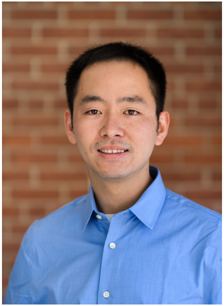
A modern maximum-likelihood theory for high-dimensional logistic regression

Pragya Sur
Dept. of Statistics
Stanford University

Workshop on Robust and High-Dimensional Statistics
Simons Institute for the Theory of Computing
2018

Collaborators



Classical Maximum-Likelihood Theory

Logistic regression in R: $n = 200$, $p = 60$

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

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

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.1836  -0.9808   0.3590   0.9770   2.4853

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.3320037  0.2029364   1.636   0.1018
X1          -0.3080503  0.1969881  -1.564   0.1179
X2           0.1707889  0.2096599   0.815   0.4153
X3          -0.1491842  0.1883217  -0.792   0.4283
X4           0.0346026  0.1987109   0.174   0.8618
X5          -0.0962019  0.1725523  -0.558   0.5772
X6           0.4634118  0.2167999   2.138   0.0326 *
...
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Logistic regression in R: $n = 200$, $p = 60$

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

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

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.1836  -0.9808   0.3590   0.9770   2.4853

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.3320037  0.2029364   1.636   0.1018
X1          -0.3080503  0.1969881  -1.564   0.1179
X2           0.1707889  0.2096599   0.815   0.4153
X3          -0.1491842  0.1883217  -0.792   0.4283
X4           0.0346026  0.1987109   0.174   0.8618
X5          -0.0962019  0.1725523  -0.558   0.5772
X6           0.4634118  0.2167999   2.138   0.0326 *
...
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Can inference be trusted?

Logistic Regression setting

- Consider n i.i.d. samples $(y_i, \boldsymbol{X}_i)$, $y_i \in \{0,1\}$, $\boldsymbol{X}_i \in \mathbb{R}^p$,

$$\mathbb{P}[y_i = 1 | \boldsymbol{X}_i] = \sigma(\boldsymbol{X}_i' \boldsymbol{\beta}) := \frac{e^{\boldsymbol{X}_i' \boldsymbol{\beta}}}{1 + e^{\boldsymbol{X}_i' \boldsymbol{\beta}}}, \quad \boldsymbol{\beta} \in \mathbb{R}^p$$

- MLE and reduced MLE

$$\begin{aligned} \hat{\boldsymbol{\beta}} &= \arg\min_{\boldsymbol{\beta} \in \mathbb{R}^p} && \ell(\boldsymbol{\beta}) \\ \hat{\boldsymbol{\beta}}_{(-j)} &= \arg\min_{\boldsymbol{\beta} \in \mathbb{R}^p, \beta_j = 0} && \ell(\boldsymbol{\beta}) \end{aligned}$$

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^{n} \{\rho(\boldsymbol{X}_i' \boldsymbol{\beta}) - (\boldsymbol{X}_i' \boldsymbol{\beta}) y_i\}, \qquad \rho(t) = \log(1 + e^t) \quad \text{(link fun.)}$$

- For testing $\mathcal{H}_{0,j} : \beta_j = 0$ vs $\mathcal{H}_1 : \beta_j \neq 0$

$$\log \mathrm{LRT}_j = \ell(\hat{\boldsymbol{\beta}}) - \ell(\hat{\boldsymbol{\beta}}_{(-j)})$$

Basic staples of classical theory

Theorem (Classical MLE distribution)

Under 'suitable regularity conditions', p fixed, $n \to \infty$, with Fisher information $\boldsymbol{I}_{\boldsymbol{\beta}}$

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{\mathrm{d}} \mathcal{N}(\mathbf{0}, \boldsymbol{I}_{\boldsymbol{\beta}}^{-1}),$$

Theorem (Wilks' theorem)

Under suitable 'regularity conditions', p fixed, $n \to \infty$

$$-2 \log \mathrm{LRT} \xrightarrow{d} \chi^2 \qquad \textit{(under null)}$$

Similar result for testing a group of k variables.

Basic staples of classical theory

Theorem (Classical MLE distribution)

Under 'suitable regularity conditions', p fixed, $n \to \infty$, with Fisher information $\boldsymbol{I}_{\boldsymbol{\beta}}$

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{\mathrm{d}} \mathcal{N}(\mathbf{0}, \boldsymbol{I}_{\boldsymbol{\beta}}^{-1}),$$

Theorem (Wilks' theorem)

Under suitable 'regularity conditions', p fixed, $n \to \infty$

$$-2 \log \mathrm{LRT} \xrightarrow{d} \chi^2 \qquad \textit{(under null)}$$

Similar result for testing a group of k variables.

- Extensions to diverging dimensions—Huber ('73), Portnoy ('88), $p = o(\sqrt{n})$.

Agenda

Classical theory used for inference all the time, and by all software packages.

1. The MLE is approximately unbiased.
2. Variance of the MLE is approximately given by inverse Fisher information.
3. LRT is approximately distributed as a χ^2.

This talk

Is classical inference accurate in modern settings where n, p are both large $(\to \infty)$ and n/p is 5 or 10?

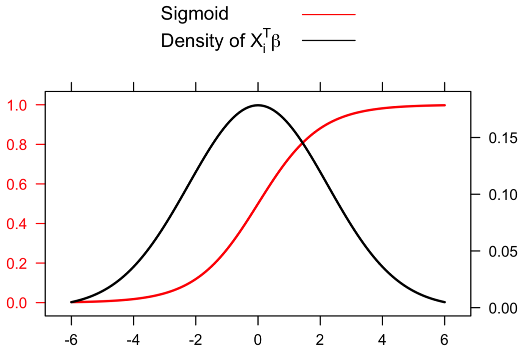
What do we see in simulation studies?

Scaling

Simulation settings:

- Gaussian covariates $\boldsymbol{X}_i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}/n)$
- Coeff. $\boldsymbol{\beta}$ scaled s.t.

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



Unbiasedness of MLE? First Example

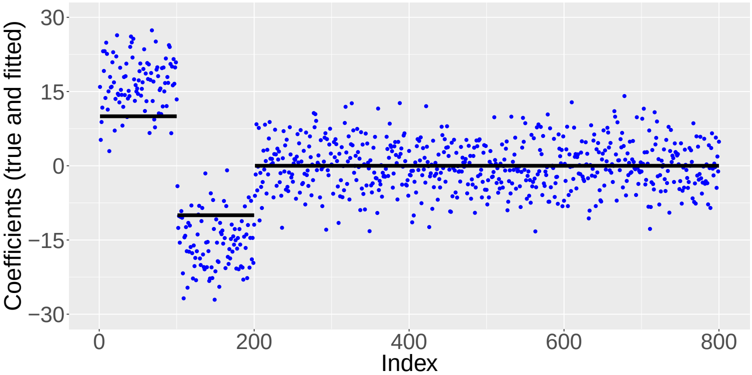


Figure: Signal (black) and MLE (blue), $n = 4000, p = 800$

Unbiasedness of MLE? First Example

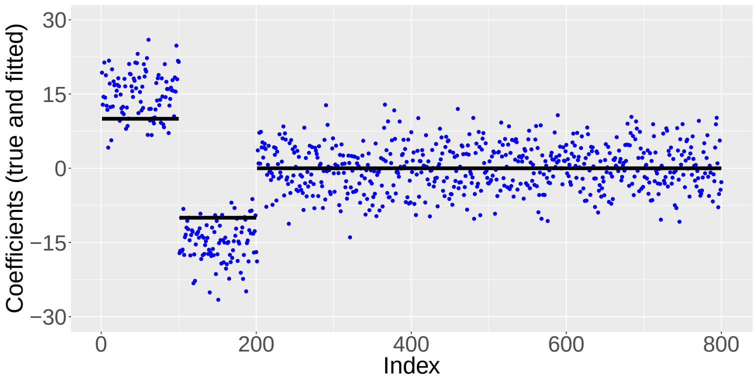


Figure: Signal (black) and MLE (blue), $n = 4000, p = 800$

Unbiasedness of MLE? Second example

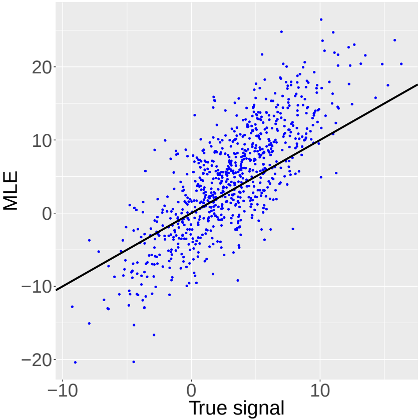


Figure: Lines with slope 1 (black) and 1.499 (red). $n = 4000$, $p = 800$

Unbiasedness of MLE? Second example

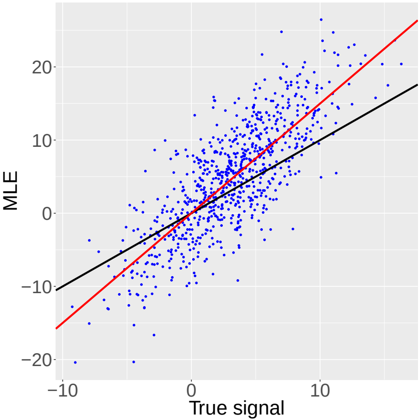
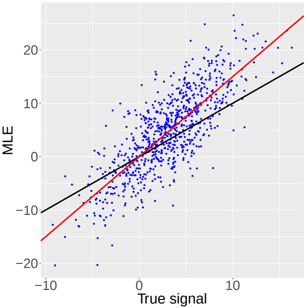


Figure: Lines with slope 1 (black) and 1.499 (red). $n = 4000$, $p = 800$

$\rightsquigarrow$ MLE seems to be over-biased

Consequence for predicted probabilities



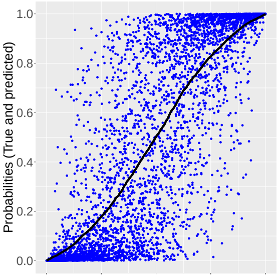


Figure: (Left) Scatterplot of $\hat{\beta}_j$ vs. β_j. (Right) True and predicted probabilities.

$\rightsquigarrow$ Predictions biased towards the extremes

Accuracy of standard errors?

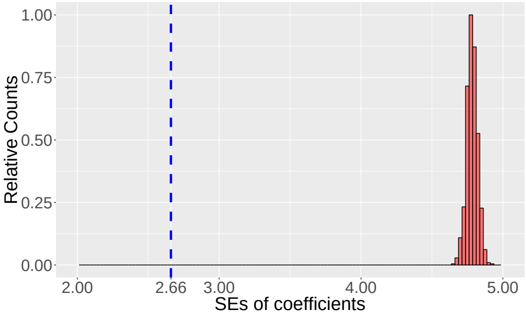


Figure: SEs of null coeff. estimates obtained via MC simulations (red) and classical value (blue)

$\rightsquigarrow$ MLE exhibits variance inflation in high dimensions

Accuracy of Wilks' theorem?

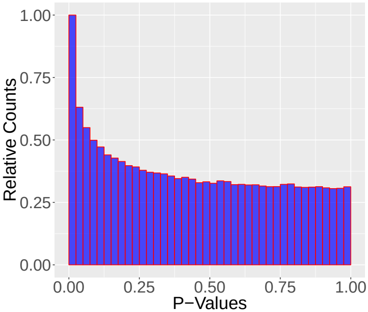


Figure: P-values (under the null) based on χ^2 approximation.

Observed earlier in Candès et al. ('16)
Studied under $\boldsymbol{\beta} = \mathbf{0}$, S., Chen and Candès ('17)

Accuracy of Wilks' theorem?

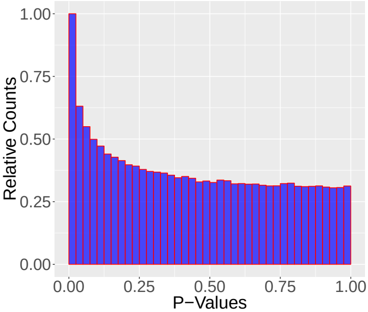


Figure: P-values (under the null) based on χ^2 approximation.

$\rightsquigarrow$ P-values far from uniform. Note, LRT distribution here is continuous.

Observed earlier in Candès et al. ('16)
Studied under $\boldsymbol{\beta} = \mathbf{0}$, S., Chen and Candès ('17)

Merely a finite sample effect?

Historically known

- MLE exhibits bias in small samples.
- LLR performs poorly in small samples.

Merely a finite sample effect?

Historically known

- MLE exhibits bias in small samples.
- LLR performs poorly in small samples.

- Several correction methods: Bartlett, Schaefer, Cordeiro, McCullagh, Firth...
- Central theme (under classical asymptotics):

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

Merely a finite sample effect?

Historically known

- MLE exhibits bias in small samples.
- LLR performs poorly in small samples.

- Several correction methods: Bartlett, Schaefer, Cordeiro, McCullagh, Firth...
- Central theme (under classical asymptotics):

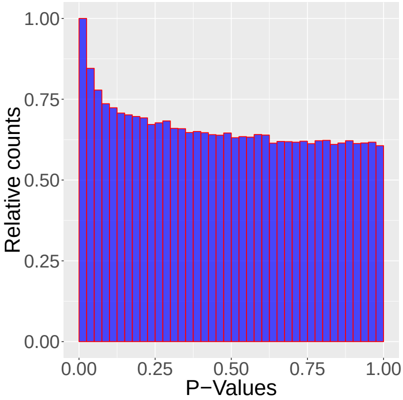
$$\mathbb{E}\left[-2\log \mathsf{LRT}\right] = 1 + \frac{\alpha}{n} + O\left(\frac{1}{n^2}\right)$$

- Plug in estimator α_n for α. Corrected statistic:

$$\frac{-2\log \mathsf{LRT}}{1+\frac{\alpha_n}{n}}$$

- Bartlett correction—specific choice for α_n.

Bartlett corrected p-values



Traditional finite sample corrections do not suffice in high dimensions

Failures of Classical Theory

1. MLE over-estimates effect magnitudes
   $\rightsquigarrow$ Predictions for risk of a disease shrunk to 0 or 1.
2. Variability of MLE is underestimated
   $\rightsquigarrow$ invalid confidence intervals.
3. P-values based on LRT far from uniform under the null
   $\rightsquigarrow$ Entirely unreliable inference.

Failures of Classical Theory

1. MLE over-estimates effect magnitudes
   $\rightsquigarrow$ Predictions for risk of a disease shrunk to 0 or 1.
2. Variability of MLE is underestimated
   $\rightsquigarrow$ invalid confidence intervals.
3. P-values based on LRT far from uniform under the null
   $\rightsquigarrow$ Entirely unreliable inference.

Serious need for a modern maximum-likelihood theory in high dimensions!

A Modern Maximum-Likelihood Theory for High Dimensions

Asymptotic Setting

Sequence of problems with $n, p \to \infty$ and covariates $\sqrt{n}\boldsymbol{X}_i \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_{p\times p})$

- Dimensionality κ:

$$p/n \to \kappa \in (0,1)$$

- Signal strength (SNR):

$$\mathsf{Var}(\boldsymbol{X}_i'\boldsymbol{\beta}) \to \gamma^2$$

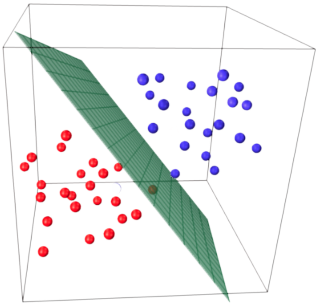
- Conditions on the signal:

$$\frac{1}{p}\sum_{i=1}^{p}\delta_{\beta_i} \xrightarrow{\mathrm{d}} \Pi, \quad \mathbb{E}\,\Pi^2 < \infty, \quad \frac{1}{p}\sum_{i=1}^{p}\beta_j^2 \to \mathbb{E}_{\Pi}(\beta^2)$$

When does the MLE exist?

Albert and Anderson (1984)

The MLE does not exist if a hyperplane separates the two groups, and exists otherwise.



Analytical characterization

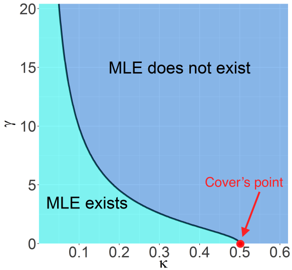
Cover (1964)

If $\boldsymbol{X}_i$ i.i.d. from continuous distribution F and $\boldsymbol{\beta} = \mathbf{0}$, MLE does not exist (asymp.) if $\kappa > 1/2$.

Analytical characterization

Cover (1964)

If $\boldsymbol{X}_i$ i.i.d. from continuous distribution F and $\boldsymbol{\beta} = \mathbf{0}$, MLE does not exist (asymp.) if $\kappa > 1/2$.



Theorem (Candès and S.('18))

- $V \stackrel{\mathrm{d}}{=} YX$, $X \sim \mathcal{N}(0,1)$, *and* $Y = \pm 1$, $\mathbb{P}(Y = 1|X) = 1/(1+\exp(-\gamma X))$
- $Z \sim \mathcal{N}(0,1) \perp\!\!\!\perp V$, $\qquad h_{\mathsf{MLE}}(\gamma) = \min_{t \in \mathbb{R}} \left\{ \mathbb{E}(tV - Z)_+^2 \right\}$

$$\begin{aligned} \kappa > h_{\mathsf{MLE}}(\gamma) \quad &\Longrightarrow \quad \lim_{n,p\to\infty} \mathbb{P}\{\text{MLE exists}\} = 0 \\ \kappa < h_{\mathsf{MLE}}(\gamma) \quad &\Longrightarrow \quad \lim_{n,p\to\infty} \mathbb{P}\{\text{MLE exists}\} = 1 \end{aligned}$$

A nonlinear system of equations

Equation system (S) in 3 unknowns $(\alpha, \sigma, \lambda)$, parametrized by (κ, γ)

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

$$(Q_1, Q_2) \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}(\alpha, \sigma))$$

$$\mathbf{\Sigma} = \begin{bmatrix} \gamma^2 & -\alpha\gamma^2 \\ -\alpha\gamma^2 & \alpha^2\gamma^2 + \kappa\sigma^2 \end{bmatrix}$$

This system holds lots of keys...

A nonlinear system of equations

Recall

- Signal strength: $\gamma^2 := \lim \mathrm{Var}(\boldsymbol{X}_i'\boldsymbol{\beta})$.
- Dimensionality: $\kappa = \lim p/n$.
- Link function: $\rho(t) = \log(1+e^t)$.
- Proximal mapping operator: $\mathsf{prox}_{\lambda\rho}(z) = \arg\min_{t\in\mathbb{R}} \left\{\lambda\rho(t) + \frac{1}{2}(t-z)^2\right\}$

Equation system (S) in 3 unknowns $(\alpha, \sigma, \lambda)$, parametrized by (κ, γ)

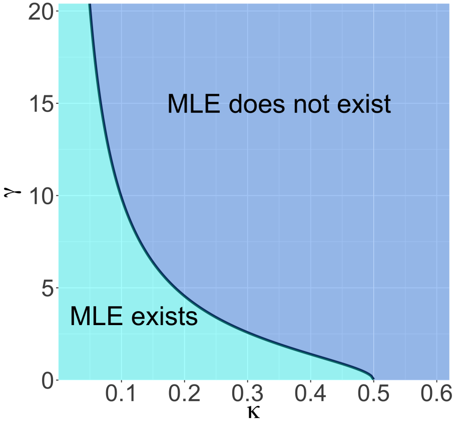
$$(S)\begin{cases} \sigma^2 = \dfrac{1}{\kappa^2}\mathbb{E}\left[2\rho'(Q_1)\left(\lambda\rho'(\mathsf{prox}_{\lambda\rho}(Q_2))\right)^2\right] \\ 0 = \mathbb{E}\left[\rho'(Q_1)Q_1\lambda\rho'(\mathsf{prox}_{\lambda\rho}(Q_2))\right] \\ 1-\kappa = \mathbb{E}\left[\dfrac{2\rho'(Q_1)}{1+\lambda\rho''(\mathsf{prox}_{\lambda\rho}(Q_2))}\right] \end{cases}$$

$$(Q_1, Q_2) \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}(\alpha,\sigma))$$

$$\mathbf{\Sigma} = \begin{bmatrix} \gamma^2 & -\alpha\gamma^2 \\ -\alpha\gamma^2 & \alpha^2\gamma^2 + \kappa\sigma^2 \end{bmatrix}$$

This system holds lots of keys...

Existence of MLE and system (S)



MLE exists (asymp.) iff (S) has a unique solution

'Average' MLE behavior

Assume MLE exists asymp. ((S) has unique solution $(\alpha_*, \sigma_*, \lambda_*)$).

$$\text{Roughly} \quad \frac{\hat{\beta}_j - \alpha_* \beta_j}{\sigma_\star} \sim \mathcal{N}(0,1)$$

'Average' MLE behavior

Assume MLE exists asymp. ((S) has unique solution $(\alpha_*, \sigma_*, \lambda_*)$).

$$\text{Roughly} \quad \frac{\hat{\beta}_j - \alpha_* \beta_j}{\sigma_\star} \sim \mathcal{N}(0,1)$$

Theorem (S. and Candès '18)

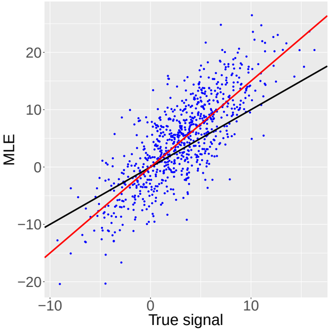
For any bivariate pseudo-Lipschitz function ψ of order 2,

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

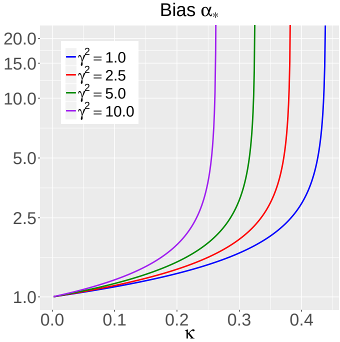
where $Z \sim \mathcal{N}(0,1)$ and $\beta \sim \Pi \perp\!\!\!\perp Z$. Recall Π is weak limit of $\frac{1}{p}\sum_{j=1}^{p} \delta_{\beta_j}$.

Consequence 1: Bias

$$\psi(t,u) = t \quad \Longrightarrow \quad \frac{1}{p}\sum_{j=1}^{p}\left(\hat{\beta}_j - \alpha_*\beta_j\right) \xrightarrow{\text{a.s.}} 0$$

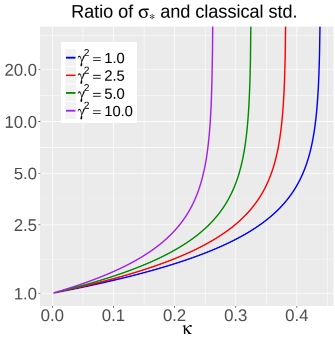


Bias $\alpha_* = 1.499$



Consequence 2: Variance

$$\psi(t,u) = t^2 \quad \Longrightarrow \quad \frac{1}{p}\sum_{j=1}^{p}\left(\hat{\beta}_j - \alpha_*\beta_j\right)^2 \xrightarrow{\text{a.s.}} \sigma_*^2$$



Consequence 3: confidence intervals

$$\psi(t,u) = 1\{-1.96 \le t/\sigma_* \le 1.96\}$$
$$\implies \quad \frac{1}{p}\sum_{j=1}^{p} 1\{-1.96 \le (\hat{\beta}_j - \alpha_*\beta_j)/\sigma_* \le 1.96\} \xrightarrow{\text{a.s.}} 0.95$$

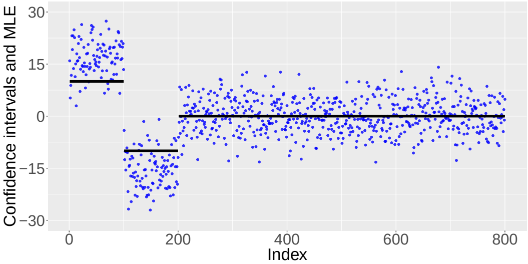
$$\mathsf{CI}_j = \left[\frac{\hat{\beta}_j \pm 1.96\sigma_*}{\alpha_*}\right]$$

Consequence 3: confidence intervals

$$\psi(t,u) = 1\{-1.96 \le t/\sigma_* \le 1.96\}$$

$$\Longrightarrow \quad \frac{1}{p}\sum_{j=1}^{p} 1\{-1.96 \le (\hat{\beta}_j - \alpha_* \beta_j)/\sigma_* \le 1.96\} \xrightarrow{\text{a.s.}} 0.95$$

$$\mathsf{CI}_j = \left[\frac{\hat{\beta}_j \pm 1.96\sigma_*}{\alpha_*}\right]$$

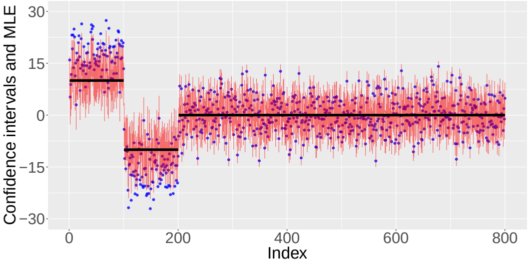


Consequence 3: confidence intervals

$$\psi(t,u) = 1\{-1.96 \le t/\sigma_* \le 1.96\}$$

$$\Longrightarrow \quad \frac{1}{p}\sum_{j=1}^{p} 1\{-1.96 \le (\hat\beta_j - \alpha_*\beta_j)/\sigma_* \le 1.96\} \xrightarrow{\text{a.s.}} 0.95$$

$$\mathsf{CI}_j = \left[\frac{\hat\beta_j \pm 1.96\sigma_*}{\alpha_*}\right]$$



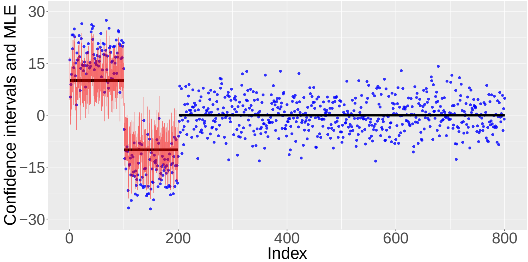
Nominal	Average coverage
95%	94.63 (0.12)
90%	89.70 (0.16)

all β_j (200 replicates)

Consequence 4: confidence intervals for which $\beta_j \neq 0$

$$\psi(t,u) = 1\{-1.96 \le t/\sigma_* \le 1.96\}1\{u \neq 0\}$$

$$\implies \quad \mathrm{Ave}_{j:\beta_j \neq 0}\{\beta_j \in \mathsf{CI}_j\} \xrightarrow{\text{a.s.}} 0.95$$



Nominal	Average coverage
95%	94.11 (0.12)
90%	88.89 (0.16)

all $\beta_j \neq 0$ (200 replicates)

Distribution of nulls

Theorem (S. and Candès, '18)

For any null variable $\beta_j = 0$,

$$\hat{\beta}_j \xrightarrow{\mathrm{d}} \mathcal{N}(0, \sigma_*^2).$$

For any k *null variables* $i_1, \ldots, i_k$, $(\hat{\beta}_{i_1}, \ldots, \hat{\beta}_{i_k})$ *is jointly asymp. independent.*

Distribution of nulls

Theorem (S. and Candès, '18)

For any null variable $\beta_j = 0$,

$$\hat{\beta}_j \xrightarrow{\mathrm{d}} \mathcal{N}(0, \sigma_*^2).$$

For any k null variables $i_1, \ldots, i_k$, $(\hat{\beta}_{i_1}, \ldots, \hat{\beta}_{i_k})$ is jointly asymp. independent.

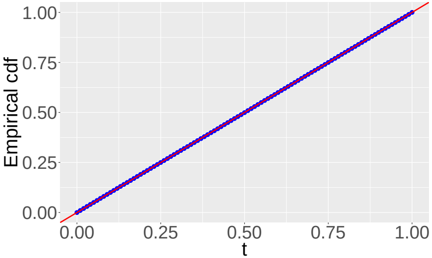


Figure: Empirical cdf of $\Phi(\hat{\beta}_j/\sigma_*)$; $n = 4000$, $p = 400$, $\gamma^2 = 5$

Distribution of the LRT

Theorem (S. and Candès '18)

Assume MLE exists asymp. ((S) has unique solution $(\alpha_*, \sigma_*, \lambda_*)$).
For a null j, $\beta_j = 0$,

$$-2\log(LRT_j) \xrightarrow{\mathrm{d}} \frac{\kappa\sigma_*^2}{\lambda_*}\chi_1^2.$$

Similar extension to groups of k variables.

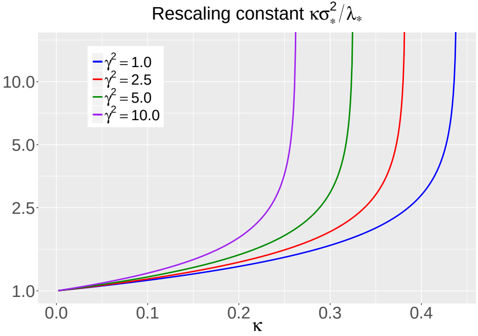
Distribution of the LRT

Theorem (S. and Candès '18)

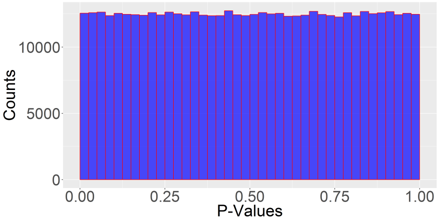
Assume MLE exists asymp. ((S) has unique solution $(\alpha_*, \sigma_*, \lambda_*)$).
For a null j, $\beta_j = 0$,

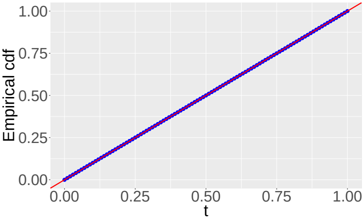
$$-2\log(LRT_j) \xrightarrow{\mathrm{d}} \frac{\kappa\sigma_*^2}{\lambda_*}\chi_1^2.$$

Similar extension to groups of k variables.



Bulk and tail asymptotics using our correction





Bulk and tail asymptotics using our correction

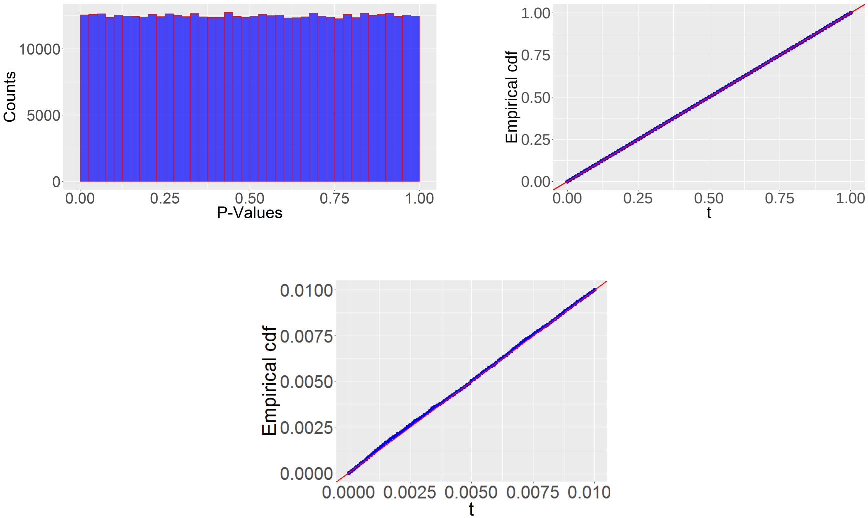
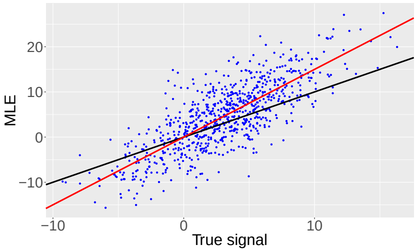


Figure: Histogram and empirical cdfs of p-values under the null

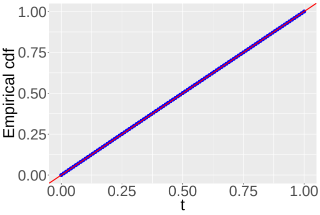
Non-Gaussian covariates

$$\mathbb{P}(X_j = 0) = p_j^2 \quad \mathbb{P}(X_j = 1) = 2p_j(1 - p_j) \quad \mathbb{P}(X_j = 2) = (1 - p_j)^2$$

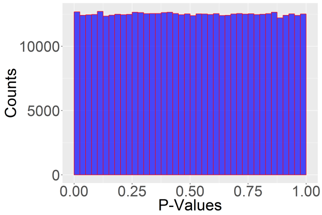
(SNPs in Hardy-Weinberg equilibrium)



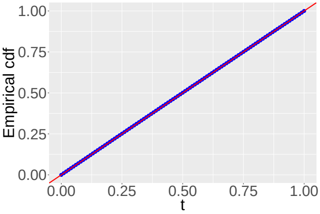
Non-Gaussian covariates



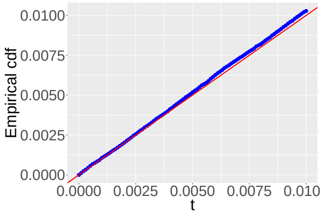
(a) Emp. cdf of $\Phi(\hat{\beta}_j/\sigma_*)$ for null β_j



(b) P-vals from LLR approx. for this null



(c) Empirical dist. of p-vals from (b)



(d) Tail behavior of (c)

Main Mathematical Ideas

Tools and Inspiration

- MLE phase transition

$$\mathbb{P}\{\text{MLE exists}\} \longrightarrow 0/1$$

- Convex geometry—Cover('65), Amelunxen et al.('13)

Tools and Inspiration

- MLE phase transition

$$\mathbb{P}\{\text{MLE exists}\} \longrightarrow 0/1$$

- Average behavior

$$\mathrm{Ave}\,\psi(\hat{\beta}_j - \alpha_* \beta_j, \beta_j) \xrightarrow{\text{a.s.}} \mathbb{E}[\psi(\sigma_\star Z, \beta)]$$

- Convex geometry—Cover('65), Amelunxen et al.('13)

- Generalized Approximate Message Passing, robust M-estimation (G-AMP)—Rangan ('10), Javanmard and Montanari ('12), Donoho and Montanari('13)

Tools and Inspiration

- MLE phase transition

$$\mathbb{P}\{\text{MLE exists}\} \longrightarrow 0/1$$

- Average behavior

$$\text{Ave}\,\psi(\hat{\beta}_j - \alpha_* \beta_j, \beta_j) \xrightarrow{a.s.} \mathbb{E}[\psi(\sigma_\star Z, \beta)]$$

- Null dist. & LRT

$$\hat{\beta}_j \xrightarrow{\mathrm{d}} \mathcal{N}(0, \sigma_*^2)$$

$$-2\log(\mathsf{LRT}_j) \xrightarrow{\mathrm{d}} \frac{\kappa\sigma_*^2}{\lambda_*}\chi_1^2$$

- Convex geometry—Cover('65), Amelunxen et al.('13)
- Generalized Approximate Message Passing, robust M-estimation (G-AMP)—Rangan ('10), Javanmard and Montanari ('12), Donoho and Montanari('13)
- Leave-one-out arguments (robust M-estimation), non-asymptotic RMT—El Karoui('13), El Karoui, Bean, Bickel, Lim, Yu('13)

Approximate message passing (AMP) for the MLE

- Consider an iterative algorithm with $\hat{\boldsymbol{\beta}}$ as fixed point
- Track iterates $\hat{\boldsymbol{\beta}}^t$ at each stage via state evolution (SE)
- Show $\hat{\boldsymbol{\beta}}^t$ converges to $\hat{\boldsymbol{\beta}}$ in an appropriate sense

DMM ('09), BM ('11), JM ('13), BLM ('15)

Approximate message passing (AMP) for the MLE

- Consider an iterative algorithm with $\hat{\boldsymbol{\beta}}$ as fixed point
- Track iterates $\hat{\boldsymbol{\beta}}^t$ at each stage via state evolution (SE)
- Show $\hat{\boldsymbol{\beta}}^t$ converges to $\hat{\boldsymbol{\beta}}$ in an appropriate sense

The algorithm

Update $\{\hat{\boldsymbol{\beta}}^t, \boldsymbol{S}^t\}$ iteratively (from init. cond. $(\hat{\boldsymbol{\beta}}^0, \boldsymbol{S}^0 = \boldsymbol{X}\boldsymbol{\beta}^0)$

$$\begin{aligned} \hat{\boldsymbol{\beta}}^t &= \hat{\boldsymbol{\beta}}^{t-1} + \kappa^{-1}\boldsymbol{X}'\Psi_t(\boldsymbol{y}, \boldsymbol{S}^{t-1}) \\ \boldsymbol{S}^t &= \boldsymbol{X}\hat{\boldsymbol{\beta}}^t - \Psi_{t-1}(\boldsymbol{y}, \boldsymbol{S}^{t-1}) \end{aligned} \tag{1}$$

Ψ_t (applied element-wise) depends on scalars $\{\lambda_t\}_{t\geq 0}$

$$\Psi_t(y,s) = \lambda_t r_t \qquad r_t = y - \rho'(\mathsf{prox}_{\lambda_t\rho}(\lambda_t y + s)) \tag{2}$$

Can be interpreted as scaled residuals

Why this algorithm?

Assume $\lambda_t \equiv \lambda$ (constant). If $\{\hat{\boldsymbol{\beta}}^\star, \boldsymbol{S}^\star\}$ fixed point

$$\boldsymbol{X}'\{\boldsymbol{y} - \rho'(\mathsf{prox}_{\lambda\rho}(\lambda\boldsymbol{y} + \boldsymbol{S}))\} = \boldsymbol{0}$$
$$(\lambda\boldsymbol{y} + \boldsymbol{S}^\star) - \lambda\rho'(\mathsf{prox}_{\lambda\rho}(\lambda\boldsymbol{y} + \boldsymbol{S})) = \boldsymbol{X}\hat{\boldsymbol{\beta}}^\star$$

Why this algorithm?

Assume $\lambda_t \equiv \lambda$ (constant). If $\{\hat{\boldsymbol{\beta}}^\star, \boldsymbol{S}^\star\}$ fixed point

$$\boldsymbol{X}'\{\boldsymbol{y} - \rho'(\mathrm{prox}_{\lambda\rho}(\lambda\boldsymbol{y} + \boldsymbol{S}))\} = \mathbf{0}$$

$$(\lambda\boldsymbol{y} + \boldsymbol{S}^\star) - \lambda\rho'(\mathrm{prox}_{\lambda\rho}(\lambda\boldsymbol{y} + \boldsymbol{S})) = \boldsymbol{X}\hat{\boldsymbol{\beta}}^\star$$

Prox properties yield

$$z - \lambda\rho'(\mathrm{prox}_{\lambda\rho}(z)) = \mathrm{prox}_{\lambda\rho}(z) \quad \Longrightarrow \quad \mathrm{prox}_{\lambda\rho}(\lambda\boldsymbol{y} + \boldsymbol{S}) = \boldsymbol{X}\hat{\boldsymbol{\beta}}^\star$$

$$\Longrightarrow \quad \boldsymbol{X}'\{\boldsymbol{y} - \rho'(\boldsymbol{X}\hat{\boldsymbol{\beta}}^\star)\} = \mathbf{0}$$

Why this algorithm?

Assume $\lambda_t \equiv \lambda$ (constant). If $\{\hat{\boldsymbol{\beta}}^\star, \boldsymbol{S}^\star\}$ fixed point

$$\boldsymbol{X}'\{\boldsymbol{y} - \rho'(\mathsf{prox}_{\lambda\rho}(\lambda\boldsymbol{y} + \boldsymbol{S}))\} = \boldsymbol{0}$$
$$(\lambda\boldsymbol{y} + \boldsymbol{S}^\star) - \lambda\rho'(\mathsf{prox}_{\lambda\rho}(\lambda\boldsymbol{y} + \boldsymbol{S})) = \boldsymbol{X}\hat{\boldsymbol{\beta}}^\star$$

Prox properties yield

$$z - \lambda\rho'(\mathsf{prox}_{\lambda\rho}(z)) = \mathsf{prox}_{\lambda\rho}(z) \quad \Longrightarrow \quad \mathsf{prox}_{\lambda\rho}(\lambda\boldsymbol{y} + \boldsymbol{S}) = \boldsymbol{X}\hat{\boldsymbol{\beta}}^\star$$
$$\Longrightarrow \quad \boldsymbol{X}'\{\boldsymbol{y} - \rho'(\boldsymbol{X}\hat{\boldsymbol{\beta}}^\star)\} = \boldsymbol{0}$$

Fixed point $\hat{\boldsymbol{\beta}}^\star$ obeys KKT conditions (MLE)

State evolution

Starting from α_0, σ_0, define for $t = 0, 1, \ldots$

(1) λ_t solution to

$$\mathbb{E}\left[\frac{2\rho'(Q_1^t)}{1 + \lambda\rho''(\mathsf{prox}_{\lambda\rho}(Q_2^t))}\right] = 1 - \kappa \qquad (Q_1^t, Q_2^t) \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}(\alpha_t, \sigma_t))$$

(2) updates $\alpha_{t+1}, \sigma_{t+1}$

$$\alpha_{t+1} = \alpha_t + \frac{1}{\kappa\gamma^2}\,\mathbb{E}\left[2\rho'(Q_1^t)Q_1^t\lambda_t\rho'(\mathsf{prox}_{\lambda_t\rho}(Q_2^t))\right]$$

$$\sigma_{t+1}^2 = \frac{1}{\kappa^2}\,\mathbb{E}\left[2\rho'(Q_1^t)\left(\lambda_t\rho'(\mathsf{prox}_{\lambda_t\rho}(Q_2^t))\right)^2\right]$$

- $\{\alpha_t, \sigma_t, \lambda_t\}$ is called the State Evolution sequence.

State evolution

Starting from α_0, σ_0, define for $t = 0, 1, \ldots$

(1) λ_t solution to

$$\mathbb{E}\left[\frac{2\rho'(Q_1^t)}{1 + \lambda\rho''(\mathsf{prox}_{\lambda\rho}(Q_2^t))}\right] = 1 - \kappa \qquad (Q_1^t, Q_2^t) \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}(\alpha_t, \sigma_t))$$

(2) updates $\alpha_{t+1}, \sigma_{t+1}$

$$\alpha_{t+1} = \alpha_t + \frac{1}{\kappa\gamma^2}\,\mathbb{E}\left[2\rho'(Q_1^t)Q_1^t\lambda_t\rho'(\mathsf{prox}_{\lambda_t\rho}(Q_2^t))\right]$$

$$\sigma_{t+1}^2 = \frac{1}{\kappa^2}\,\mathbb{E}\left[2\rho'(Q_1^t)\left(\lambda_t\rho'(\mathsf{prox}_{\lambda_t\rho}(Q_2^t))\right)^2\right]$$

- $\{\alpha_t, \sigma_t, \lambda_t\}$ is called the State Evolution sequence.

Claim

(κ, γ)-region where MLE exists is where (1)–(2) converge to unique fixed point $(\alpha_*, \sigma_*, \lambda_*)$/solution to our system

Marginals via approximate message passing (AMP)

- Consider an iterative algorithm with $\hat{\boldsymbol{\beta}}$ as fixed point
- Track iterates $\hat{\boldsymbol{\beta}}^t$ at each stage via state evolution (SE)
- Show $\hat{\boldsymbol{\beta}}^t$ converges to $\hat{\boldsymbol{\beta}}$ in an appropriate sense

Correctness of SE

Set $\alpha_0 = \alpha_*, \sigma_0 = \sigma_*$ $\quad\rightsquigarrow\quad$ $\alpha_t = \alpha_*, \sigma_t = \sigma_*, \lambda_t = \lambda_*$ for all t

Theorem

Assume $\hat{\boldsymbol{\beta}}^0$ is s.t.

$$\lim_{n,p\to\infty} \frac{1}{p}\|\hat{\boldsymbol{\beta}}^0 - \alpha_* \boldsymbol{\beta}\|^2 = \sigma_*^2, \qquad \frac{1}{\gamma^2} \lim_{n\to\infty} \frac{\langle \hat{\boldsymbol{\beta}}^0, \hat{\boldsymbol{\beta}} \rangle}{n} = \alpha_*$$

In region where MLE exists, for any pseudo-Lipschitz ψ, AMP trajectory obeys

$$\frac{1}{p}\sum_{j=1}^{p} \psi(\hat{\beta}_j^t - \alpha_* \beta_j, \beta_j) \xrightarrow{\text{a.s.}} \mathbb{E}\left[\psi(\sigma_* Z, \beta)\right] \tag{3}$$

$Z \sim \mathcal{N}(0,1) \perp\!\!\!\perp \beta \sim \Pi$ $\qquad$ *(recall $\sum_{j=1}^{p} \delta_{\beta_j}/p \xrightarrow{\mathrm{d}} \Pi$)*

Correctness of SE

Set $\alpha_0 = \alpha_*, \sigma_0 = \sigma_* \quad \rightsquigarrow \quad \alpha_t = \alpha_*, \sigma_t = \sigma_*, \lambda_t = \lambda_*$ for all t

Theorem

Assume $\hat{\boldsymbol{\beta}}^0$ is s.t.

$$\lim_{n,p\to\infty} \frac{1}{p}\|\hat{\boldsymbol{\beta}}^0 - \alpha_* \boldsymbol{\beta}\|^2 = \sigma_*^2, \qquad \frac{1}{\gamma^2} \lim_{n\to\infty} \frac{\langle \hat{\boldsymbol{\beta}}^0, \hat{\boldsymbol{\beta}} \rangle}{n} = \alpha_*$$

In region where MLE exists, for any pseudo-Lipschitz ψ, AMP trajectory obeys

$$\frac{1}{p}\sum_{j=1}^{p} \psi(\hat{\beta}_j^t - \alpha_* \beta_j, \beta_j) \xrightarrow{\text{a.s.}} \mathbb{E}\left[\psi(\sigma_* Z, \beta)\right] \tag{3}$$

$Z \sim \mathcal{N}(0,1) \perp\!\!\!\perp \beta \sim \Pi \qquad$ *(recall $\sum_{j=1}^{p} \delta_{\beta_j}/p \xrightarrow{\mathrm{d}} \Pi$)*

$\rightsquigarrow$ Takeaway: $\{\alpha_*, \sigma_*\}$ tracks bias and variance of AMP iterates

Marginals via approximate message passing (AMP)

- Consider an iterative algorithm with $\hat{\boldsymbol{\beta}}$ as fixed point
- Track iterates $\hat{\boldsymbol{\beta}}^t$ at each stage via state evolution (SE)
- Show $\hat{\boldsymbol{\beta}}^t$ converges to $\hat{\boldsymbol{\beta}}$ in an appropriate sense

Convergence to the MLE

Theorem

Assume $\hat{\boldsymbol{\beta}}^0$ is s.t.

$$\lim_{n,p\to\infty} \frac{1}{p}\|\hat{\boldsymbol{\beta}}^0 - \alpha_* \boldsymbol{\beta}\|^2 = \sigma_*^2, \qquad \frac{1}{\gamma^2} \lim_{n\to\infty} \frac{\langle \hat{\boldsymbol{\beta}}^0, \hat{\boldsymbol{\beta}} \rangle}{n} = \alpha_*$$

In region where MLE exists

$$\lim_{t\to\infty} \lim_{n\to\infty} \frac{1}{p} \sum_{j=1}^{p} \psi(\hat{\beta}_j^t - \alpha_* \beta_j, \beta_j) = \lim_{n\to\infty} \frac{1}{p} \sum_{j=1}^{p} \psi(\hat{\beta}_j - \alpha_* \beta_j, \beta_j)$$

Convergence to the MLE

Theorem

Assume $\hat{\boldsymbol{\beta}}^0$ is s.t.

$$\lim_{n,p\to\infty} \frac{1}{p}\|\hat{\boldsymbol{\beta}}^0 - \alpha_* \boldsymbol{\beta}\|^2 = \sigma_*^2, \qquad \frac{1}{\gamma^2} \lim_{n\to\infty} \frac{\langle \hat{\boldsymbol{\beta}}^0, \hat{\boldsymbol{\beta}} \rangle}{n} = \alpha_*$$

In region where MLE exists

$$\lim_{t\to\infty} \lim_{n\to\infty} \frac{1}{p} \sum_{j=1}^{p} \psi(\hat{\beta}_j^t - \alpha_* \beta_j, \beta_j) = \lim_{n\to\infty} \frac{1}{p} \sum_{j=1}^{p} \psi(\hat{\beta}_j - \alpha_* \beta_j, \beta_j)$$

All info. about large sample bias & variance of $\hat{\boldsymbol{\beta}}^t$ may be transferred to MLE

$\Longrightarrow$ Bias $= \alpha_*$ Variance $= \sigma_*^2$

Analysis of LRT

$$-\log \mathsf{LRT}_j = \ell(\hat{\boldsymbol{\beta}}_{(-j)}) - \ell(\hat{\boldsymbol{\beta}}) =: Q_2 + Q_3.$$

$$Q_2 = \frac{1}{2}\sum_{i=1}^{n} \rho''\left(\boldsymbol{X}_i'\hat{\boldsymbol{\beta}}\right)\left(\boldsymbol{X}_{i,-j}'\hat{\boldsymbol{\beta}}_{-j} - \boldsymbol{X}_i'\hat{\boldsymbol{\beta}}\right)^2$$

$$Q_3 = \frac{1}{6}\sum_{i=1}^{n} \rho'''(\gamma_i)\left(\boldsymbol{X}_{i,-j}'\hat{\boldsymbol{\beta}}_{-j} - \boldsymbol{X}_i'\hat{\boldsymbol{\beta}}\right)^3 \quad \gamma_i \in (\boldsymbol{X}_{i,-j}'\hat{\boldsymbol{\beta}}_{-j}, \boldsymbol{X}_i'\hat{\boldsymbol{\beta}})$$

Analysis of LRT

$$-\log \mathsf{LRT}_j = \ell(\hat{\boldsymbol{\beta}}_{(-j)}) - \ell(\hat{\boldsymbol{\beta}}) =: Q_2 + Q_3.$$

$$Q_2 = \frac{1}{2}\sum_{i=1}^{n} \rho''\left(\boldsymbol{X}_i'\hat{\boldsymbol{\beta}}\right)\left(\boldsymbol{X}_{i,-j}'\hat{\boldsymbol{\beta}}_{-j} - \boldsymbol{X}_i'\hat{\boldsymbol{\beta}}\right)^2$$

$$Q_3 = \frac{1}{6}\sum_{i=1}^{n} \rho'''(\gamma_i)\left(\boldsymbol{X}_{i,-j}'\hat{\boldsymbol{\beta}}_{-j} - \boldsymbol{X}_i'\hat{\boldsymbol{\beta}}\right)^3 \quad \gamma_i \in (\boldsymbol{X}_{i,-j}'\hat{\boldsymbol{\beta}}_{-j}, \boldsymbol{X}_i'\hat{\boldsymbol{\beta}})$$

- $\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\beta}}_{-j}$ high-dimensional and dependent.
- How do we track the differences $\boldsymbol{X}_{i,-j}'\hat{\boldsymbol{\beta}}_{-j} - \boldsymbol{X}_i'\hat{\boldsymbol{\beta}}$?

Leave-one-out representation

- Replace $\hat{\boldsymbol{\beta}}$ by a surrogate, starting from $\hat{\boldsymbol{\beta}}_{-j}$. For instance, if $j = 1$ is null,

$$\hat{\boldsymbol{\beta}} \approx \begin{bmatrix} 0 \\ \hat{\boldsymbol{\beta}}_{-1} \end{bmatrix} + \begin{bmatrix} b_1 \\ \boldsymbol{w} \end{bmatrix}$$

- Call RHS leave-one-out (L-O-O) representation of $\hat{\boldsymbol{\beta}}$.
- Carefully tailored choice of surrogate required—problem specific.

Inspired by El Karoui, Bean, Bickel, Lim and Yu ('13), El Karoui('13)
See also cavity method from statistical physics (Zhou's talk)

Leave-one-out representation

- Replace $\hat{\boldsymbol{\beta}}$ by a surrogate, starting from $\hat{\boldsymbol{\beta}}_{-j}$. For instance, if $j = 1$ is null,

$$\hat{\boldsymbol{\beta}} \approx \begin{bmatrix} 0 \\ \hat{\boldsymbol{\beta}}_{-1} \end{bmatrix} + \begin{bmatrix} b_1 \\ \boldsymbol{w} \end{bmatrix}$$

- Call RHS leave-one-out (L-O-O) representation of $\hat{\boldsymbol{\beta}}$.
- Carefully tailored choice of surrogate required—problem specific.

Consequence

If $\beta_j = 0$ and $\boldsymbol{b_{-j}}$ is the L-O-O representation of $\hat{\boldsymbol{\beta}}$, starting from $\hat{\boldsymbol{\beta}}_{-j}$, w.h.p.

$$\|\hat{\boldsymbol{\beta}} - \boldsymbol{b_{-j}}\| \leq Cn^{-1/2+o(1)}$$

$$\sup_{1 \leq i \leq n} |\boldsymbol{X}_{i,-j}' \hat{\boldsymbol{\beta}}_{-j} - \boldsymbol{X}_i' \hat{\boldsymbol{\beta}}| \leq Cn^{-1/2+o(1)}$$

Inspired by El Karoui, Bean, Bickel, Lim and Yu ('13), El Karoui('13)
See also cavity method from statistical physics (Zhou's talk)

Main steps: LRT

(1) Recall $-\log\ \mathsf{LRT}_j = Q_2 + Q_3$, $\sup_{1 \le i \le n} |\boldsymbol{X}_{i,-j}' \hat{\boldsymbol{\beta}}_{-j} - \boldsymbol{X}_i' \hat{\boldsymbol{\beta}}| \le C n^{-1/2+o(1)}$

$$\implies Q_3 = o_P(1)$$

Main steps: LRT

(1) Recall $-\log\ \mathsf{LRT}_j = Q_2 + Q_3$, $\sup_{1 \le i \le n} |\boldsymbol{X}'_{i,-j}\hat{\boldsymbol{\beta}}_{-j} - \boldsymbol{X}'_i\hat{\boldsymbol{\beta}}| \le Cn^{-1/2+o(1)}$

$$\implies Q_3 = o_P(1)$$

(2) Use L-O-O representation to simplify Q_2

$$2Q_2 := (\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{-j})'\nabla^2\ell(\hat{\boldsymbol{\beta}})(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{-j}) = \frac{\kappa\,\hat{\beta}_j^2}{\lambda_{[-j]}} + o_P(1)$$

Main steps: LRT

(1) Recall $-\log\ \mathsf{LRT}_j = Q_2 + Q_3$, $\sup_{1\le i\le n} |\boldsymbol{X}_{i,-j}'\hat{\boldsymbol{\beta}}_{-j} - \boldsymbol{X}_i'\hat{\boldsymbol{\beta}}| \le Cn^{-1/2+o(1)}$

$$\Longrightarrow\ Q_3 = o_P(1)$$

(2) Use L-O-O representation to simplify Q_2

$$2Q_2 := (\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{-j})'\nabla^2\ell(\hat{\boldsymbol{\beta}})(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{-j}) = \frac{\kappa\,\hat{\beta}_j^2}{\lambda_{[-j]}} + o_P(1)$$

(3) Analysis of scaling $\lambda_{[-j]}$

$$\lambda_{[-j]} := \mathrm{Tr}\left[\left(\nabla^2\ell_{-j}(\hat{\boldsymbol{\beta}}_{-j})\right)^{-1}\right] \xrightarrow{\mathbb{P}} \lambda_*$$

Main steps: LRT

(1) Recall $-\log\ \mathsf{LRT}_j = Q_2 + Q_3$, $\sup_{1\le i\le n} |\boldsymbol{X}_{i,-j}'\hat{\boldsymbol{\beta}}_{-j} - \boldsymbol{X}_i'\hat{\boldsymbol{\beta}}| \le Cn^{-1/2+o(1)}$

$$\implies Q_3 = o_P(1)$$

(2) Use L-O-O representation to simplify Q_2

$$2Q_2 := (\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{-j})'\nabla^2\ell(\hat{\boldsymbol{\beta}})(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{-j}) = \frac{\kappa\,\hat{\beta}_j^2}{\lambda_{[-j]}} + o_P(1)$$

(3) Analysis of scaling $\lambda_{[-j]}$

$$\lambda_{[-j]} := \mathrm{Tr}\left[\left(\nabla^2\ell_{-j}(\hat{\boldsymbol{\beta}}_{-j})\right)^{-1}\right] \xrightarrow{\mathbb{P}} \lambda_*$$

(4) Use L-O-O representation to analyze null marginals: $\hat{\beta}_j \overset{\mathrm{d}}{\to} \mathcal{N}(0, \sigma_*^2)$

Main steps: LRT

(1) Recall $-\log\ \mathsf{LRT}_j = Q_2 + Q_3$, $\sup_{1\le i\le n} |\boldsymbol{X}_{i,-j}'\hat{\boldsymbol{\beta}}_{-j} - \boldsymbol{X}_i'\hat{\boldsymbol{\beta}}| \le Cn^{-1/2+o(1)}$

$$\implies Q_3 = o_P(1)$$

(2) Use L-O-O representation to simplify Q_2

$$2Q_2 := (\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{-j})'\nabla^2\ell(\hat{\boldsymbol{\beta}})(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{-j}) = \frac{\kappa\,\hat{\beta}_j^2}{\lambda_{[-j]}} + o_P(1)$$

(3) Analysis of scaling $\lambda_{[-j]}$

$$\lambda_{[-j]} := \mathrm{Tr}\left[\left(\nabla^2\ell_{-j}(\hat{\boldsymbol{\beta}}_{-j})\right)^{-1}\right] \xrightarrow{\mathbb{P}} \lambda_*$$

(4) Use L-O-O representation to analyze null marginals: $\hat{\beta}_j \xrightarrow{\mathrm{d}} \mathcal{N}(0, \sigma_*^2)$

(5) Variance inflation σ_*^2 and spread in eigenvalues of Hessian λ_*
$\rightsquigarrow$ rescaling factor $\kappa\sigma_*^2/\lambda_*$

Recap

System solutions interpretations

- Introduced system of equations with solutions $(\alpha_*, \sigma_*, \lambda_*)$
- System was parametrized by κ, γ.
- Bias of MLE: α_*
- Variance of MLE: σ_*
- LRT distribution: (σ_*, λ_*).

Recap

System solutions interpretations

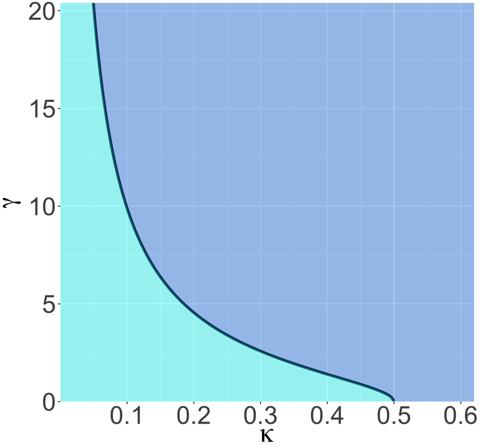
- Introduced system of equations with solutions $(\alpha_*, \sigma_*, \lambda_*)$
- System was parametrized by κ, γ.
- Bias of MLE: α_*
- Variance of MLE: σ_*
- LRT distribution: (σ_*, λ_*).

But, SNR γ is unknown in applications!

ProbeFrontier: Towards Accurate Inference

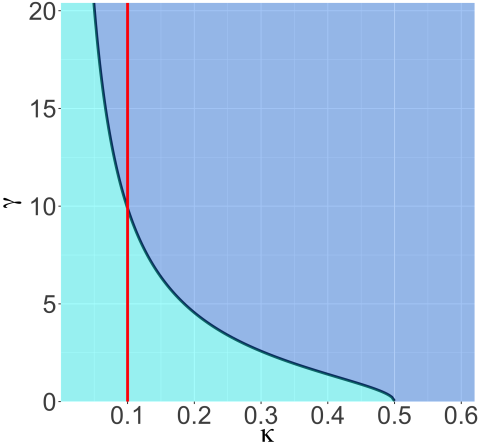
ProbeFrontier

Would need to know dimensionality κ and SNR γ



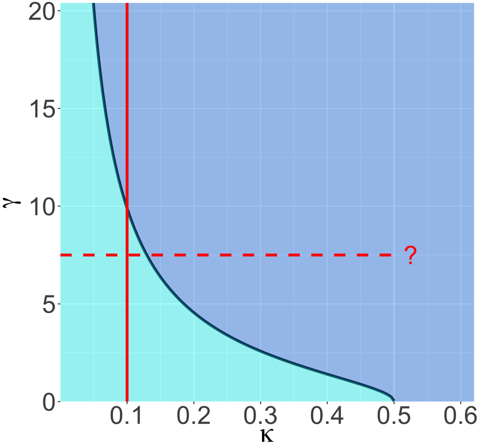
ProbeFrontier

Would need to know dimensionality κ and SNR γ



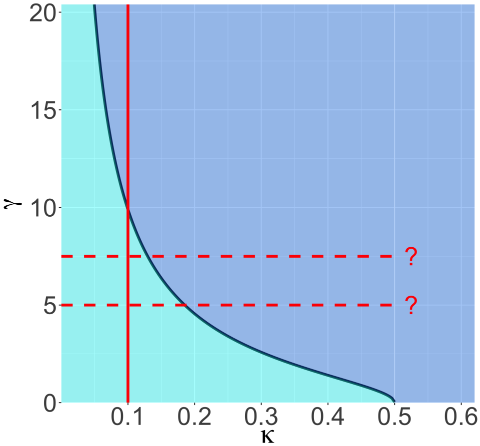
ProbeFrontier

Would need to know dimensionality κ and SNR γ



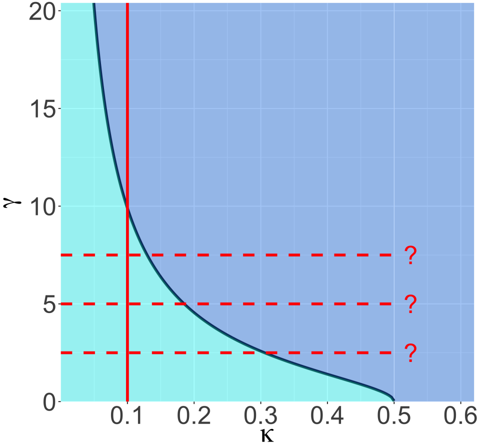
ProbeFrontier

Would need to know dimensionality κ and SNR γ



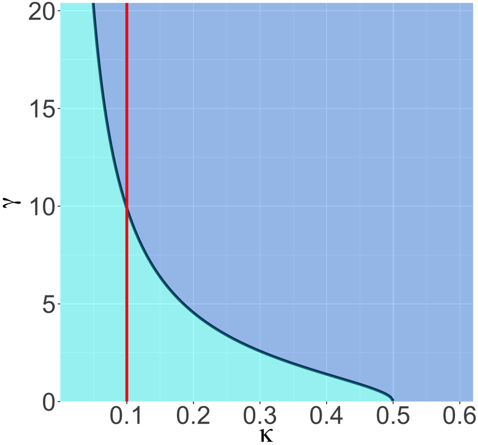
ProbeFrontier

Would need to know dimensionality κ and SNR γ



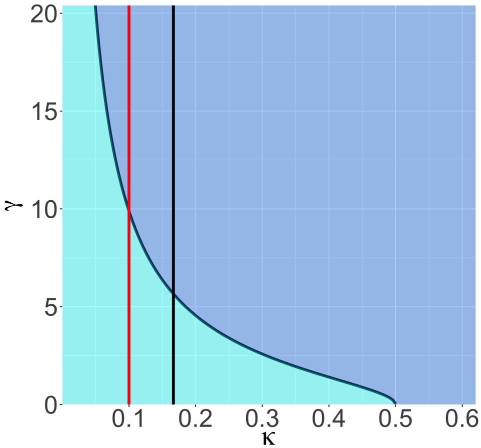
ProbeFrontier

Would need to know dimensionality κ and SNR γ



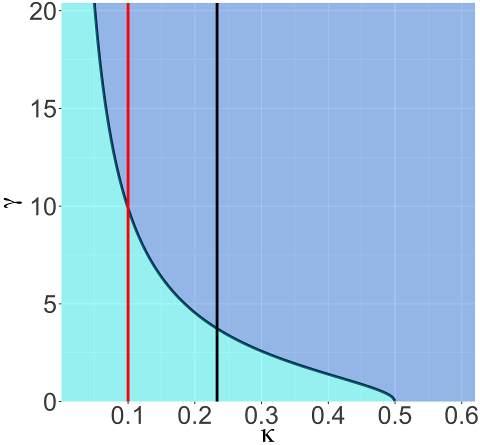
ProbeFrontier

Would need to know dimensionality κ and SNR γ



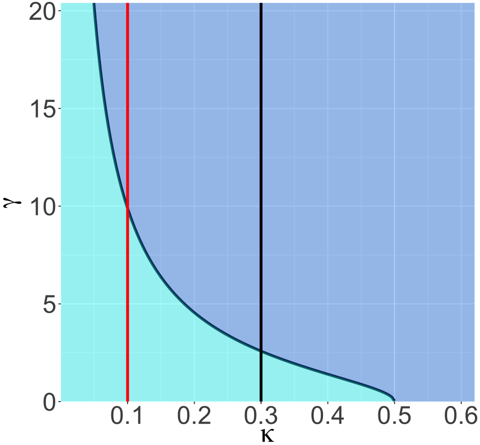
ProbeFrontier

Would need to know dimensionality κ and SNR γ



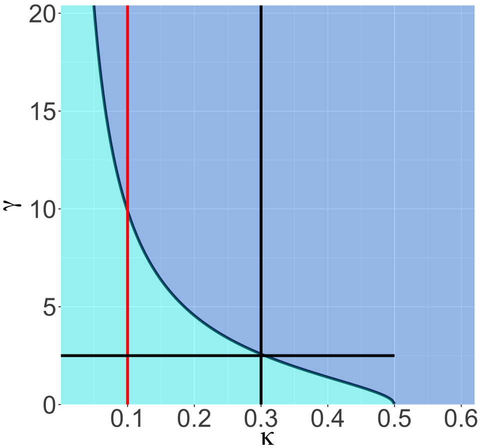
ProbeFrontier

Would need to know dimensionality κ and SNR γ



ProbeFrontier

Would need to know dimensionality κ and SNR γ



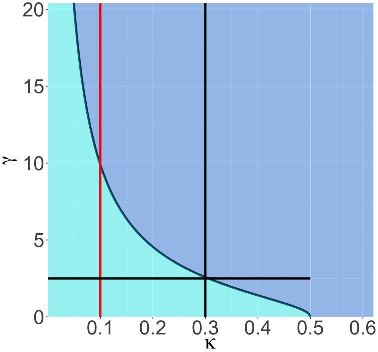
ProbeFrontier

(1) Subsample

(2) Test whether MLE exists via LP

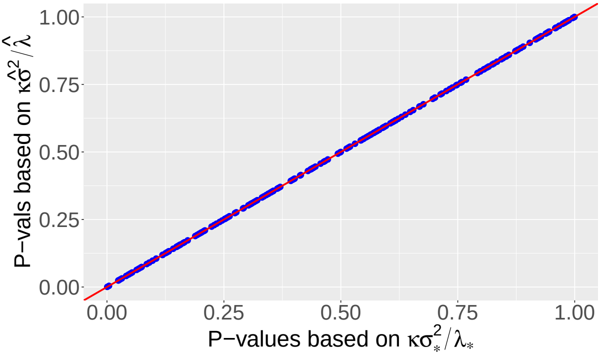
(3) Record transition point

(4) Read off $\hat{\gamma}$

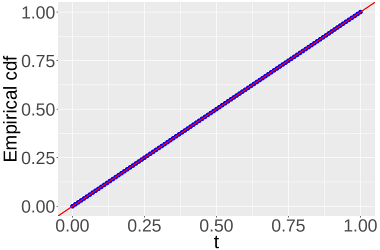


Acknowledgement: Discussions between E. Candès, R. Barber and B. Nadler (Oberwolfach, March 2018)

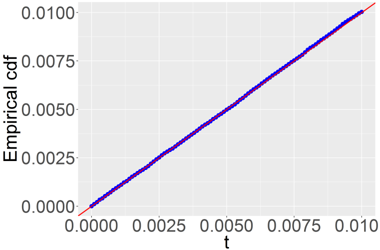
Empirical performance: null LLR p-values



Empirical performance: null LLR p-values

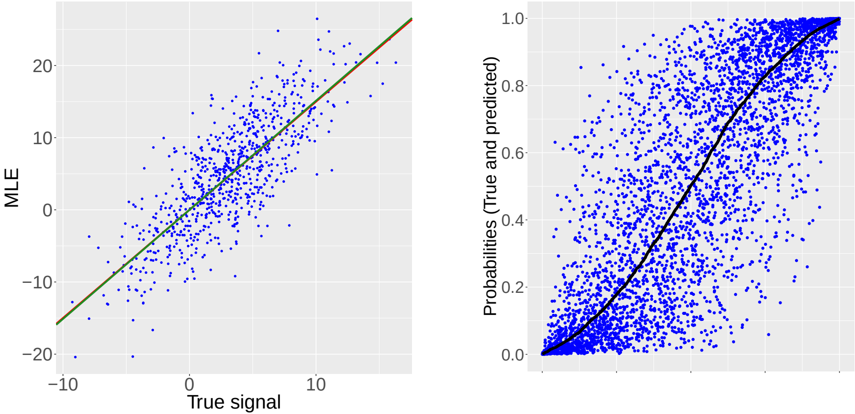


Bulk



Extreme tail

Empirical performance: de-biasing the MLE



$\alpha_* = 1.499$ (red line) and ProbeFrontier gives $\hat{\alpha} = 1.511$ (green line)

Summary and future research

- Asymptotic normality of MLE marginals
- Asymptotically exact quantification of MLE bias and variance
- Asymptotic distribution of the LRT, valid p-values
- Extremely accurate in finite samples
- Estimation of unknown parameters for practical applications

Summary and future research

- Asymptotic normality of MLE marginals
- Asymptotically exact quantification of MLE bias and variance
- Asymptotic distribution of the LRT, valid p-values
- Extremely accurate in finite samples
- Estimation of unknown parameters for practical applications

Open questions

- Penalized estimators? (Ongoing)
- Correlated covariates?
- Other GLMs ?

	Decorr.	Corr.
$\boldsymbol{\beta} = 0$ (SCC, '17)	✓	✓
$\boldsymbol{\beta} \neq 0$ (SC, '18)	✓	

Thank You!

All papers available at: `https://web.stanford.edu/~pragya/`