Failures of classical maximum-likelihood theory in high-dimensional logistic regression

Pragya Sur
Ph.D. Candidate
Stanford Statistics
(based on joint work with Emmanuel Candès and Yuxin Chen)
Staples of Classical Maximum-Likelihood Theory

- Classical asymptotics: $p$ fixed, $n \to \infty$

\[ \sqrt{n}(\hat{\beta} - \beta) \xrightarrow{d} \mathcal{N}(0, I^{-1}_\beta) \]

- Under the null,

\[ -2 \log \text{LRT} \xrightarrow{d} \chi^2 \]

Is classical inference accurate in modern settings where $n, p$ are both large and $n/p$ is 5 or 10?
Consider \( n \) i.i.d. samples \((y_i, X_i)\) from a logistic model

\[
P(y_i = 1 | X_i) = (1 + \exp(-X_i'\beta))^{-1}
\]

Covariates are generated from a Gaussian distribution

\[
\sqrt{n}X_i \overset{i.i.d.}{\sim} \mathcal{N}(0, I_{p \times p})
\]

Coefficients \( \beta \in \mathbb{R}^p \) scaled s.t.

\[
\gamma^2 := \text{Var}(X_i'\beta) = 5
\]

Dimensionality factor \( n/p = 5 \)
Is the MLE asymptotically unbiased?

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

- Dimensions sufficiently large so possibly not merely a finite sample effect.
- Same feature seen on several repetitions and for other choices of dimensions.
Unbiasedness of MLE? Second example

Figure: Scatterplot of $(\hat{\beta}_j, \beta_j)$. Line with slope 1 (black), $n = 4000, p = 800$.

〜〜 MLE seems to be over-biased even in large samples.

The bias has been noted before in small sample problems. Traditionally this has been attributed to a finite sample effect.
What about standard errors?

Figure: SEs of null coeff. estimates obtained via MC simulations (red). Classical inverse Fisher information value (blue)

MLE exhibits variance inflation in high dimensions.
Accuracy of Wilks’ theorem?

Particularly problematic for multiple testing applications!

Figure: P-values (under the null) based on $\chi^2$ approximation to the LRT

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

Observed earlier in Candès et. al. ('16)