On the Likelihood Ratio Test in High Dimensional Logistic Regression

Pragya Sur

Ph.D. Candidate
Department of Statistics
Stanford University

Algorithmic Challenges in Protecting Privacy for Biomedical Data
IPAM
Jan, 2018

Collaborators

- Emmanuel Candès (Stanford)
- Yuxin Chen (Princeton and formerly Stanford)

P-values in Logistic Regression

Logistic regression in R: $n = 100$, $p = 30$

```
> fit = glm(y ~ X, family = binomial)
> summary(fit)

Call:
glm(formula = y ~ X, family = binomial)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.7727  -0.8718   0.3307   0.8637   2.3141

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.086602   0.247561   0.350  0.72647
X1           0.268556   0.307134   0.874  0.38190
X2           0.412231   0.291916   1.412  0.15790
X3           0.667540   0.363664   1.836  0.06642 .
X4          -0.293916   0.331553  -0.886  0.37536
X5           0.207629   0.272031   0.763  0.44531
X6           1.104661   0.345493   3.197  0.00139 **
...
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Logistic regression in R: $n = 100$, $p = 30$

```
> fit = glm(y ~ X, family = binomial)
> summary(fit)

Call:
glm(formula = y ~ X, family = binomial)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.7727  -0.8718   0.3307   0.8637   2.3141

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.086602   0.247561   0.350  0.72647
X1           0.268556   0.307134   0.874  0.38190
X2           0.412231   0.291916   1.412  0.15790
X3           0.667540   0.363664   1.836  0.06642 .
X4          -0.293916   0.331553  -0.886  0.37536
X5           0.207629   0.272031   0.763  0.44531
X6           1.104661   0.345493   3.197  0.00139 **
...
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Can inference be trusted?

Likelihood Ratio Test

- Independent observations $(y_i, \boldsymbol{X}_i)$, $y_i \in \{0,1\}, \boldsymbol{X}_i \in \mathbb{R}^p$

$$\mathbb{E}\left[y_i | \boldsymbol{X}_i\right] = \text{logit}^{-1}(\boldsymbol{X}_i^\top \boldsymbol{\beta})$$

- Likelihood function:

$$\mathcal{L}(\boldsymbol{\beta}) = \prod_{i=1}^{n} \frac{e^{(\boldsymbol{X}_i^\top \boldsymbol{\beta}) y_i}}{1 + e^{(\boldsymbol{X}_i^\top \boldsymbol{\beta})}}$$

- Testing problem $\mathcal{H}_0 : \beta_j = 0$ vs $\mathcal{H}_1 : \beta_j \neq 0$

$$\text{LRT}_j = \frac{\mathcal{L}\big(\hat{\boldsymbol{\beta}}_{(-j)}\big)}{\mathcal{L}\big(\hat{\boldsymbol{\beta}}\big)} \qquad \begin{array}{lcll} \hat{\boldsymbol{\beta}} & = & \arg\max_{\boldsymbol{\beta} \in \mathbb{R}^p} & \mathcal{L}(\boldsymbol{\beta}) \\ \hat{\boldsymbol{\beta}}_{(-j)} & = & \arg\max_{\boldsymbol{\beta} \in \mathbb{R}^p, \boldsymbol{\beta}_j = 0} & \mathcal{L}(\boldsymbol{\beta}) \end{array}$$

Wilks' phenomenon

Theorem (Wilks' theorem)

Under suitable 'regularity conditions', under $\mathcal{H}_0$, p fixed and $n \to \infty$

$$-2 \log LRT_j \xrightarrow{d} \chi^2_1$$

Samuel Wilks (1906–1964)

Wilks' phenomenon

Theorem (Wilks' theorem)

Under suitable 'regularity conditions', under $\mathcal{H}_0$, p fixed and $n \to \infty$

$$-2 \log LRT_j \xrightarrow{d} \chi_1^2$$

- If $Y \sim \chi_1^2$, p-val $:= 1 - F_{\chi_1^2}(-2 \log \mathrm{LRT}_j)$

$$\implies \text{p-val} \sim \mathrm{U}(0,1)$$

- Testing problem $\beta_1, \ldots, \beta_k = 0$ vs at least one non-zero, same conditions,

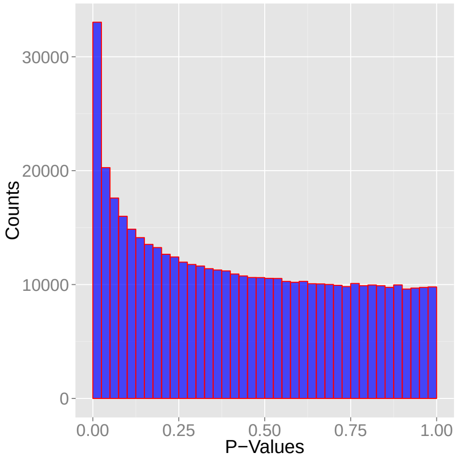
$$-2 \log \mathrm{LRT} \xrightarrow{d} \chi_k^2$$

Samuel Wilks (1906–1964)

Classical p-values in high dimensions

$\boldsymbol{X}$ has i.i.d $\mathcal{N}(0,1)$ entries, $\boldsymbol{\beta} = \mathbf{0}$, $y_i \overset{\text{i.i.d.}}{\sim}$ Bernoulli$(1/2)$

$n = 4000 \quad p = 1200$



Observed earlier in Candès et. al. '16

Grossly problematic

- Large number of false discoveries.
- Estimates of p-value probabilities (take 1):

	Classical p-values (Accuracy)
$\mathbb{P}\{$p-value $\leq 5\%\}$	$11.1044\%(0.0668\%)$
$\mathbb{P}\{$p-value $\leq 1\%\}$	$3.6383\%(0.038\%)$
$\mathbb{P}\{$p-value $\leq 0.5\%\}$	$2.2477\%(0.0292\%)$
$\mathbb{P}\{$p-value $\leq 0.1\%\}$	$0.7519\%(0.0155\%)$
$\mathbb{P}\{$p-value $\leq 0.05\%\}$	$0.4669\%(0.0112\%)$
$\mathbb{P}\{$p-value $\leq 0.01\%\}$	$0.1575\%(0.0064\%)$

What went wrong?

Theorem (Recall...)

Under suitable 'regularity conditions', under $\mathcal{H}_0$, p fixed and $n \to \infty$

$$-2 \log \mathit{LRT}_j \xrightarrow{\mathrm{d}} \chi^2_1$$

- Satisfies all regularity conditions.
- Finite sample effect?

Merely a finite sample effect?

$$\mathbb{E}\left[2\Lambda_j\right] = 1 + \frac{\alpha}{n} + O\left(\frac{1}{n^2}\right) \qquad \Lambda_j = -\log(\mathsf{LRT}_j)$$

Corrected statistic

$$\frac{2\Lambda_j}{1 + \frac{\alpha_n}{n}}$$

Bartlett (1937), Cordeiro (1983), Moulton et. al. (1993)

Merely a finite sample effect?

$$\mathbb{E}\left[2\Lambda_j\right] = 1 + \frac{\alpha}{n} + O\left(\frac{1}{n^2}\right) \qquad \Lambda_j = -\log(\mathsf{LRT}_j)$$

Corrected statistic

$$\frac{2\Lambda_j}{1 + \frac{\alpha_n}{n}}$$

$$\alpha_n = \frac{n}{2}\left[\mathrm{Tr}(\boldsymbol{D}_p^2 - \boldsymbol{D}_{p-1}^2)\right]$$

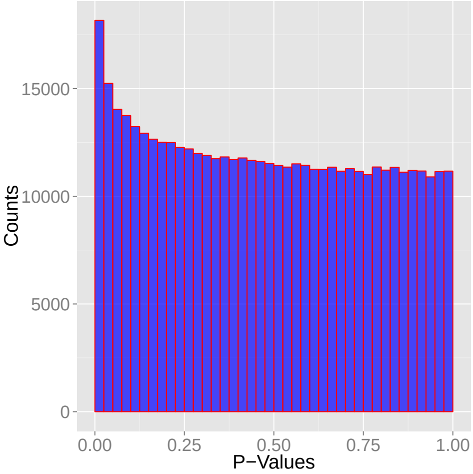
$$\boldsymbol{D}_p = \mathrm{Diag}\left(\boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top\right)$$

$$\boldsymbol{D}_{p-1} = \mathrm{Diag}\left(\boldsymbol{X}_{(-j)}\left({\boldsymbol{X}_{(-j)}}^\top \boldsymbol{X}_{(-j)}\right)^{-1}\boldsymbol{X}_{(-j)}^\top\right)$$

Bartlett (1937), Cordeiro (1983), Moulton et. al. (1993)

Bartlett corrected p-values

$$\boldsymbol{\beta} = \mathbf{0}, \quad n = 4000, \quad p = 1200$$



Estimates of p-value probabilities (take 2)

	Classical	Bartlett-corrected
$\mathbb{P}\{\text{p-value} \leq 5\%\}$	$11.1044\%(0.0668\%)$	$6.9592\%(0.0534\%)$
$\mathbb{P}\{\text{p-value} \leq 1\%\}$	$3.6383\%(0.038\%)$	$1.6975\%(0.0261\%)$
$\mathbb{P}\{\text{p-value} \leq 0.5\%\}$	$2.2477\%(0.0292\%)$	$0.9242\%(0.0178\%)$
$\mathbb{P}\{\text{p-value} \leq 0.1\%\}$	$0.7519\%(0.0155\%)$	$0.2306\%(0.0078\%)$
$\mathbb{P}\{\text{p-value} \leq 0.05\%\}$	$0.4669\%(0.0112\%)$	$0.124\%(0.0056\%)$
$\mathbb{P}\{\text{p-value} \leq 0.01\%\}$	$0.1575\%(0.0064\%)$	$0.0342\%(0.0027\%)$

What went wrong?

Theorem (Recall...)

Under suitable 'regularity conditions', under $\mathcal{H}_0$, p fixed and $n \to \infty$

$$-2 \log LRT_j \xrightarrow{\mathrm{d}} \chi^2_1$$

- Traditional finite sample corrections do not work.
- Recall in simulations $n = 4000, p = 1200 \implies p/n = 0.3$.

What happens in High Dimensions ?

Asymptotic setup

- Sequence of problems with $n, p \to \infty$ and

$$p/n \to \kappa \in (0, 0.5)$$

MLE exists iff $\kappa < 1/2$ - S., Chen, Candès ('17), Amelunxen et. al. ('14), Cover ('65)

Asymptotic setup

- Sequence of problems with $n, p \to \infty$ and

$$p/n \to \kappa \in (0, 0.5)$$

- $\boldsymbol{X}$ has i.i.d. rows $\boldsymbol{X}_i \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}), \boldsymbol{\Sigma} \in \mathbb{R}^{p \times p}$
- $y_i | \boldsymbol{X} \overset{\text{ind}}{\sim}$ Bernoulli $\left(e^{\boldsymbol{X}_i^\top \boldsymbol{\beta}} / \left(1 + e^{\boldsymbol{X}_i^\top \boldsymbol{\beta}} \right) \right)$

MLE exists iff $\kappa < 1/2$ - S., Chen, Candès ('17), Amelunxen et. al. ('14), Cover ('65)

Asymptotic setup

- Sequence of problems with $n, p \to \infty$ and

$$p/n \to \kappa \in (0, 0.5)$$

- $\boldsymbol{X}$ has i.i.d. rows $\boldsymbol{X}_i \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}), \boldsymbol{\Sigma} \in \mathbb{R}^{p \times p}$
- $y_i | \boldsymbol{X} \overset{\text{ind}}{\sim}$ Bernoulli $\left(e^{\boldsymbol{X}_i^\top \boldsymbol{\beta}} / \left(1 + e^{\boldsymbol{X}_i^\top \boldsymbol{\beta}} \right) \right)$
- $\boldsymbol{\beta} = \mathbf{0}$

MLE exists iff $\kappa < 1/2$ - S., Chen, Candès ('17), Amelunxen et. al. ('14), Cover ('65)

Asymptotic setup

- Sequence of problems with $n, p \to \infty$ and
$$p/n \to \kappa \in (0, 0.5)$$
- $\boldsymbol{X}$ has i.i.d. rows $\boldsymbol{X}_i \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}), \boldsymbol{\Sigma} \in \mathbb{R}^{p \times p}$
- $y_i | \boldsymbol{X} \overset{\text{ind}}{\sim}$ Bernoulli $\left(e^{\boldsymbol{X}_i^\top \boldsymbol{\beta}} / \left(1 + e^{\boldsymbol{X}_i^\top \boldsymbol{\beta}}\right)\right)$
- $\boldsymbol{\beta} = \mathbf{0}$

Goal: characterize asymptotic distribution of $2\Lambda_j = -2\log(\mathsf{LRT}_j)$

MLE exists iff $\kappa < 1/2$ - S., Chen, Candès ('17), Amelunxen et. al. ('14), Cover ('65)

Main result

Theorem (S., Chen and Candès ('17))

If $p/n \to \kappa \in (0, 0.5)$, *Log-Likelihood Ratio (LLR) obeys*

$$2\Lambda_j \stackrel{\mathrm{d}}{\to} \frac{\tau_*^2}{\lambda_*}\chi_1^2$$

Main result

Theorem (S., Chen and Candès ('17))

If $p/n \to \kappa \in (0, 0.5)$*, Log-Likelihood Ratio (LLR) obeys*

$$2\Lambda_j \;\stackrel{\mathrm{d}}{\to}\; \frac{\tau_*^2}{\lambda_*}\chi_1^2$$

where $(\tau_*, \lambda_*) \in \mathbb{R}_+^2$ *is the unique solution to the system of equations*

$$\begin{cases} \kappa\tau^2 = \mathbb{E}\left[\left(\Psi\left(\tau Z;\ \lambda\right)\right)^2\right] \\ \kappa = \mathbb{E}\left[\Psi'\left(\tau Z;\ \lambda\right)\right] \end{cases} \qquad Z \sim \mathcal{N}(0,1)$$

Main result

Theorem (S., Chen and Candès ('17))

If $p/n \to \kappa \in (0, 0.5)$, *Log-Likelihood Ratio (LLR) obeys*

$$2\Lambda_j \overset{\mathrm{d}}{\to} \frac{\tau_*^2}{\lambda_*}\chi_1^2$$

where $(\tau_*, \lambda_*) \in \mathbb{R}_+^2$ *is the unique solution to the system of equations*

$$\begin{cases} \kappa\tau^2 = \mathbb{E}\left[\left(\Psi\left(\tau Z;\ \lambda\right)\right)^2\right] \\ \kappa = \mathbb{E}\left[\Psi'\left(\tau Z;\ \lambda\right)\right] \end{cases} \qquad Z \sim \mathcal{N}(0, 1)$$

LLR for testing any subset of variables of finite cardinality k *converges to* $\frac{\tau_*^2}{\lambda_*}\chi_k^2$

Main result

Theorem (S., Chen and Candès ('17))

If $p/n \to \kappa \in (0, 0.5)$, *Log-Likelihood Ratio (LLR) obeys*

$$2\Lambda_j \overset{\mathrm{d}}{\to} \frac{\tau_*^2}{\lambda_*}\chi_1^2$$

where $(\tau_*, \lambda_*) \in \mathbb{R}_+^2$ *is the unique solution to the system of equations*

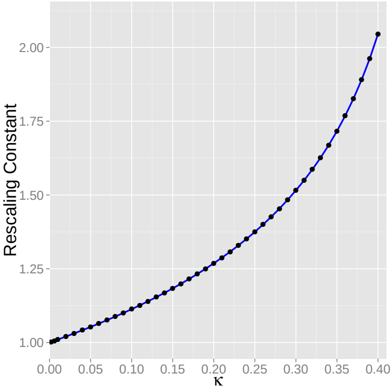
$$\begin{cases} \kappa\tau^2 = \mathbb{E}\left[\left(\Psi\left(\tau Z;\ \lambda\right)\right)^2\right] \\ \kappa = \mathbb{E}\left[\Psi'\left(\tau Z;\ \lambda\right)\right] \end{cases} \qquad Z \sim \mathcal{N}(0,1)$$

LLR for testing any subset of variables of finite cardinality k *converges to* $\frac{\tau_*^2}{\lambda_*}\chi_k^2$

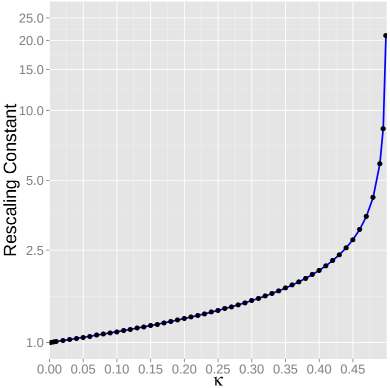
$$\rho(t) = \log(1 + e^t), \quad \Psi(z; \lambda) := \lambda\rho'(\mathsf{prox}_{\lambda\rho}(z))$$

$$\mathsf{prox}_{\lambda\rho}(z) := \arg\min_{x\in\mathbb{R}} \left\{\lambda\rho(x) + \frac{1}{2}(x - z)^2\right\}$$

The rescaling constant



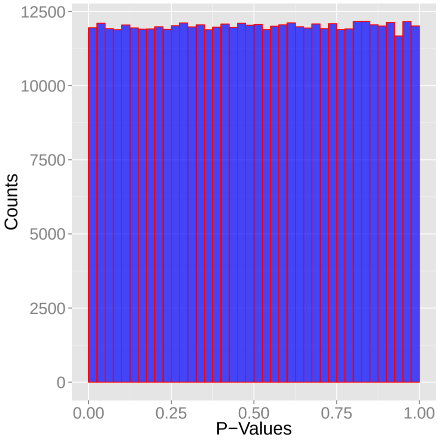
κ ranges from 0 to 0.4
y-axis: linear scale



κ ranges from 0 to 0.5
y-axis: log scale

S.-Chen-Candès (SCC) corrected p-values

$$\boldsymbol{\beta} = \mathbf{0}, \quad n = 4000, \quad p = 1200$$

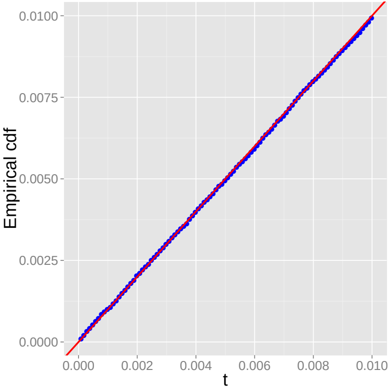


Estimates of pvalue probabilities (take 3)

	Bartlett-corrected	SCC-corrected
$\mathbb{P}\{$p-value $\leq 5\%\}$	$6.9592\%(0.0534\%)$	$5.0110\%(0.0453\%)$
$\mathbb{P}\{$p-value $\leq 1\%\}$	$1.6975\%(0.0261\%)$	$0.9944\%(0.0186\%)$
$\mathbb{P}\{$p-value $\leq 0.5\%\}$	$0.9242\%(0.0178\%)$	$0.4952\%(0.0116\%)$
$\mathbb{P}\{$p-value $\leq 0.1\%\}$	$0.2306\%(0.0078\%)$	$0.1008\%(0.0051\%)$
$\mathbb{P}\{$p-value $\leq 0.05\%\}$	$0.124\%(0.0056\%)$	$0.0542\%(0.0036\%)$
$\mathbb{P}\{$p-value $\leq 0.01\%\}$	$0.0342\%(0.0027\%)$	$0.0104\%(0.0014\%)$

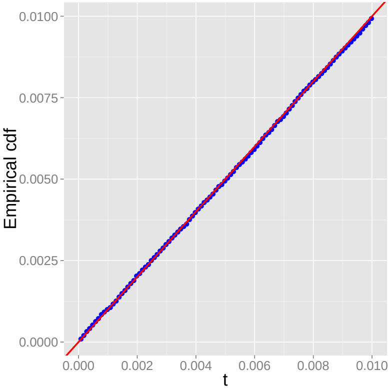
Efficacy in finite samples

Efficacy in finite samples

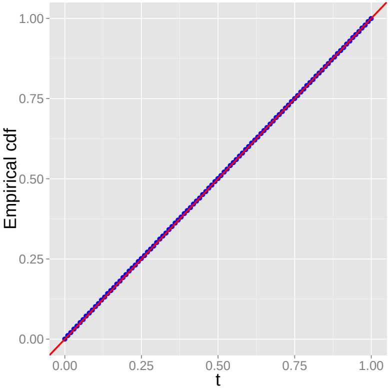


(a) $n = 4000, p = 1200$
$0 \leq t \leq 0.01$

Efficacy in finite samples



(a) $n = 4000, p = 1200$
$0 \leq t \leq 0.01$



(b) $n = 200, p = 60$
$0 \leq t \leq 1$

Log-Likelihood Ratio Analysis

- Class labels $y_i \in \{0,1\}$ or $\tilde{y}_i \in \{-1,1\}$, Negative log-likelihood:

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^{n} \rho(-\tilde{y}_i \boldsymbol{X}_i^\top \boldsymbol{\beta}) \stackrel{\mathrm{d}}{=} \rho(\boldsymbol{X}_i^\top \boldsymbol{\beta}), \quad \rho(t) = \log(1+e^t)$$

- Denote $\tilde{\boldsymbol{\beta}} := \hat{\boldsymbol{\beta}}_{(-j)}$. LLR simplifies:

$$2\Lambda_j \;=\; \ell(\tilde{\boldsymbol{\beta}}) - \ell(\hat{\boldsymbol{\beta}}) \;=\; \sum_{i=1}^{n} \left\{ \rho(\boldsymbol{X}_i^\top \tilde{\boldsymbol{\beta}}) - \rho(\boldsymbol{X}_i^\top \hat{\boldsymbol{\beta}}) \right\} = Q_2 + Q_3$$

Log-Likelihood Ratio Analysis

- Class labels $y_i \in \{0,1\}$ or $\tilde{y}_i \in \{-1,1\}$, Negative log-likelihood:

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^{n} \rho(-\tilde{y}_i \boldsymbol{X}_i^\top \boldsymbol{\beta}) \overset{\mathrm{d}}{=} \rho(\boldsymbol{X}_i^\top \boldsymbol{\beta}), \quad \rho(t) = \log(1+e^t)$$

- Denote $\tilde{\boldsymbol{\beta}} := \hat{\boldsymbol{\beta}}_{(-j)}$. LLR simplifies:

$$2\Lambda_j \;=\; \ell(\tilde{\boldsymbol{\beta}}) - \ell(\hat{\boldsymbol{\beta}}) \;=\; \sum_{i=1}^{n} \left\{ \rho(\boldsymbol{X}_i^\top \tilde{\boldsymbol{\beta}}) - \rho(\boldsymbol{X}_i^\top \hat{\boldsymbol{\beta}}) \right\} = Q_2 + Q_3$$

$$Q_2 = \sum_{i=1}^{n} \rho''\left(\boldsymbol{X}_i^\top \hat{\boldsymbol{\beta}}\right)\left(\boldsymbol{X}_i^\top \tilde{\boldsymbol{\beta}} - \boldsymbol{X}_i^\top \hat{\boldsymbol{\beta}}\right)^2 = (\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}})^\top \nabla^2 \ell(\hat{\boldsymbol{\beta}})(\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}})$$

$$Q_3 = \frac{1}{3}\sum_{i=1}^{n} \rho'''(\gamma_i)\left(\boldsymbol{X}_i^\top \tilde{\boldsymbol{\beta}} - \boldsymbol{X}_i^\top \hat{\boldsymbol{\beta}}\right)^3, \gamma_i \in (\boldsymbol{X}_i^\top \tilde{\boldsymbol{\beta}}, \boldsymbol{X}_i^\top \hat{\boldsymbol{\beta}})$$

Difference from Classical Theory

Recall $2\Lambda_j = Q_2 + Q_3$.

Classical Theory	**High Dimensional Regime**

Difference from Classical Theory

Recall $2\Lambda_j = Q_2 + Q_3$.

Classical Theory	**High Dimensional Regime**
$Q_3 = o_P(1)$	$Q_3 = o_P(1)$

Difference from Classical Theory

Recall $2\Lambda_j = Q_2 + Q_3$.

Classical Theory

$Q_3 = o_P(1)$

$Q_2 = (\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}})^\top \nabla^2 \ell(\hat{\boldsymbol{\beta}})(\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}})$

$= (\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}})^\top E[\nabla^2 \ell(\hat{\boldsymbol{\beta}})](\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}) + o_P(1)$

High Dimensional Regime

$Q_3 = o_P(1)$

Difference from Classical Theory

Recall $2\Lambda_j = Q_2 + Q_3$.

Classical Theory

$Q_3 = o_P(1)$

$Q_2 = (\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}})^\top \nabla^2 \ell(\hat{\boldsymbol{\beta}})(\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}})$

$= (\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}})^\top E[\nabla^2 \ell(\hat{\boldsymbol{\beta}})](\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}) + o_P(1)$

$= p\hat{\boldsymbol{\beta}}_j^2/4\kappa + o_P(1)$

High Dimensional Regime

$Q_3 = o_P(1)$

Difference from Classical Theory

Recall $2\Lambda_j = Q_2 + Q_3$.

Classical Theory

$Q_3 = o_P(1)$

$Q_2 = (\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}})^\top \nabla^2 \ell(\hat{\boldsymbol{\beta}})(\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}})$

$= (\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}})^\top E[\nabla^2 \ell(\hat{\boldsymbol{\beta}})](\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}) + o_P(1)$

$= p\hat{\boldsymbol{\beta}}_j^2/4\kappa + o_P(1)$

Classical MLE theory:

$$\sqrt{p}\hat{\boldsymbol{\beta}}_j \to \mathcal{N}(0, 4\kappa)$$
$$\implies Q_2 \xrightarrow{\mathrm{d}} \chi_1^2.$$

High Dimensional Regime

$Q_3 = o_P(1)$

Difference from Classical Theory

Recall $2\Lambda_j = Q_2 + Q_3$.

Classical Theory

$Q_3 = o_P(1)$

$Q_2 = (\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}})^\top \nabla^2 \ell(\hat{\boldsymbol{\beta}})(\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}})$

$= (\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}})^\top E[\nabla^2 \ell(\hat{\boldsymbol{\beta}})](\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}) + o_P(1)$

$= p\hat{\boldsymbol{\beta}}_j^2/4\kappa + o_P(1)$

Classical MLE theory:

$$\sqrt{p}\hat{\boldsymbol{\beta}}_j \to \mathcal{N}(0, 4\kappa)$$
$$\implies Q_2 \xrightarrow{\text{d}} \chi_1^2.$$

High Dimensional Regime

$Q_3 = o_P(1)$

$Q_2 = (\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}})^\top \nabla^2 \ell(\hat{\boldsymbol{\beta}})(\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}})$

$\neq (\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}})^\top E[\nabla^2 \ell(\hat{\boldsymbol{\beta}})](\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}) + o_P(1)$

Difference from Classical Theory

Recall $2\Lambda_j = Q_2 + Q_3$.

Classical Theory

$Q_3 = o_P(1)$

$Q_2 = (\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}})^\top \nabla^2 \ell(\hat{\boldsymbol{\beta}})(\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}})$

$= (\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}})^\top E[\nabla^2 \ell(\hat{\boldsymbol{\beta}})](\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}) + o_P(1)$

$= p\hat{\boldsymbol{\beta}}_j^2/4\kappa + o_P(1)$

Classical MLE theory:

$$\sqrt{p}\hat{\boldsymbol{\beta}}_j \to \mathcal{N}(0, 4\kappa)$$
$$\implies Q_2 \stackrel{\mathrm{d}}{\to} \chi_1^2.$$

High Dimensional Regime

$Q_3 = o_P(1)$

$Q_2 = (\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}})^\top \nabla^2 \ell(\hat{\boldsymbol{\beta}})(\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}})$

$\neq (\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}})^\top E[\nabla^2 \ell(\hat{\boldsymbol{\beta}})](\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}) + o_P(1)$

$= p\hat{\boldsymbol{\beta}}_j^2/\mathrm{Tr}((\nabla^2 \ell(\tilde{\boldsymbol{\beta}}))^{-1}) + o_P(1)$

Difference from Classical Theory

Recall $2\Lambda_j = Q_2 + Q_3$.

Classical Theory

$Q_3 = o_P(1)$

$Q_2 = (\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}})^\top \nabla^2 \ell(\hat{\boldsymbol{\beta}})(\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}})$

$= (\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}})^\top E[\nabla^2 \ell(\hat{\boldsymbol{\beta}})](\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}) + o_P(1)$

$= p\hat{\boldsymbol{\beta}}_j^2 / 4\kappa + o_P(1)$

Classical MLE theory:

$$\sqrt{p}\hat{\boldsymbol{\beta}}_j \to \mathcal{N}(0, 4\kappa)$$
$$\implies Q_2 \xrightarrow{\mathrm{d}} \chi_1^2.$$

High Dimensional Regime

$Q_3 = o_P(1)$

$Q_2 = (\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}})^\top \nabla^2 \ell(\hat{\boldsymbol{\beta}})(\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}})$

$\neq (\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}})^\top E[\nabla^2 \ell(\hat{\boldsymbol{\beta}})](\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}) + o_P(1)$

$= p\hat{\boldsymbol{\beta}}_j^2 / \mathrm{Tr}((\nabla^2 \ell(\tilde{\boldsymbol{\beta}}))^{-1}) + o_P(1)$

S., Chen and Candès ('17):

$$\sqrt{p}\hat{\boldsymbol{\beta}}_j \to \mathcal{N}(0, \tau_*^2), \quad \mathrm{Tr}((\nabla^2 \ell(\tilde{\boldsymbol{\beta}}))^{-1}) \xrightarrow{\mathbb{P}} \lambda_*$$
$$\implies Q_2 \xrightarrow{\mathrm{d}} \left(\tau_*^2/\lambda_*\right) \chi_1^2.$$

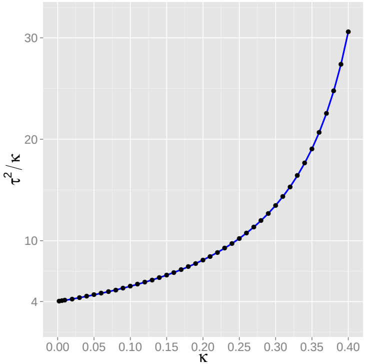
Variance Comparison

- **Classical Theory**

$$\lim_{n\to\infty} \text{Var}(\sqrt{p}\hat{\beta}_1) = 4\kappa$$

- **High Dimensional Regime**

$$\lim_{n\to\infty} \text{Var}(\sqrt{p}\hat{\beta}_1) = \tau_*^2$$



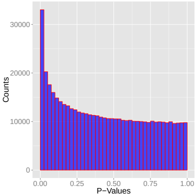
κ ranges from 0 to 0.4
y-axis: τ_*^2/κ

Summary and future research

For a class of logistic regression and other models: probit,...

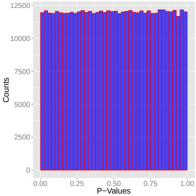
Wilks' theorem (1938)

$$-2 \log \mathsf{LRT} \overset{\mathrm{d}}{\to} \chi^2_{\mathrm{df}}$$



Sur, Chen, Candès (2017)

$$-2 \log \mathsf{LRT} \overset{\mathrm{d}}{\to} \gamma(p/n)\chi^2_{\mathrm{df}}$$



Ongoing and future work

- How about $\boldsymbol{\beta} \neq \mathbf{0}$?
- Universality (or non-universality) of limit law for different designs?
- Inference for penalized logistic regression?

Reference

"The Likelihood Ratio Test in High-Dimensional Logistic Regression Is Asymptotically a Rescaled Chi-Square." https://arxiv.org/abs/1706.01191

Thank You!