Project Report: A Review of High Dimensional Bootstrap Applied to Linear Models

Arka Sinha (MB2304) Arunsoumya Basu (MB2306)

Samprit Chakraborty (MB2323) Sirsha Dey (MB2330)

Indian Statistical Institute, Kolkata

May 19, 2024

Contents

1 Introduction **3**

1.1 Problems in high dimensional fixed asymptotic set up 3

1.2 Aim of the paper 3

2 High-dimensional Maximum Likelihood Theory **4**

3 An example with non-Gaussian covariates **5**

4 Why Bootstrap Fails **6**

5 Signal Strength Estimation **7**

5.1 Estimation of γ 8

5.1.1 The Probe Frontier Method 8

5.1.2 η vs γ Curve Method 9

5.2 Estimation of η: The SLOE Algorithm 9

5.3 Estimation of α_j and σ_j 11

5.4 Construction of Confidence Interval 12

5.4.1 Bootstrap-g 12

5.4.2 Bootstrap-t 13

5.5 Coverage proportion 13

6 Simulations **14**

7 Conclusion **19**

8 Acknowledgements **19**

1 Introduction

We have studied the papers *A modern maximum-likelihood theory for high-dimensional logistic regression* by Pragya Sur and Emmanuel J Candès and *An adaptively resized parametric bootstrap for inference in high-dimensional generalized linear models* by Qian Zhao and Emmanuel J Candès. Here we mainly focus on the paper by Zhao and Candès. This work [3] specifically addresses the accuracy of bootstrap methods in logistic regression models when both p (the number of parameters) and n (the sample size) are very large and increase at a fixed ratio and deal with a resized bootstrap method to infer model parameters in arbitrary dimensions.

1.1 Problems in high dimensional fixed asymptotic set up

In linear regression for example, while the residual bootstrap is weakly consistent if p is fixed and $n \to \infty$, it is inconsistent when $n, p \to \infty$ in such a way that $p/n \to \kappa > 0$. Motivated by results from high-dimensional maximum likelihood theory it is proposed to use corrected residuals to achieve correct inference. Another example is: although the nonparametric bootstrap can be used to construct a valid confidence region for the spectrum of a covariance matrix when the problem dimension is fixed, it yields incorrect estimates of the distribution of the largest eigenvalue if $p/n \to \kappa > 0$.

1.2 Aim of the paper

The paper study the bootstrap for inferring the distribution of the maximum likelihood estimator (MLE) in high-dimensional logistic regression models. It is found that the standard parametric bootstrap and the pairs bootstrap are both incorrect. It was also shown that recent high-dimensional

maximum likelihood theory (HDT) developed for multivariate Gaussian covariates does not correctly predict the distribution of the MLE when the covariates are heavy tailed. Both these failures call for solutions and in this paper, a novel resized bootstrap by combining the bootstrap method with insights from HDT is designed.

2 High-dimensional Maximum Likelihood Theory

The High-dimensional theory (HDT) generalizes the classical asymptotic setting and offers a more accurate characterization of the distribution of M-estimators when both n and p are large.

Consider a logistic model in which the covariates $X \in \mathbb{R}^p$ are mutivariate Gaussian and $\mathbb{P}(Y = 1 \mid X) = \sigma\left(X^\top \beta\right)$, where $\sigma(t) = 1/\left(1 + e^{-t}\right)$ is the usual sigmoid function. It was shown [2] that if $\hat{\beta}$ denotes the MLE, then

$$\frac{\sqrt{n}\left(\hat{\beta}_j - \alpha_\star \beta_j\right)}{\sigma_\star / \tau_j} \xrightarrow{d} \mathcal{N}(0, 1)$$

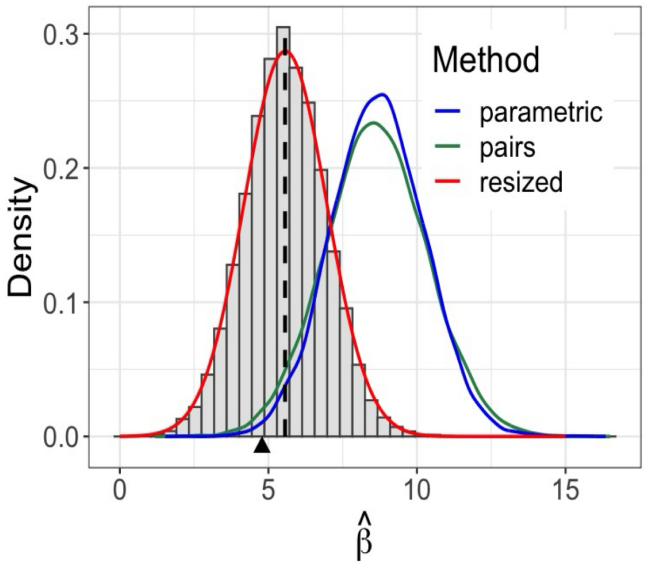
where β_j (resp. $\hat{\beta}_j\Big)$ is the j th (resp. estimated) model coefficient. In contrast to classical asymptotic theory, which states that the MLE is unbiased, the MLE is centered at $\alpha_\star \beta_j$, for some $\alpha_\star > 1$ whenever κ is positive. The standard deviation is $\sigma_\star / \tau_j$; here, τ_j is the conditional standard deviation of the j th variable given all the other variables whereas the parameters $\alpha_\star$ and $\sigma_\star$ are determined by κ and the signal strength γ^2 defined as $\gamma^2 = \mathrm{Var}\left(X^\top \beta\right)$. The parameters $\alpha_\star$ and $\sigma_\star$ both increase as either the dimensionality κ increases or the signal-to-noise ratio γ increases.

Now, the above asymptotic distribution does not hold when the covariates follow a general distribution. For instance, suppose the covariates $X \in \mathbb{R}^p$ are sampled from a multivariate t-distribution. Then we expect that $\alpha_\star$ and $\sigma_\star$ would depend on the degrees of freedom of the t-distribution. In the next section we will describe an example with non-Gaussian covariates.

3 An example with non-Gaussian covariates

Here [3] they have simulated a high-dimensional logistic regression model with $n = 4000$ and $p = 400$. They sample covariates from a multivariate t-distribution and standardize each variable so that $\mathrm{Var}(X_j) = 1/p$. They pick 50 non-null variables and sample their coefficients from a mixture of Gaussians $\mathcal{N}(5, 1)$ and $\mathcal{N}(-5, 1)$ with equal weights.

In the figure below, a histogram of the logistic MLE of a randomly chosen coefficient in 10,000 repeated experiments is shown. Here, the covariates are sampled from a multivariate t-distribution with 8 degrees of freedom. The bootstrap MLE densities are displayed for the parametric bootstrap (blue), the pairs bootstrap (green) and the proposed resized bootstrap (red). The triangle indicates the true coefficient and the dashed line indicates the average MLE.



We can see that the MLE is approximately Gaussian. But the MLE is biased upward, since the value of the true coefficient under study is less than the average MLE. Due to both a poor centering

and a poor assessment of variability, the classical Wald confidence interval would significantly undercover β_j.

For the parametric bootstrap, they generate samples by fixing the covariates at the observed values and sample responses from a logistic model whose coefficients equal the MLE; put another way, they have chosen $\hat{F} = F_{\hat{\beta}}$. The parametric bootstrap (blue) does not begin to describe the MLE distribution since the average value is almost twice that of the true coefficient.

The pairs bootstrap generates bootstrap samples by sampling with replacement from the observed data, i.e., $\hat{F}$ to be the empirical distribution is chosen. The pairs bootstrap also fails to approximate the MLE distribution since the green curve shifts to the right and is much wider than the histogram.

The red curve shows the accuracy of the proposed resized bootstrap. We can see that this best describes the MLE distribution; for instance, both the mean and standard deviation are close to the true values.

4 Why Bootstrap Fails

The reason why paired bootstrap fails is because the number of unique samples available is $(1-\frac{1}{e})n$ for large n. Thus the effective dimensionality is inflated and becomes $\frac{\kappa e}{e-1}$. Now since Bias and Variance of MLE increase with increase in dimensionality, the pair bootstrap overestimates both.

The parametric bootstrap however overestimates the signal strength as it can be shown when the covariates follow $N(0,1)$, then according to [2] Theorem 2,

$$\lim_{n,p\to\infty} \mathrm{Var}(X_{\mathrm{new}}^{\top}\hat{\beta}) \xrightarrow{\text{a.s.}} \alpha_*^2\gamma^2 + \kappa\sigma_*^2 > \gamma^2$$

Also, a problem that can occur is that the MLE ceases to exist. That happens when the two classes

are perfectly linearly separable. Intuitively, data gets more linearly separable when it is projected to higher dimensional spaces and thus increasing dimensionality should cause a problem. Also if the signal strength is high, the two classes would be quite separated from each other. As shown in the figure below, there is a phase transition curve for MLE.

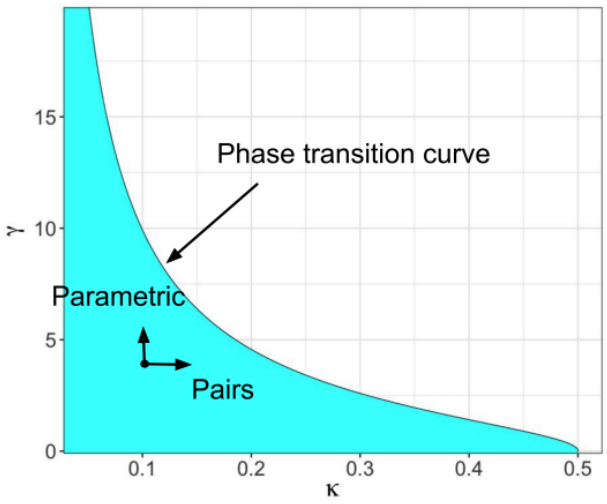


Figure 1: According to the high-dimensional theory, the asymptotic distribution of the MLE depends on the problem dimension κ and the signal strength γ. The pairs bootstrap over-estimates κ whereas the parametric bootstrap over-estimates γ. Therefore, both methods lead to incorrect estimates of the MLE distribution. The blue region shows pairs of values of (κ, γ) where the MLE exists.

5 Signal Strength Estimation

As noted previously, signal strength is one of the important parameters that are required. In fact as claimed in [2], all the predictions of their HDT depend on β only through the signal strength γ. Essentially, a high signal strength means a high predictive power of the co-variates. The original

paper [2] on HDT uses the Probe Frontier Method to estimate the signal strength, which uses the idea that the MLE ceases to exist asymptotically beyond a certain threshold, dimensionality kept fixed.

It is known from [2] Theorem 1 that for each γ, there is a maximum dimensionality $g_{\text{MLE}}^{-1}(\gamma)$ at which the MLE ceases to exist. We propose an estimate $\hat{\kappa}$ of $g_{\text{MLE}}^{-1}(\gamma)$ and set $\hat{\gamma} = g_{\text{MLE}}(\hat{\kappa})$.

5.1 Estimation of γ

5.1.1 The Probe Frontier Method

Given a data sample (y_i, X_i), we begin by choosing a fine grid of values $\kappa \leq \kappa_1 \leq \kappa_2 \leq \ldots \leq \kappa_K$. For each κ_j, we execute the following procedure:

Subsample Sample $n_j = \frac{p}{\kappa_j}$ observations from the data without replacement, rounding to the nearest integer. Ignoring the rounding, the dimensionality of this subsample is $\frac{p}{n_j} = \kappa_j$.

Check whether MLE exists For the subsample, check whether the MLE exists or not. This is done by solving a linear programming feasibility problem; if there exists a vector $b \in \mathbb{R}^p$ such that $X_i^\top b$ is positive when $y_i = 1$ and negative otherwise, then perfect separation between cases and controls occurs and the MLE does not exist. Conversely, if the linear program is infeasible, then the MLE exists.

Repeat Repeat the two previous steps B times and compute the proportion of times $\hat{\pi}(\kappa_j)$ the MLE does not exist.

We next find (κ_{j-1}, κ_j), such that κ_j is the smallest value in K for which $\hat{\pi}(\kappa_j) \geq 0.5$. By linear

interpolation between κ_{j-1} and κ_j, we obtain $\hat{\kappa}$ for which the proportion of times the MLE does not exist would be 0.5. We set $\hat{\gamma} = g_{\text{MLE}}(\hat{\kappa})$

The problem with this method is that this works for gaussian but for other distributions, the phase transition curve depends also on other parameters related to covariate distribution, and since we invert the phase transition curve at the value of $\hat{\kappa}$, this parameter dependence makes it hard to do so.

5.1.2 η vs γ Curve Method

We are interested in the relation between the following two quantities

$$\gamma^2 = \text{Var}(X_{\text{new}}^\top \beta)$$

$$\eta^2 = \text{Var}(X_{\text{new}}^\top \hat{\beta})$$

As mentioned in [3], both the bias and variance of the MLE increases with increasing signal strength γ. Thus we expect a monotonically increasing curve when η is plotted against γ.

The one-to-one relation between γ and $\eta(\gamma)$ suggests that, if $\text{Var}(X^\top \beta^*) \approx \text{Var}(X^\top \beta)$, then $\text{Var}(X^\top \hat{\beta}^*) \approx \text{Var}(X^\top \hat{\beta})$, where $\hat{\beta}^*$ denotes the MLE when the true coefficient is β^*. Thus, we estimate γ^2 by $\text{Var}(X^\top \beta^*)$, where β^* obeys : $\text{Var}(X_{\text{new}}^\top \hat{\beta}^*) = \eta^2$.
Computing η^2 is not possible here, since it is evaluated at a new data point, so the SLOE algorithm is used to estimate it.

5.2 Estimation of η: The SLOE Algorithm

In this section, we describe the Signal Strength Leave-One-Out Estimation (SLOE) Algorithm in brief. [1] introduced this technique for finding η in a high dimensional setup for a logistic regression

model, where the MLE $\hat{\beta}$ satisfies

$$\hat{\beta} = \underset{\beta \in \mathbb{R}^p}{\operatorname{argmax}} \frac{1}{n} \sum_{i=1}^{n} y_i \log(g(X_i^\top \beta)) + (1 - y_i) \log(1 - g(X_i^\top \beta))$$

with $g(z) = \dfrac{e^z}{1 + e^z}$ being the CDF of the logistic distribution. It is to be noted that this log likelihood cannot be maximized analytically and relies on an iterative (Newton-Raphson or Fisher Scoring) method instead. Now, a simple estimator of η^2 in this scenario is the leave one out estimator (which has been shown to be consistent for η in [1]) $\hat{\eta}^2_{\text{LOO}}$ given by

$$\hat{\eta}^2_{\text{LOO}} = \frac{1}{n} \sum_{i=1}^{n} (X_i^\top \hat{\beta}_{-i})^2 - \left(\frac{1}{n} \sum_{i=1}^{n} (X_i^\top \hat{\beta}_{-i}) \right)^2$$

where $\hat{\beta}_{-i}$ denotes the MLE calculated without the pair $\left(y_i, X_i^\top\right)$. However, this is computationally inefficient, since it requires maximizing the log likelihood n times. So the authors used S_i, a computationally cheaper alternative of $X_i^\top \hat{\beta}_{-i}$ where

$$S_i = X_i^\top \hat{\beta} + q_i f'_{y_i}(X_i^\top \hat{\beta})$$

with $q_i = \dfrac{X_i^\top \mathbf{H}^{-1} X_i}{1 - (X_i^\top \mathbf{H}^{-1} X_i) f''_{y_i}(X_i^\top \hat{\beta})}$ and $f_y(t) = \log(1 + e^{-(2y-1)t})$, $y \in \{0, 1\}$ denoting the negative log likelihood for the response y and linear predictor t.

Here $\mathbf{H}$ denotes the Hessian of the negative log likelihood evaluated at the MLE $\hat{\beta}$. So, the new estimator $\hat{\eta}^2_{\text{SLOE}}$ is given by

$$\hat{\eta}^2_{\text{SLOE}} = \frac{1}{n} \sum_{i=1}^{n} S_i^2 - \left(\frac{1}{n} \sum_{i=1}^{n} S_i \right)^2$$

The main idea behind this proposal is a first Taylor expansion of g around each of the n quantities $X_i^\top \hat{\beta}$ and the observation that the leave-one-out Hessian matrices (arising amidst the derivation) are rank 1 updates of the original Hessian matrix, so the inverse of the former can be computed through the Sherman-Morrison identity.

The proposed estimator $\hat{\eta}^2_{\text{SLOE}}$ has been shown to be consistent for η^2 in [1], but we skip the details of the proof. The algorithm used thereafter to get the curve and thereby the value of signal strength γ is given as follows:

Algorithm 1 Estimating Signal Strength

Require: Observed data (x_i, y_i), $1 \leq i \leq n$, and a GLM formula.

Estimate $\tilde{\eta} = \text{Var}(X^\top \hat{\beta})$ via leave-one-out techniques;

Pick a sequence $\{0 = s_1, \ldots, s_I = 1\}$;

for $i = 1, \ldots, I$ **do**

  Set $\beta_{s_i} = s_i \times \hat{\beta}$ and $\gamma_i = \text{sd}(X\beta_{s_i})$;

  for $j = 1, \ldots, J$ **do**

    Simulate Y_i^j given x_i using β_{s_i} as model coefficients;

    Fit a GLM for (x_i, Y_i^j) to estimate $\hat{\eta}^j(\gamma(s_i))$;

  end for

end for

Fit a smooth curve $\hat{\eta}(\gamma)$;

Estimate $\hat{\gamma}$ by solving $\hat{\eta}(\hat{\gamma}) = \tilde{\eta}$;

Output: Estimated $\hat{\gamma}$.

5.3 Estimation of α_j and σ_j

In this section, we describe how to estimate the bias α_j and the standard deviation σ_j. To estimate σ_j, the standard deviation of the bootstrap MLE is used, i.e.,

$$\hat{\sigma}_j^2 = \frac{1}{B-1} \sum_{b=1}^{B} \left(\hat{\beta}_j^b - \bar{\beta}_j \right)^2, \quad \text{where} \quad \bar{\beta}_j = \frac{1}{B} \sum_{b=1}^{B} \hat{\beta}_j^b.$$

We estimate α_j by weighted regression: that is, we regress $\bar{\beta}^b$ onto $\beta_\star$ by assigning to each MLE coordinate a weight inversely proportional to its estimated variance $\hat{\sigma}_j^2$. We assume a common bias factor because all the α_j 's are equal when the covariates are multivariate Gaussian.

Algorithm 2 Resized Bootstrap Procedure

Require: Observed data (x_i, y_i), $1 \leq i \leq n$, and a GLM formula.

Compute resized coefficients $\beta_\star$;

for $b = 1, \ldots, B$ **do**

Simulate Y_i^b given x_i using $\beta_\star$ as model coefficients;

Fit a GLM for (x_i, Y_i^b) to obtain the bootstrap MLE $\hat{\beta}^b$;

end for

Estimate the standard deviation of the MLE $\hat{\sigma}_j$;

Estimate a common factor $\hat{\alpha}$ by regressing $\bar{\beta}$ onto $\beta_\star$ with weights proportional to $1/\hat{\sigma}_j^2$;

Output: $\hat{\alpha}$ and $\hat{\sigma}_j$.

5.4 Construction of Confidence Interval

In this section, we discuss methods to construct the confidence interval.

5.4.1 Bootstrap-g

It is assumed that the MLE is approximately Gaussian, i.e.,

$$\frac{\hat{\beta}_j - \alpha_j \beta_j}{\sigma_j} \approx \mathcal{N}(0, 1)$$

where α_j and σ_j denote the bias and standard deviation, this approximation yields the following $(1 - q)$ CI for β_j :

$$\left[\frac{1}{\hat{\alpha}_j} \left(\hat{\beta}_j - z_{1-q/2} \hat{\sigma}_j \right), \frac{1}{\hat{\alpha}_j} \left(\hat{\beta}_j - z_{q/2} \hat{\sigma}_j \right) \right].$$

Here, z_q is the quantile of a standard Gaussian, while $\hat{\alpha}_j$ and $\hat{\sigma}_j$ refer to estimates of α_j and σ_j.

5.4.2 Bootstrap-t

When the normal approximation is inadequate, the following approximation is used

$$\frac{\hat{\beta}_j - \alpha_j \beta_j}{\sigma_j} \overset{d}{\approx} \frac{\hat{\beta}_j^b - \alpha_j \beta_{\star,j}}{\sigma_j},$$

where the right-hand side refers to the distribution of $\hat{\beta}_j^b$ conditional on the observed covariates. After plugging in the estimated $\hat{\alpha}_j$ and $\hat{\sigma}_j$, we obtain a $(1-q)$ CI as

$$\left[\frac{1}{\hat{\alpha}_j}\left(\hat{\beta}_j - t_j^b[1-q/2]\hat{\sigma}_j\right), \frac{1}{\hat{\alpha}_j}\left(\hat{\beta}_j - t_j^b[q/2]\hat{\sigma}_j\right)\right]$$

where $t_j^b[q]$ denotes the quantile of the right-hand side of $\frac{\hat{\beta}_j^b - \alpha_j \beta_{\star,j}}{\sigma_j}$.This confidence interval is referred as the "Bootstrap-t" confidence interval.

5.5 Coverage proportion

We implemented both the approaches of finding estimated coverage probabilities of a non-null component of the parameter vector, as described in the previous section. However, both the estimates, arising from large sample normal approximation and empirical bootstrap distribution, involved $\hat{\alpha}$ and $\hat{\sigma}$, which were computationally expensive to calculate across multiple samples. So we calculated the MLEs $\hat{\beta}$ across 100 replications while the other terms of the CI were calculated once and for all. Thus we obtained a crude estimate of coverage proportion, nonetheless the values came out to be 0.94 in both cases, indicating that we would have obtained good coverage proportions, had we not recycled those values.

6 Simulations

Here we present the results of the simulations mentioned in [3] that we could reproduce. The most important of them is the following revelation, again cited in [3].

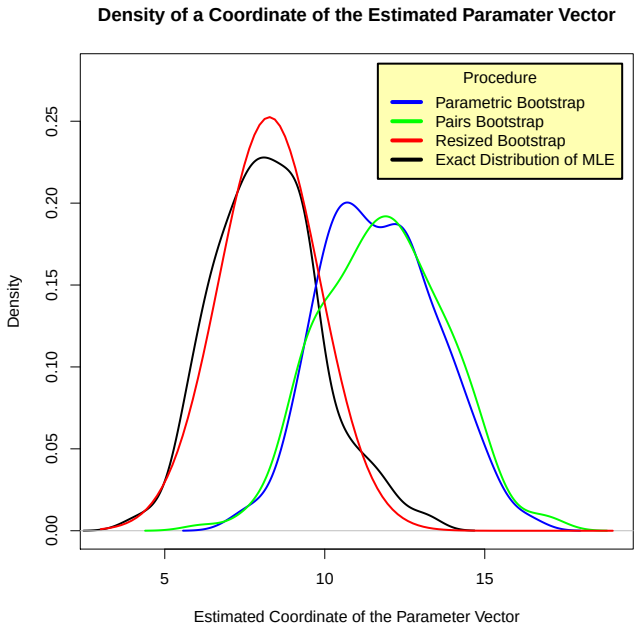


Figure 2: The density of a non-null coordinate of the parameter vector obtained by parametric bootstrap, pairs bootstrap and resized bootstrap compared to the exact distribution of the MLE

We chose $n = 2000, p = 200$ (so that $\kappa = 0.1$) and 20 non-null parameters, each of them simulated from an equiprobable mixture of $\mathcal{N}(-5, 1)$ and $\mathcal{N}(5, 1)$. The rows of the covariate matrix were chosen to be iid samples from multivariate t distribution with 8 degrees of freedom. As revealed in [3], both pairs bootstrap and parametric bootstrap perform poorly, having large positive biases and large variances.

	Parametric	Pairs	Resized	Actual MLE
Mean	11.70	11.83	8.27	8.22
S.D.	1.77	1.91	1.58	1.66

The parametric bootstrap, pairs bootstrap and actual distribution of a selected coordinate of the MLE were all simulated using 200 replications. We observe that the mean and s.d. of the actual MLE is best captured by the resized bootstrap. In fact, resized bootstrap gives a decent approximation for the *distribution* of the selected coordinate of the MLE.

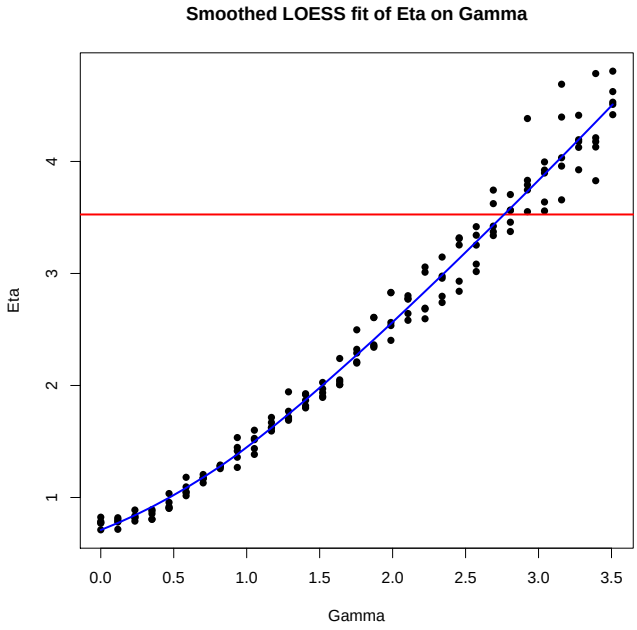


Figure 3: The η vs γ curve from our simulations

This shows the $\gamma - \eta$ curve, which, as expected, is monotonic. The MLE was scaled by an equispaced sequence in $[0, 1]$ with 31 terms, and in each case the γ values were calculated. A LOESS smoothing was done to obtain a continuous curve and the η available at hand, i.e., $\tilde{\eta}$, was

used to solve for $\hat{\gamma}$, as prescribed in Algorithm 1.

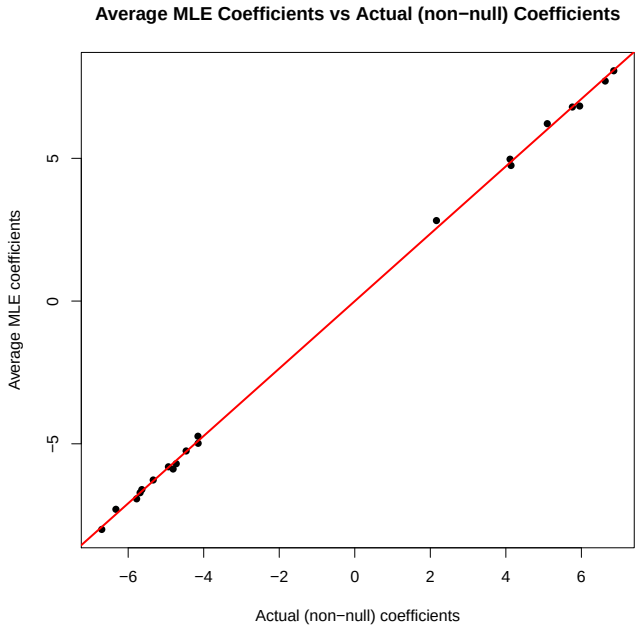


Figure 4: The non-null coefficients and their average MLE counterparts

This figure shows the maximum likelihood estimates of the 20 non-null parameters, averaged across 200 replications versus the actual values of the corresponding parameters. Even with only 200 replications, the linear fit is excellent, with an intercept very close to 0 and a slope of 1.16. This indicates that the biases of the MLEs for different coordinates do not depend much on the magnitude of the actual coefficients.

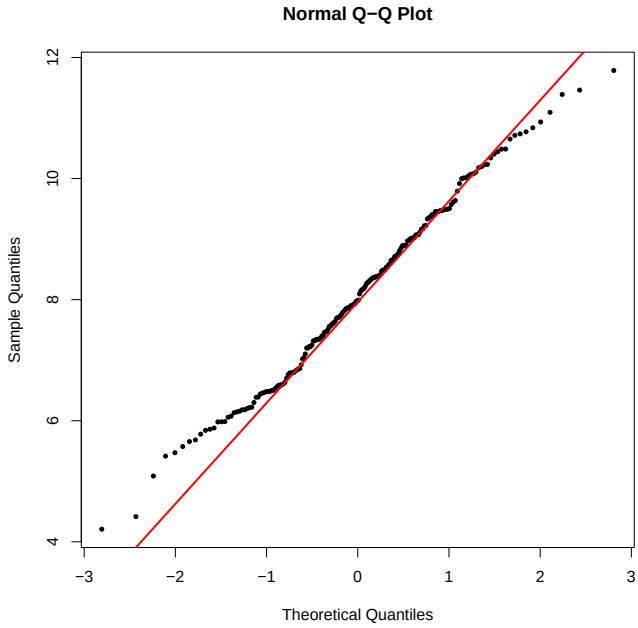


Figure 5: Q-Q Plot of the MLE of a selected non-null parameter coordinate

A Q-Q plot of the MLE of a non-null coordinate against the quantiles of the standard normal distribution indicates that the MLE is more or less normally distributed. This contains two layers of approximation: the sample size ($n = 2000$) required for the asymptotics to kick in as well as the number of MLEs sampled (200), so we can safely attribute the deviation to the fact that only 200 MLEs were found.

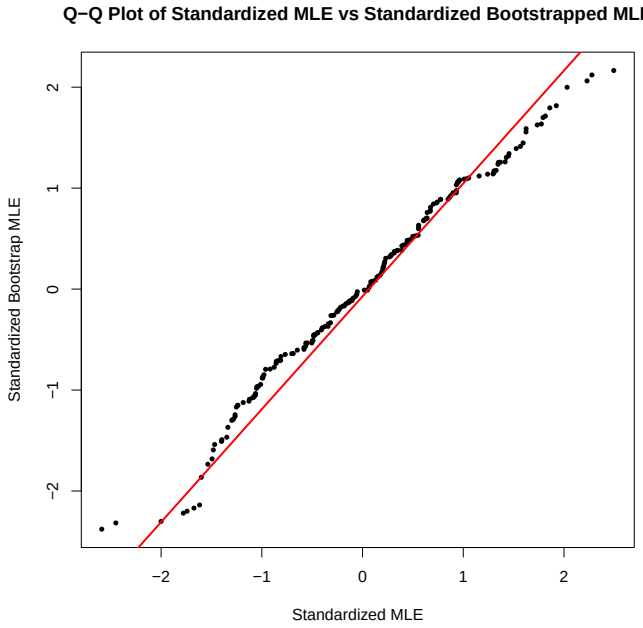


Figure 6: Q-Q Plot of the standardized MLE and the standardized bootstrap MLE for the selected (non-null) parameter coordinate

A further Q-Q plot of the standardized MLE with the standardized bootstrap MLE (using resized bootstrap for parameter estimation followed by simulation from that model) highlights the efficacy of resized bootstrap. Again, both the simulations were with 200 replications only.

We also tried to investigate the performance of resized bootstrap when the sample size is not too large, say, 400 for $\kappa = 0.1$ using covariates from the Pareto distribution with unit scale and shape parameter 5, as prescribed in [3]. However, we encountered the issue of non existence of MLE in quite a large number of cases which probably occurred due to the linear separability of the data points in $\mathbb{R}^p$. Curiously enough, the authors of [3] have not made any reference to this phenomenon.

7 Conclusion

In this project, we reviewed a modern technique of performing statistical inference in linear models when the ratio of the dimension of the parameter space and the sample size is bounded away from 0. The failure of the conventional bootstrap techniques in high dimensional setup called for an alternative computationally feasible technique which could be used for inference. We saw that the resized bootstrap technique introduced in [3] combining ideas from [1] and [2] is a reasonable alternative, successfully capturing the distribution of the MLE and giving good estimates of the coverage probabilities, although it is yet to be tested extensively. We implemented the simulation schemes suggested in [3] to verify some of the plots and statements made there. However, we could not explore the realm of small sample sizes since the problem of non-existence of MLE in low sample size logistic regression settings left us a bit puzzled, and unfortunately we could not come up with a way to tackle it.

All the R codes can be found here.

8 Acknowledgements

We would like to express our gratitude towards Prof. Soumendu Sundar Mukherjee, Indian Statistical Institute, Kolkata, for presenting us with the opportunity to work on this project. The topic, which has been quite recently introduced in the literature, was interesting to work on and made the learning experience enjoyable. We extend our thanks to the authors of the paper.

References

[1] Steve Yadlowsky et al. SLOE: A faster method for statistical inference in high-dimensional logistic regression. 2021.

[2] Pragya Sur and Emmanuel J Candès. A modern maximum-likelihood theory for high-dimensional logistic regression. *Proceedings of the National Academy of Sciences*, 116(29):14516–14525, 2019.

[3] Qian Zhao and Emmanuel J Candes. An adaptively resized parametric bootstrap for inference in high-dimensional generalized linear models. *arXiv preprint arXiv:2208.08944*, 2022.