# Likelihood Ratio Test in High-Dimensional Logistic Regression Is Asymptotically a *Rescaled* Chi-Square

Yuxin Chen

Electrical Engineering, Princeton University

# Coauthors

Pragya Sur
Stanford Statistics

Emmanuel Candès
Stanford Statistics & Math

# In memory of Tom Cover (1938 - 2012)

Tom @ Stanford EE

"*We all know the feeling that follows when one investigates a problem, goes through a large amount of algebra, and finally investigates the answer to find that the entire problem is illuminated not by the analysis but by the inspection of the answer*"

# Inference in regression problems

Example: logistic regression

$$y_i \sim \text{logistic-model}\left( \boldsymbol{x}_i^\top \boldsymbol{\beta} \right), \qquad 1 \le i \le n$$

# Inference in regression problems

Example: logistic regression

$$y_i \sim \text{logistic-model}\left( \boldsymbol{x}_i^\top \boldsymbol{\beta} \right), \qquad 1 \le i \le n$$

One wishes to determine which covariate is of importance, i.e.

$$\beta_j = 0 \qquad \text{vs.} \qquad \beta_j \neq 0 \qquad (1 \le j \le p)$$

# Classical tests

$$\beta_j = 0 \qquad \text{vs.} \qquad \beta_j \neq 0 \qquad (1 \leq j \leq p)$$

Standard approaches (widely used in R, Matlab, etc): use asymptotic distributions of certain statistics

# Classical tests

$$\beta_j = 0 \qquad \text{vs.} \qquad \beta_j \neq 0 \qquad (1 \leq j \leq p)$$

Standard approaches (widely used in R, Matlab, etc): use asymptotic distributions of certain statistics

- Wald test: Wald statistic $\rightarrow \chi^2$
- Likelihood ratio test: log-likelihood ratio statistic $\rightarrow \chi^2$
- Score test: score $\rightarrow \mathcal{N}(\mathbf{0}, \text{Fisher Info})$
- ...

# Example: 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 these inference calculations (e.g. p-values) be trusted?

# This talk: likelihood ratio test (LRT)

$$\beta_j = 0 \qquad \text{vs.} \qquad \beta_j \neq 0 \qquad (1 \leq j \leq p)$$

Log-likelihood ratio (LLR) statistic

$$\mathsf{LLR}_j := \ell(\widehat{\boldsymbol{\beta}}) - \ell(\widehat{\boldsymbol{\beta}}_{(-j)})$$

- $\ell(\cdot)$: log-likelihood
- $\widehat{\boldsymbol{\beta}} = \arg\max_{\boldsymbol{\beta}} \ell(\boldsymbol{\beta})$: unconstrained MLE

# This talk: likelihood ratio test (LRT)

$$\beta_j = 0 \qquad \text{vs.} \qquad \beta_j \neq 0 \qquad (1 \le j \le p)$$

Log-likelihood ratio (LLR) statistic

$$\mathsf{LLR}_j := \ell(\widehat{\boldsymbol{\beta}}) - \ell(\widehat{\boldsymbol{\beta}}_{(-j)})$$

- $\ell(\cdot)$: log-likelihood
- $\widehat{\boldsymbol{\beta}} = \arg\max_{\boldsymbol{\beta}} \ell(\boldsymbol{\beta})$: unconstrained MLE
- $\widehat{\boldsymbol{\beta}}_{(-j)} = \arg\max_{\boldsymbol{\beta}:\beta_j=0} \ell(\boldsymbol{\beta})$: constrained MLE

# Wilks' phenomenon '1938

Samuel Wilks, Princeton

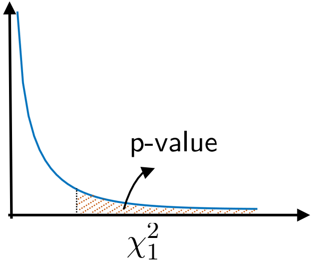
$$\beta_j = 0 \qquad \text{vs.} \qquad \beta_j \neq 0 \qquad (1 \leq j \leq p)$$

LRT asymptotically follows chi-square distribution (under null)

$$2\,\mathsf{LLR}_j \;\xrightarrow{\mathrm{d}}\; \chi_1^2 \quad (p \text{ fixed}, n \to \infty)$$

# Wilks' phenomenon '1938

Samuel Wilks, Princeton




*assess significance of coefficients*

$$\beta_j = 0 \qquad \text{vs.} \qquad \beta_j \neq 0 \qquad (1 \leq j \leq p)$$

LRT asymptotically follows chi-square distribution (under null)

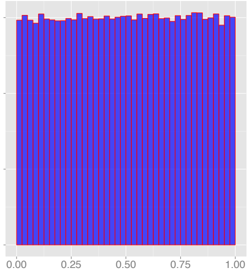
$$2\,\mathsf{LLR}_j \xrightarrow{\mathrm{d}} \chi^2_1 \quad (p \text{ fixed}, n \to \infty)$$

# Classical LRT in high dimensions

$$p/n \ \in (1, \infty)$$

Linear regression

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \underbrace{\boldsymbol{\eta}}_{\text{i.i.d. Gaussian}}$$



classical p-values are uniform

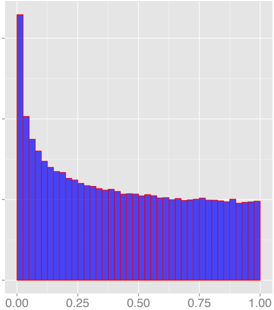
For linear regression (with Gaussian noise) in high dimensions, $2\mathrm{LLR}_j \ \sim \ \chi^2_1$ (classical test always works)

# Classical LRT in high dimensions

$$p = 1200, n = 4000$$

Logistic regression

$$\boldsymbol{y} \sim \text{logistic-model}(\boldsymbol{X}\boldsymbol{\beta})$$



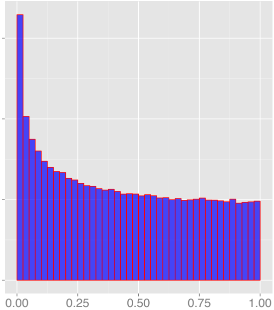

classical p-values are highly nonuniform

# Classical LRT in high dimensions

$$p = 1200, n = 4000$$

Logistic regression

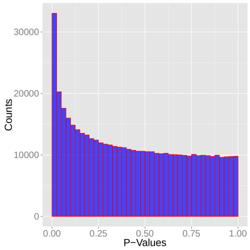
$$\boldsymbol{y} \sim \text{logistic-model}(\boldsymbol{X}\boldsymbol{\beta})$$



classical p-values are highly nonuniform

Wilks' theorem seems inadequate in accommodating logistic regression in high dimensions

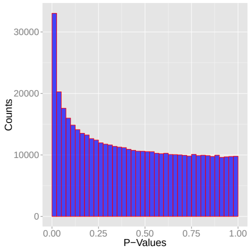
# Bartlett correction? ($n = 4000$, $p = 1200$)



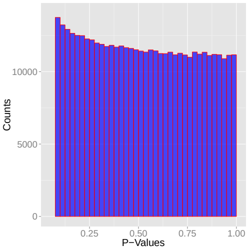

classical Wilks

- Bartlett correction (finite sample effect): $\frac{2\mathsf{LLR}_j}{1+\alpha_n/n} \sim \chi^2_1$

# Bartlett correction? ($n = 4000$, $p = 1200$)



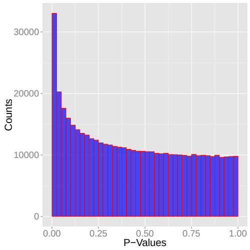

classical Wilks



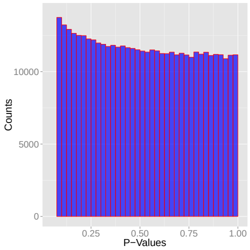

Bartlett-corrected

- Bartlett correction (finite sample effect): $\frac{2\mathsf{LLR}_j}{1+\alpha_n/n} \sim \chi^2_1$
  - p-values are still non-uniform $\longrightarrow$ this is NOT finite sample effect

# Bartlett correction? ($n = 4000$, $p = 1200$)




classical Wilks

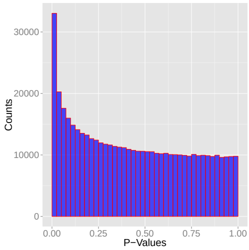



Bartlett-corrected

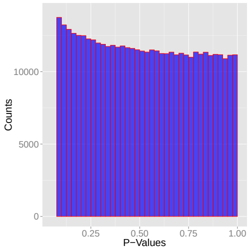
- Bartlett correction (finite sample effect): $\frac{2\mathsf{LLR}_j}{1+\alpha_n/n} \sim \chi_1^2$
  - p-values are still non-uniform $\longrightarrow$ this is NOT finite sample effect

What happens in high dimensions?

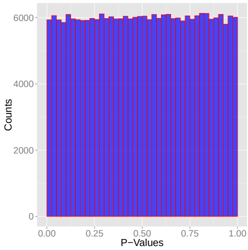
# Our findings




classical Wilks



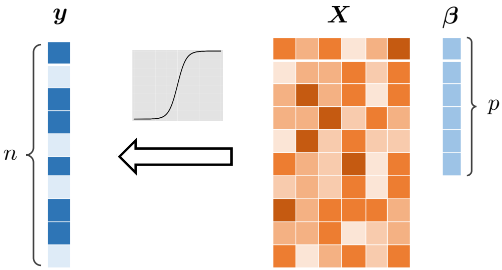

Bartlett-corrected




rescaled $\chi^2$

- Bartlett correction (finite sample effect): $\frac{2\mathrm{LLR}_j}{1+\alpha_n/n} \sim \chi^2_1$
  - p-values are still non-uniform $\longrightarrow$ this is NOT finite sample effect
- A glimpse of our theory: LRT follows a rescaled $\chi^2$ distribution

# Problem formulation (formal)



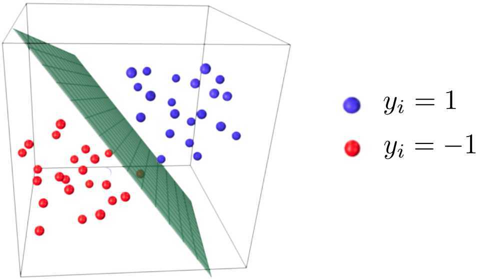

- Gaussian design: $\boldsymbol{X}_i \overset{\text{ind.}}{\sim} \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$
- Logistic model:

$$y_i = \begin{cases} 1, & \text{with prob. } \frac{1}{1+\exp(-\boldsymbol{X}_i^\top \boldsymbol{\beta})} \\ -1, & \text{with prob. } \frac{1}{1+\exp(\boldsymbol{X}_i^\top \boldsymbol{\beta})} \end{cases} \qquad 1 \le i \le n$$

- Proportional growth: $p/n \to$ constant
- Global null: $\boldsymbol{\beta} = \mathbf{0}$

# When does MLE exist?

$$\text{(MLE)} \quad \text{maximize}_{\boldsymbol{\beta}} \quad \underbrace{\ell(\boldsymbol{\beta}) = -\sum_{i=1}^{n} \log\left\{1 + \exp(-y_i \boldsymbol{X}_i^\top \boldsymbol{\beta})\right\}}_{\leq 0}$$



**MLE is unbounded if $\exists$ perfect separating hyperplane**

# When does MLE exist?

$$\text{(MLE)} \quad \text{maximize}_{\boldsymbol{\beta}} \quad \underbrace{\ell(\boldsymbol{\beta}) = -\sum_{i=1}^{n} \log\left\{1 + \exp(-y_i \boldsymbol{X}_i^\top \boldsymbol{\beta})\right\}}_{\leq 0}$$

If $\exists$ a hyperplane that perfectly separates $\{y_i\}$, i.e.

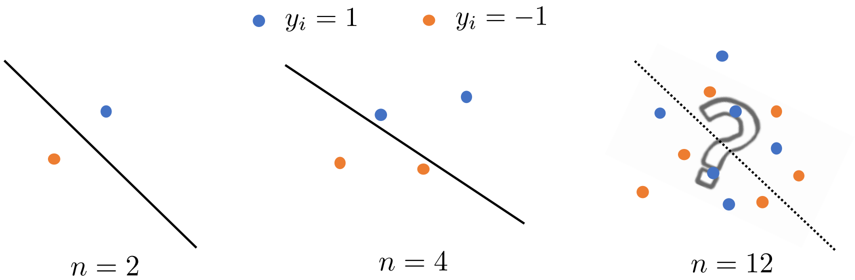
$$\exists\, \widehat{\boldsymbol{\beta}} \quad \text{s.t.} \quad y_i \boldsymbol{X}_i^\top \widehat{\boldsymbol{\beta}} > 0 \text{ for all } i$$

then MLE is unbounded

$$\lim_{a \to \infty} \ell( \underbrace{a\widehat{\boldsymbol{\beta}}}_{\text{unbounded}} ) = 0$$

# When does MLE exist?

**Separating capacity (Tom Cover, Ph. D. thesis '1965)**

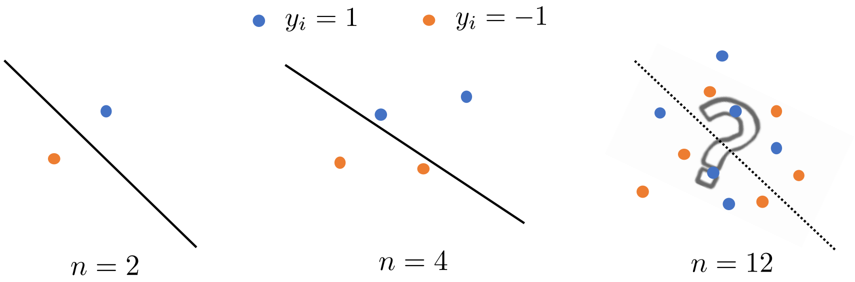



number of samples $n$ increases

$\Longrightarrow$ more difficult to find separating hyperplane

# When does MLE exist?

**Separating capacity (Tom Cover, Ph. D. thesis '1965)**




**Theorem 1 (Cover '1965)**

*Under i.i.d. Gaussian design, a separating hyperplane exists with high prob. iff $n/p < 2$ (asymptotically)*

# Main result: asymptotic distribution of LRT

**Theorem 2 (Sur, Chen, Candès '2017)**

*Suppose $n/p > 2$. Under i.i.d. Gaussian design and global null,*

$$2\,\mathsf{LLR}_j \;\xrightarrow{\ \mathrm{d}\ }\; \underbrace{\alpha\left(\frac{p}{n}\right)\chi_1^2}_{\textit{rescaled } \chi^2}$$

# Main result: asymptotic distribution of LRT

**Theorem 2 (Sur, Chen, Candès '2017)**

*Suppose $n/p > 2$. Under i.i.d. Gaussian design and global null,*

$$2\,\mathsf{LLR}_j \xrightarrow{\mathrm{d}} \underbrace{\alpha\Big(\frac{p}{n}\Big)\,\chi^2_1}_{\textit{rescaled } \chi^2}$$

- $\alpha(p/n)$ can be determined by solving a system of 2 nonlinear equations and 2 unknowns

$$\tau^2 = \frac{n}{p}\mathbb{E}\left[(\Psi(\tau Z; b))^2\right]$$
$$\frac{p}{n} = \mathbb{E}\left[\Psi'(\tau Z; b)\right]$$

where $Z \sim \mathcal{N}(0,1)$, $\Psi$ is some operator, and $\alpha(p/n) = \tau^2/b$

# Main result: asymptotic distribution of LRT

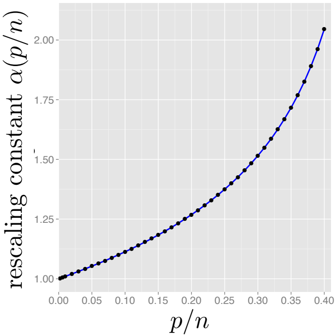
**Theorem 2 (Sur, Chen, Candès '2017)**

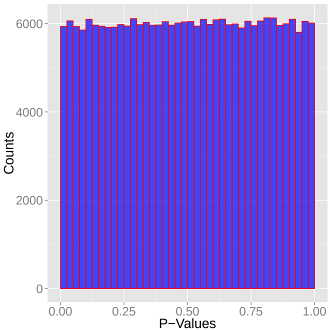
*Suppose $n/p > 2$. Under i.i.d. Gaussian design and global null,*

$$2\,\mathsf{LLR}_j \xrightarrow{\mathrm{d}} \underbrace{\alpha\Big(\frac{p}{n}\Big)\,\chi^2_1}_{\textit{rescaled } \chi^2}$$

- $\alpha(p/n)$ can be determined by solving a system of 2 nonlinear equations and 2 unknowns
  - $\alpha(\cdot)$ depends only on aspect ratio $p/n$
  - It is not a finite sample effect
  - $\alpha(0) = 1$: matches classical theory

# Our adjusted LRT theory in practice

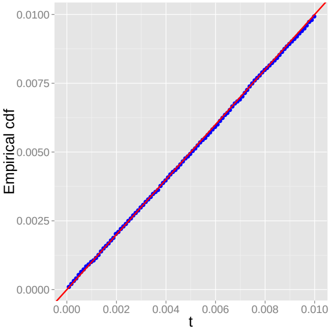



rescaling constant for logistic model    empirical p-values $\approx$ Unif$(0, 1)$

Empirically, LRT $\approx$ rescaled $\chi^2_1$ (as predicted)

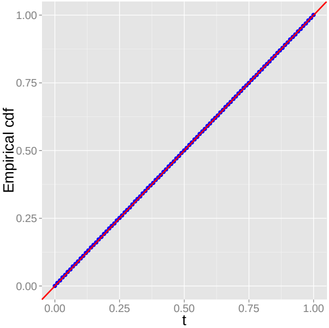
# Validity of tail approximation




Empirical CDF of adjusted pvalues ($n = 4000$, $p = 1200$)

Empirical CDF is in near-perfect aggreement with diagonal, suggesting that our theory is remarkably accurate even when we zoom in

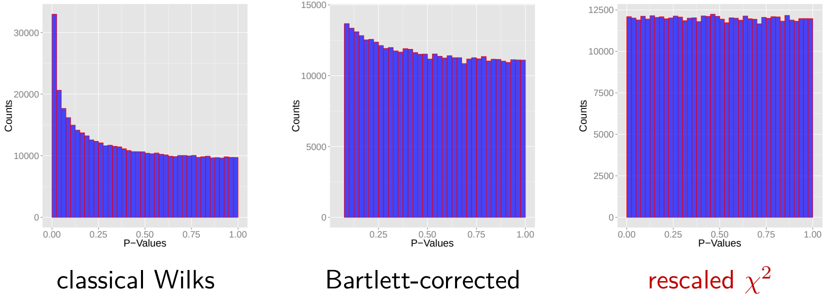
# Efficacy under moderate sample sizes




Empirical CDF of adjusted pvalues ($n = 200$, $p = 60$)

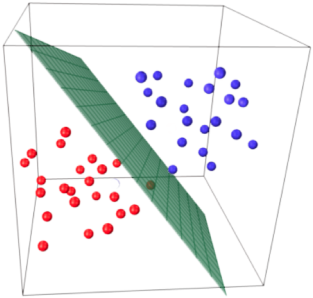
Our theory seems adequete for moderately large samples

# Universality: non-Gaussian covariates

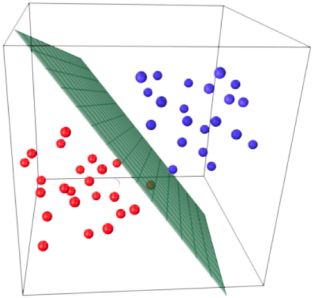


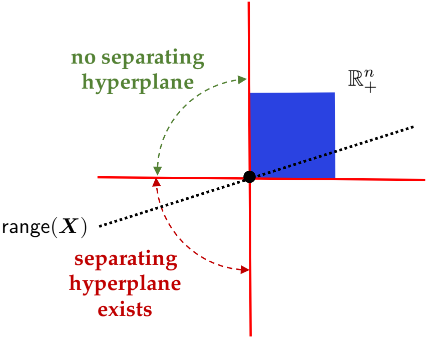

i.i.d. Bernoulli design, $n = 4000$, $p = 1200$

# Connection to convex geometry



# Connection to convex geometry






WLOG, if $y_1 = \cdots = y_n = 1$, then

$$\text{separability} \quad = \quad \underbrace{\left\{\text{range}(\boldsymbol{X}) \cap \mathbb{R}^n_+ \neq \{\mathbf{0}\}\right\}}_{\text{can be analyzed via convex geometry (e.g. Amelunxen et al.)}}$$

# Connection to robust M-estimation

Since $y_i = \pm 1$ and $\boldsymbol{X}_i \overset{\text{ind.}}{\sim} \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma})$,

$$\begin{aligned}
\text{maximize}_{\boldsymbol{\beta}} \quad \ell(\boldsymbol{\beta}) &= -\sum_{i=1}^{n} \log\left\{1 + \exp(-y_i \boldsymbol{X}_i^\top \boldsymbol{\beta})\right\} \\
&\overset{\mathrm{d}}{=} \underbrace{-\sum_{i=1}^{n} \log\left\{1 + \exp(-\boldsymbol{X}_i^\top \boldsymbol{\beta})\right\}}_{:= \sum_{i=1}^{n} \rho(-\boldsymbol{X}_i \boldsymbol{\beta})}
\end{aligned}$$

# Connection to robust M-estimation

Since $y_i = \pm 1$ and $\boldsymbol{X}_i \overset{\text{ind.}}{\sim} \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma})$,

$$\text{maximize}_{\boldsymbol{\beta}} \quad \ell(\boldsymbol{\beta}) = -\sum_{i=1}^{n} \log\left\{1 + \exp(-y_i \boldsymbol{X}_i^\top \boldsymbol{\beta})\right\}$$

$$\overset{\mathrm{d}}{=} \underbrace{-\sum_{i=1}^{n} \log\left\{1 + \exp(-\boldsymbol{X}_i^\top \boldsymbol{\beta})\right\}}_{:=\sum_{i=1}^{n} \rho(-\boldsymbol{X}_i \boldsymbol{\beta})}$$

$$\Longrightarrow \quad \text{MLE } \widehat{\boldsymbol{\beta}} = \underbrace{\arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \rho(y_i - \boldsymbol{X}_i \boldsymbol{\beta})}_{\text{robust M-estimation}} \quad \text{with } \boldsymbol{y} = \boldsymbol{0}$$

# Connection to robust M-estimation

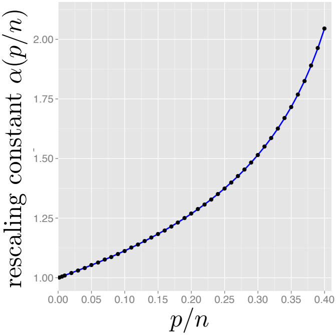
$$\text{MLE } \widehat{\boldsymbol{\beta}} = \underbrace{\arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \rho(y_i - \boldsymbol{X}_i \boldsymbol{\beta})}_{\text{robust M-estimation}}$$

**Variance inflation** as $p/n \downarrow$ (El Karoui et al. '13, Donoho, Montanari '13)

# Connection to robust M-estimation

$$\text{MLE } \widehat{\boldsymbol{\beta}} = \underbrace{\arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \rho(y_i - \boldsymbol{X}_i \boldsymbol{\beta})}_{\text{robust M-estimation}}$$

**Variance inflation** as $p/n \downarrow$ (El Karoui et al. '13, Donoho, Montanari '13)




variance inflation

$\longrightarrow$ increasing rescaling factor

# This is not just about logistic regression

Our theory is applicable to

- logistic model
- probit model
- linear model (under Gaussian noise)
  - rescaling const $\alpha(p/n) \equiv 1$ (consistent with classical theory)
- linear model (under non-Gaussian noise)
- $\cdots$

# Key ingredients in establishing our theory

Key step is to realize that

$$2\,\mathsf{LLR}_j \xrightarrow{\mathrm{d}} \underbrace{\frac{p}{b(p/n)}\widehat{\beta}_j^2}_{\text{rescaled } \chi^2}$$

where $b(\cdot)$ depends only on $\frac{p}{n}$, $\widehat{\beta}_j \sim \mathcal{N}\Big(0, \frac{\sqrt{\alpha\left(\frac{p}{n}\right)b\left(\frac{p}{n}\right)}}{\sqrt{p}}\Big)$

# Key ingredients in establishing our theory

Key step is to realize that

$$2\,\mathsf{LLR}_j \xrightarrow{\mathrm{d}} \underbrace{\frac{p}{b(p/n)}\widehat{\beta}_j^2}_{\text{rescaled } \chi^2}$$

where $b(\cdot)$ depends only on $\frac{p}{n}$, $\widehat{\beta}_j \sim \mathcal{N}\Big(0, \frac{\sqrt{\alpha\left(\frac{p}{n}\right)b\left(\frac{p}{n}\right)}}{\sqrt{p}}\Big)$

1. **Convex geometry:** show $\|\widehat{\boldsymbol{\beta}}\| = O(1)$
2. **Approximate message passing:** determine asymptotic distribution of $\|\widehat{\boldsymbol{\beta}}\|$
3. **Leave-one-out analysis:** characterize marginal distribution of $\widehat{\beta}_j$ (*rescaled Gaussian*) and ensure higher-order terms vanish

# A small sample of prior work

- **Candès, Fan, Janson, Lv '16:** observed empirically nonuniformity of p-values in logistic regression
- **Fan, Demirkaya, Lv '17:** classical asymptotic normality of MLE (basis of Wald test) fails to hold in logistic regression when $p \asymp n^{2/3}$
- **El Karoui, Beana, Bickel, Limb, Yu '13, El Karoui '13, Donoho, Montanari '13, El Karoui '15**: robust M-estimation for linear models (assuming strong convexity)

# Summary

- Caution needs to be exercised when applying classical statistical procedures in a high-dimensional regime — a regime of growing interest and relevance
- What shall we do under non-null ($\boldsymbol{\beta} \neq \mathbf{0}$)?

**Paper**: "The likelihood ratio test in high-dimensional logistic regression is asymptotically a *rescaled* chi-square", Pragya Sur, Yuxin Chen, Emmanuel Candès, 2017.