

High dimensional logistic regression and its limitations

Erik Bergqvist, Yuhan Chen, Yiren Fang, Redmond Bamford, Andreea Patra

Contents

1 Abstract	2
2 Defining the logistical regression model	3
3 A comparison between <code>glm()</code> and <code>lm()</code> in R	4
4 The behaviour of the bias of the MLE coefficients for large n and p	6
5 Some preliminary calculations relating γ^2 , β , kappa, and X	10
6 Relation between κ and the bias	12
7 An investigation of the bias α_* in relation to varying the β_0 parameter	15
8 An investigation of the bias α_* in relation to the distribution of the eigenvalues of the empirical covariance matrix	18

1 Abstract

Logistic regression is a widely used model in statistics for classification problems. For example, credit card companies use it to identify odd transactions on their customers' bank accounts. In logistic regression n independent response variables are modeled against p predictor variables, these responses are listed as a matrix of covariates $\mathbf{X} \in \mathbb{R}^{n \times p}$. The Maximum Likelihood Estimator (MLE) is a common and useful tool for estimating the unknown predictor parameters $\beta_0, \beta_1, \dots, \beta_p$. Pragya Sur and Emmanuel Candes developed theorems predicting the behaviour of the bias and the correlation of MLE estimates for large numbers of observations, n , dependent on large numbers of parameters, p , such that $p < n$.

Candes and Sur predicted a threshold¹, dependent on a signal strength γ^2 and the dimensionality $\kappa = \frac{p}{n}$ of \mathbf{X} , which determines whether the MLE exists in theory when $n, p \rightarrow \infty$ such that $\kappa \in (0, 1)$. This project uses R to investigate their theorems by variation of κ and γ^2 for large n and p . It establishes that when it is not probable that the MLE exists in theory, the MLE bias greatly increases. The project also delves into the impact of sampling from different β_0 parameters on the MLE bias as well their correlation to the true parameters. The behaviour of the bias of the MLE is explained through consideration of the distribution of eigenvalues of the covariance matrix of \mathbf{X} . Building on the analysis of the distribution of the eigenvalues, the effect of randomly scaling the columns of \mathbf{X} and sampling the entries of \mathbf{X} from a Bernoulli distribution instead of the Gaussian distribution is explored.

2 Defining the logistical regression model

Introduction to the aim of the model

For n independent observations y_i of a response variable Y_i such that $y_i \in \{0, 1\}$, we are given a vector of predictor variables $\mathbf{X}_i \in \mathbb{R}^p$ and the aim is to relate $\pi_i = \Pr(Y_i = 1)$ to a set predictors values (x_{i1}, \dots, x_{ip}) from the vector \mathbf{X}_i . The value of 1 is used to indicate a "success" and the 0 to show a "failure". The linear model has a design matrix and a vector β of parameters that needs to be estimated. The design matrix has n rows and $p+1$ columns where p is the number of parameters and n is the number of observations. The first element of each row is 1 and it is corresponding to the intercept β_0 . There is one corresponding parameter in β for each of the independent vectors in the design matrix.²

Linear predictor

A linear predictor is defined as: $\mu_i = \beta_0 + \sum_{j=1}^p \beta_j x_{ij}$ ²

Logit function

A monotone function such that $g^{-1}(\mu_i) \mapsto p_i$ is required to connect the mean to the linear predictor. Since $\mu_i \in \mathbb{R}$ and $y_i \in \{0, 1\}$, the inverse of ρ satisfies these properties. $\rho'(t) : \mathbb{R} \rightarrow [0, 1] : t \mapsto \frac{e^t}{1+e^t}$.³

Logistic regression model

In logistic regression p_i is conditioned on the covariates so that: $\Pr(Y_i = 1 | \mathbf{X}) = \rho'(\mathbf{X}'\beta)$, where $\rho'(t) = \frac{e^t}{1+e^t}$. Therefore, we get the form of the logistic regression model as:

$$\log\left(\frac{\pi_i}{1 - \pi_i}\right) = \mu_i$$

The left hand side is called the log odds and represent the probability of a positive event divided by a negative event. Therefore, when the odds are greater than 1, they favor a "success". Hence the linear predictor $\mu_i = \mathbf{X}_i\beta$ can also be defined as log odds.²

Estimation of β

For $\beta \in \mathbb{R}^{p+1}$ to estimate the $p+1$ unknown parameters. The joint probability density on observing \mathbf{y} is given by:

$$f(\mathbf{y}|\beta) = \prod_{i=1}^n \pi_i^{y_i} (1 - \pi_i)^{1-y_i}$$

The likelihood function is given by setting $L(\mathbf{y}|\beta) = f(\mathbf{y}|\beta)$. The maximum likelihood estimator (MLE) for \mathbf{Y} is obtained by finding the combination of entries for β that maximizes the likelihood function. The MLE is equivalent to maximizing the following log-likelihood function:

$$l(\beta|\mathbf{y}) = \sum_{i=1}^n \left(y_i \left(\sum_{j=0}^p x_{ij}\beta_j \right) - \log(1 + e^{\sum_{j=0}^p x_{ij}\beta_j}) \right)$$

If one computes $\frac{\partial l(\beta|\mathbf{y})}{\partial \beta_j}$ for $j = 0 \rightarrow p$, the MLE exists if a solution for these combinations of entries of β are set equal to 0 and the matrix formed by calculating $\frac{\partial l(\beta|\mathbf{y})}{\partial \beta_i \partial \beta_j}$ for i and j as before is negative definite.²

To calculate this solution R uses a iteratively reweighted least squares algorithm.²

3 A comparison between `glm()` and `lm()` in R

To fit the logistic regression model, we estimate the β parameters using the `glm()` function that is provided by the R software package. One important question arises of why we cannot use the ordinary regression function `lm()` to estimate the mean response $\Pr(Y_i = 1) = \pi_i$. To investigate

the difference between them, we will simulate data from this model:

$$\log\left(\frac{\pi(x)}{1 - \pi(x)}\right) = -2 + 5x$$

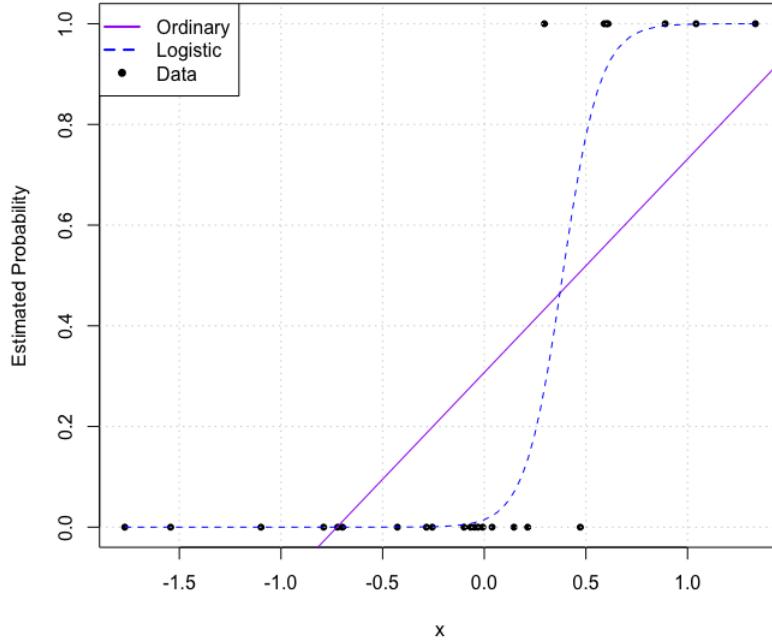


Figure 1: Ordinary vs Logistic regression function

For both of the functions, we plot the estimated mean. The data, with the response Y taking values of 0 and 1 is pictured through the black scattered points. We observe that the ordinary function produces estimates that are less than 0 or greater than 1 and since the estimated mean π is a probability that is not possible. What `glm()` does differently is that it uses the logit function so that it keeps the values between a specified region of 0 and 1. So, throughout this paper we will continue to use the `glm()` to fit the logistic regression model.

4 The behaviour of the bias of the MLE coefficients for large n and p

Throughout this section, we set $n=4000$ and $p=800$, hence the dimensionality $\kappa = \frac{p}{n}$ equals $\frac{1}{5}$. The variance of the log-odds ratio $\mathbf{X}'_i\boldsymbol{\beta}$, the signal strength, is set to be fixed at 5. Furthermore, for all of the models, it is assumed that β_0 is equal to 0.

$$\gamma^2 := \text{Var}(\mathbf{X}'_i\boldsymbol{\beta}) = 5$$

Setting

For a random matrix \mathbf{X} with entries X_{ij} i.i.d $\mathcal{N}(0, 1/n)$ entries and a column vector $\boldsymbol{\beta}$ with 800 entries, we set the first eighth of the entries of $\boldsymbol{\beta}$ to have the constant value of 10 and the second eighth of the $\boldsymbol{\beta}$ are set to -10, the remaining 600 entries of $\boldsymbol{\beta}$ are set to zero. We plot the calculated MLE coefficients resulting from the `glm` function for this setting.

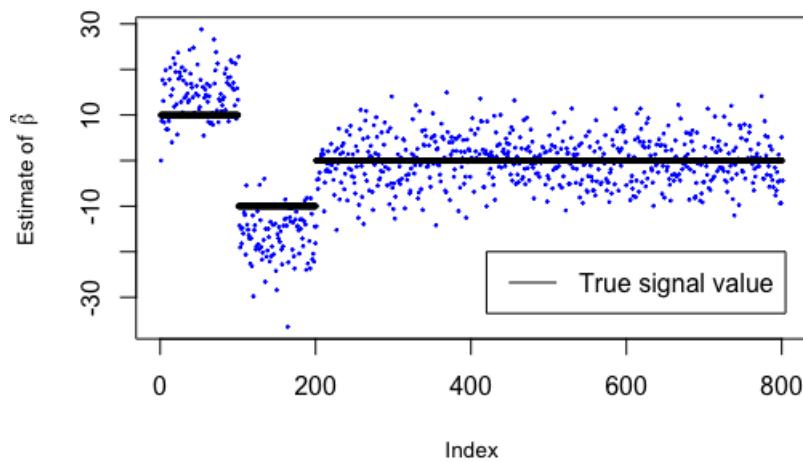


Figure 2: Scatter plot of the MLE estimates compared to the true $\boldsymbol{\beta}$ parameter value

Evaluation

The black line represents the true β parameter values, whilst the blue scatter correspond to the calculated MLE coefficient. When the coefficients of β are positive with the constant value of 10, most of the MLE estimates lie above the black line which indicates a strong bias. The next set of coefficient exhibit the same behaviour when the coefficients of β are negative, i.e. most of the blue scatter lie below the black line. Overall, the plot shows that the MLE estimate has a tendency of overestimating a non-zero parameter value β . For null parameters of β the MLE estimates are distributed around 0.

Setting

To show that the behaviour of the bias of the MLE coefficients is not unique to the first example, we considered a different choice for the parameters of β . Instead β_j is sampled from a Gaussian distribution with mean equal to 3 and variance of 16 which satisfies the criteria of γ^2 .

We then plot the pairs $(\beta_j, \hat{\beta}_j)$ on a single graph, and obtain a line of best fit to these scatter points.

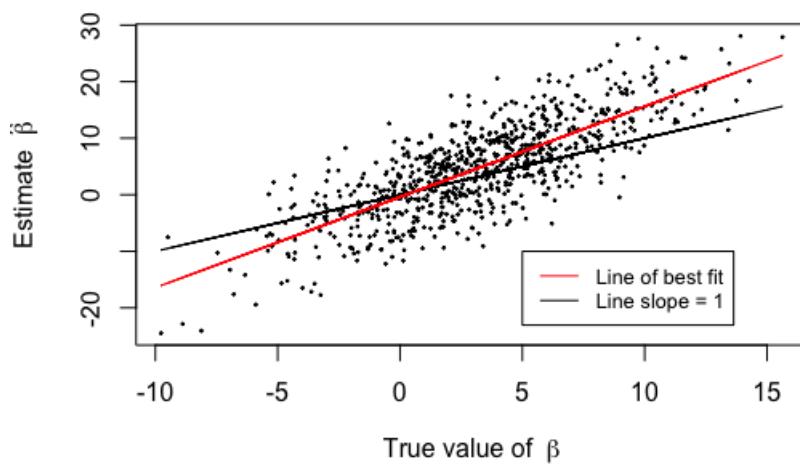


Figure 3: Scatterplot of the pairs $(\beta_j, \hat{\beta}_j)$

Evaluation

The red line represents a linear best fit of the black scatter. Its slope α^* is calculated as being roughly equal to 1.5 which is what Candes and Sur predict in their theorem. In classical linear regression, you would expect the MLE coefficients to be distributed around the black line that has slope 1, however Figure 3 shows an important consequence of the bias induced by the large n and p setting. The scattered points are decorrelated around $\alpha^*\beta$ which implies that $E(\hat{\beta}) = 1.5\beta$ which is also predicted by Candes and Sur.

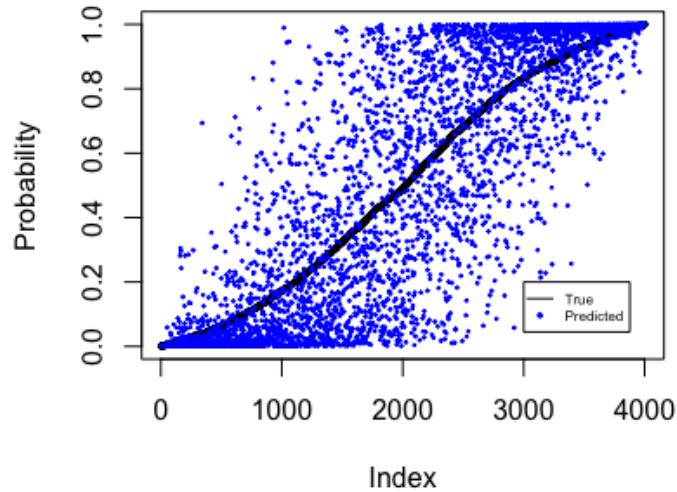


Figure 4: True and estimated conditional probabilities

On Figure 4, the black sigmoid curve shows the true conditional probability $\Pr(Y_i = 1|X) = \rho'(\mathbf{X}'_*\boldsymbol{\beta})$ and scatter points shows corresponding estimated probability $\rho'(\mathbf{X}'_*\hat{\boldsymbol{\beta}})$. We could observe dramatic shrinkage of estimated probability towards end points, specifically, when $\Pr(Y_i = 1|X) < 1/2$, many points are located near 0 and symmetrically when $\Pr(Y_i = 1|X) > 1/2$, many points are close to 1. Since the MLE coefficients tend to overestimate the effect magnitudes, the estimated probability also tend to overestimate or underestimate the true probability depending on whether the true probability is greater or less than 1/2. In practice, it shows that the MLE coefficients are not completely reliable as they may predict that an outcome is almost a certain

”success”, but in reality is not.⁴

Standard errors of the null parameters

In this setting, half of the β_j 's are sampled from a Gaussian distribution with a mean equal to 7 and variance of 1, whilst the other half vanish. As stated in the paper written by Candes and Sur, the standard deviation can be calculated mathematically using the inverse Fisher Information matrix. The classically predicted value of the standard error for null estimates is 2.33¹ but from Figure 5. we can see that the standard deviation is concentrated around 4. Therefore it is also clear that for large n and p, the standard error of the MLE null estimates is larger than in classical theory.

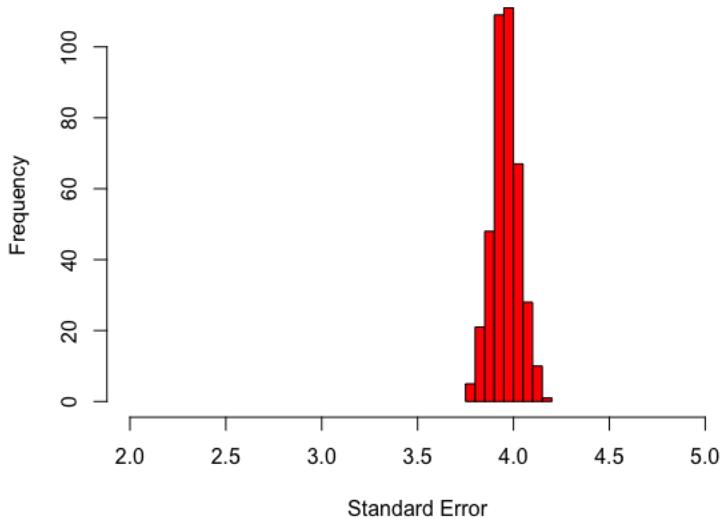


Figure 5: Standard error distribution of the 400 null parameters of the MLE fit

5 Some preliminary calculations relating γ^2 , β , kappa, and \mathbf{X}

Throughout the rest of this paper we will simulate examples in order to understand the behaviour of the MLE estimates for the logistic regression. The following results were calculated for the purpose of specifying the variance and mean of the β coefficients in such a way that theoretically it is highly probable that there exists an MLE.

Note that:

$$\text{Var}(X'_i \beta) = \gamma^2$$

We need to choose β so that this is satisfied when sampling from a covariance matrix X with entries such that $X_{ij} \sim \mathcal{N}(0, 1/n)$. To investigate properties in the same context as Candes and Sur we also let the β coefficients be Gaussian distributed i.e. $\beta_j \sim \mathcal{N}(\mu, \sigma^2) \forall j = 1 \rightarrow p$. As X and β are independently distributed and using the law of total variance, we get:

$$\begin{aligned} \text{Var}(X'_i \beta) &= \sum_{j=1}^p \text{Var}(X_{ij} \beta_j) \\ &= \sum_{j=1}^p \left(E_{\beta_j} [\text{Var}_{X_{ij} | \beta_j}(X_{ij} \beta_j)] + \text{Var}_{\beta_j} [E_{X_{ij} | \beta_j}(X_{ij} \beta_j)] \right) \\ &= \sum_{j=1}^p \left(E_{\beta_j} [\beta_j^2 \text{Var}_{X_{ij} | \beta_j}(X_{ij})] + \text{Var}_{\beta_j} [\beta_j E_{X_{ij} | \beta_j}(X_{ij})] \right) \end{aligned}$$

Sampling when $\beta_j \sim \mathcal{N}(\mu, \sigma^2)$ for fixed γ^2 and κ

Given the conditions above $\text{Var}_{X_{ij}|\beta_j}(X_{ij}) = \frac{1}{n}$ and $\text{E}_{X_{ij}|\beta_j}(X_{ij}) = 0$. The following holds:

$$\begin{aligned}\text{Var}(X_i' \boldsymbol{\beta}) &= \frac{1}{n} \sum_{j=1}^p \text{E}(\beta_j^2) = \frac{1}{n} \sum_{j=1}^p (\sigma^2 + \mu^2) \\ &= \frac{p}{n} (\sigma^2 + \mu^2) = \kappa (\sigma^2 + \mu^2)\end{aligned}$$

So for $\mu = 3$, σ^2 has to satisfy: $\sigma^2 = \frac{\gamma^2}{\kappa} - 9$, where $\kappa = \frac{p}{n}$. Therefore for the R code the standard deviation for these cases were sampled from, $\sigma = \sqrt{\frac{\gamma^2}{\kappa} - 9}$.

Randomly scaling the columns of \mathbf{X}

For \mathbf{X} as above let $Z = XD$, where $D \in \mathbb{R}^{p \times p}$ is a diagonal matrix of the following form

$$\begin{pmatrix} d_1 & & 0 \\ & \ddots & \\ 0 & & d_p \end{pmatrix}$$

where d_i realizations of a distribution $\mathcal{N}(0, \sigma_d^2)$. Hence $Z_{ij} = d_i X_{ij}$. As X_i and d_i are independently distributed, using law of iterated expectations and law of total variance, we first calculate the expectation and variance of Z_{ij} :

$$\begin{aligned}\text{E}(Z_{ij}) &= \text{E}_{d_i}[\text{E}(d_i X_{ij})] \\ &= \text{E}_{d_i}[d_i \text{E}(X_{ij})] = 0\end{aligned}$$

$$\begin{aligned}
\text{Var}(Z_{ij}) &= \text{E}_{d_i}[\text{Var}_{X_{ij}|d_j}(d_i X_{ij})] + \text{Var}_{d_j}[\text{E}_{X_{ij}|d_i}(d_i X_{ij})] \\
&= \text{E}_{d_i}[d_i^2 \text{Var}_{X_{ij}|d_i}(X_{ij})] = \frac{1}{n} \text{E}_{d_i}(d_i^2) \\
&= \frac{1}{n} \left[\text{E}_{d_i}(d_i)^2 + \text{Var}_{d_i}(d_i) \right] = \frac{\sigma_d^2}{n}
\end{aligned}$$

Repeating the same calculations as before for $\sum_{j=1}^p \text{Var}(Z_{ij}\beta_j)$, we get:

$$\begin{aligned}
\sum_{j=1}^p \text{Var}(Z_{ij}\beta_j) &= \sum_{j=1}^p \left(\text{E}_{\beta_j}[\beta_j^2 \text{Var}(Z_{ij})] + \text{Var}_{\beta_j}[\beta_j \text{E}(Z_{ij})] \right) \\
&= p \text{E}_{\beta_j}[\beta_j^2 \text{Var}(Z_{ij})] + \text{Var}_{\beta_j}[\beta_j \text{E}(Z_{ij})] \\
&= \kappa \sigma_d^2 (\sigma^2 + \mu^2) = \gamma^2
\end{aligned}$$

Sampling \mathbf{X} from Bernoulli distribution

Instead of the Gaussian distribution let $X_{ij} \sim \text{Bernoulli}(1/2)$. Trivially, $\text{E}(X_{ij}) = \frac{1}{2}$ and $\text{Var}(X_{ij}) = \frac{1}{4}$. The entries of $\boldsymbol{\beta}$ can be Gaussian iid distributed as before. Then using the expression for total variance from before:

$$\begin{aligned}
\text{Var}(X'_i \boldsymbol{\beta}) &= \sum_{j=1}^p \left(\text{E}_{\beta_j}[\beta_j^2 \text{Var}_{X_{ij}|\beta_j}(X_{ij})] + \text{Var}_{\beta_j}[\beta_j \text{E}_{X_{ij}|\beta_j}(X_{ij})] \right) \\
&= \sum_{j=1}^p \left(\text{E}_{\beta_j}[\frac{\beta_j^2}{4}] + \text{Var}_{\beta_j}[\frac{\beta_j}{2}] \right) \\
&= \frac{p}{4} (2\text{Var}_{\beta_j}[\beta_j] + \mu^2) = \frac{p}{4} (2\sigma^2 + \mu^2)
\end{aligned}$$

6 Relation between κ and the bias

An important property of the logistical regression is that the MLE does not exist in all situations. Many statisticians including Silvapulle, Santner and Duffy studied the conditions for the existence of the MLE for large n and p. It has been proved that the MLE does not exist when the data

points are completely separated and it does exist and is unique when the data points overlap. The complete separation implies that there is a vector $\mathbf{b} \in \mathbb{R}^p$ such that $y_i x_i b > 0$ for all $i = 1 \rightarrow n$. The n data points overlap if for every vector β there is at least one point such that $y_i x_i b > 0$ and one point such that $y_k x_k b < 0$.² In the paper written by Emmanuel J. Candes and Pragya Sur there is a precise characterization of the region where the MLE exists.⁵

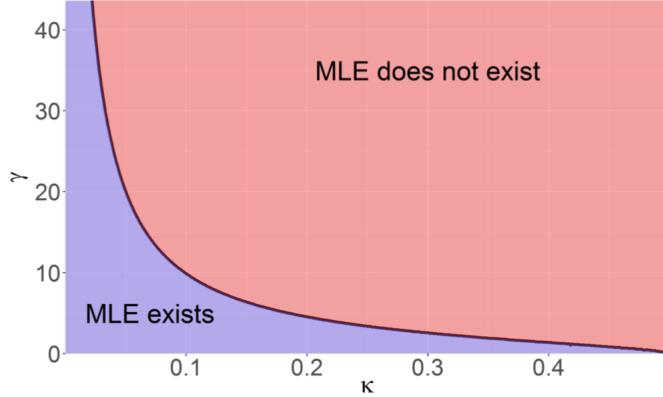


Figure 6: A precise characterization of the region where MLE exists¹

The theorem developed by Candes and Pragya Sur predicts a curve that separates the $k - \gamma$ plane into two regions.¹ The MLE does not asymptotically exist if the signal strength γ exceeds a certain function $g_{MLE}(k)$ for $g_{MLE}^{-1}(\gamma) = \min_{t \in \mathbb{R}} \{E(Z - tV)_+^2\}$ where Z is the standard Gaussian variable with density $\varphi(t)$ and V is an independent continuous random variable with density $2\rho'(\gamma t)\varphi(t)$. We begin the analysis by looking at how the bias α_* varies when κ is changing. We assume a setting when p and n both go to infinity in such a way that $\frac{p}{n} \rightarrow k$. We plot the behaviour of the kappa against the bias for two values of the γ ². We keep the β parameters with a mean equal to 3 and a variance equal to $\sqrt{\frac{\gamma^2}{k} - \mu^2}$. In this way, we can understand that the characteristics hold in other settings as well.

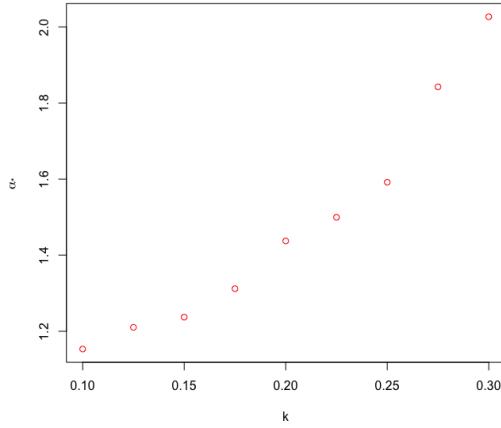


Figure 7: $\gamma^2 = 3$

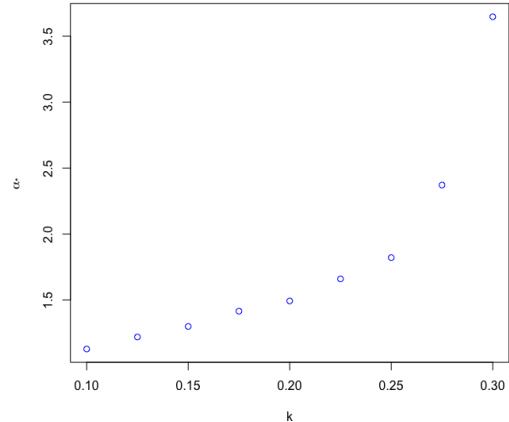


Figure 8: $\gamma^2 = 5$

The plots from Figure 7 and 8 were generated from the average of 10 random samples of α_* at each particular κ . We can deduce that the larger the dimensionality, κ , or the larger the signal strength γ , the larger the slope. Moreover as κ approaches zero, the slope approaches 1 which implies that it becomes asymptotically unbiased. After a certain point specific for each of the variances, α_* begins to diverge in Figure 6. If we plot the true value of β against the estimated coefficients as we did in the Gaussian distributed example we can see that the red line comes closer and closer to the black line as we decrease κ .

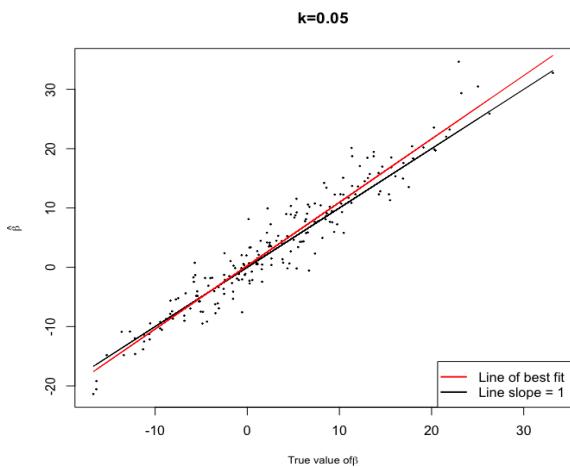


Figure 9: Scatterplot of the pairs $(\beta_j, \hat{\beta}_j)$

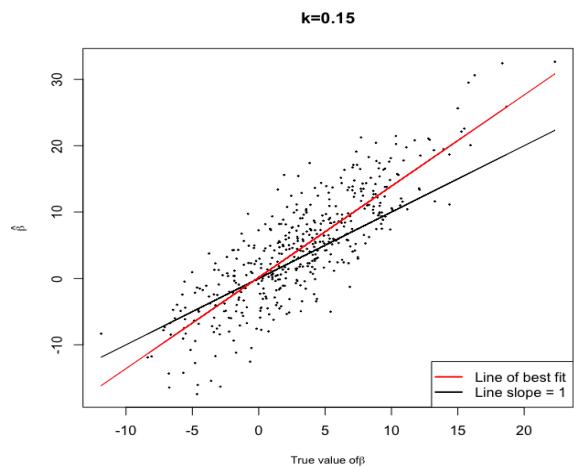


Figure 10: Scatterplot of the pairs $(\beta_j, \hat{\beta}_j)$

In their paper, Pragya Sur and Emmanuel Candes developed a theorem that statistically quantifies the behaviour of the MLE.¹ For any pseudo-Lipschitz function of order 2, the marginal distribution of the MLE obey when Z is a standard Gaussian variable

$$\frac{1}{p} \sum_{j=1}^p \psi(\hat{\beta}_j - \alpha_* \beta_j, \beta_j) \xrightarrow{\text{a.s.}} E(\psi(\sigma_* Z, \beta))$$

If we choose the the pseudo-Lipschitz function to be $\psi(t, x) = t$ then the result will become

$$\frac{1}{p} \sum_{j=1}^p (\hat{\beta}_j - \alpha_* \beta_j) \xrightarrow{\text{a.s.}} E(\sigma_* Z)$$

Because Z has the mean equal to zero, the relation becomes $\frac{1}{p} \sum_{j=1}^p (\hat{\beta}_j - \alpha_* \beta_j) \xrightarrow{\text{a.s.}} 0$.

Therefore, the estimate is centered around $\alpha_* \beta$ rather than β exactly as it is shown in our previous examples.

7 An investigation of the bias α_* in relation to varying the β_0 parameter

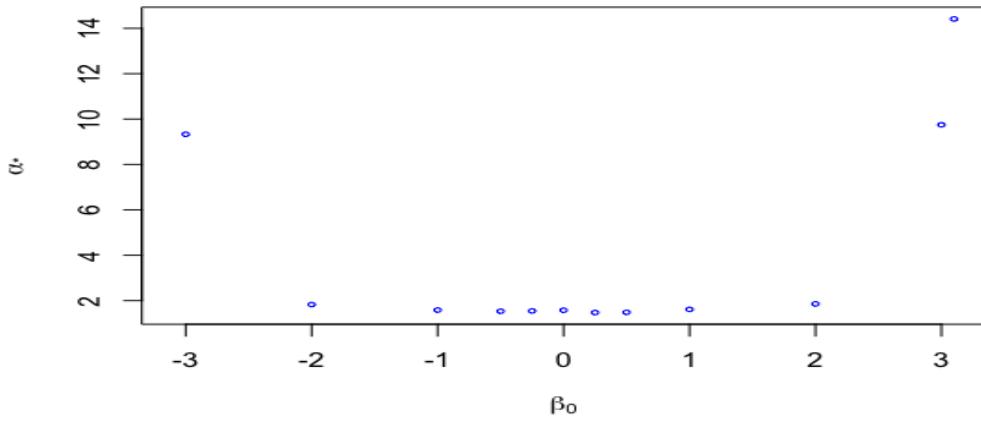


Figure 11: Scatterplot of the bias α_* against the corresponding β_0

This setting has $n = 4000$ and $p = 800$. The dimensionality is therefore, $\kappa = 0.2$. For the logistic regression the entries of $\beta \in \mathbb{R}^p$ were distributed with a mean of 3. In order to make the signal strength $\gamma^2 = 5$ the standard deviation of the entries are set to $\sqrt{\frac{5}{\kappa} - 9} = 4$. So far in our modelling the constant β_0 has not been clearly specified and the glm function has modelled the data with $\beta_0=0$. We could have chosen β_0 to be any real number.

This plot was the sample averages of 10 samples of α^* at each β_0 value for different covariate matrices. It is clear from the figure that changing between $\beta_0 = -2$ to $\beta_0 = 2$ has little influence on the behaviour of the bias α^* . Near $\beta_0 = \pm 3$ the terms seem to diverge. Even though in theory the signal strength remains unchanged as the β_0 term is changed. Hence the bias occurs even though the glm fit is being made in a well behaved neighbourhood for κ and γ^2 . This is caused by the shrinkage effect

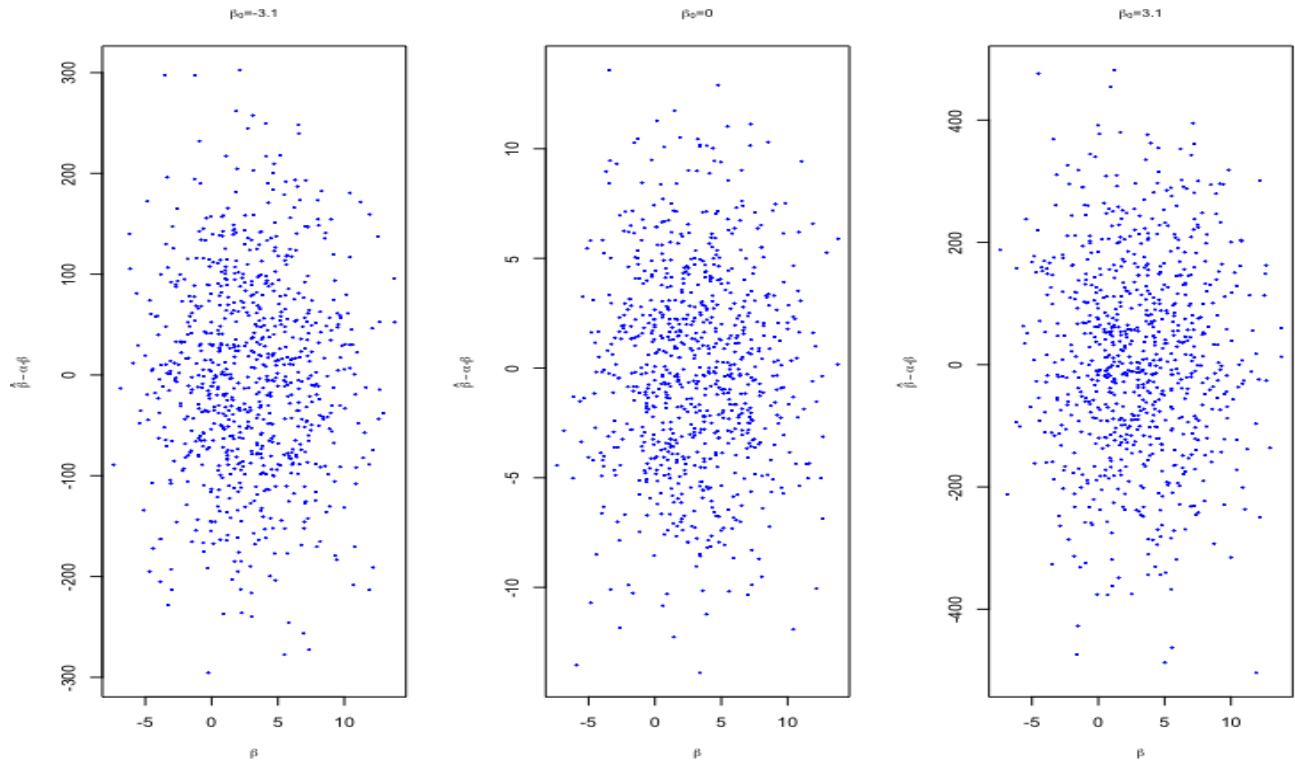


Figure 12: Scatterplot of the bias-adjusted residuals against β

As is clear from Figure 12, the fitted coefficients remain decorrelated against the true parameters when adjusted for the bias. It is clear that for $\beta_0 = \pm 3.1$ there are serious issues with the model as the terms magnitude of the error of the MLE estimates are in the order of 300-400. This is much larger than the magnitude in the $\beta_0 = 0$ case.

To explain the behaviour of the bias induced by increasing the magnitude of β_0 is a sensible result. Changing the β_0 parameter makes it so that we nearly only sample 0 or 1 from the Bernoulli distribution. This forces the glm to fit the data for the MLE estimates in the extreme regions. This means that when the glm function interpolates the data it does so in a region of the sigma curve at which there is little information on the behaviour of the curvature. This effect is because n and p are finite and so very few observations are made in the middle of, and opposite end of the S curve and therefore fails to incorporate important characteristics of the logistic model.

With regards to choosing an extreme β_0 value, consider $\beta_0 \rightarrow \pm\infty$ then the values of the other beta parameters would not be important as the data sampled from the binomial distribution would be only 0's or 1's and a sufficient model for predicting the results is trivial. The MLE would not be needed to begin with. Since the variance is quite small and the number of n is finite it is very likely that the model attempts to model this type of case. ones to the design matrix and calculated the MLE with $\beta_0 = 0$.

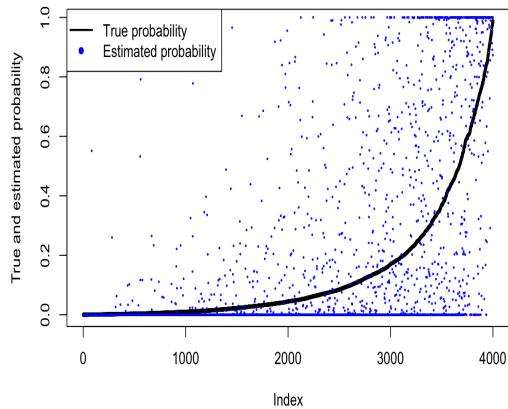


Figure 13: $\beta_0 = -3.1$

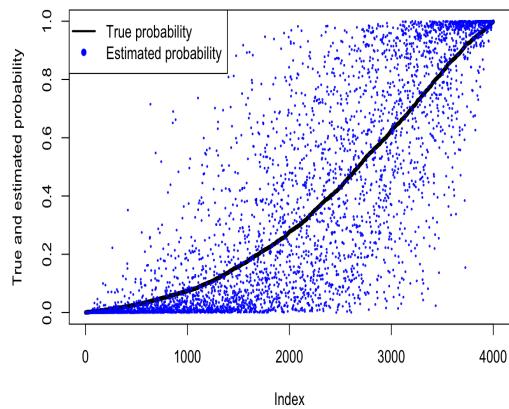


Figure 14: $\beta_0 = -1$

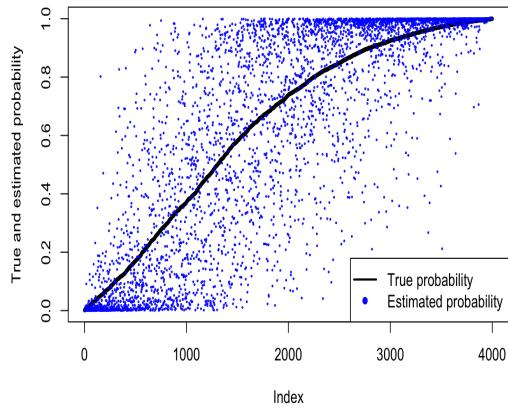


Figure 15: $\beta_0 = 1$

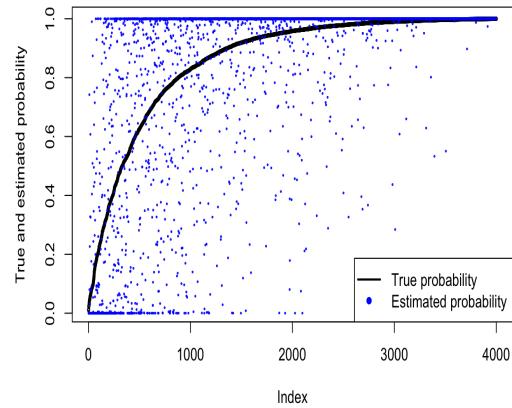


Figure 16: $\beta_0 = 3.1$

Figures 13-16. illustrate this effect clearly. For the $\beta_0 = \pm 1$ cases the curves are slightly warped toward the ends where more estimated probabilities are accounted for. Since the magnitude of β_0 is smaller, the shape of the sigmoid curve is still present. For the $\beta_0 = \pm 3.1$ the glm fit has altered the predicted probabilities so they do not follow the normal behaviour of the sigmoid curve.

8 An investigation of the bias α_* in relation to the distribution of the eigenvalues of the empirical covariance matrix

Marchenko-Pastur distribution

For a random $m \times n$ matrix X with entries $(X)_{ij} \sim N(0, \sigma^2)$, $\sigma^2 < \infty$. For $G := \frac{XX^t}{n}$, the spectrum, $\Sigma := \{\text{eigenvalues of } G\}$ is a set of random variables. Given that $\frac{m}{n} \rightarrow \kappa \in [0, 1]$, under the assumption that $n, m \rightarrow \infty$, G has eigenvalues distributed from $\rho(\lambda) = \frac{\kappa}{2\pi\sigma^2} \frac{\sqrt{(\lambda_+ - \lambda)(\lambda - \lambda_-)}}{\lambda}$ where λ_+, λ_- are the maximum and minimum eigenvalues respectively. They are calculated as $\lambda_{\pm} = \sigma^2(1 \pm \sqrt{1/Q})^2$ for the eigenvalues. Where $Q = \frac{1}{\kappa}$.

The analysis of the distribution of the spectrum is relevant to explaining the behaviour of the bias, α_* as we changed the κ and γ^2 parameters. It was noted that at a certain region predicted by the function g^{-1} ⁶ the probability that the MLE exists approaches either 0 or 1 as n and p go to infinity. This explained why very often when calculating the MLE in the vicinity of these zones the glm would diverge as it is very probable that there is no true MLE in these regions.⁵

The relationship between the covariance matrix and the approximation of the MLE

If $n >> p$ for an $n \times p$ matrix X , then a good approximation for the covariance matrix is the empirical covariance matrix $E = \frac{1}{n}XX^t$.⁷

If $\hat{\beta}$ is an MLE conditioned on X , then under certain assumptions the MLE follows a multivariate normal distribution with the true parameter as its mean and a covariance matrix of the form $\frac{1}{n}(\text{Var}(\nabla_{\beta}\ln(f_X(X; \beta))))^{-1}$. Furthermore for i.i.d. samples it can be shown that $\text{Var}(\nabla_{\beta}\ln(f_X(X; \beta))) = E(\nabla_{\beta\beta}\ln(f_X(X; \beta)))$. The covariance matrix calculated from the second order partial derivatives is the expectation hessian matrix of the log likelihood and is also referred to as the Fisher Information. This quantity is unknown due to its dependence on β .³

For a random matrix X described as before, the covariance matrix E of X and I is the fisher information of X are related in the following way; if both E and I are positive definite then $I \geq E^{-1}$ where \geq implies that $I - E^{-1}$ is positive semi definite.⁸

Variation of γ^2

These results were collected for varying signal strengths and and $\kappa = 0.2$. The γ^2 values were chosen so that the MLE estimates were collected in regions where in theory it is highly probable that the MLE exist.

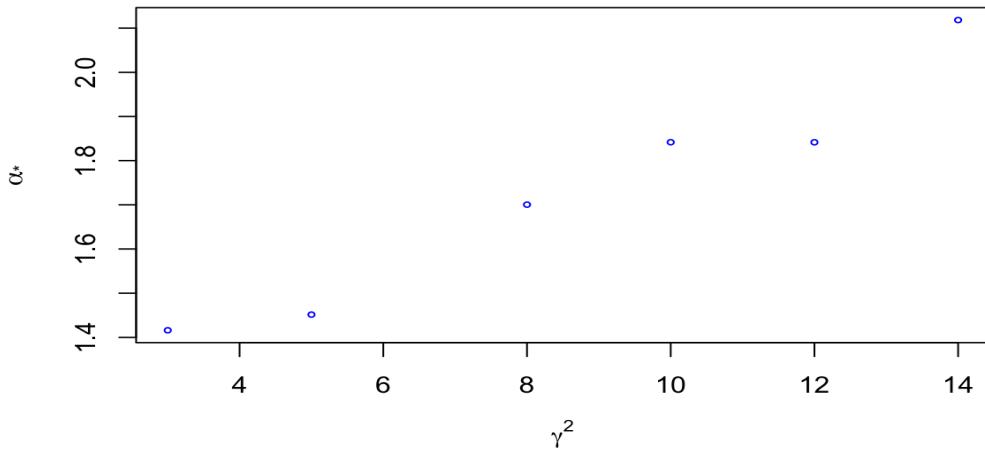


Figure 17: Relationship between bias α_* and γ^2

Figure 17 shows the behaviour of the sample average of 5 samples α_* for different values of γ^2 . As we approach the region which it is less probable a MLE exists, we see that the bias increases. This is likely caused by the increasing signal strength as it makes the individual observations less reliable.

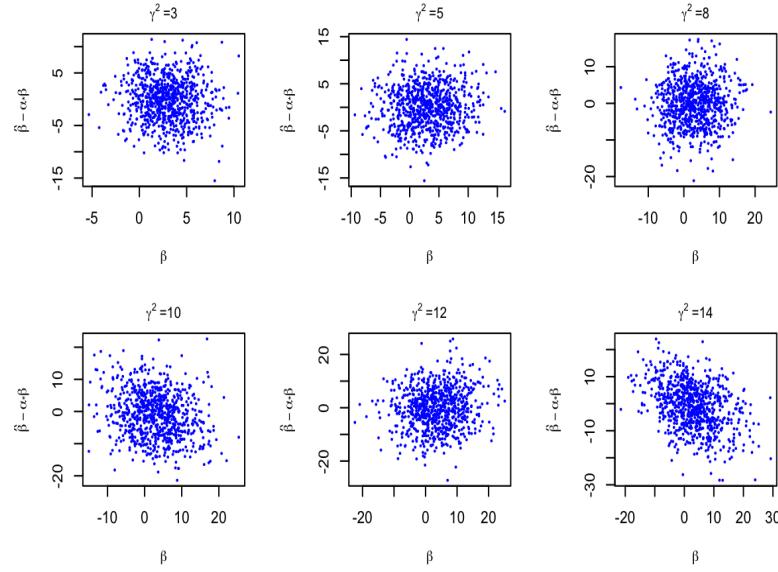


Figure 18: Scatterplot of the bias-adjusted residuals for different values of γ^2

Figure 18 models the decorrelated fit of MLE for the different signal strength values and it behaves in accordance to what Sur and Candes predicted in their theorems.

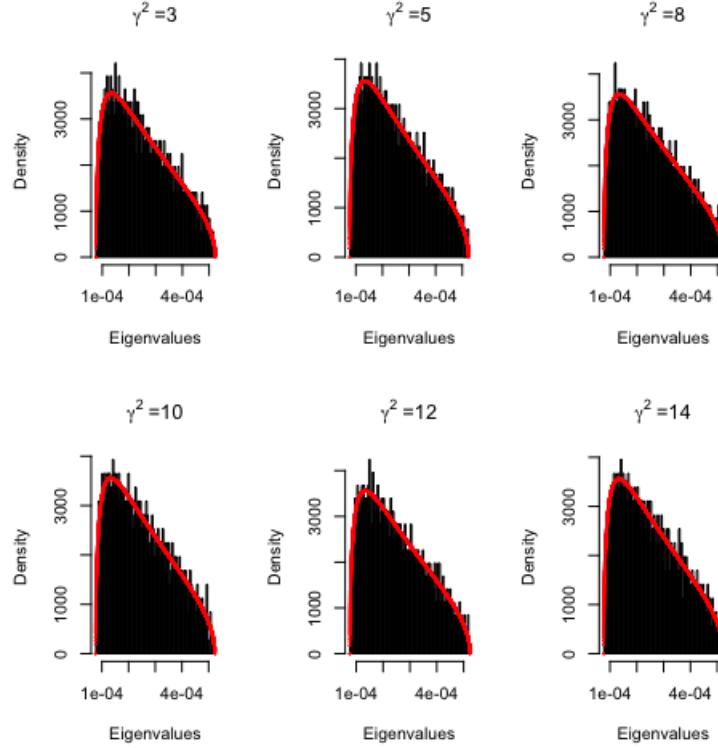


Figure 19: Distribution of eigenvalues for different values of γ^2 and the predicted Marchenko-Pastur distribution (in red)

Because only the entries of β were changed in order to alter the signal strength, it is not surprising that the eigenvalues of the empirical covariate matrix all have roughly the same distribution. The histograms slightly different shapes but this is probably due to noise. But they all behave as the Marchenko Pastur distribution predicts them to behave. This indicates that for the range of n and p that we have worked with the distribution of the eigenvalues have are not very susceptible to random noise.

Variation of κ for fixed $\gamma^2 = 5$

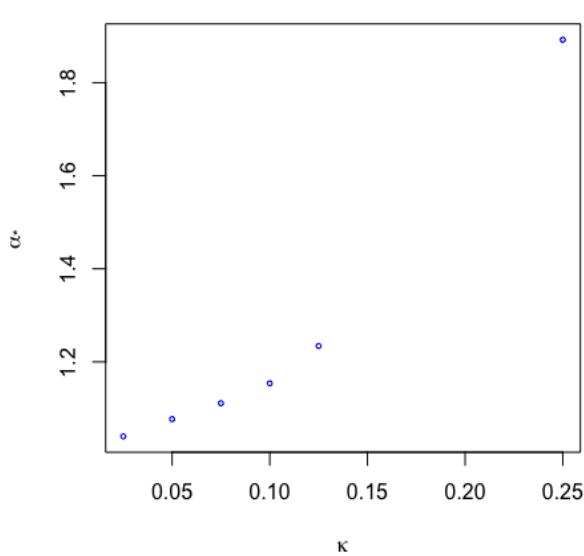


Figure 20: Relationship between α_* and κ

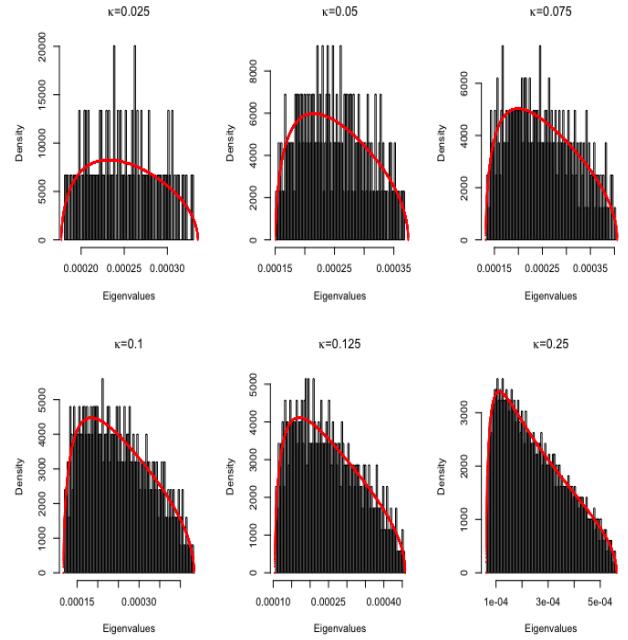


Figure 21: Distribution of eigenvalues and the predicted Marchenko-Pastur distribution (in red)

Figure 20 is a plot of the bias related to the number of parameters that were used to calculate κ . This case have already been discussed, however, it is useful to analyze how the distribution of eigenvalues changes as kappa increases. Figure 20 shows the sample average of the α_* samples. Figure 21 is a plot of the distribution of the eigenvalues for the empirical covariate matrix. Except for the first plot corresponding to $\kappa = 0.025$, the histograms exhibit the same shape as predicted by the Marchenko-Pastur distribution. This problematic plot is caused by having a very low number of parameters, $p=100$. Which my limit the effect induced by large n and p . In this example, the bias also increases as the eigenvalues begins to center near 0.

Investigation of the randomly scaled columns of \mathbf{X}

For this section the setting as is as discussed in section 5. for randomly scaling the columns by d_i and $\sigma_d^2 = 1$. In this section $n=5$ and κ is varied with entries of β chosen to have a Gaus-

sian distribution with mean,0, and as standard deviation to keep $\gamma^2 = 1, 5$ for all values of κ .

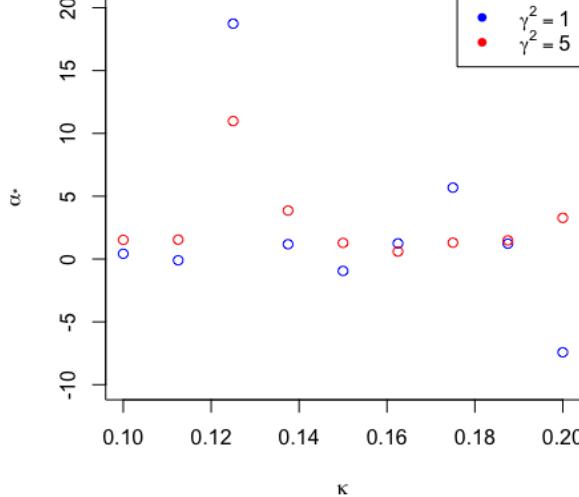


Figure 22: Relationship between α_* and κ

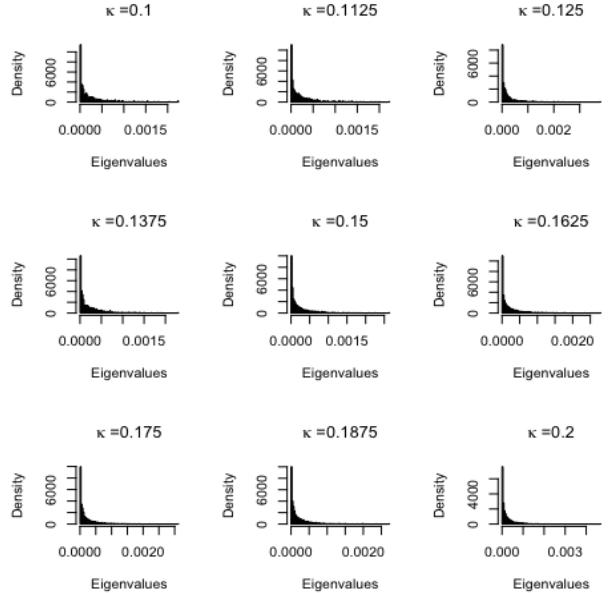


Figure 23: Distribution of eigenvalues

Figure 22. shows the sample mean of 10 samples for each α_* . The MLE is calculated in a region in which, in theory, it is likely to exist. This is why the α_* tends to stay in the region of 1. There is some noise but this could perhaps be reduced on increasing the number of samples. From Figure 22. suggests that MLE properties remain on re scaling the columns of the covariate matrix. This is in accordance to predictions made by Sur and Candes stating that the properties of the MLE can be transferred on re-scaling the covariate matrix. Another behaviour pattern which is interesting is that the $\gamma^2 = 5$ is clearly less susceptible to random noise.

Figure 23. is a plot of the eigenvalues corresponding to one of the samples for each κ . In this case it is clear that the eigenvalues are centered heavily near 0. This is strange because α_* is well behaved in this region.

Investigation of the \mathbf{X} being a randomly generated matrix entries sampled from Bernoulli(1/2)

In order to compare a Bernoulli random matrix to the matrix generated from the Gaussian distribution, it is necessary to change the β parameters so that the signal strength is $\gamma^2 = 5$. Using the results from section 5, for the entries of β following a Gaussian distribution if one sets $\mu = 0$, then $\frac{p\sigma^2}{2} = \gamma^2$. So when $\gamma^2 = 5$, we get $\sigma = \sqrt{\frac{10}{p}}$, when $\gamma^2 = 1$, $\sigma = \sqrt{\frac{2}{p}}$ and when $\gamma^2 = 3$, $\sigma = \sqrt{\frac{6}{p}}$.

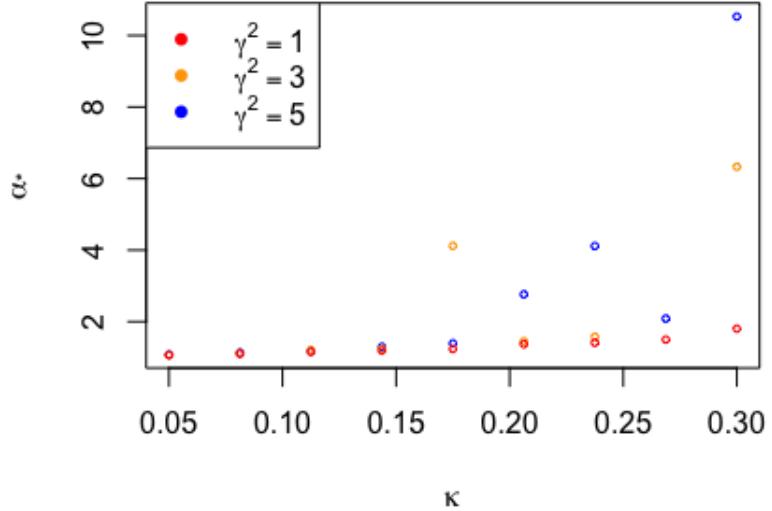


Figure 24: Relationship between α_* and κ

Figure 24 shows the sample averages for 20 samples of α_* as κ is varied. The $\gamma^2 = 1$ case is well behaved and is similar to the plot generated for the Gaussian case. The only difference is that the rate at which the bias increases is slower. The $\gamma^2 = 3$ and $\gamma^2 = 5$ cases behave as $\gamma^2 = 1$ however after $\kappa = 0.15$ it is clear that they are still susceptible to noise.

An interesting observation from the result is that for the case where $\gamma^2 = 5$ it is clear that there MLE fit begins to diverge as it has a bias of 4.5. when $\kappa = 0.3$. This suggests that this region is near the threshold at which the MLE perhaps ceases to exist when $n, p \rightarrow \infty$

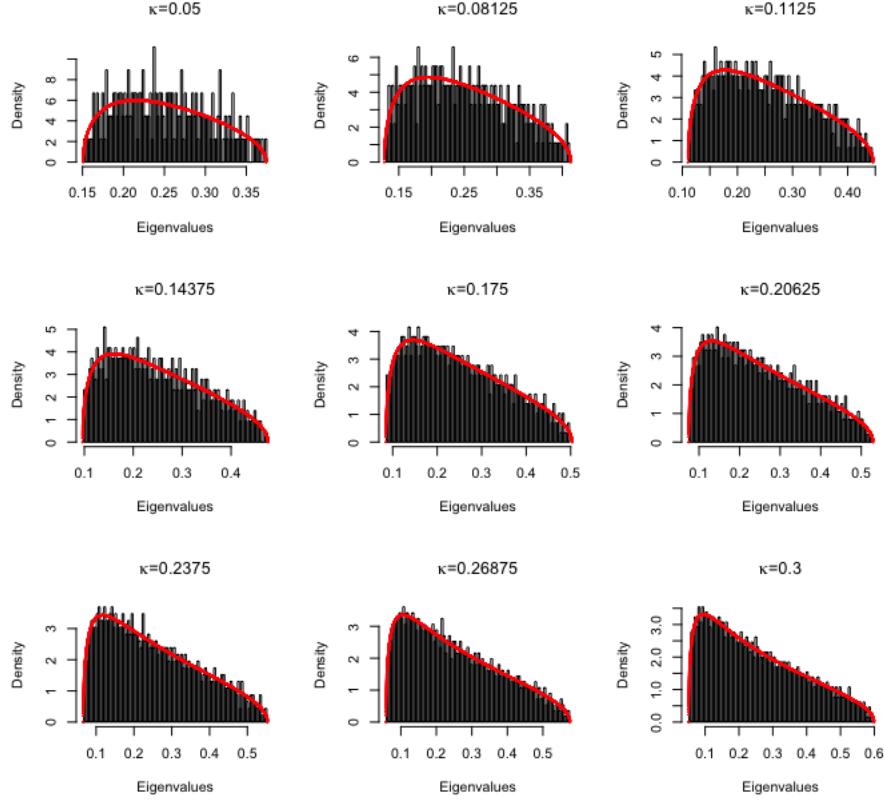


Figure 25: Distribution of eigenvalues and the predicted Marchenko-Pastur distribution (in red)

Figure 25 shows the distribution of the eigenvalues of the empirical covariance matrices as before. Here it is important to note that the eigenvalues are evenly distributed. Similar to the matrix X generated from the Gaussian distribution as κ increases, the eigenvalues begin to center near 0. The rate at which the eigenvalues begin to center near 0 is slower than for the Gaussian case which corresponds well to the rate at which the bias increases being slower. This behaviour indicates that there is likely a threshold where the MLE ceases to exist for this case as well. This would be in accordance to speculations made by Sur and Candes on how their results can be generalized for other distributions. The MLE appears to exist for the Bernoulli distribution for signal strengths where the MLE for Gaussian distributed X matrices were not well defined. This may indicate that the area κ, γ^2 region where the theoretical MLE exists as $n, p \rightarrow \infty$.

Evaluation of the results

The covariance of each matrix was calculated numerically in R to give a value for G. It is clear that the relation holds when comparing the theoretical and calculated distribution. In the regions where the R `glm` function was observed to diverge the distribution of the eigenvalues are more concentrated around 0. In one instance Figure 25 the theoretical distribution and the histogram are nearly uniformly distributed and there are very few eigenvalues near 0, and the corresponding gradient for this value is very close to 1 . For the latter part of this figure, which was in the zone in which it becomes more likely that the `glm` does not exist there is a significant number of eigenvalues near 0 here the gradient jumped to 6 and 10 for the the larger signal strengths. Between these cases a gradual build of the bias can be observed. In summary this implies that there is possibly a relation between there being a significant portion of eigenvalues of the covariance matrix being numerically near 0 and the MLE being less well defined.

From having analyzed the behaviour of the MLE as we approach the region at which in theory stops existing there is a heavy concentration of 0 eigenvalues in several cases. This bias induced could be related to the variance of the MLE estimates from the Fisher information matrix which may diverge numerically when considering the discussion of the relationship between the covariance matrix and the approximation of the MLE. The Hessian matrix of second partial derivatives is the covariance matrix of the MLE parameter estimates for β . This could be because the the covariance matrix is essentially singular for the cases where many of the eigenvalues are approximately 0. This will at least guarantee that caluclating the inverse of the covariance matrix will be computationally heavy. This explains why the `glm` function often diverges in the zones where it becomes highly probably that there is no true MLE.¹⁰

Evaluation of expanding into other distributions of random matrices

Our investigation implies that the way γ^2 and κ determine the existence of the MLE for Gaussian distributions, applies to Bernoulli distributed cases as well. The Marchenko-Pastur distribution

only requires the entries to be randomly distributed under certain conditions.⁷ Hence the distribution can still be estimated as before. Both the numerical histograms and theoretical predictions overlap which indicates that the empirical covariance matrix is a good estimate in this case also. The eigenvalues being more uniformly distributed for this covariance matrix i.e. not concentrating near 0, explains why α_* behaved in the manner previously described.

References

1.

Pragya Sur, Emmanuel J. Candès (2018) *A Modern Maximum-Likelihood Theory for High-dimensional Logistic Regression*

Available from: <http://statweb.stanford.edu/~candes/papers/LogisticAMP.pdf>

2.

David Dalpiaz (2019) *Applied Statistics with R*

Available from: <https://daviddalpiaz.github.io/appliedstats/>.

3.

Scott A. Czepiel (2002) *Maximum Likelihood Estimation of Logistic Regression Models: Theory and Implementation*.

Available from: <https://czep.net/stat/mlelr.pdf>.

4.

Pragya Sur (2018) *Presentation at Robust and High-Dimensional Statistics Event in University of California, Berkeley*.

Available from: <https://www.youtube.com/watch?v=Qph2TEWU7Og>

5.

Pragya Sur, Emmanuel J. Candès (2018) *The Phase Transition for the Existence of the Maximum*

Likelihood Estimate in High-dimensional Logistic Regression.

Available from: <https://statweb.stanford.edu/~candes/papers/MLELogistic.pdf>

6.

Jim Gatheral (2008) *Random Matrix Theory and Covariance Estimation.*

Available from: <https://mfe.baruch.cuny.edu/wp-content/uploads/2015/02/RandomMatrixBaruch2015.pdf>

7.

Miriam Huntley (2013) *Numerical Solutions to the General Marcenko Pastur Equation.*

Available from: http://www.mit.edu/~18.338/2013s/projects/mh_report.pdf

8.

Abram, Kagan, Zinoviy, Landsman(1999) *Statistics Probability Letters. Relation between the covariance and Fisher information matrices.*

Available from: <https://www.journals.elsevier.com/statistics-and-probability-letters/>

9.

Marco Taboga (2010) *Maximum likelihood - Covariance matrix estimation*

Avialable from: <https://www.statlect.com/fundamentals-of-statistics/maximum-likelihood-covariance-matrix-estimation>

10.

Christoph Molnar (2019) *Interpretable Machine Learning*

Available from: <https://christophm.github.io/interpretable-ml-book/index.html>