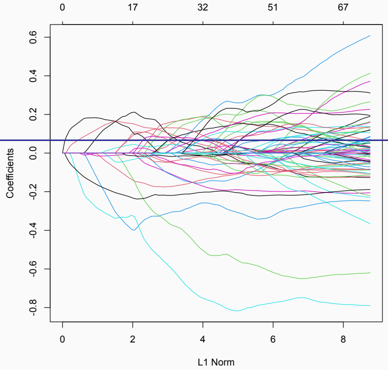
Topics in Likelihood Inference

STA4508H



Nancy Reid

University of Toronto

February 16, 2022

Various ‘types’ of likelihood

1. likelihood, marginal and conditional likelihood, profile likelihood, adjusted profile
2. semi-parametric likelihood, partial likelihood
3. quasi-likelihood, composite likelihood — misspecified models
4. empirical likelihood, penalized likelihood
5. likelihood inference in high dimensions
6. simulated likelihood, indirect inference
7. bootstrap likelihood, *h*-likelihood, weighted likelihood, pseudo-likelihood, local likelihood, sieve likelihood

Presentations

Feb 16	Angela: Cox (2013) Robert: Barndorff-Nielsen and Cox (1979) Shiki: Solomon and Cox (1992)
Feb 23	Hengchao: Rotnitzky et al. (2000) Siyue: De Stavola and Cox (2008) Manuel: Battey and Cox (2018) Ziang: Cox (1975)

Feb 16 in SS 1087; Feb 23 online

exercises Jan 26 has details about report structure

High-dimensional inference

- $f(y;\theta), y \in \mathbb{R}^n; \theta \in \mathbb{R}^p$, p large relative to n, or $p > n$
- Partial likelihood has $p = n - 1 + d$, yet usual asymptotic theory applies
- Empirical likelihood has $p = n - 1$, yet usual asymptotic theory applies
- "Neyman-Scott problems" have $y_{ij} \sim f(\cdot;\psi,\lambda_i), j = 1,\ldots,m; i = 1,\ldots,k$, so $n = km$ and $p = 1 + k$ i.e. $p/n = O(1)$ if $m \to \infty$, k fixed; usual theory does not apply

Empirical likelihood

- $Y_1, \ldots, Y_n$ i.i.d. F $\qquad \mu = \mathrm{E}(Y_i) = \int y dF(y)$
- profile likelihood: maximize $\prod_{i=1}^n p_i,$ $\quad$ subject to $p_i \geq 0, \Sigma p_i = 1, \Sigma p_i Y_i = \mu$
- solution

$$\hat{p}_i(\mu) = \frac{1}{n}\frac{1}{1+\lambda(Y_i-\mu)}, \text{ where } \lambda \text{ solves}$$
$$0 = \frac{1}{n}\sum_{i=1}^n \frac{Y_i-\mu}{1+\lambda(Y_i-\mu)}$$

- Theorem: $R(Y) = \{\mu : g(Y;\mu) \leq k_\alpha\}$ is an approximate $1-\alpha$ confidence interval for μ, where $\mathrm{pr}\{\chi_1^2 \geq k_\alpha\} = \alpha$
- if $E|Y_i|^3 < \infty$, under $H_0 : \mu = \mu_0$: $\qquad$ actually $\mathrm{var}(Y_i) < \infty$

$$-2\sum_{i=1}^n \log\{n\hat{p}_i(\mu_0)\} \xrightarrow{d} \chi_1^2, \quad n \to \infty$$

- proof first shows

$$\lambda = O_p(n^{-1/2})$$

- then

$$\lambda \doteq \frac{\bar{Y} - \mu}{S(\mu)}, \quad S(\mu) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \mu)^2$$

- then Taylor series expansion:

$$-2\sum_{i=1}^{n}\log\{n\hat{p}_i(\mu_0)\} \doteq \frac{n(\bar{Y} - \mu_0)^2}{S(\mu_0)}$$

- see Owen (2001) *Empirical Likelihood* for many generalizations, including proportional hazards model

Neyman-Scott problems

- $Y_{ij} \sim N(\mu_i, \sigma^2), \quad j = 1, \ldots, m; \quad i = 1, \ldots, q; \qquad \theta = (\mu_1, \ldots, \mu_q, \sigma^2)$ Sartori's notation
- $Y_{ij} \sim N(\mu, \sigma_i^2), \quad j = 1, \ldots, m; \quad i = 1, \ldots, q; \qquad \theta = (\mu, \sigma_1^2, \ldots, \sigma_q^2)$
- $Y_{ij} \sim Bern(p_{ij}), \quad j = 1, 2; \quad i = 1, \ldots, q; \qquad \psi = \log\{\frac{p_{i1}(1-p_{i2})}{p_{i2}(1-p_{i1})}\}; \quad \lambda_1, \ldots, \lambda_q$
- $Y_{ij} \sim Gamma(\psi, \lambda_i), \quad j = 1, \ldots, m; \quad i = 1, \ldots, q; \qquad \theta = (\psi, \lambda_1, \ldots, \lambda_q)$
- $Y_{i1} \sim Gamma(m, \psi/\lambda_i), Y_{i2} \sim Gamma(m, \psi\lambda_i); \qquad \theta = (\psi, \lambda_1, \ldots, \lambda_q)$
- sample size $n = mq$; $\qquad m \to \infty, q$ fixed $\qquad$ or m fixed , $q \to \infty$ $\qquad$ or $m, q \to \infty$
- "methods based on the profile likelihood may fail unless $q = o(n^{1/2})$, ... based on modified profile likelihoods still perform accurately, provided that $q = o(n^{3/4})$"

- Gamma example:
  $Y_{ij} \sim Gamma(\psi, \lambda_i), \quad j = 1, \dots, m; \quad i = 1, \dots, q; \qquad \theta = (\psi, \lambda_1, \dots, \lambda_q)$

- there is an exact conditional log-likelihood for ψ — linear exponential family
- $\ell_C(\psi) = \psi s + q \log \Gamma(m\psi) - mq \log \Gamma(\psi)$

- profile log-likelihood gives poor estimates for large q
- $\ell_P(\psi) = \psi s + qm\psi \log(m\psi) - mq\psi - mq \log \Gamma(\psi)$

- modified profile log-likelihood is very close to conditional
- $\ell_M(\psi) = \psi s + q(m\psi - 1/2) \log(m\psi) - mq\psi - mq \log \Gamma(\psi)$

Table 1: *Example 2. Inference about common shape parameter in q gamma samples of size m. Probabilities* $\Phi\{r_P(\psi)\}$ *and* $\Phi\{r_M(\psi)\}$ *with* ψ *such that* $\Phi\{r_C(\psi)\} = 0{\cdot}05$ *in samples with* $\hat{\psi}_C = 1$. *For each q, values in bold face correspond to the smallest m, for* r_P *and* r_M, *such that the probability is within* 0·01 *of* 0·05.

m		$q=4$	$q=8$	$q=16$	$q=64$	$q=128$
3	r_P	0·190	0·299	0·487	0·952	0·999
	r_M	**0·053**	**0·055**	**0·058**	0·070	0·080
4	r_P	0·159	0·239	0·383	0·866	0·988
	r_M	0·052	0·053	0·055	0·062	0·068
5	r_P	0·141	0·206	0·322	0·777	0·962
	r_M	0·052	0·052	0·054	**0·058**	0·062
6	r_P	0·129	0·184	0·281	0·698	0·924
	r_M	0·051	0·052	0·053	0·056	**0·059**
200	r_P	**0·060**	0·064	0·071	0·098	0·126
	r_M	0·050	0·050	0·050	0·050	0·050
400	r_P	0·056	**0·059**	0·064	0·081	0·098
	r_M	0·050	0·050	0·050	0·050	0·050
800	r_P	0·054	0·057	**0·059**	0·071	0·080
	r_M	0·050	0·050	0·050	0·050	0·050
3000	r_P	0·052	0·053	0·055	**0·059**	0·064
	r_M	0·050	0·050	0·050	0·050	0·050
6000	r_P	0·051	0·053	0·053	0·057	**0·060**
		[illegible]	[illegible]	[illegible]	[illegible]	[illegible]

Increasing dimension asymptotics

- classical: $p/n \to 0$, p fixed, $n \to \infty$ — θ has dimension p or p_n
- semi-classical $p_n/n \to 0$ or $p_n^{3/2}/n \to 0$ or $p_n^2/n \to 0$ — Portnoy, Sartori
- moderate dimensions $p_n/n \to \beta$ — Sur & Candes '17
- high dimensions $p_n/n \to \infty$
- ultra-high dimensions $p_n \sim e^n$
- Portnoy 1988
  - MLE "will tend to be consistent" if $p/n \to 0$
  - asymptotic approximations okay if $p^{3/2}/n \to 0$
  - and fail if $p^2/n \to 0$ — Portnoy 1984, 1985

High-dimensional linear regression

- $y = X\beta + \epsilon, \quad E(\epsilon) = 0, \ \text{cov}(\epsilon) = \sigma^2 I \qquad p >> n$ — running example, $n = 71, p = 4088$

$$\hat\beta_{ridge} = \arg\min_\beta\{\frac{1}{n}(y - X\beta)^T(y - X\beta) + \lambda\sum_{j=1}^{p}\beta_j^2,$$

$$\hat\beta_{lasso} = \arg\min_\beta\{\frac{1}{n}(y - X\beta)^T(y - X\beta) + \lambda\sum_{j=1}^{p}|\beta_j|$$

- usual to assume $\Sigma_{i=1}^n x_{ij} = 0, \Sigma_{i=1}^n x_{ij}^2 = 1$ — so units are comparable

  $\hat\beta_0 = \bar y$ is not 'shrunk'
- Lasso penalty leads to several $\hat\beta_k = 0$ — sparse solutions
- there are many variations on the penalty term
- λ is a tuning parameter — often selected by cross-validation

... high-dimensional inference

- Inferential goals (§2.2)
  - (a) prediction of surface $X\beta$ or $y_{new} = x_{new}^{\mathrm{T}}\beta$
  - (b) estimation of β
  - (c) estimation of $S = \{j : \beta_j \neq 0)$ ‘support set’

- (a): “identifiability of β is not necessary; from this perspective, prediction is often a much easier problem”

- (b): “an identifiability assumption on the design X is required, for example, a restricted eigenvalue condition” not checkable (?)

- (c): “ideally, ... $\hat{S} = S$ with high probability” $\hat{S}$ estimate of S e.g. $\{j; \hat{\beta}_{j,Lasso} \neq 0\}$
  - (c): requires β_{min} condition: $\min|\beta_j| > c$, $c \sim \sqrt{\log p/n}$ $\simeq 0.34|S|$
  - often replaced by (c’): ‘screening’ $\hat{S} \supset S$ with high probability

also needs conditions on X

… high-dimensional inference

- Inferential goals (§2.2)
  - (a) prediction of surface $X\beta$ or $y_{new} = x_{new}^{\mathrm{T}}\beta$
  - (b) estimation of β
  - (c) estimation of $S = \{j : \beta_j \neq 0\}$ 'support set'

- can solve (a) and (b), if "the underlying truth is sparse"
- for example, if $|S| << n/\log p$ then just need $\log p << n$ $\log(4088) \doteq 8$
- "if the true underlying model is not sparse, them high-dimensional inference is ill posed and uninformative"
- re (a) prediction accuracy can be assessed by cross-validation
- note that (a) can be estimable even if $p > n$, as long as $X\beta$ of small enough dimension

Inference about β, $p > n$

- re (b): inference for β: e.g. p-values for testing $H_{0,j} : \beta_j = 0$
- three methods suggested: multi-sample splitting, debiasing, stability selection
- multi-sample splitting: fit the model on random half, say, of observations, leads to $\hat{S}$
- use $X^{(\hat{S})}$ in fitting to the other half
- $P_j = p$-value for t-test of H_j if $j \in \hat{S}$, o.w. 1
- $P_{corr,j} = P_j \times |\hat{S}|, j \in \hat{S}$, o.w. 1
- Repeat B times and aggregate $P^b_{corr,j}$

$P^b_{corr,j}$ not independent

Inference about β, $p > n$

- de-biasing $\hat{\beta}_{ridge/Lasso,corr,j} = \hat{b}_j - \text{bias}$ — see paper
- Can show resulting estimate $\hat{\beta}_{ridge/Lasso,corr,j} \sim N(\beta_j, \sigma^2_\epsilon w_j)$ — w_j known
- $\hat{\beta}_{ridge/Lasso,corr,j} \neq 0$, any j, so need multiplicity correction — $p = 4088$
- stability selection — Meinshausen & B 2010; flexible
- on their example
  - Lasso selects 30 features;
  - multi-sample selects 1;
  - bias-corrected Ridge selects 0;
  - stability selection selects 3 — implemented in `hdi`

Non-linear models

- example y_i independent, $\mathrm{E}(y_i) = \mu_i(\beta)$; $g(\mu_i) = \beta_0 + x_i^{\mathrm{T}}\beta$
- Lasso-type 'mle': $\arg\min\{-\frac{1}{n}\ell(\beta, \beta_0; y) + \lambda\Sigma_{j=1}|\beta_j|\}$ $\quad \beta = (\beta_1, \dots, \beta_p)$
- can use multi-sample splitting or stability selection
- a version of de-biasing applies to GLMs, based on weighted least squares
- a marginal approach would fit $y = \alpha_0 + \alpha_j x^{(j)}$, one feature at a time
- leading to 4088 p-values, and then need techniques for controlling FWER or FDR

- Model: $y_i = x_i^{\mathrm{T}}\beta + Z_i, \quad i = 1, \dots, n$ independent
- M-estimation:

$$\sum_{i=1}^{n} x_i \psi(y_i - x_i^{\mathrm{T}}\hat{\beta}) = 0 \tag{1}$$

- result: if ψ is monotone, and $p\log(p)/n \to 0$, and conditions on X, then

there is a solution of (1) satisfying $||\hat{\beta} - \beta||^2 = O(p/n)$

- "rows of X behave like a sample from a distribution in $\mathbb{R}^p$"

- if $p^{3/2}\log n/n \to 0$, then

$$\max|x_i^{\mathrm{T}}(\hat{\beta} - \beta)| \xrightarrow{p} 0$$

- and

$$a_n^{\mathrm{T}}(\hat{\beta} - \beta) \xrightarrow{d} N(0, \sigma^2)$$

$\sigma^2 = a_n^{\mathrm{T}}(X^{\mathrm{T}}X)^{-1}a_n \mathrm{E}\psi^2(Z)/\{E\psi'(Z)\}^2$