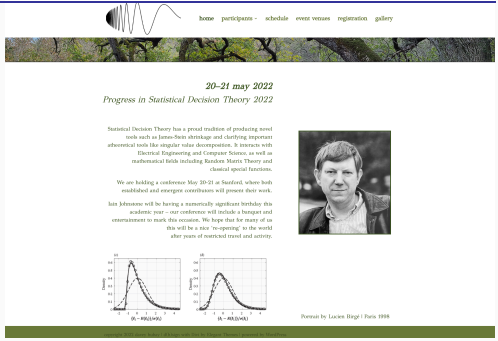
Likelihood inference in high dimensions

Nancy Reid
University of Toronto

joint with Heather Battey, Yanbo Tang



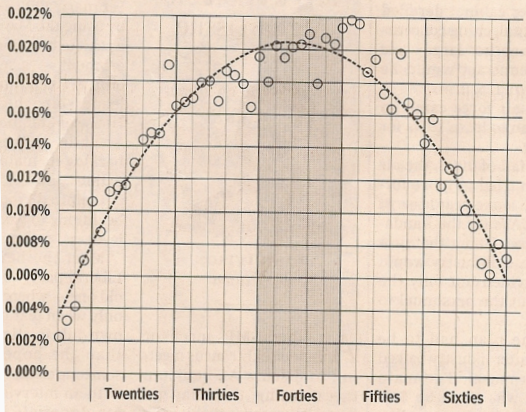

Portrait by Lucien Birgé | Paris 1998

HAVING A MID-LIFE CRISIS? YOU'RE NOT ALONE

A study involving two million people in 72 countries found men and women were less happy in their 40s but that improved in later life.

PROBABILITY OF DEPRESSION BY AGE

PERCENTAGE LIKELIHOOD

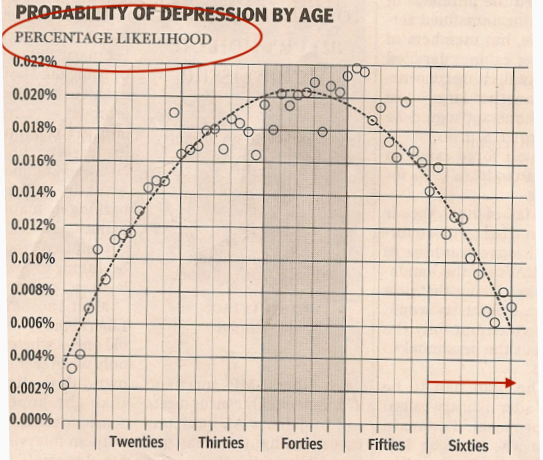


SOURCES: IS WELL-BEING U-SHAPED OVER THE LIFE CYCLE?

RICHARD JOHNSON / NATIONAL POST

HAVING A MID-LIFE CRISIS? YOU'RE NOT ALONE

A study involving two million people in 72 countries found men and women were less happy in their 40s but that improved in later life.



SOURCES: IS WELL-BEING U-SHAPED OVER THE LIFE CYCLE?

RICHARD JOHNSON / NATIONAL POST

Inference in high dimensions

Motivated by design of experiments and likelihood inference

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

Can ideas from factorial experiments inform inference on "main effects" in high-dimensional settings?

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

How fast can p_n grow with n using higher-order approximations?

Linear model: set-up

$$\begin{aligned} y_{n\times 1} &= X_{n\times p}\beta_{p\times 1} + \epsilon_{n\times 1}, \\ &= x_j\beta_j + X_{-j}\beta_{-j} + \epsilon \end{aligned} \qquad p > n$$

- scalar parameter of interest β_j; nuisance parameter $\beta_{-j} \in \mathbb{R}^{p-1}$
- if column j is orthogonal to all other columns, univariate regression; assume column sums = 0

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

- if $p < n$, can arrange this by regression of y on the univariate residual after regression of x_j on X_{-j} $x_j - \hat{x}_j$

... transformation

- β has same interpretation if

A is $n \times n$

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

- suppose we can choose $A = A^j$ to make $\tilde{x}_j$ and $\tilde{X}_{(-j)}$ nearly orthogonal

super-saturated factorials

- $$\tilde{\beta}_j = \frac{\tilde{x}_j^T \tilde{y}}{\tilde{x}_j^T \tilde{x}_j} = \frac{x_j^{jT} \tilde{y}^j}{\tilde{x}_j^{jT} \tilde{x}_j^j} \qquad \tilde{x}_j = A^j x_j$$

LS estimate from univariate regression

for each j

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

signal variables $\mathcal{S}$

... calculations

- $\tilde{\beta}_j = \dfrac{\tilde{x}_j^T \tilde{y}}{\tilde{x}_j^T \tilde{x}_j}$ LS estimate from univariate regression $\tilde{x}_j = A^j x_j$
- $\mathbb{E}(\tilde{\beta}_j) = \beta_j + \underbrace{\sum_{k \in \mathcal{S}} \beta_k \vartheta_k}_{\text{bias}}, \quad \text{var}(\tilde{\beta}_j) = \sigma^2 V_{jj}$ $V_{jj} = \dfrac{\tilde{x}_j^T A^j A^{jT} \tilde{x}_j}{(\tilde{x}_j^T \tilde{x}_j)^{-2}}$
- bias and variance depend on transformation A^j $\vartheta_k = \dfrac{\tilde{x}_j^T \tilde{x}_k}{\tilde{x}_j^T \tilde{x}_j}$
- Choose A^j to minimize $V_{jj} + \sum_{k \in \mathcal{S}} \vartheta_k^2$
- upper bound on total mean-squared error $\text{bias}^2 \leq ||\beta_{(-j)}||^2 \sum_{k \in \mathcal{S}} \vartheta_k^2$

... strategy

- Choose A^j, linear transformation, to minimize
  upper bound on cumulative MSE over all parameters $V_{jj} + \sum_{k \in \mathcal{S}} \vartheta_k^2$
- Define $q_j = A^{jT}\tilde{x}_j = A^{jT}A^j x_j$ solve

$$\underset{q \in \mathbb{R}^n}{\operatorname{argmin}} \frac{q^{\mathrm{T}}\left(\delta I_n + X_{-j}X_{-j}^{\mathrm{T}}\right)q}{(q^{\mathrm{T}}x_j)^2}$$

- Explicit solutions $a \neq 0 \in \mathbb{R}$

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

Prop. 1 (Battey & R 2021)

- Eigenvalue condition required for minimum

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

- $P(\delta,\delta)$ is ridge regression projection

$$I_n - X_{-j}(\delta I_{p-1} + X_{-j}^{\mathrm{T}}X_{-j})^{-1}X_{-j}^{\mathrm{T}}$$

... now what?

- find optimal q_j, hence $A^j \longrightarrow$ transform regression
- estimate each β_j by simple linear regression $\tilde{y}$ on $\tilde{x}_j$ $\tilde{y} = A^j y$
- as if jth column of X orthogonal to all the others
- 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
- bias in $\tilde{\beta}_j$ is $O(s/n)$; coverage of confidence intervals off by $O(s/\sqrt{n})$

Prop 3 (Battey & R 2021)

- assume sparsity in columns of X — induce sparsity
- subtract estimate of bias in OLS estimator — ignore bias $O(s/n)$

- Z&Z equation for β_j:

$$z_j(y - x_j\beta_j) = 0$$

z_j to be chosen

- recommend using residuals from Lasso regression of x_j on X_{-j}

- B&R equation for β_j:

$$q_j(y - x_j\beta_j) = 0$$

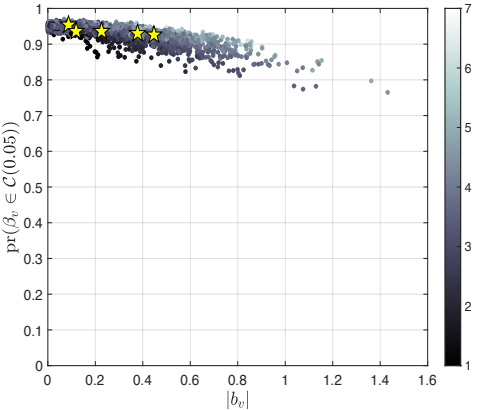
$q_j = A^{jT}A^j x_j$

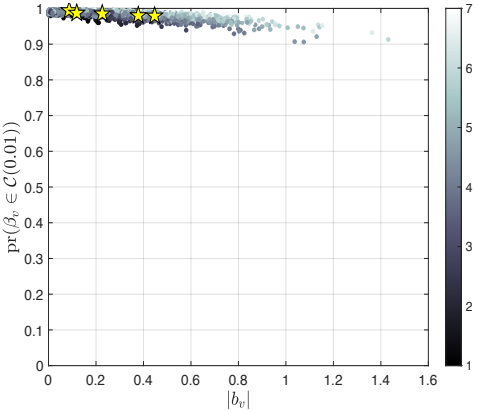
- leads to residuals from ridge regression of x_j on X_{-j}

- X from gene expression data $n = 71, p = 4088$
- y simulated with Gaussian noise and $s = 5$ signal variables
- 4088 estimates $\tilde{\beta}_j$ and confidence intervals
- compare to debiased lasso
- OLS computation faster
- lasso version requires handling $p - 1$ nuisance parameters for each estimate and CI

SIMULATIONS BASED ON REAL COVARIATE DATA

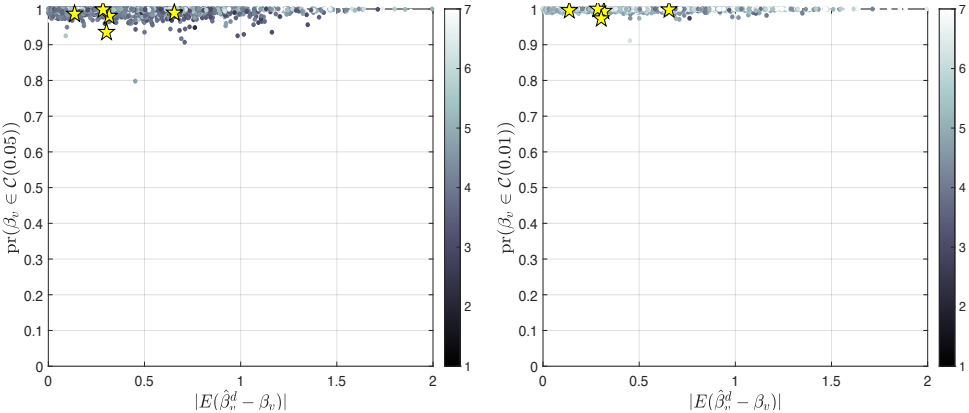
nominal level	modal coverage	median coverage	proportion with coverage > 0.9	proportion with coverage > 0.95	median length
0.05	0.951	0.942	0.939	0.228	3.32
0.01	0.989	0.987	1	0.994	4.37





EQUIVALENT RESULTS FOR THE DEBIASED LASSO

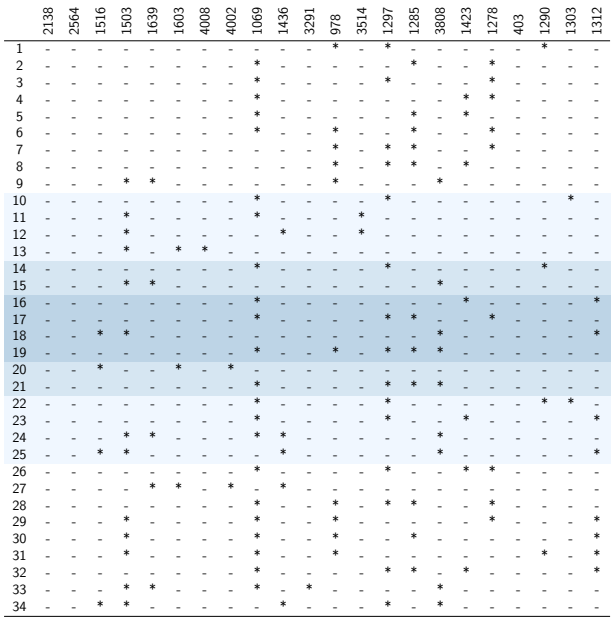
nominal level	modal coverage	median coverage	proportion with coverage > 0.9	proportion with coverage > 0.95	median length
0.05	1	0.998	1	0.996	4.29
0.01	1	1	1	1	5.64



Illustration

- 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 selection with confidence sets $\mathcal{M}$, with confidence limits for the associated regression coefficients.

variable index v	proportion of models in $\mathcal{M}$	$\widetilde{\beta}_v$	lower limit (0.05)	upper limit (0.05)	lower limit (0.01)	upper limit (0.01)
2138	0.218	-0.062	-0.382	0.259	-0.483	0.359
$2564^{L,E}$	0.218	-1.481	-1.801	-1.160	-1.901	-1.060
$1516^{L,E}$	0.218	0.343	0.022	0.663	-0.079	0.763
$1503^{L,E}$	0.217	-0.325	-0.646	-0.0050	-0.746	0.096
$1639^{L,E}$	0.217	-0.406	-0.726	-0.086	-0.827	0.015
1603	0.217	-1.048	-1.368	-0.728	-1.469	-0.627
4008^{E}	0.217	-0.366	-0.686	-0.046	-0.787	0.055
$4002^{L,E}$	0.216	-0.505	-0.825	-0.185	-0.926	-0.084
1069^{E}	0.216	-0.398	-0.718	-0.078	-0.819	0.023
1436^{E}	0.215	-0.463	-0.783	-0.143	-0.884	-0.042
3291	0.215	-0.640	-0.960	-0.320	-1.061	-0.220
978	0.214	-0.259	-0.580	0.061	-0.680	0.162
$3514^{L,E}$	0.214	1.373	1.053	1.694	0.953	1.794
$1297^{L,E}$	0.213	0.219	-0.102	0.539	-0.202	0.640
1285	0.213	0.172	-0.148	0.493	-0.249	0.593
3808^{E}	0.213	0.677	0.356	0.997	0.256	1.098
1423	0.212	0.043	-0.277	0.363	-0.378	0.464
$1278^{L,E}$	0.212	0.147	-0.173	0.467	-0.274	0.568
403	0.211	0.902	0.582	1.223	0.481	1.323
1290	0.211	0.189	-0.131	0.510	-0.232	0.610
1303^{E}	0.211	0.187	-0.133	0.507	-0.234	0.608



- Proposed usage is as an adjunct to, and means of refining, confidence sets of models (Cox and Battey, 2017, 2018)
- Reported is the central region of a confidence "distribution" of models in the sense of Fisher (1930) and Cox (1958) from a real example with $p = 4088$ and $n = 71$.

Likelihood 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: $p_n/n \to \kappa$ — Candès/Sur, Lei/Bickel/El Karoui, Alexei (?)
- "high dimension": $p_n/n \to \infty$
- where is/what causes the 'breakpoint' between semi-classical and moderate?
- higher-order approximations seem to be very accurate, even with many nuisance parameters

... likelihood asymptotics

- $f(y;\theta)$, $y \in \mathbb{R}^n$, $\theta \in \mathbb{R}^p$
- parameter of interest and nuisance parameters $\theta = (\psi, \lambda)$
- low-dimensional high-dimensional
- log-likelihood function $\ell(\theta; y) = \log f(y;\theta), \quad \theta \in \mathbb{R}^p, \quad y \in \mathbb{R}^n$
- profile log-likelihood function $\ell_p(\psi; y) = \ell(\hat{\theta}_\psi) = \ell(\psi, \hat{\lambda}_\psi)$ $\theta = (\psi, \lambda)$
- classical limit theory p fixed, $n \to \infty$

$$w = 2\{\ell_p(\hat{\psi}) - \ell_p(\psi)\} \xrightarrow{d} \chi^2_q, \qquad r = \pm w^{1/2} \xrightarrow{d} N(0,1), \qquad (q = 1)$$

... likelihood asymptotics, $p = p_n$

- log-likelihood function $\ell(\theta; y) = \log f(y; \theta), \quad \theta \in \mathbb{R}^p, \quad y \in \mathbb{R}^n$
- profile log-likelihood function $\ell_p(\psi; y) = \ell(\hat{\theta}_\psi) = \ell(\psi, \hat{\lambda}_\psi)$ $\theta = (\psi, \lambda)$
- classical limit theory $n \to \infty$, p fixed, q=1

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

- 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 $\lim_{n\to\infty} p_n/n$

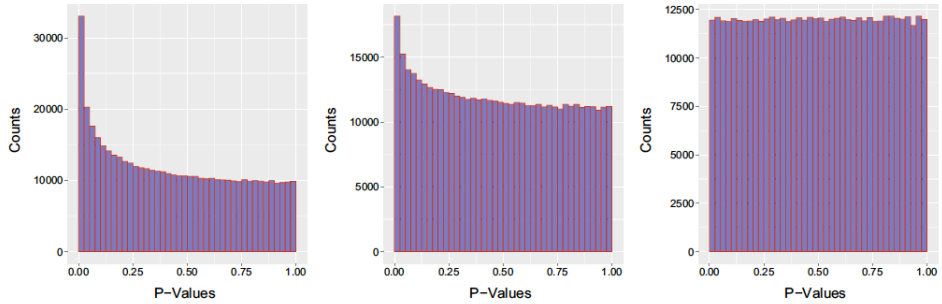


Fig. 1 Histogram of p-values for logistic regression under i.i.d. Gaussian design, when $\boldsymbol{\beta} = \mathbf{0}$, $n = 4000$, $p = 1200$, and $\kappa = 0.3$: **a** classically computed p-values; **b** Bartlett-corrected p-values; **c** adjusted p-values by comparing the LLR to the rescaled chi square $\alpha(\kappa)\chi^2_1$ (27)

Improvements to likelihood inference, p fixed

1. adjust the profile log-likelihood function for estimation of nuisance parameters

   - $\ell_p(\psi) = \ell(\psi, \hat{\lambda}_\psi) \quad \longrightarrow \quad \ell_{mp}(\psi) = \ell(\psi, \hat{\lambda}_\psi) - \frac{1}{2} \log |j_{\lambda\lambda}(\psi, \hat{\lambda}_\psi)|$ $\qquad j_{\lambda\lambda}$: 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

$$r = \text{sign}(\hat{\psi} - \psi)[2\{\ell_p(\hat{\psi}) - \ell_p(\psi)\}]^{1/2} \qquad r \sim N(0,1) + O_p(n^{-1/2})$$

$$r^* = r \quad + \underbrace{r_{np}}_{\text{nuisance parameters}} + \underbrace{r_{inf}}_{\text{distribution}}$$

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

Barndorff-Nielsen 90; Fraser 90; Pierce & Peters 92

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

$p = O(n^\alpha), \alpha < 0.5$

- nuisance parameter adjustment involves Fisher information for λ:

$$r_{np} \simeq \frac{1}{r} \log \left\{ \frac{|j_{\lambda\lambda}(\hat{\psi}, \hat{\lambda})|^{1/2}}{|j_{\lambda\lambda}(\psi, \hat{\lambda}_\psi)|^{1/2}} \right\}$$

- distribution adjustment compares likelihood root to Wald or Rao (e.g.) asy. equiv.

$$r_{inf} \simeq \frac{1}{r} \log \left(\frac{t}{r} \right)$$

- t is one of the classical test statistics

$$t = (\hat{\psi} - \psi)/\hat{\sigma}, \quad \text{or} \quad t = \ell'_p(\psi)/\hat{\tau}, \quad \text{or ...}$$

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

$p = O(n^\alpha), \alpha < 0.5$

- Result 1

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

- Result 2

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

- restriction semi-classical

$$p = O(n^\alpha), \alpha < 0.5$$

- this explains 'folk theorem' about higher order approximations

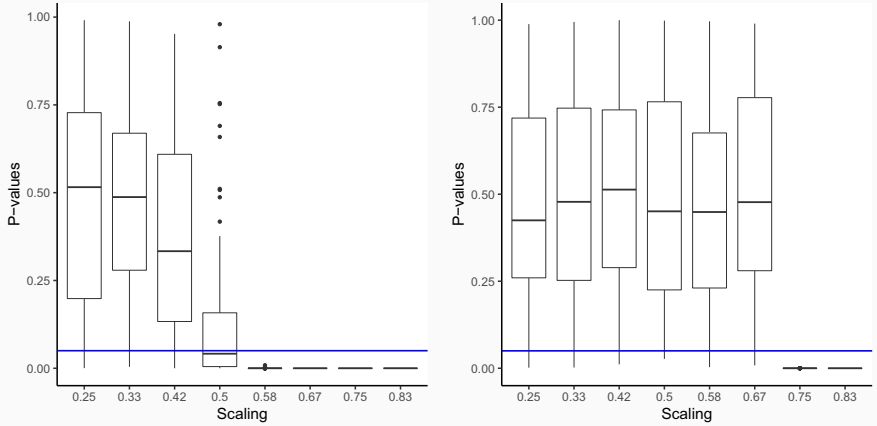


Fig. 3. Plots for logistic regression illustrating the difference in the breakdown point of uniformity of the p-value distribution based on the standard normal approximation to the distribution of (a) r and of (b) r^*: we see that p-values based on the r^*-approximation appear to be uniformly distributed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

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 London A* **474**, 20170631.

Bühlmann, P., Kalisch, M. and Meir, L. (2014). High-dimensional statistics with a view toward applications in biology. *Annu. Rev. Stat. Appl.* **1**, 255–278.

Cox, D. R. and Battey, H. S. (2017). Large numbers of explanatory variables, a semi-descriptive analysis. *PNAS* **114**, 8592–8595.

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. *Ann. Statist.* **42**, 1166–1202.

Zhang, C-H. and Zhang, S. S. (2014). Confidence intervals for low dimensional parameters in high dimensional linear models. *JRSS B* **76**, 217–242.

References ii

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* **52**, 485–496.

Fraser, D.A.S. (1990). Tail probabilities from observed likelihoods. *Bka* **77**, 65–76.

Kosmidis, I. and Firth, D. (2021). Jeffreys' prior penalty, finiteness and shrinkage in binomial response models. *Biometrika* **108**, 71–82.

Lei, L., Bickel, P.J. and El-Karoui, N.E. (2016). Asymptotics for high-dimensional *M*-estimates: fixed design results. *Prob. Th. Rel. Fields* **172**, 983–1079. Pierce, D.A. and Peters, D. (1992). Practical use of higher order asymptotics for multiparameter exponential families. *JRSS B* **54**, 701–737.

Sartori, N. (2003). Modified profile likelihoods with stratum nuisance parameters. *Bimetrika* **90**, 533–549.

Sur, P. and Candès, E. J. (2019) A modern maximum likelihood theory for high-dimensional logistic regression. *PNAS* **116**, 14516–14525.

Sur, P., Chen, Y. and Candès, E. J. (2019) The likelihood ratio test in high-dimensional logistic regression is asymptotically a rescaled chi-square. *Prob. Th. Rel. Fields* **175**, 487–558.