have replicated these studies [3], and unless otherwise stated, the plots provided here are my own. For every simulation in this section the covariate matrix $\mathbf{X}$ has i.i.d. $\mathrm{N}(0, \frac{1}{n})$ entries, $\boldsymbol{\beta}$ is chosen such that $\gamma^2 = \mathrm{Var}(\mathbf{x}_i^T\boldsymbol{\beta}) = 5$, and the dimensionality and sample sizes are chosen as $p = 800$ and $n = 4000$, respectively. Figure 1 considers $\boldsymbol{\beta}$ with 600 zero entries, 100 entries with $\beta_j = 10$ and 100 with $\beta_j = -10$, and the figure displays that in this case, the MLEs are upwardly biased in magnitude.

Figure 2 now considers all of the entries of $\boldsymbol{\beta}$ to be an i.i.d. sample from $\mathrm{N}(3, 16)$. Figure 2a shows that the true regression coefficients are again overestimated in magnitude. In this specific case, the theory proposed by [1] actually quantifies this bias, predicting that $\mathbb{E}\{\hat{\beta}_j\} = 1.499\beta_j$. Figure 2b shows that the predicted probabilities $\rho'(\mathbf{x}_i^T\hat{\boldsymbol{\beta}})$ are often very close to 0 or 1, when the corresponding true predicted probability is not. Figures 3 and 4 consider a simulation where half of the coefficients are zero and half are i.i.d. $\mathrm{N}(7, 1)$. In this simulation, the covariate matrix and then the response vector is sampled 1,000 times, each time finding the MLE estimates of the coefficients as well as the estimate of the standard error based on the fisher information approximation (7). Figure 3a is a histogram of the monte-carlo estimates of the standard errors for each *null* coefficient, whereas figure 3b is a histogram of the classically estimated standard errors for one of the null coefficients. It is expected that if asymptotic normality holds approximately, then these histograms would both be centered about the same value. However this is clearly not the case. Moreover, [1] finds analytically the expected value of the expected fisher information (the second expectation is over the covariates themselves, while the first is over the responses) to be 2.66, drastically lower than estimated. Figure 4 shows a histogram of the p-values from each of the samples, corresponding to a χ^2 test of the null hypothesis $\beta_1 = 0$. Due to the relatively small sample size of 1000, the results aren't very definitive. However, of note is that there seems to be a higher concentration of estimated p-values near 0. This is consistent with the finding of [1] that the LLR follows a rescaled form of the χ^2 distribution, with the scaling factor greater than 1. In such a scenario, it is expected that p-values based on the unscaled χ^2 distribution will be artificially shrunk. Since significance of predictors relies on p-values being less than a certain threshold, such a discrepancy can impact the quality of coefficient inference greatly.

[5] $\rho'(\cdot)$ denotes the logistic function, defined by $\rho'(x) = \frac{d}{dx}\rho(x) = \frac{d}{dx}\log(1+e^x) = \frac{e^x}{1+e^x} = \frac{1}{1+e^{-x}}$

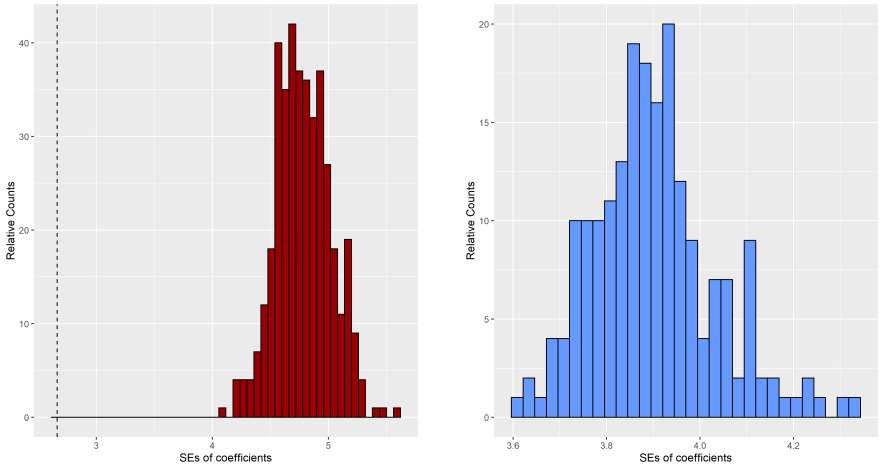
[6] When n is sufficiently large for the approximation to hold

[7] Again, approximately, for n sufficiently large



(a) (b)

Figure 2: (a) MLE estimates of β_j's vs true β_j's (blue). Black line has slope 1, while the red line has slope 1.499, which adjusts for the upward bias of the MLE estimates. (b) Predicted probabilities using the MLE estimates of the β_j's (blue) compared to the true probabilities modelled (black)



(a) (b)

Figure 3: Results from the Monte Carlo simulation. (a) Histogram for a 1000-sample Monte-carlo estimation of the standard error of each null β_j, i.e. those satisfying $\beta_j = 0$. (b) Histogram of the estimated standard errors for $\hat{\beta}_1 =$ using asymptotic normality for each of the 1000 samples

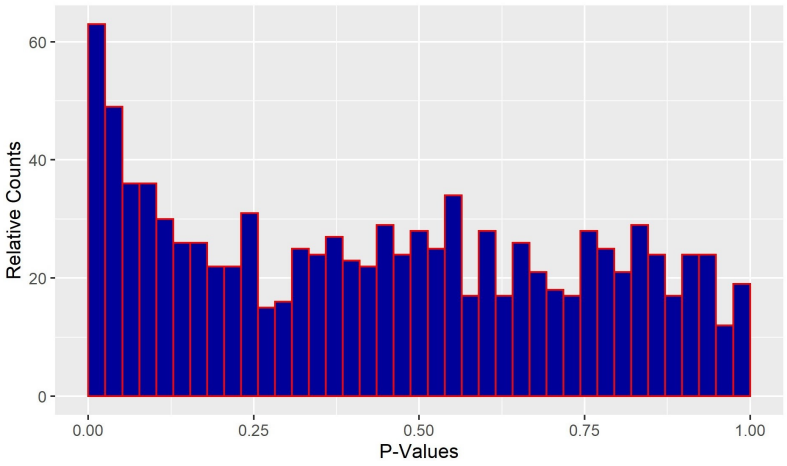


Figure 4: In the same setting as 3, this is a histogram of the estimated p-values using the χ^2_1 approximation for testing the null hypothesis $\beta_1 = 0$, which follows from Wilks' theorem.

3.3 A modern maximum-likelihood theory for high-dimensional logistic regression

Let us clarify the assumptions made by this theory first. The first is that $\mathbf{x}_i \overset{\text{iid}}{\sim} \mathrm{N}(\mathbf{0}, a\boldsymbol{I}_p)$, where $\boldsymbol{I}_p$ is the $p \times p$ identity matrix, and $a > 0$ is some scaling factor which we set to $\frac{1}{n}$[8]. Let us denote by $\lim_{p\to\infty}$ the limit

$$n \to \infty, \quad p \to \infty, \quad \frac{p}{n} \to \kappa \in \left(0, \frac{1}{2}\right).$$

The next assumption is that the β_j are scaled such that

$$\lim_{p\to\infty} Var\{\mathbf{x}_i^T \boldsymbol{\beta}\} = \gamma^2 \tag{9}$$

The final assumption is then that we restrict ourselves to cases when the MLE exists. [1] cites [4] as a reference for the following theorem, which characterises the region where the MLE exists as the region satisying $\gamma < g_{\text{MLE}}(\kappa)$:

Theorem 3.1. *Let $Z \sim N(0,1)$ with pdf $\phi(t)$, and let V be an independent continuous random variable with pdf $2\rho'(\gamma t)\phi(t)$ Denoting $x_+ = \max(x, 0)$, define*

$$g^{-1}_{MLE}(\gamma) = \min_{t\in\mathbb{R}}\{\mathbb{E}\{((Z - tV)_+)^2\}\} \tag{10}$$

Then, in the setting described above,

$$\gamma > g_{MLE}(\kappa) \implies \lim_{p\to\infty} \mathbb{P}\{MLE \text{ exists}\} = 0,$$
$$\gamma < g_{MLE}(\kappa) \implies \lim_{p\to\infty} \mathbb{P}\{MLE \text{ exists}\} = 1.$$

The parameter γ can be thought of as the signal strength, in the sense that a low γ corresponds to $\mathbf{x}_i^T\boldsymbol{\beta}$ being more tightly clumped around zero, and therefore many of the predicted probabilities $\mathbb{P}(y_i \mid \mathbf{x}_i) = 1/(1 + e^{-\mathbf{x}_i^T\boldsymbol{\beta}})$ being approximately 0.5. Therefore for a low γ one cannot predict either of the binary outcomes $y_i = 0$ and $y_i = 1$ with much certainty - the model is not very different to a model that chooses y_i based on a coin flip. So the signal strength of the model is low in the sense that the model doesn't provide us with much information.

[8] We do this for convenience, to retrieve the correct result from the following results in the case where $a \neq \frac{1}{n}$, you simply have to scale $\boldsymbol{\beta}$ and $\hat{\boldsymbol{\beta}}$ appropriately

The theory introduces 3 further parameters α, σ, λ which characterise the $\lim_{p\to\infty}$ behaviour of the MLEs and the LLR. These are defined as the solution to the system of equations (13), where

$$\begin{pmatrix} Q_1 \\ Q_2 \end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \gamma^2 & -\alpha\gamma^2 \\ -\alpha\gamma^2 & \alpha^2\gamma^2 + \kappa\sigma^2 \end{pmatrix} \right) \tag{11}$$

and, defining $\rho(t) = \log(1 + e^t)$,

$$\mathrm{prox}_{\lambda\rho}(z) = \underset{t\in\mathbb{R}}{\mathrm{argmin}} \left\{ \lambda\rho(t) + \frac{1}{2}(t - z)^2 \right\} \tag{12}$$

$$\begin{cases} \sigma^2 = \dfrac{1}{\kappa^2}\mathbb{E}\left\{ 2\left(\lambda\rho'(\mathrm{prox}_{\lambda\rho}(Q_2))\right)^2 \rho'(Q_1) \right\} \\ 0 = \mathbb{E}\left\{ Q_1\rho'(Q_1)\lambda\rho'(\mathrm{prox}_{\lambda\rho}(Q_2)) \right\} \\ 1 - \kappa = \mathbb{E}\left\{ \dfrac{2\rho'(Q_1)}{1 + \lambda\rho''(\mathrm{prox}_{\lambda\rho}(Q_2))} \right\} \end{cases}, \tag{13}$$

The first theorem which justifies solving (13) is

Theorem 3.2. *Assume the dimensionality and signal strength parameters κ and γ are such that $\gamma < g_{MLE}(\kappa)$, i.e. the MLE exists. Assume the logistic regression model described above where the empirical distribution of $\{\beta_j\}$ converges weakly to a distribution Π with finite second moment. Suppose further that the second moment converges in the sense that as $p \to \infty, Ave_j(\beta_j^2) \to \mathbb{E}\left\{\beta^2\right\}, \beta \sim \Pi$. Then, for any pseudo-Lipschitz function ψ of order 2, the marginal distributions of the MLE coordinates obey*

$$\frac{1}{p}\sum_{j=1}^{p} \psi(\hat{\beta}_j - \alpha_\star\beta_j, \beta_j) \overset{a.s.}{\to} \mathbb{E}\left\{\psi(\sigma_\star Z, \beta)\right\}, \quad Z \sim N(0,1) \tag{14}$$

where $\beta \sim \Pi$, independent of Z.

I state without justification that $\psi(t,u) = t, \psi(t,u) = t^2$, and $\psi(t,u) = tu$ are indeed pseudo-Lipschitz functions of order 2. Therefore the following statements are direct consequences of theorem[9] 3.2:

$$\frac{1}{p}\sum_{j=1}^{p}(\hat{\beta}_j - \alpha_\star\beta_j) \overset{\text{a.s.}}{\to} 0 \tag{15}$$

$$\frac{1}{p}\sum_{j=1}^{p}(\hat{\beta}_j - \alpha_\star\beta_j)^2 \overset{\text{a.s.}}{\to} \sigma_\star^2 \tag{16}$$

$$\frac{1}{p}\sum_{j=1}^{p}(\hat{\beta}_j - \alpha_\star\beta_j)\beta_j \overset{\text{a.s.}}{\to} 0 \tag{17}$$

(15) quantifies the bias of the MLE estimates, showing that they are biased by a multiplicative factor, and that therefore the estimator $\frac{\hat{\boldsymbol{\beta}}_{\text{MLE}}}{\alpha_\star}$ is unbiased. (16) quantifies the variance of the MLE estimates and (17) states that the rescaled(unbiased) MLE estimates of the regression coefficients $\frac{\hat{\beta}_j}{\alpha_\star}$ are uncorrelated with the true coefficients β_j

The next theorem presented in [1] is

[9]The following looks like some sort of result analogous to the law of large numbers, where an empirical mean converges to some quantity. Note, however, the subtlety that the sums are taken over the components of $\boldsymbol{\beta}$ instead of being taken over a number of samples. These results therefore quantify the bulk behaviour of the components of $\boldsymbol{\beta}$ as opposed to the average behaviour of every individual component. That being said, while this is a subtlety that is necessary to clarify in the proof of these results, I judge that this distinction is not of any practical significance. Hence, I simply explain these results in what follows as if they were indeed considering empirical sample means.

Theorem 3.3. *Let j be any variable such that $\beta_j = 0$. Then in the setting of 3.2, the MLE obeys*

$$\hat{\beta}_j \xrightarrow{d} N(0, \sigma_\star^2) \tag{18}$$

and furthermore, for any finite subset of null variables $\{j_1, \ldots, j_k\}$, the components of $(\hat{\beta}_{j_1}, \ldots, \hat{\beta}_{j_k})$ are asymptotically independent.

Theorem 3.3 gives the asymptotic distributions of the components of $\hat{\boldsymbol{\beta}}$ under the same type of null hypothesis as was assumed in the application of Wilks' theorem in equation (8). The following, final, theorem gives the analogue of Wilks' theorem in this theory:

Theorem 3.4. *Consider twice the log-likelihood ratio as in the left hand side of (8). In the setting of theorem 3.2 and under the null hypothesis $\beta_{j_1} = \beta_{j_2} = \cdots = \beta_{j_k} = 0$, the following holds in the limit $p \to \infty$*

$$2\Lambda \xrightarrow{d} \frac{\kappa\sigma_\star^2}{\lambda_\star}\chi_k^2 \tag{19}$$

Therefore twice the log-likelihood ratio is simply a rescaled χ^2 distribution, as opposed to a χ^2 distribution in Wilks' theorem. As is seen in [1], the multiplicative factor $\frac{\kappa\sigma_\star^2}{\lambda_\star}$ is greater than one whenever $\kappa > 0$. This rougly means that higher values of the LLR are more likely than the classical approximation suggests, meaning that p-values based on the classical approximation will be shrunk compared to their true values, increasing the type I error of the hypothesis which tests the null hypothesis $\beta_{j_1} = \beta_{j_2} = \cdots = \beta_{j_k} = 0$.

3.4 Limitations

The iid covariates assumption is quite a strong assumption, and therefore care should be taken when implementing this method in practice. A common classification problem with high p is one which applies a one-hot encoding to one (or more) categorical variables which have many levels. Applying a one-hot encoding to a single categorical variables with m levels will result in a covariate matrix with a $p = m - 1$ dataset[10]. Applying a one-hot encoding to two categorical variables with m_1 and m_2 levels, respectively, gives $p = m_1 + m_2 - 2$. But, if one includes all possible interaction terms, then the number of parameters becomes $p = m_1 + m_2 - 2 + m_1 m_2$. In all of these cases, the dummy variables and their components are neither independent nor are they normally distributed. This leads to a dataset with high multicollinearity, which is known to cause variance inflation and affecting values of the regression coefficients, thus making statistical inference untrustworthy. Moreover, such a high-dimensional dataset obtained by a one-hot encoding will inevitably have lots of zeros and therefore be quite sparse, which may result in overfitting.

4 Linear Discriminant Analysis

4.1 Linear discriminant analysis

Linear discriminant analysis (LDA) is a supervised classification method. Say we have n independent observations $\{X_1, X_2, \ldots, X_n\}$, where each observation belongs to one of K groups/classes. LDA assumes that each observation $X_i \in \mathbb{R}^p$ belonging to class k is distributed as $\mathrm{N}(\mu_k, \Sigma_p)$. That is, the distribution of X_i is a multivariate normal distribution with a mean vector which depends on the class to which X_i belongs, and with a covariance matrix that is the same for all classes. Consider the case of only two classes Π_0 and Π_1, where $n = n_0 + n_1$ observations are observed, with a set of n_0 of these belonging to Π_0 and n_1 belonging to Π_1. Label first set as $\{X_1, \ldots, X_{n_0}\}$ and the second as $\{X_{n_0+1}, \ldots, X_n\}$. The probability of observing a single point X given that it belongs to group i is

$$\mathbb{P}(X \mid X \in \Pi_i) = \frac{1}{\sqrt{(2\pi)^p \det \Sigma}} \exp\left(-\frac{1}{2}(X - \mu_i)^T \Sigma^{-1} (X - \mu_i)\right)$$

[10] To prevent perfect collinearity in the variables, one has to drop one of the variables, hence why we have $p = m - 1$

So the log probability is

$$\log \mathbb{P}(X \mid X \in \Pi_i) = -\frac{1}{2}\log((2\pi)^p \det \Sigma) - \frac{1}{2}(X-\mu_i)^T\Sigma^{-1}(X-\mu_i)$$

Assuming that $\mathbb{P}(X \mid X \in \Pi_i) > \mathbb{P}(X \mid X \in \Pi_j)$ gives[11]

$$\begin{aligned}
-\frac{1}{2}\log((2\pi)^p \det \Sigma) - \frac{1}{2}(X-\mu_i)^T\Sigma^{-1}(X-\mu_i) &> -\frac{1}{2}\log((2\pi)^p \det \Sigma) - \frac{1}{2}(X-\mu_j)^T\Sigma^{-1}(X-\mu_j) \\
\implies -\frac{1}{2}(X-\mu_i)^T\Sigma^{-1}(X-\mu_i) &> -\frac{1}{2}(X-\mu_j)^T\Sigma^{-1}(X-\mu_j) \\
\implies -\frac{1}{2}X^T\Sigma^{-1}X + X^T\Sigma^{-1}\mu_i - \frac{1}{2}\mu_i^T\Sigma^{-1}\mu_i &> -\frac{1}{2}X^T\Sigma^{-1}X + X^T\Sigma^{-1}\mu_j - \frac{1}{2}\mu_j^T\Sigma^{-1}\mu_j \\
\implies X^T\Sigma^{-1}\mu_i - \frac{1}{2}\mu_i^T\Sigma^{-1}\mu_i &> X^T\Sigma^{-1}\mu_j - \frac{1}{2}\mu_j^T\Sigma^{-1}\mu_j \\
\implies X^T\Sigma^{-1}(\mu_i-\mu_j) &> \frac{1}{2}\left(\mu_i^T\Sigma^{-1}\mu_i - \mu_j^T\Sigma^{-1}\mu_j\right) \\
\implies X^T\Sigma^{-1}(\mu_i-\mu_j) &> \frac{1}{2}(\mu_i+\mu_j)^T\Sigma^{-1}(\mu_i-\mu_j)
\end{aligned}$$

which implies

$$\mathbb{P}(X \mid X \in \Pi_i) > \mathbb{P}(X \mid X \in \Pi_j) \iff \left(X - \frac{\mu_i+\mu_j}{2}\right)^T \Sigma^{-1}(\mu_i-\mu_j) > 0 \tag{20}$$

Define

$$W(\mu_i, \mu_j, X) \coloneqq \left(X^T - \frac{\mu_i+\mu_j}{2}\right)^T \Sigma^{-1}(\mu_i-\mu_j) \tag{21}$$

Use the sample means as estimators of μ_0 and μ_1:

$$\hat{\mu}_0 = \bar{X}_0 \coloneqq \frac{1}{n_0}\sum_{i=1}^{n_0} X_i$$

$$\hat{\mu}_1 = \bar{X}_1 \coloneqq \frac{1}{n_1}\sum_{i=n_0+1}^{n} X_i$$

Substituting these estimators for their true values in (20), and estimating the class membership of an observation X by that which maximises the likelihood gives the following classification rule:

$$\begin{aligned}
X \in \Pi_0 \text{ if } W(\bar{X}_0, \bar{X}_1, X) > 0, \\
X \in \Pi_1 \text{ if } W(\bar{X}_0, \bar{X}_1, X) \leq 0,
\end{aligned}$$

Where $W(\bar{X}_0, \bar{X}_1, X)$ is called Anderson's W statistic. Consider the classification error, i.e. the probability of making an incorrect class assignment(classification) of a given observation X is:

$$\begin{aligned}
\epsilon =& \mathbb{P}\{\text{Classify X incorrectly} \mid \bar{X}_0, \bar{X}_1\} \\
=& \mathbb{P}\{X \in \Pi_0\}\mathbb{P}\{W(\bar{X}_0, \bar{X}_1, X) \leq 0 \mid X \in \Pi_0, \bar{X}_0, \bar{X}_1\} \qquad \text{Using the partition rule} \\
&+ \mathbb{P}\{X \in \Pi_1\}\mathbb{P}\{W(\bar{X}_0, \bar{X}_1, X) > 0 \mid X \in \Pi_1, \bar{X}_0, \bar{X}_1\} \\
=& \alpha_0\epsilon^0 + \alpha_1\epsilon^1
\end{aligned}$$

[11] The consideration this type of inequality for classification can be justified by the Neyman Pearson lemma, in that if membership of a given data point to group 0 is the null hypothesis, and membership to group 1 is the alternative hypothesis, this type of inequality gives lowest possible type II error for the given type I error that is attained by the test.

Where I have denoted $\alpha_i \coloneqq \mathbb{P}\{X \in \Pi_i\}$, $\epsilon^0 \coloneqq \mathbb{P}\{W(\bar{X}_0, \bar{X}_1, X) \leq 0 \mid X \in \Pi_0, \bar{X}_0, \bar{X}_1\}$, and $\epsilon^1 \coloneqq \mathbb{P}\{W(\bar{X}_0, \bar{X}_1, X) > 0 \mid X \in \Pi_1, \bar{X}_0, \bar{X}_1\}$. α_i is considered to be the a-priori mixing probability for class i. Some calculation gives

$$\begin{aligned}\epsilon =& \alpha_0 \Phi \left\{ -\frac{\left(\mu_0 - \frac{1}{2}\left(\bar{X}_0 + \bar{X}_1\right)\right)^T \Sigma^{-1}\left(\bar{X}_0 - \bar{X}_1\right)}{\sqrt{\left(\bar{X}_0 - \bar{X}_1\right)^T \Sigma^{-1}\left(\bar{X}_0 - \bar{X}_1\right)}} \right\} \\ &+ \alpha_1 \Phi \left\{ \frac{\left(\mu_1 - \frac{1}{2}\left(\bar{X}_0 + \bar{X}_1\right)\right)^T \Sigma^{-1}\left(\bar{X}_0 - \bar{X}_1\right)}{\sqrt{\left(\bar{X}_0 - \bar{X}_1\right)^T \Sigma^{-1}\left(\bar{X}_0 - \bar{X}_1\right)}} \right\},\end{aligned} \tag{22}$$

where $\Phi(\cdot)$ is the cumulative distribution function (cdf) of $\mathrm{N}(0, 1)$. Taking the expected value of this over the training data gives

$$\begin{aligned}\mathbb{E}(\epsilon) =& \alpha_0 \mathbb{E}(\epsilon^0) + \alpha_1 \mathbb{E}(\epsilon^1) \\ =& \alpha_0 \mathbb{P}\{W(\bar{X}_0, \bar{X}_1, X) \leq 0 \mid X \in \Pi_0\} + \alpha_1 \mathbb{P}\{W(\bar{X}_0, \bar{X}_1, X) > 0 \mid X \in \Pi_1\}\end{aligned}$$

This is a useful quantity because it allows us to evaluate the overall performance of LDA together with the classification rule we have chosen, and therefore it can allow us to compare LDA to other classification methods. However, in practice we do not have access to $\alpha_0, \alpha_1, \mu_0, \mu_1$ or Σ, which are required to calculate $\mathbb{E}(\epsilon)$. Moreover, the Bayes' error[12], which is the asymptotic $n \to \infty$ error rate of this classifier, is not the asymptotic error rate in the limit where the dimension of the problem is not negligible to the sample size. Therefore, the goal of [2] is to consider the properties of various estimators of ϵ and particularly their asymptotic expected values when the dimensionality is not negligible, which we could compare to $\mathbb{E}(\epsilon)$ and the Bayes' error. [2] does this by considering the following limit, which it calls the *Kolmogorov asymptotic conditions (k.a.c.)*:

$$\begin{aligned} n_0 \to \infty \quad & \frac{p}{n_0} \to J_0 \\ n_1 \to \infty \quad & \frac{p}{n_1} \to J_1 \\ p \to \infty \quad & \delta_p \to \delta, \end{aligned} \tag{23}$$

Where $\delta_p \to \delta$ is defined by considering the following sequence of LDA problems in the setup as discussed at the start of this section (namely that there are two classes Π_0 and Π_1, whose members are realisations of $\mathrm{N}(\mu_0, \Sigma)$ and $\mathrm{N}(\mu_1, \Sigma)$, respectively):

$$\{\mu_{p,0}, \mu_{p,1}, \Sigma_p, n_{p,0}, n_{p,1}\}, \quad p = 1, 2, \ldots \tag{24}$$

For each of these problems δ_p is the Mahalanobis distance between the means of the two respective groups with respect to their common covariance matrix:

$$\delta_p = \sqrt{(\mu_1 - \mu_0)^T \Sigma^{-1} (\mu_1 - \mu_0)}$$

and δ is simply assumed to be the value that δ_p converges to in (24). I use the notation $\lim_{k.a.c.} \coloneqq \lim_{(23)}$. An important point to note is that [2] only considers cases where the common covariance matrix Σ is known. Proving the asymptotic properties of estimators where Σ has to be estimated is a more difficult (albeit important) task that is outside the scope of [2] and therefore also of this report.

[12] This is the lowest possible error rate for a given classifier, which can be found to be $\Phi(-\delta/2) \coloneqq \Phi(-\frac{1}{2}(\mu_0 - \mu_1)^T \Sigma (\mu_0 - \mu_1))$ for LDA, by applying the central limit theorem and the continuous mapping theorem to (22)

4.2 Error estimators

The *resubstitution error estimator* $\hat{\epsilon}_r$ is

$$\begin{aligned}\hat{\epsilon}_r &= \frac{1}{n}\left[\sum_{i=1}^{n_0} I\left\{W\left(\bar{X}_0, \bar{X}_1, X_i\right) \leq 0\right\} + \sum_{i=n_0+1}^{n_0+n_1} I\left\{W\left(\bar{X}_0, \bar{X}_1, X_i\right) > 0\right\}\right] \\ &= \hat{\alpha}_0 \hat{\epsilon}_r^0 + \hat{\alpha}_1 \hat{\epsilon}_r^1\end{aligned} \tag{25}$$

Where $\hat{\alpha}_i = \frac{n_i}{n}$[13], $\hat{\epsilon}_r^0 = \frac{1}{n_0}\sum_{i=1}^{n_0} I\left\{W\left(\bar{X}_0, \bar{X}_1, X_i\right) \leq 0\right\}$, and $\hat{\epsilon}_r^1 = \frac{1}{n_1}\sum_{i=n_0+1}^{n_0+n_1} I\left\{W\left(\bar{X}_0, \bar{X}_1, X_i\right) \leq 0\right\}$. Its expected value is

$$\begin{aligned}\mathrm{E}\left(\hat{\epsilon}_r\right) &= \hat{\alpha}_0 \mathrm{E}\left(\hat{\epsilon}_r^0\right) + \hat{\alpha}_1 \mathrm{E}\left(\hat{\epsilon}_r^1\right) \\ &= \hat{\alpha}_0 P\left\{W\left(\bar{X}_0, \bar{X}_1, X_1\right) \leq 0\right\} + \hat{\alpha}_1 P\left\{W\left(\bar{X}_0, \bar{X}_1, X_{n_0+1}\right) > 0\right\},\end{aligned} \tag{26}$$

The *plug-in error estimator* $\hat{\epsilon}_p$ is the error obtained by simply substituting $\hat{X}_0, \hat{X}_1, \hat{\alpha}_0, \hat{\alpha}_1$ for $\mu_0, \mu_1, \alpha_0, \alpha_1$ in (22). Straightforward calculation yields

$$\begin{aligned}\hat{\epsilon}_p &= \Phi\left\{-\frac{1}{2}\sqrt{(\bar{X}_0 - \bar{X}_1)^T \Sigma^{-1} (\bar{X}_0 - \bar{X}_1)}\right\} \\ &\coloneqq \Phi\left\{-\hat{\delta}/2\right\}.\end{aligned} \tag{27}$$

The *smoothed resubstitution error estimator* is given by replacing the indicator function in (25) by some smooth function g, with the idea being that this would hopefully reduce the variance in the corresponding error estimator

$$\begin{aligned}\hat{\epsilon}_{sr} &= \frac{1}{n}\left(\sum_{i=1}^{n_0} g(W(\bar{X}_0, \bar{X}_1, X_i)) + \sum_{i=n_0+1}^{n} \left[1 - g(W(\bar{X}_0, \bar{X}_1, X_i))\right]\right) \\ &= \hat{\alpha}_0 \hat{\epsilon}_{sr}^0 + \hat{\alpha}_1 \hat{\epsilon}_{sr}^1,\end{aligned} \tag{28}$$

where

$$\begin{aligned}\hat{\epsilon}_{sr}^0 &= \frac{1}{n_0}\sum_{i=1}^{n_0} g(W(\bar{X}_0, \bar{X}_1, X_i)) \\ \hat{\epsilon}_{sr}^1 &= \frac{1}{n_1}\sum_{i=n_0+1}^{n} [1 - g(W(\bar{X}_0, \bar{X}_1, X_i))].\end{aligned} \tag{29}$$

We choose $g(x) = \Phi(-x/(b\hat{\delta}))$ for some (yet unspecified) parameter $b > 0$.

4.3 Asymptotic results

[2] states and proves a theorem which they call the *fundamental theorem of known covariance LDA*. While it is a vital component of proving the other theorems in the paper (which provide asymptotic results for the error estimators discussed), I omit it, as the focus of this report is on exploring the results themselves, and not getting caught up in the rigorous justifications thereof. Note also that while theorems 4.1 and 4.2 are proved in [2], they had already been proved before, and this paper only provides a simpler proof of those results using the aforementioned *fundamental theorem of known covariance LDA*

Theorem 4.1. *In the sequence of LDA problems given by (24) and under the Kolmogorov asymptotic conditions and for $X \in \Pi_0$:*

$$W(\bar{X}_0, \bar{X}_1, X) \xrightarrow{d} N\left(\frac{1}{2}(\delta^2 + J_1 - J_0), \delta^2 + J_0 + J_1\right)$$

[13] Given that it is the probability of class membership without considering $W(\bar{X}_0, \bar{X}_1, X)$, it is natural to estimate it as the proportion of the training data which comprises the corresponding class, i.e. $\hat{\alpha}_i = \frac{n_i}{n}$.

$$\lim_{k.a.c.} \mathbb{E}\{\epsilon\} = \alpha_0 \Phi\left(-\frac{1}{2}\frac{\delta^2 + J_1 - J_0}{\sqrt{\delta^2 + J_0 + J_1}}\right) + \alpha_1 \Phi\left(-\frac{1}{2}\frac{\delta^2 + J_0 - J_1}{\sqrt{\delta^2 + J_0 + J_1}}\right) \tag{30}$$

Theorem 4.2. *In the sequence of LDA problems given by (24) and under the Kolmogorov asymptotic conditions,*

$$W(\bar{X}_0, \bar{X}_1, X) \xrightarrow{d} N\left(\frac{1}{2}\left(\delta^2 + J_0 + J_1\right), \delta^2 + J_0 + J_1\right)$$

$$\lim_{k.a.c.} \mathbb{E}\{\hat{\epsilon}_r\} = \Phi\left(-\frac{1}{2}\frac{\delta^2 + J_0 + J_1}{\sqrt{\delta^2 + J_0 + J_1}}\right) \tag{31}$$

Theorem 4.3. *In the sequence of LDA problems given by (24) and under the Kolmogorov asymptotic conditions,*

$$\lim_{k.a.c.} \mathbb{E}\{\hat{\epsilon}_p\} = \Phi\left(-\frac{1}{2}\frac{\delta^2 + J_0 + J_1}{\sqrt{\delta^2 + J_0 + J_1}}\right) \tag{32}$$

Theorem 4.4. *In the sequence of LDA problems given by (24) and under the Kolmogorov asymptotic conditions,*

$$\lim_{k.a.c.} \mathbb{E}\{\hat{\epsilon}_{sr}\} = \Phi\left(-\frac{1}{2}\frac{\delta^2 + J_0 + J_1}{\sqrt{(1+b^2)(\delta^2 + J_0 + J_1)}}\right) \tag{33}$$

From which it follows that

$$\lim_{k.a.c.} \mathbb{E}\{\hat{\epsilon}_{sr}\} > \lim_{k.a.c.} \mathbb{E}\{\hat{\epsilon}_r\} = \lim_{k.a.c.} \mathbb{E}\{\hat{\epsilon}_p\}$$

Notice that all of the error estimators are, in general, asymptotically biased. However, the smoothed-resubstitution error estimator has a free parameter b which we have yet to specify. Therefore, given $\alpha_0, \alpha_1, J_1, J_0$, and δ, we can set the RHS of (33) equal to that of (30) and solve the resulting equation for b. The solution b_{opt} to this equation would correspond to $\hat{\epsilon}_{sr}$ being asymptotically unbiased, so that $\lim_{\text{k.a.c.}} \mathbb{E}\{\hat{\epsilon}_{sr}\} = \lim_{\text{k.a.c.}} \mathbb{E}\{\epsilon\}$.

4.4 Simulations

I ran the same simulations done in [2] to analyse the accuracy of these asymptotic results. For a known covariance matrix and various dimensionalities p, I computed the true error and its estimators, while making the simplifying assumption that $n_0 = n_1 = \frac{n}{2}$ and therefore $J_0 = J_1$. Straightforward calculation shows that in this case, the aforementioned b_{opt} which makes $\hat{\epsilon}_{sr}$ asymptotically unbiased is in fact $b_{\text{opt}} = \frac{2}{\delta^2}\sqrt{J_0(J_0 + \delta^2)}$. I further choose μ_0, μ_1, Σ such that $\delta^2 = 4$. I use this true value of δ to calculate b_{opt} for the purposes of evaluating $\hat{\epsilon}_{sr}$ and the estimator $\hat{\delta}$ of δ elsewhere. For the above setup, one can show:

$$\begin{aligned}
\lim_{k.a.c.} \mathbb{E}\{\hat{\epsilon}_r\} = \lim_{k.a.c.} \mathbb{E}\{\hat{\epsilon}_p\} &= \Phi\left(-\frac{1}{2}\frac{\delta^2 + 2J_0}{\sqrt{\delta^2 + 2J_0}}\right) \\
\lim_{k.a.c.} \mathbb{E}\{\hat{\epsilon}_{sr}\} = \lim_{k.a.c.} \mathbb{E}\{\epsilon\} &= \Phi\left(-\frac{1}{2}\frac{\delta^2}{\sqrt{\delta^2 + 2J_0}}\right)
\end{aligned} \tag{34}$$

The following plots are those from the simulations just discussed. Figures 5, 6, 7 and 8 plot the approximation of the expected values, as well as a Monte-Carlo estimate thereof, for various sample sizes, and dimensions. It can be seen from 5 and 6 that the resubstitution error estimator $\hat{\epsilon}_r$, as well as the plug-in error estimators $\hat{\epsilon}_p$ are negatively biased, as their expected values are all below the Bayes' error line, while every approximation of the expected value of the true error ϵ is above this line. On the other hand, the smoothed-resubstitution error estimator $\hat{\epsilon}_{sr}$ with $b = b_{opt}$ appears to be an unbiased estimator of ϵ, as is seen by the striking similarity of their graphs. This is consistent with what we expected theoretically by our choice of b.

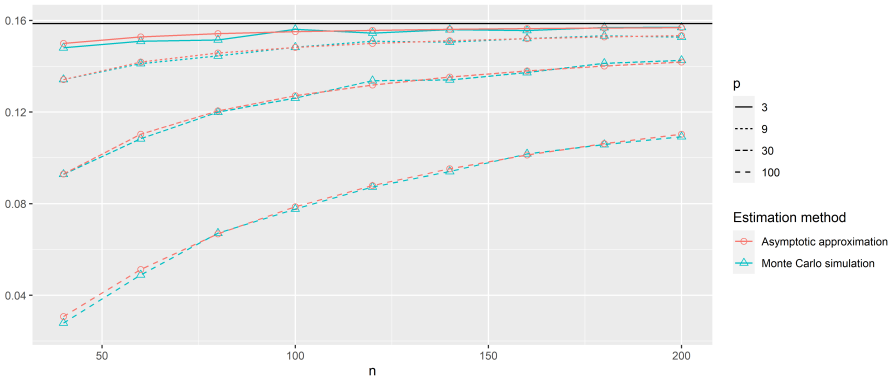


Figure 5: Estimation of $\mathbb{E}\{\hat{\epsilon}_r\}$ (y-axis) vs the number of observations (x-axis) for each approximation (differentiated by colour) and dimensionality of each observation (differentiated by various levels of dotting). The black line is the Bayes' error corresponding to two-class LDA problems with $\delta^2 = 4$, which is 0.1586. The setup of the problem and the approximations used are as discussed in section 4.4

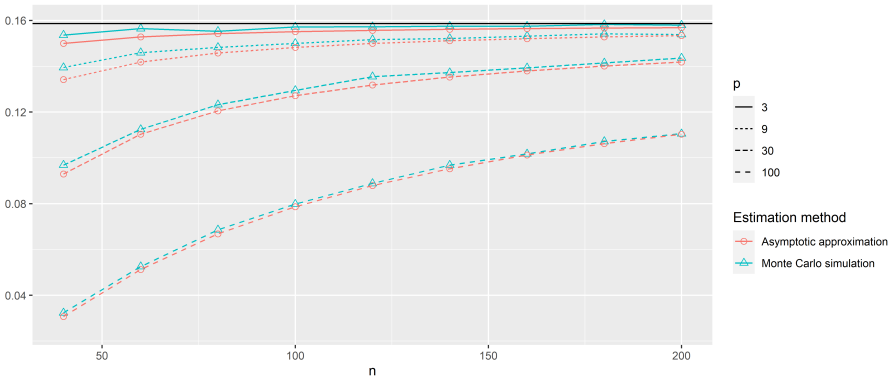


Figure 6: Estimation of $\mathbb{E}\{\hat{\epsilon}_p\}$ (y-axis) vs the number of observations (x-axis) for each approximation (differentiated by colour) and dimensionality of each observation (differentiated by various levels of dotting). The black line is the Bayes' error corresponding to two-class LDA problems with $\delta^2 = 4$, which is 0.1586. The setup of the problem and the approximations used are as discussed in section 4.4

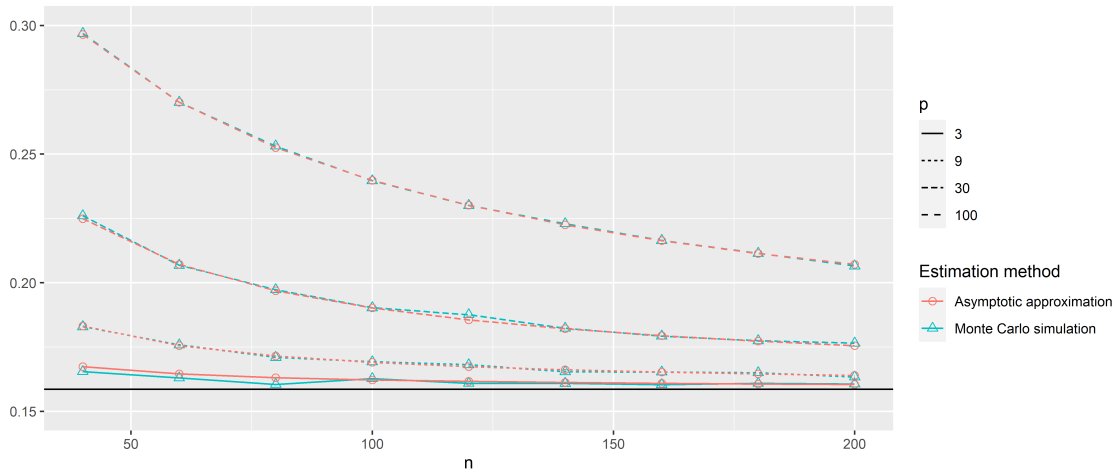


Figure 7: Estimation of $\mathbb{E}\{\hat{\epsilon}_{sr}\}$ (y-axis) vs the number of observations (x-axis) for each approximation (differentiated by colour) and dimensionality of each observation (differentiated by various levels of dotting). The black line is the Bayes' error corresponding to two-class LDA problems with $\delta^2 = 4$, which is 0.1586. The setup of the problem and the approximations used are as discussed in section 4.4

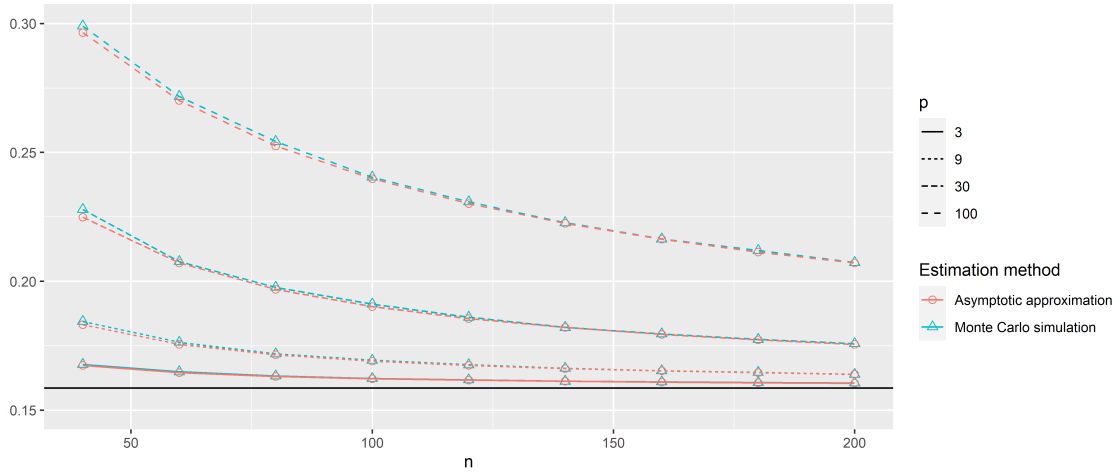


Figure 8: Estimation of $\mathbb{E}\{\epsilon\}$ (y-axis) vs the number of observations (x-axis) for each approximation (differentiated by colour) and dimensionality of each observation (differentiated by various levels of dotting). The black line is the Bayes' error corresponding to two-class LDA problems with $\delta^2 = 4$, which is 0.1586. The setup of the problem and the approximations used are as discussed in section 4.4

References

[1] P. Sur and E. J. Candès, "A modern maximum-likelihood theory for high-dimensional logistic regression," *Proceedings of the National Academy of Sciences*, vol. 116, no. 29, pp. 14516–14525, 2019.

[2] A. Zollanvari and M. G. Genton, "On kolmogorov asymptotics of estimators of the misclassification error rate in linear discriminant analysis," *Sankhyā: The Indian Journal of Statistics, Series A (2008-)*, vol. 75, no. 2, pp. 300–326, 2013.

[3] https://github.com/SebastianChk/....

[4] E. Candès and P. Sur, "The phase transition for the existence of the maximum likelihood estimate in high-dimensional logistic regression," *The Annals of Statistics*, vol. 48, pp. 27–42, 02 2020.