Mathematical Statistics II

STA2212H S LEC9101

Week 2

January 17 2023

Bookmark Sign in

Two-thirds of world's glaciers expected to disappear by end of the century, study in Science journal says

SETH BORENSTEIN

But if the world can limit future warming to just a few more tenths of a degree and fulfill international goals – technically possible but unlikely according to many scientists – then slightly less than half the globe's glaciers will disappear, said the same study. Mostly small but wellknown glaciers are marching to extinction, study authors said.

In an also unlikely worst-case scenario of several degrees of warming, 83 per cent of the world's glaciers would likely disappear by the year 2100, study authors said.

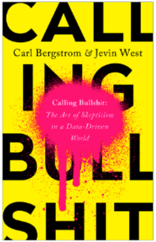
The study, published Thursday in the journal Science, examined all of the globe's 215,000 landbased glaciers – not counting those on ice sheets in Greenland and Antarctica – in a more comprehensive way than past studies. Scientists

Tourists hike to visit the Nigardsbreen glacier in Jostedal, Norway, last August. Scientists project the planet will lose between 38.7 trillion and 64.4 trillion tonnes of glacial ice by the end of the century.

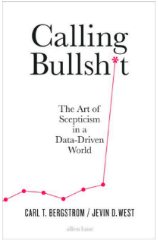
1. Recap
2. Nonparametric Likelihood MS 5.6
3. Profile Likelihood
4. Bayesian Estimation MS 5.8

Upcoming seminars of interest

- January 23 11.00 –12.00 Jevin West Details
- “The Art of Skepticism in a Data-Driven World”
- 140 St. George St., 4th floor



Penguin Random House



Now available! *Calling Bullshit: The Art of Skepticism in a Data-Driven World*, by Carl Bergstrom and Jevin West. Available here.

Recap

- data $x_1, \ldots, x_n$ independent observations; model $f(\boldsymbol{x};\theta) = \prod f(x_i;\theta), \quad \theta \in \mathbb{R}$
- limit theorem $\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} N(0, I_1^{-1}(\theta))$
- approximation $\hat{\theta} \dot\sim N\{\theta, I^{-1}(\hat{\theta})\}$, or $\hat{\theta} \dot\sim N\{\theta, J^{-1}(\hat{\theta})\}$ $\qquad I(\theta) = nI_1(\theta), \quad J(\theta) = -\ell''(\theta;\boldsymbol{x})$

Recap

- data $x_1, \ldots, x_n$ independent observations; model $f(\boldsymbol{x};\theta) = \prod f(x_i;\theta), \quad \theta \in \mathbb{R}$
- limit theorem $\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} N(0, I_1^{-1}(\theta))$
- approximation $\hat{\theta} \dot{\sim} N\{\theta, I^{-1}(\hat{\theta})\}$, or $\hat{\theta} \dot{\sim} N\{\theta, J^{-1}(\hat{\theta})\}$ $\quad I(\theta) = nI_1(\theta), \quad J(\theta) = -\ell''(\theta;\boldsymbol{x})$

- data $x_1, \ldots, x_n$ independent observations; model $f(\boldsymbol{x};\boldsymbol{\theta}) = \prod f(x_i;\boldsymbol{\theta}), \quad \boldsymbol{\theta} \in \mathbb{R}^p$
- limit theorem $\sqrt{n}\{I(\boldsymbol{\theta})\}^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \xrightarrow{d} N_p(\mathbf{0}, I_d)$
- approximation $\hat{\boldsymbol{\theta}} \dot{\sim} N_p\{\boldsymbol{\theta}, I^{-1}(\hat{\boldsymbol{\theta}})\}$, or $\hat{\boldsymbol{\theta}} \dot{\sim} N_p\{\boldsymbol{\theta}, J^{-1}(\hat{\boldsymbol{\theta}})\}$

Recap

- data $x_1, \ldots, x_n$ independent observations; model $f(\boldsymbol{x};\theta) = \prod f(x_i;\theta), \quad \theta \in \mathbb{R}$
- limit theorem $\sqrt{n}(\hat\theta - \theta) \xrightarrow{d} N(0, I_1^{-1}(\theta))$
- approximation $\hat\theta \mathrel{\dot\sim} N\{\theta, I^{-1}(\hat\theta)\}$, or $\hat\theta \mathrel{\dot\sim} N\{\theta, J^{-1}(\hat\theta)\}$ $\qquad I(\theta) = nI_1(\theta), \quad J(\theta) = -\ell''(\theta;\boldsymbol{x})$

- data $x_1, \ldots, x_n$ independent observations; model $f(\boldsymbol{x};\boldsymbol{\theta}) = \prod f(x_i;\boldsymbol{\theta}), \quad \boldsymbol{\theta} \in \mathbb{R}^p$
- limit theorem $\sqrt{n}\{I(\boldsymbol{\theta})\}^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \xrightarrow{d} N_p(\boldsymbol{0}, I_d)$
- approximation $\hat{\boldsymbol{\theta}} \mathrel{\dot\sim} N_p\{\boldsymbol{\theta}, I^{-1}(\hat{\boldsymbol{\theta}})\}$, or $\hat{\boldsymbol{\theta}} \mathrel{\dot\sim} N_p\{\boldsymbol{\theta}, J^{-1}(\hat{\boldsymbol{\theta}})\}$

- data $x_1, \ldots, x_n$ independent observations
  true model $F(\boldsymbol{x}) = \prod F(x_i), \quad \boldsymbol{\theta} \in \mathbb{R}^p$ $\qquad$ assumed model $\ell(\boldsymbol{\theta};\boldsymbol{x}), \quad \ell'(\hat{\boldsymbol{\theta}};\boldsymbol{x}) = 0$
- limit theorem $\sqrt{n}\{\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}(F)\} \xrightarrow{d} N\{\boldsymbol{0}, J^{-1}(F)I(F)J^{-1}(F)\}$ $\qquad \boldsymbol{\theta}(F), I(F), J(F)$

... Recap

- proof requires many smoothness conditions on underlying model
- i.i.d. can often be weakened to independent (not i.d.) observations,
  or even dependent — need WLLN and CLT
- MS Theorem 5.3, p.253 has a careful proof for $\theta \in \mathbb{R}$
  see also MSI, Nov 29, likelihood handout
- key step is

$$\sqrt{n}(\hat{\theta} - \theta) = \frac{-n^{-1/2}\Sigma_{i=1}^{n}\ell'(X_i;\theta)}{n^{-1}\Sigma_{i=1}^{n}\ell''(X_i;\theta) + (\hat{\theta}-\theta)(2n)^{-1}\Sigma_{i=1}^{n}\ell'''(X_i;\theta^*)}$$

... Recap

- proof requires many smoothness conditions on underlying model
- i.i.d. can often be weakened to independent (not i.d.) observations,
  or even dependent

need WLLN and CLT

- MS Theorem 5.3, p.253 has a careful proof for $\theta \in \mathbb{R}$

see also MSI, Nov 29, likelihood handout

- key step is

$$\sqrt{n}(\hat{\theta} - \theta) = \frac{-n^{-1/2}\Sigma_{i=1}^{n}\ell'(X_i;\theta)}{n^{-1}\Sigma_{i=1}^{n}\ell''(X_i;\theta) + (\hat{\theta} - \theta)(2n)^{-1}\Sigma_{i=1}^{n}\ell'''(X_i;\theta^*)}$$

- vector version is

$$\sqrt{n}\Sigma_{k=1}^{p}(\hat{\theta}_k - \theta_k)\{n^{-1}\ell''_{jk}(\hat{\boldsymbol{\theta}}) + (2n)^{-1}\Sigma_{l=1}^{p}(\hat{\theta}_l - \theta_l)\ell'''_{jkl}(\boldsymbol{\theta}^*)\} = -n^{-1/2}\ell'_j(\boldsymbol{\theta}),$$

$j = 1, \ldots, p$

Loose ends

see also MS I Nov 29

- proof of consistency (Thms 5.1,2) uses WLLN applied to

$$\phi_n(t) = \frac{1}{n}\sum_{i=1}^{n} \log \frac{f(X_i;t)}{f(X_i;\theta)}$$

$$M_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} \log \frac{f(X_i;\theta)}{f(X_i;\theta_{true})}$$

$$\phi(t) = \mathrm{E}_\theta \log \frac{f(X_i;t)}{f(X_i;\theta)} \equiv -K(f_t : f_\theta)$$

$$\mathrm{E}_{\theta_{true}} \log \frac{f(X_i;\theta)}{f(X_i;\theta_{true})} \equiv -D(\theta_{true}, \theta)$$

Loose ends

- proof of consistency (Thms 5.1,2) uses WLLN applied to

$$\phi_n(t) = \frac{1}{n}\sum_{i=1}^{n} \log \frac{f(X_i;t)}{f(X_i;\theta)} \qquad M_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} \log \frac{f(X_i;\theta)}{f(X_i;\theta_{true})}$$

$$\phi(t) = \mathrm{E}_\theta \log \frac{f(X_i;t)}{f(X_i;\theta)} \equiv -K(f_t : f_\theta) \qquad \mathrm{E}_{\theta_{true}} \log \frac{f(X_i;\theta)}{f(X_i;\theta_{true})} \equiv -D(\theta_{true}, \theta)$$

- by Jensen's $\phi(t)$ maximized at θ, which suggests $\hat{\theta} \to \theta$ — but functions are tricky
- need $\sup$ condition, see Thm 5.1 (a)

Loose ends

see also MS I Nov 29

- proof of consistency (Thms 5.1,2) uses WLLN applied to

$$\phi_n(t) = \frac{1}{n}\sum_{i=1}^{n} \log \frac{f(X_i;t)}{f(X_i;\theta)} \qquad M_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} \log \frac{f(X_i;\theta)}{f(X_i;\theta_{true})}$$

$$\phi(t) = \mathrm{E}_\theta \log \frac{f(X_i;t)}{f(X_i;\theta)} \equiv -K(f_t : f_\theta) \qquad \mathrm{E}_{\theta_{true}} \log \frac{f(X_i;\theta)}{f(X_i;\theta_{true})} \equiv -D(\theta_{true}, \theta)$$

- by Jensen's $\phi(t)$ maximized at θ, which suggests $\hat{\theta} \to \theta$ — but functions are tricky
- need sup condition, see Thm 5.1 (a)

- $K(f : f_0)$ Kullback-Leibler divergence measures 'closeness' of densities f and f_0

  $E_0\{\log f_0(X)/f(X)\}$

- maximum likelihood estimator minimizes K-L divergence between empirical cdf and model

  $E_{F_n} \log\{dF_n(\boldsymbol{x})/f_\theta(\boldsymbol{x})\}$

Nonparametric MLE

- sample $x_1, \ldots, x_n$ independent, identically distributed, with cdf F

no parametric model assumed

- likelihood function $L(F) = \prod f(x_i)$

- assume solution puts mass only at $x_1, \ldots, x_n$
- log-likelihood function $\ell(p) = \sum_{i=1}^{n} \log(p_i)$

Nonparametric MLE

- sample $x_1, \dots, x_n$ independent, identically distributed, with cdf F

no parametric model assumed

- likelihood function $L(F) = \prod f(x_i)$

- assume solution puts mass only at $x_1, \dots, x_n$
- log-likelihood function $\ell(p) = \sum_{i=1}^{n} \log(p_i)$

- maximized at $p_i = 1/n, i = 1, \dots, n$

Lagrange

- gives empirical cdf

$$\hat{F}_n(x) = \frac{1}{n} \sum_{i=1}^{n} 1(X_i \leq x)$$

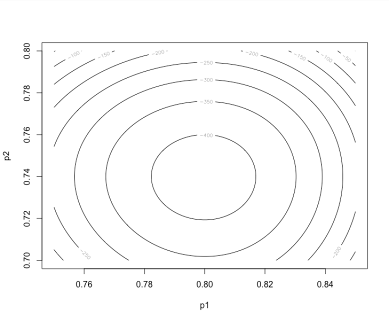
Multi-parameter example: logistic regression

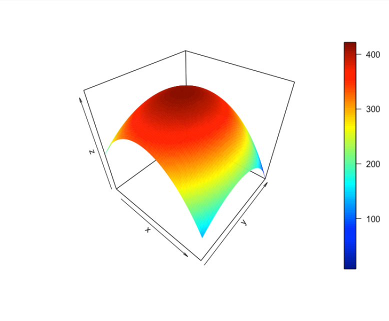
```
Boston$crim2 <- Boston$crim > median(Boston$crim) # define binary response
Boston.glm <- glm(crim2 ~ . - crim, family = binomial,
                  data = Boston) #fit logistic regression
summary(Boston.glm)

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -34.103704   6.530014  -5.223 1.76e-07 ***
zn           -0.079918   0.033731  -2.369  0.01782 *
indus        -0.059389   0.043722  -1.358  0.17436
chas          0.785327   0.728930   1.077  0.28132
nox          48.523782   7.396497   6.560 5.37e-11 ***
rm           -0.425596   0.701104  -0.607  0.54383
age           0.022172   0.012221   1.814  0.06963 .
```

```
Boston.glm <- glm(crim2 ~ . - crim, family = binomial,
                  data = Boston) #fit logistic regression
confint(Boston.glm)
Waiting for profiling to be done...
                    2.5 %       97.5 %
(Intercept) -47.480389822 -21.699753794
zn           -0.152359922  -0.020567540
indus        -0.149113408   0.024168460
chas         -0.646429219   2.233443233
nox          34.967619055  64.088411260
rm           -1.811639107   0.950196261
age          -0.001231256   0.046865843
dis           0.280762523   1.140619391
rad           0.376833861   0.975898274
tax          -0.012038221  -0.001324887
```

Multi-parameter setting





$Y_1 \sim Binom(n_1, p_1)$, $Y_2 \sim Binom(n_2, p_2)$, independently

observed values $y_1 = 160, n_1 = 200, y_2 = 180, n_2 = 200$

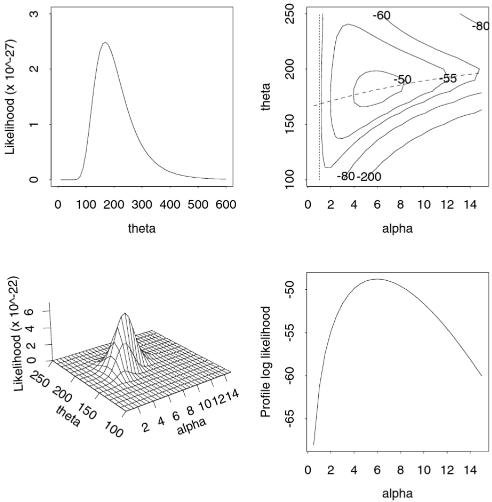
… Multi-parameter setting

Profile likelihood function

... Profile likelihood function

... Profile likelihood function

Figure 4.1 Likelihoods for the spring failure data at stress 950 N/mm^2. The upper left panel is the likelihood for the exponential model, and below it is a perspective plot of the likelihood for the Weibull model. The upper right panel shows contours of the log likelihood for the Weibull model; the exponential likelihood is obtained by setting $\alpha = 1$. that is, slicing L along the vertical dotted line. The lower right panel shows the profile log likelihood for α, which corresponds to the log likelihood values along the dashed line in the panel above, plotted against α.



model

prior

posterior

sample

Frequentist and Bayesian contrast

Frequentist:

- There is a fixed parameter (unknown) we are trying to learn
- Our methods are evaluated using probabilities based on $f(x; \theta)$

Bayesian:

- The parameter can be treated as a random variable
- We model its distribution $\pi(\theta)$
- Combine this with a model $f(x \mid \theta)$
- Update prior belief on the basis of the data

Example: Binomial

$X_1, \ldots, X_n$ i.i.d. Bernoulli (θ)

$$\pi(\theta; \alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha-1}(1-\theta)^{\beta-1}, 0 < \theta < 1$$

posterior mean, mode

Example: censored exponential

$X_1, \ldots, X_n$ i.i.d. Exponential (λ) $\qquad \pi(\lambda) \sim \text{Exp}(\alpha)$

censored at r smallest x; let $Y_i = X_{(i)}, i = 1, \ldots, r$

$$f(\boldsymbol{y} \mid \lambda) = \prod_{i=1}^{r} \lambda^r \exp(-\lambda y_i) \prod_{i=r+1}^{n} \exp(-\lambda y_r) = \lambda^r \exp\{-\lambda \Sigma_{i=1}^r y_i + (n-r) y_r\}$$

Exponential families and conjugate priors

$$f(x;\theta) = \exp\{c(\theta)T(x) - d(\theta) + S(x)\}; \qquad \pi(\theta;\alpha,\beta) = K(\alpha,\beta)\exp\{\alpha c(\theta) - \beta d(\theta)\}$$

Exponential families and conjugate priors

$f(x;\theta) = \exp\{c(\theta)T(x) - d(\theta) + S(x)\};$ $\qquad \pi(\theta;\alpha,\beta) = K(\alpha,\beta)\exp\{\alpha c(\theta) - \beta d(\theta)\}$

Example: $f(x;\theta) = \theta(1-\theta)^x, x = 0, 1, \ldots; 0 < \theta < 1$

Exponential families and conjugate priors

$$f(x;\theta) = \exp\{c(\theta)T(x) - d(\theta) + S(x)\}; \qquad \pi(\theta;\alpha,\beta) = K(\alpha,\beta)\exp\{\alpha c(\theta) - \beta d(\theta)\}$$

Example: $f(x;\theta) = \theta(1-\theta)^x, x = 0, 1, \ldots; 0 < \theta < 1$

Example: $f(x;\mu) = \frac{1}{\sqrt{2\pi}}\exp\{-\frac{1}{2}(x-\mu)^2\}$

Example: Bivariate normal

Table 3.1 *Scores from two tests taken by 22 students,* `mechanics` *and* `vectors`.

	1	2	3	4	5	6	7	8	9	10	11
`mechanics`	7	44	49	59	34	46	0	32	49	52	44
`vectors`	51	69	41	70	42	40	40	45	57	64	61

	12	13	14	15	16	17	18	19	20	21	22
`mechanics`	36	42	5	22	18	41	48	31	42	46	63
`vectors`	59	60	30	58	51	63	38	42	69	49	63

Table 3.1 shows the scores on two tests, `mechanics` and `vectors`, achieved by $n = 22$ students. The sample correlation coefficient between the two scores is $\hat{\theta} = 0.498$,

$$\hat{\theta} = \sum_{i=1}^{22}(m_i - \bar{m})(v_i - \bar{v}) \Bigg/ \left[\sum_{i=1}^{22}(m_i - \bar{m})^2 \sum_{i=1}^{22}(v_i - \bar{v})^2\right]^{1/2}, \quad (3.10)$$

with m and v short for `mechanics` and `vectors`, $\bar{m}$ and $\bar{v}$ their averages. We wish to assign a Bayesian measure of posterior accuracy to the true correlation coefficient θ, "true" meaning the correlation for the hypothetical population of all students, of which we observed only 22.

If we assume that the joint (m, v) distribution is bivariate normal (as

$$f(\hat{\theta} \mid \theta) = \frac{1}{\pi}(n-2)(1-\theta^2)^{(n-1)/2}(1-\hat{\theta}^2)^{(n-4)/2} \int_0^\infty \frac{1}{\cosh(w) - \theta\hat{\theta}} dw$$

Example: Bivariate normal

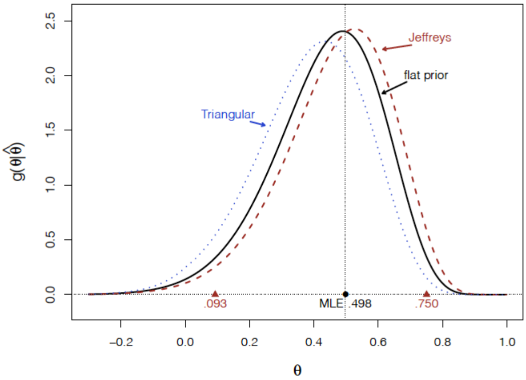


Figure 3.2 Student scores data; posterior density of correlation θ for three possible priors.

Example: Binomial

SM Ex.11.11

11.2 · Inference 579

Table 11.2 Mortality rates r/m from cardiac surgery in 12 hospitals (Spiegelhalter *et al.*, 1996b, p. 15). Shown are the numbers of deaths r out of m operations.

A	0/47	B	18/148	C	8/119	D	46/810	E	8/211	F	13/196
G	9/148	H	31/215	I	14/207	J	8/97	K	29/256	L	24/360

provided the mode lies inside the parameter space. Here $\tilde{J}(\theta)$ is the second deriva-
tive matrix of $-\tilde{\ell}(\theta)$. This expansion corresponds to a posterior multivariate normal

prior for hospital A $Beta(1, 1)$

posterior mean

580 *11 · Bayesian Models*

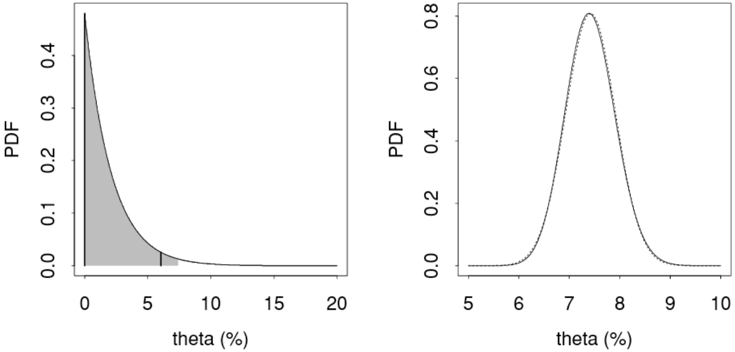


Figure 11.1 Cardiac surgery data. Left panel: posterior density for θ_A, showing boundaries of 0.95 highest posterior credible interval (vertical lines) and region between posterior 0.025 and 0.975 quantiles of $\pi(\theta_A \mid y)$ (shaded). Right panel: exact posterior beta density for overall mortality rate θ (solid) and normal approximation (dots).

put all hospitals together; 208 failures ‘

“Flat” priors

Marginalization

Not all likelihood functions are regular

Example: $X_1, \ldots, X_n$ i.i.d. $U(0, \theta)$

… Not all likelihood functions are regular

MS Exercise 5.1

$X_1, \ldots, X_n$ i.i.d. $f(x; \boldsymbol{\theta}) = a(\theta_1, \theta_2)h(x), \quad \theta_1 \leq x \leq \theta_2$