Likelihood inference in high dimensions

Nancy Reid
University of Toronto

joint with Heather Battey, Yanbo Tang

Introduction

Inference in high dimensions

... motivated by design of experiments and likelihood inference

Part 1: linear models with $p > n$ — Heather Battey & NR

Part 2: likelihood asymptotics with $p = p_n$ — Yanbo Tang & NR

1. Linear model:

$$y_{n\times 1} = X_{n\times p}\beta_{p\times 1} + \epsilon_{n\times 1}, \quad p > n$$

$$\begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & & \vdots \\ x_{n1} & \cdots & x_{np} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \vdots \\ \vdots \\ \beta_p \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

$$\underset{\sim}{y} = X\underset{\sim}{\beta} + \underset{\sim}{\varepsilon}$$

Motivation?

- data $y = (y_1, \ldots, y_n)$
- model $f(y; \theta), \quad \theta \in \mathbb{R}^p;$ or $f(y \mid x; \beta)$ $\qquad y = X\beta + \epsilon$
- parameter of interest and nuisance parameters $\theta = (\psi, \lambda)$
- low-dimensional high-dimensional
- for example factorial and fractional factorial designs e.g. design matrix X is orthogonal
- for example adjustments to profile log-likelihood e.g. $\hat{\sigma}^2 = \frac{RSS}{n} \longrightarrow \tilde{\sigma}^2 = \frac{RSS}{n-p}$

Part I: Linear Model, $p > n$

$$y_{n\times 1} = X_{n\times p}\beta_{p\times 1} + \epsilon_{n\times 1}$$

assume column sums = 0

- parameter of interest $\beta_j = \psi$; nuisance parameter $\beta_{(-j)} = \lambda$
- if column j is orthogonal to all other columns, univariate regression

$$\hat{\beta}_j = \frac{\sum_{i=1}^n y_i x_{ij}}{\sum_{i=1}^n x_{ij}^2}$$

- could arrange this by regressing column j on other columns
- and then regressing y on the univariate residual $x_j - \hat{x}_j$
- this only works for $p < n$

- first note that β has same interpretation if

A is $n \times n$

$$Ay_{n\times 1} = AX_{n\times p}\beta + A\epsilon_{n\times 1}; \quad \tilde{y} = \tilde{X}\beta + \tilde{\epsilon}$$

- suppose we can choose $A = A^j$ to make x_j and $X_{(-j)}$ nearly orthogonal

super-saturated factorials

- $$\tilde{\beta}_j = \frac{\tilde{x}_j^T \tilde{y}}{\tilde{x}_j^T \tilde{x}_j} = \frac{\sum_i \tilde{y}_i \tilde{x}_{ij}}{\sum_i \tilde{x}_{ij}^2} \quad = \quad \frac{x_j^{jT} \tilde{y}^j}{\tilde{x}_j^{jT} \tilde{x}_j^j} = \frac{\sum_i \tilde{y}_i^j \tilde{x}_{ij}^j}{\sum_i \tilde{x}_{ij}^{j2}}$$

LS estimate from univariate regression

for each j

- $$\mathbb{E}(\tilde{\beta}_j) = \beta_j + \underbrace{\sum_{k\neq j} \beta_k \vartheta_k}_{\text{bias}} = \beta_j + \sum_{k\in\mathcal{S}} \beta_k \vartheta_k = \beta_j + \sum_{k\in\mathcal{S}} \beta_k \underbrace{\frac{\tilde{x}_j^T \tilde{x}_k}{\tilde{x}_j^T \tilde{x}_j}}_{\vartheta_k}$$

$\tilde{x}_j = A^j x_j$

... calculations

- $$\tilde{\beta}_j = \frac{\tilde{x}_j^T \tilde{y}}{\tilde{x}_j^T \tilde{x}_j} = \frac{\sum_i \tilde{y}_i \tilde{x}_{ij}}{\sum_i \tilde{x}_{ij}^2}$$

LS estimate from univariate regression

- $$\mathbb{E}(\tilde{\beta}_j) = \beta_j + \underbrace{\sum_{k \in \mathcal{S}} \beta_k \vartheta_k}_{\text{bias}}, \qquad \text{var}(\tilde{\beta}_j) = \sigma^2 V_{jj}, \qquad V_{jj} = V_{jj}^j = (\tilde{x}_j^T \tilde{x}_j)^{-2} \tilde{x}_j^T A^j A^{jT} \tilde{x}_j$$

- Choose A^j to minimize mean-squared error = bias^2 + variance $\leq ||\beta_{(-j)}||^2 \sum_{k \in \mathcal{S}} \vartheta_k^2$
- Can we find A^j to minimize

$$V_{jj} + \sum_{k \in \mathcal{S}} \vartheta_k^2$$

... strategy

- Choose A^j, linear transformation, to minimize
  cumulative MSE over all parameters
- actually upper bound of cumulative MSE $V_{jj} + \sum_{k \in \mathcal{S}} \vartheta_k^2$

- then estimate each β_j by simple linear regression
- as if the jth column of X was orthogonal to all the others

- Proposition 1 (Battey & R 2021): $q_j = A^{jT}\tilde{x}_j = A^{jT}A^j x_j$

$$q_j = (\delta I_n + X_{(-j)}^T X_{(-j)})^{-1} x_j,$$

- condition for minimum: eigenvalues of a related matrix are non-negative

$$L_\delta = (\delta I_n + X_{(-j)}^T X_{(-j)}) - \{x_j^T (\delta I_n + X_{(-j)}^T X_{(-j)})^{-1} x_j\}^{-1} x_j x_j^T$$

... now what?

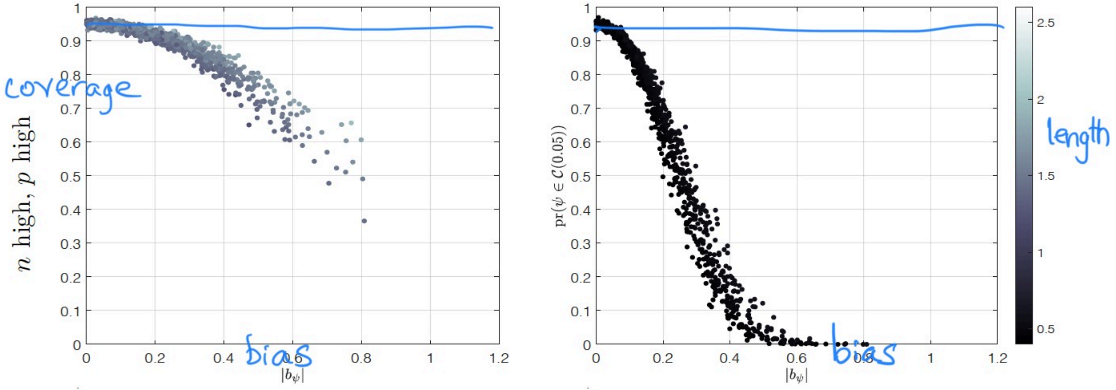
- Example: 70 observations with 2250 covariates; five covariates have non-zero β
- Compute 2250 A^j's (transformations), leading to — no model selection
- 2250 $\tilde{\beta}_j$'s and 2250 confidence intervals — $\tilde{\beta}_j \pm (\tilde{\sigma}^2 V_{jj})^{1/2} z_{1-\alpha/2}$

- asymptotically valid, if orthogonalization was successful — conditions on distribution of ϵs
- but algorithm only tries to minimize total non-orthogonality
- for some j, this non-orthogonality might be larger for the signal variables,
  leading to larger bias

- many other approaches that instead focus on identifying the signal variables first

  under some sparsity assumptions
- e.g. Zhang & Zhang (2014) estimate the bias term using the Lasso — "debiased Lasso"

$\rho = 0.9$ ← Correlation X columns → $\rho = 0.1$



coverage is better

Length is shorter

but not great

but coverage poor

ρ	n	p	modal coverage	median coverage	median length	median ϑ_θ^2	95th p.c. ϑ_θ^2
0.9	70	2450	0.941	0.921	1.509	0.0065	0.056
0.9	70	1225	0.941	0.923	1.534	0.0063	0.055
0.9	35	2450	0.941	0.909	2.127	0.0135	0.120
0.9	35	1225	0.947	0.910	2.134	0.0133	0.117
0.1	70	2450	0.939	0.732	0.504	0.0065	0.056
0.1	70	1225	0.942	0.745	0.511	0.0063	0.055
0.1	35	2450	0.948	0.715	0.707	0.0134	0.118
0.1	35	1225	0.942	0.696	0.717	0.0133	0.118

estimated main effect	ρ	n	p
modal coverage	0.995	0.933	0.986
median coverage	4.185	1.166	1.005
median length	1.216	-0.407	-0.032

… summary

- no model selection and no adjustment for multiplicity
- inference based on simple linear regression
- ignoring bias in non-orthogonality
- simple, fast, … ?useful?

- we applied it to the selection of "confidence sets for models" Battey & Cox, 2018, 2019
- in high-dimensional situations, many models may be equally informative
- Battey & Cox method is to identify these collections of models
- we used confidence intervals described here to refine this process

variable index	proportion	$\widetilde{\psi}$	lower limit	upper limit
$1516^{L,E}$	0.272	0.343	0.022	0.663
$2564^{L,E}$	0.272	-1.481	-1.801	-1.160
$1503^{L,E}$	0.251	-0.325	-0.646	-0.0050
2138	0.249	-0.062	-0.382	0.259
4008^{E}	0.240	-0.366	-0.686	-0.046
$4002^{L,E}$	0.240	-0.505	-0.825	-0.185
$1639^{L,E}$	0.235	-0.406	-0.726	-0.086
1603	0.228	-1.048	-1.368	-0.728
403	0.225	0.902	0.582	1.223
3291	0.222	-0.640	-0.960	-0.320
978	0.222	-0.259	-0.580	0.061
3808^{E}	0.222	0.677	0.356	0.997
1069^{E}	0.221	-0.398	-0.718	-0.078
$3514^{L,E}$	0.217	1.373	1.053	1.694
1436^{E}	0.199	-0.463	-0.783	-0.143
$1278^{L,E}$	0.190	0.147	-0.173	0.467

Part 2

- data $y = (y_1, \ldots, y_n)$ — Slide 2 recap
- model $f(y;\theta), \quad \theta \in \mathbb{R}^p;$ or $f(y \mid x; \beta)$ — $y = X\beta + \epsilon$
- parameter of interest and nuisance parameters $\theta = (\psi, \lambda)$
- low-dimensional high-dimensional
- for example factorial and fractional factorial designs — e.g. design matrix X is orthogonal
- for example adjustments to profile log-likelihood — e.g. $\hat{\sigma}^2 = \frac{RSS}{n} \longrightarrow \tilde{\sigma}^2 = \frac{RSS}{n-p}$

... likelihood methods, $p = O(n)$

- log-likelihood function $\quad \ell(\theta; y) = \log f(y; \theta), \quad \theta \in \mathbb{R}^p, \quad y \in \mathbb{R}^n$
- profile log-likelihood function $\quad \ell_\mathrm{p}(\psi; y) = \ell(\hat\theta_\psi) = \ell(\psi, \hat\lambda_\psi)$ $\quad \theta = (\psi, \lambda)$
- good enough if p fixed, $n \to \infty$
- for example

$n \to \infty$, p fixed

$$w = 2\{\ell_\mathrm{p}(\hat\psi) - \ell_\mathrm{p}(\psi)\} \xrightarrow{d} \chi^2_1, \qquad r = \pm w^{1/2} \xrightarrow{d} N(0, 1)$$

- fails if $p = p_n$:

$$w \xrightarrow{d} \frac{\sigma_*^2}{\lambda_*} \chi^2_1$$

Sur, Chen, Candès 2019; logistic regression, $\psi = \beta_j$

- (σ_*, λ_*) characterized as the solution of two equations the optimization path

also depends on $\lim_{n\to\infty} p_n/n$

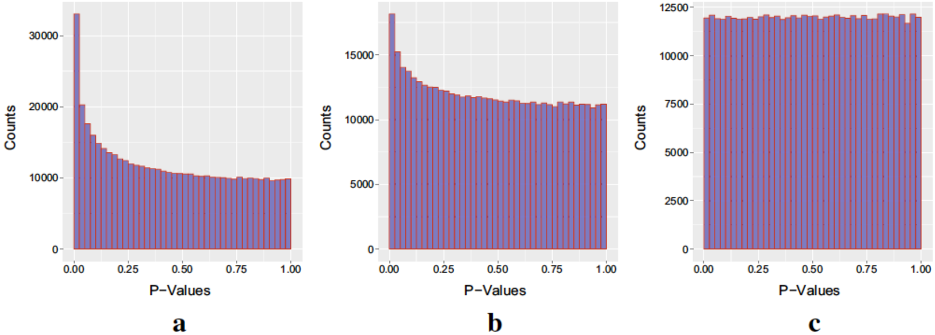


Fig. 1 Histogram of p-values for logistic regression under i.i.d. Gaussian design, when $\boldsymbol{\beta} = \mathbf{0}$, $n = 4000$, $p = 1200$, and $\kappa = 0.3$: **a** classically computed p-values; **b** Bartlett-corrected p-values; **c** adjusted p-values by comparing the LLR to the rescaled chi square $\alpha(\kappa)\chi_1^2$ (27)

Improvements to likelihood

- 1. adjust the profile log-likelihood function for estimation of nuisance parameters
- $\ell_{\mathrm{p}}(\psi) = \ell(\psi, \hat{\lambda}_\psi) \quad \longrightarrow \quad \ell_{\mathrm{mp}}(\psi) = \ell(\psi, \hat{\lambda}_\psi) - \frac{1}{2}\log|j_{\lambda\lambda}(\psi, \hat{\lambda}_\psi)|$ $j_{\lambda\lambda}$: Fisher info
- can lead to improved inference in finite samples

e.g. Kosmidis & Firth 2019 *Bka* for logistic regression

e.g. Sartori 2003 *Bka* for stratified models

- 2. adjust the log-likelihood ratio statistic

$$w = 2\{\ell_p(\hat{\psi}) - \ell_p(\psi)\}$$

- or its signed square root $\quad r = \text{sign}(\hat{\psi} - \psi)[2\{\ell_p(\hat{\psi}) - \ell_p(\psi)\}]^{1/2}$

$$r^* = r + r_{np} + r_{inf}, \qquad r^* \sim N(0,1) + O_p(n^{-3/2})$$

- Barndorff-Nielsen, 1990, *JRSS B*; Fraser, 1990, *Bka*; Pierce & Peters, 1992 *JRSS B*

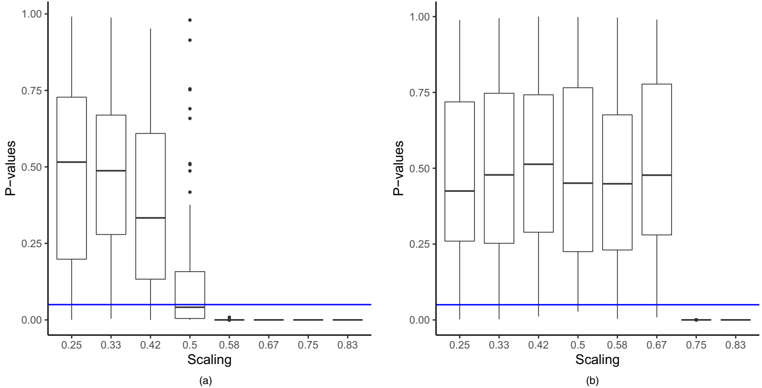


Fig. 3. Plots for logistic regression illustrating the difference in the breakdown point of uniformity of the p-value distribution based on the standard normal approximation to the distribution of (a) r and of (b) r^*: we see that p-values based on the r^*-approximation appear to be uniformly distributed up to about $p = O(n^{2/3})$, whereas those based on the normal approximation to the distribution of r begin to exhibit non-uniformity at about $p = O(n^{1/2})$

- $$r^* = r + r_{np} + r_{inf} \sim N(0,1)$$

- Tang & R Theorem 1:

$$r_{np} = O_p(p^{3/2}/n^{1/2}), \qquad \text{can be as small as} \quad O_p(p/n^{1/2})$$

- Tang & R Theorem 2:

$$r_{inf} = O_p(p/n^{1/2}), \qquad \text{can be as small as} \quad O_p(1/n^{1/2})$$

- $$r_{np} \simeq \frac{1}{r}\log\left\{\frac{|j_{\lambda\lambda}(\hat\psi,\hat\lambda)|^{1/2}}{|j_{\lambda\lambda}(\psi,\hat\lambda_\psi)|^{1/2}}\right\}, \quad r_{inf} \simeq \frac{1}{r}\log\left(\frac{t}{r}\right), \qquad t = (\hat\psi - \psi)/\hat\sigma$$

Summary

1. Linear regression, one variable at a time, no corrections for multiplicity

   Relies on isolating each variable from the others by approximate orthogonalization

2. Likelihood inference and improvements

   Relies on adjusting for estimation of nuisance parameters, and

   (less important) fine-tuning the distribution approximation

3. Classical theory impacting modern problems — much more work needed on comparisons and extensions

References i

Battey, H. and Reid, N. (2021). Inference in high-dimensional linear regression. https://arxiv.org/abs/2106.12001

Battey, H. S. and Cox, D. R. (2018). Large numbers of explanatory variables: a probabilistic assessment. *Proc Roy Soc A*

Bühlmann, P., Kalisch, M. and Meir, L. (2014). High-dimensional statistics with a view toward applications in biology. *Annu. Rev. Stat. Appl.*

Cox, D. R. and Battey, H. S. (2017). Large numbers of explanatory variables, a semi-descriptive analysis. *PNAS*

Shah, R. D. and Bühlmann, P. (2019). Double-estimation-friendly inference for high-dimensional misspecified models. arXiv:1909.10828v1. van de Geer, S., Bühlmann, P., Ritov, Y., Dezeure, R. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *AoS*

Zhang, C-H. and Zhang, S. S. (2014). Confidence intervals for low dimensional parameters in high dimensional linear models. *JRSS B*

Tang, Y. and Reid, N. (2020). Modified likelihood root in high dimensions. *J. R. Statist. Soc. B* **62**, 1349 – 1369

Barndorff-Nielsen, O.E. (1990). Approximate interval probabilities. *JRSS B*
Fan, Y., Demirkaya, E. and Lv, J. (2019) Non-uniformity of p-values can occur early in diverging dimensions. *JMLR*
Fraser, D.A.S. (1990). Tail probabilities from observed likelihoods. *Bka* Kosmidis, I. and Firth, D. (2020). Jeffreys' prior penalty, finiteness and shrinkage in binomial response models. *Biometrika*
Pierce, D.A. and Peters, D. (1992). Practical use of higher order asymptotics for multiparameter exponential families. *JRSS B*
Sartori, N. (2003). Modified profile likelihoods with stratum nuisance parameters. *Bka*
Sur, P. and Candès, E. J. (2019) A modern maximum likelihood theory for high-dimensional logistic regression. *PNAS*
Sur, P., Chen, Y. and Candès, E. J. (2019) The likelihood ratio test in high-dimensional logistic regression is asymptotically a rescaled chi-square. *PTRS*

Moderate dimensional inference – asymptotics

- $f(y;\theta)$, $y \in \mathbb{R}^n$, $\theta \in \mathbb{R}^p$
- classical: p fixed, $n \to \infty$
- semi-classical: $p_n/n \to 0$, or $p_n^{3/2}/n \to 0$

Huber, Portnoy; Sartori, Lunardon, ...

- moderate dimension $p_n/n \to \beta$

Candes, Lei/Bickel/El Karoui, ...

- "high dimension" $p_n/n \to \infty$

Semi-classical, HOA

Sartori 03, Lunardon 18

- $y_{ij} \sim f(\cdot; \psi, \lambda_i), \quad i = 1, \ldots, q; \; j = 1, \ldots, m; \;\; n = mq$

  Neyman-Scott problems
- $q \to \infty$, m fixed: classical likelihood inference fails
- $q \to \infty$, $m \to \infty$: can recover if $q = o(n^{1/2})$
- using modified likelihood from HOA, can recover if $q = o(n^{3/4})$

  Sartori
- using bias-adjusted score equation , can recover if $q = o(n^{3/4})$

  Lunardon
- HOA elimination of nuisance parameters gives large improvements in asymptotic theory and finite-sample approximations

Moderate dimension

Lei, El-Karoui, Bickel 18

- $\hat{\beta}(\rho) = \arg\min \frac{1}{n}\sum_{i=1}^{n}\rho(y_i - x_i^{\mathrm{T}}\beta)$

- coordinate-wise asymptotic normality

$$max_j d_{TV}\left\{\mathcal{L}\left(\frac{\hat{\beta}_j - \beta_j^*}{\sqrt{\mathrm{var}(\hat{\beta}_j)}}\right), N(0,1)\right\} = o(1)$$

- “For instance for least-squares, standard degrees of freedom adjustments effectively take care of many dimensionality-related problems”

- ?perhaps HOA adjustments for nuisance parameters (= ‘standard degrees of freedom adjustments’) can be effective as using $p/n \to \kappa$ asymptotics? when? why not? Kosmidis