Review 1

Focus on likelihood based inference

$$L(\theta|x^n) = \prod_{i=1}^{n} p_\theta(x_i) \quad \left[\text{can generalize to } p_\theta(x_i|g(x_i))\right]$$

: function over θ for fixed observation x^n

Shape of $L(\theta|x^n)$ captures information about θ in this particular x^n and uncertainty about θ in this x^n

$\Rightarrow$ can try to convert to direct probabilistic statements or large-sample approximations

Compare — frequentist methods

- often a loss $L(\hat{\theta}(x^n) - \theta)$ 'averaged' out over repeated sampling from $F_\theta \Rightarrow \underbrace{E_{F_\theta}[L(\hat{\theta}(x^n), \theta)]}_{\text{Risk}}$
- discuss / base inference on long-run properties.
- 1% loss from θ $1-2\%$ of F_θ-samples
- Bayesian methods
- pair $\pi(\theta) \Rightarrow$ posterior $\pi(\theta|x^n) \propto \pi(\theta) L(\theta|x^n)$
- now get probabilistic statements : $CI(x^n)$ covers θ w. prob $1-\alpha$.

- Big + w likelihood based inference $\rightarrow$ invariant to reparameterizations $\eta = g(\theta) \Rightarrow \hat{\eta} = g(\hat{\theta})$
- Also, to compare $L(\theta|x^n)$ for different values of $\theta \Rightarrow$ LR is invariant to reparameterizations and variable transformations.

$L(\theta|x^n)$ can have a 'complex' shape with multiple local maxima, skewness etc.

However, we are only really interested in θ's near maximum of $L(\theta)$, or $l(\theta)$ itself near maximum since other θ's not supported by data.

Focus on shape of $L(\theta|x^n)$ near $\hat{\theta}_n = \underset{\theta}{\operatorname{argmax}}\, L(\theta|x^n)$

* If $\log L(\theta|x^n) = l(\theta|x^n)$ is near quadratic near $\hat{\theta}$ we call $l(\theta|x^n)$ regular

  → lots of inference used in practice based on this assumption!

Really check — plot $l(\theta|x^n)$ and its quadratic approximation and compare.

* Big t as $n \to \infty$ most $l(\theta|x^n)$ look near quadratic near $\hat{\theta}_n$

. . . . . .

The concepts we have used for derivation of tests etc.

⇒ Score, information

- $l(\theta|x^n)$ = log-likelihood = $l(\theta)$ (but remember dependence on x^n)
- $S(\theta) = \frac{\partial}{\partial \theta} l(\theta|x^n)$ = score
- Definition: $\hat{\theta}_n = \hat{\theta}_{MLE} = \underset{\theta}{\operatorname{argmax}}\, l(\theta|x^n) = \{\theta : S(\theta) = 0\}$

- (curvature of) $\ell(\theta|x^n)$

$$J(\theta) = -\frac{\partial^2}{\partial\theta^2}\ell(\theta)$$

- $I(\hat{\theta})$ = curvature at $\hat{\theta}$ = positive quantity by def of $\hat{\theta}$

= "observed" Fisher information

(compare with expected Fisher information

$$E_\theta\left[-\frac{\partial^2}{\partial\theta^2}\ell(\theta|x)\right]$$

- Quadratic approximation near $\hat{\theta}$

$$\ell(\theta) \approx \ell(\hat{\theta}) + \underbrace{S(\hat{\theta})(\theta-\hat{\theta})}_{=0} - \frac{1}{2}I(\hat{\theta})(\theta-\hat{\theta})^2$$

$$\Rightarrow \boxed{\log\frac{L(\theta)}{L(\hat{\theta})} \approx -\frac{1}{2}I(\hat{\theta})(\theta-\hat{\theta})^2}$$

normalized log likelihood — quadratic approximation

- Derivative of LH $\Rightarrow$ $\boxed{S(\theta) \approx I(\hat{\theta})(\hat{\theta}-\theta)}$

This is important

$\hat{\theta}$ inherits all its properties from $S(\theta)$!

So if we understand $S(\theta)$ behaviour

$\Rightarrow$ understand $\hat{\theta}$.

Looking at the score in more detail

- $S(\theta) = \frac{\partial}{\partial\theta} \log L(\theta)$, $S(\hat{\theta}) = 0$ by definition of $\hat{\theta}$

- Take expectation of $S(\theta)$ under the model p_θ

$$E_\theta[S(\theta)] = E_\theta\left[\frac{\partial}{\partial\theta} \log L(\theta|X)\right]$$ (can focus on one X since $S(\theta)$ is additive on X_i iid p_θ)

$$= \int \frac{\partial}{\partial\theta} \log L(\theta|x)\, p_\theta(x)\, dx$$

$$= \int \frac{\frac{\partial}{\partial\theta} L(\theta|x)}{L(\theta|x)} p_\theta(x)\, dx \qquad \int \frac{\partial}{\partial\theta} p_\theta(x)\, dx = \frac{\partial}{\partial\theta} \int p_\theta(x)\, dx = 0$$

[for nice p_θ where $\frac{\partial}{\partial\theta}$ and $\int$ can change order]

- $\forall \theta \quad S(\theta)$ $E_\theta[S(\theta)^2] = \int \left(\frac{\partial \log L(\theta)}{\partial\theta}\right)^2 p_\theta(x)\, dx$

$$= \int \left(\frac{\partial \log L(\theta|x)}{\partial\theta}\right) \frac{\frac{\partial}{\partial\theta} L(\theta|x)}{L(\theta|x)} p_\theta(x)\, dx = \int \left(\frac{\partial \log L(\theta|x)}{\partial\theta}\right)\left(\frac{\partial}{\partial\theta} L(\theta|x)\right) dx$$

integration by parts $= \frac{\partial \log L(\theta)}{\partial\theta}\left(\int \frac{\partial}{\partial\theta} p_\theta(x)\, dx\right) - \int \frac{\partial^2 \log L(\theta|x)}{\partial\theta^2} p_\theta(x)\, dx = -E_\theta\left[\frac{\partial^2}{\partial\theta^2} \log L(\theta|X)\right]$

$= 0$

$= I(\theta)$

Expected Fisher information

Now, we have

$$S(\theta) \approx I(\hat{\theta})(\hat{\theta}-\theta) \qquad \text{(quadratic approx of } l(\theta)\text{)}$$

$$\rightarrow \underbrace{I(\hat{\theta})^{-1/2} S(\theta)}_{\text{sampling distribution of this?}} \approx I(\hat{\theta})^{1/2}(\hat{\theta}-\theta)$$

- $S(\theta) = \sum_i \frac{\partial}{\partial\theta} \log p_\theta(x_i) = \sum_i y_i$ where $y_i = \frac{\partial}{\partial\theta}(\log p_\theta(x_i))$

- the y_i's are i.i.d
- we know $E_\theta(y_i) = 0$ and $Var_\theta(y_i) = FI(\theta) = E_\theta\left[-\frac{\partial^2}{\partial\theta^2}(\log L(\theta|X))\right]$

$\Rightarrow E_\theta(S(\theta)) = 0$, $Var_\theta(S(\theta)) = n\,FI(\theta)$

- (central limit theorem applied to $\sum y_i = S(\theta)$

$$\Rightarrow \frac{1}{\sqrt{n}} S(\theta) \rightarrow N(0, FI(\theta))$$

or $$\boxed{(n\,FI(\theta))^{-1/2} S(\theta) \rightarrow N(0,1)}$$

but we have $(I(\hat{\theta}))^{-1/2} S(\theta) \approx I(\hat{\theta})^{1/2}(\hat{\theta}-\theta)$

but $FI(\theta) = E_\theta\left[-\frac{\partial^2}{\partial\theta^2}(\log L(\theta|X))\right]$ and $I(\hat{\theta}) = -\frac{\partial^2}{\partial\theta^2}\log L(\theta|x^n)\Big|_{\hat{\theta}}$

(can we just replace $I(\hat{\theta})$ by $FI(\theta)$?

WLLN $\frac{1}{n}\sum_{i=1}^{n} \frac{\partial^2}{\partial\theta^2} \log f_\theta(x_i) \xrightarrow{p} E_\theta\left[\frac{\partial^2}{\partial\theta^2}\log f_\theta(X)\right]$

$\Rightarrow \frac{I(\theta)}{n} \xrightarrow{p} \mathcal{I}(\theta) \quad \forall\theta$

Slutsky's theorem If $X_n \xrightarrow{d} X$ and $A_n \xrightarrow{p} a$

then $A_n X_n \xrightarrow{d} aX$

Here $I(\hat\theta)^{-1/2} S(\theta) = \underbrace{(n\mathcal{I}(\theta))^{-1/2} S(\theta)}_{\xrightarrow{d} N(0,1)} \underbrace{\left(\frac{I(\hat\theta)}{n}\right)^{-1/2} (\mathcal{I}(\theta))^{1/2}}_{\xrightarrow{p} 1}$

$\Rightarrow$ by $\sqrt{n}(\hat\theta - \theta) \xrightarrow{} N(0, \mathcal{I}(\theta)^{-1})$

$\Leftrightarrow$

$(\hat\theta - \theta) \xrightarrow{} N(0, \hat{\mathcal{I}}(\hat\theta)^{-1})$ ← used in practice since $\hat{\mathcal{I}}(\hat\theta)$ computable from data

Inference $CI = \left[\hat\theta \pm z_{1-\alpha/2}\sqrt{I(\hat\theta)^{-1}}\right]$

Other types of interval estimates used in likelihood

$\{\theta : \frac{L(\theta)}{L(\hat\theta)} > c\}$ i.e. θ's within reasonable likelihood range from the maximum $L(\hat\theta)$

What should c be?

Tricky! Data dependent $L(\theta)$ and dimension dependency makes 'standard' c difficult to find.

From quadratic approximation $\hat\theta \sim AN \Rightarrow \hat\theta^2 \sim \chi^2$

$\Rightarrow W = 2\log\frac{L(\hat\theta)}{L(\theta)} \sim \chi^2_1$, under null $H_0: \theta$

(This is an approximation based on normal model)

$\Rightarrow \{\theta : 2\log\frac{L(\hat\theta)}{L(\theta)} < -2\log c\}$ $\Rightarrow$ c should be $e^{-\frac{1}{2}\chi^2_1(1-\alpha)}$ for $1-\alpha$ coverage

Both $CI = \{\theta : 2\log\frac{L(\hat\theta)}{L(\theta)} > \chi^2_1(1-\alpha)\}$ (Wilks)

and

$CI : \{\hat\theta \pm z_{1-\alpha/2} I(\hat\theta)^{-\frac{1}{2}}\}$

are based on quadratic approximation (Wald)

In practice, Wilks works much better — allows for asymmetric CI

Multiparameter case

$$\underset{1\times 1}{\log L(\theta)} = \underset{1\times 1}{\log L(\hat\theta)} + \underset{1\times p}{S(\hat\theta)'} \underset{p\times 1}{(\theta-\hat\theta)} - \frac{1}{2}\underbrace{\underbrace{(\theta-\hat\theta)'}_{1\times p}\underbrace{I(\hat\theta)}_{p\times p}\underbrace{(\theta-\hat\theta)}_{p\times 1}}_{1\times 1}$$

$\theta = (\theta_1, \ldots, \theta_p)$

$p\times p$ matrix

$$I(\hat\theta)_{ij} = \frac{-\partial^2}{\partial\theta_i \partial\theta_j} \log L(\theta|x^n)\Big|_{\hat\theta}$$

Solve $\{\theta : S(\theta) = 0\}$ now a multiparameter problem — can be difficult

- Let's say we found solutions $\hat\theta$

  — inference?

Wilks $W = 2\log\frac{L(\hat\theta)}{L(\theta)} \sim \chi^2_{(p)}$ now (approx)

$$\Rightarrow C_1 = \left\{\theta : \frac{L(\theta)}{L(\hat\theta)} > \exp\left(-\tfrac{1}{2}\chi^2_p(1-\alpha)\right)\right\}$$

but we rarely construct confidence regions like this — more often look at marginal sampling distribution for one $\hat\theta_j$ at a time

Testing an hypothesis like: $\underline{\theta} = \underline{\theta}_0$ is done using Wilks theorem (see GLM later)

Exponential families, UMVUE, MLE

Data x^n → reduce to a summary statistic $T(x^n)$

→ often of a form $\bar{x}$ or $\bar{x}^2$ etc

$T(x^n)$ is <u>sufficient</u> for θ if the distribution $p_\theta(x^n)$ factorizes as $p_\theta(x^n) = g(T(x^n), \theta) h(x^n)$

- $g(T(x^n), \theta)$: depends on x^n only through $T(x^n)$
- $h(x^n)$: does not depend on θ

Exponential families

$$p_\theta(x) = \exp\{\eta(\theta) T(x) - A(\theta) + c(x)\}$$

$$p_\theta(x^n) = \exp\{\eta(\theta) T(x^n) - A_n(\theta) + c(x^n)\}$$

where $T(x^n) = \sum_{i=1}^n T(x_i)$, $A_n(\theta) = nA(\theta)$

this is a nice property — easy to work with

If $\eta(\theta) = \theta$ — canonical form

$$p_\theta(x^n) = \exp\{\theta T(x^n) - nA(\theta) + c(x^n)\}$$

(can show that $E(T(X)) = A'(\theta)$, $V(T(X)) = A''(\theta)$) → mean-variance relationship

Multiparameter model $p_{\underset{\sim}{\theta}}(x) = \exp\{\sum_j \eta_j(\underset{\sim}{\theta}) T_j(x) - A(\underset{\sim}{\theta}) + c(x)\}$

{ functions of $\underset{\sim}{\theta}$

p sufficient statistics T_j

For exponential families

→ $l(\theta|x) = \eta(\theta)T(x) - A(\theta) + c(x)$

$S(\theta) = \eta'(\theta)T(x) - A'(\theta)$ [$S(\theta) = T(x) - A'(\theta)$ for canonical]

→ MLE $\Rightarrow \{\theta : \eta'(\theta)T(x) = A'(\theta)\}$

→ Curvature $I(\theta) = -\eta''(\theta)T(x) + A''(\theta)$ [$I(\theta) = A''(\theta)$ canonical]

For exponential families then

$\underset{\theta}{E}(S(\theta)) = \eta'(\theta)\underline{E_\theta T(x)} - A'(\theta) = 0$

$\mathcal{I}(\theta) = A''(\theta) - \eta''(\theta)\underline{E_\theta T(x)}$

* At $\hat{\theta}$: $I(\hat{\theta}) = \mathcal{I}(\hat{\theta})$

• For canonical model : $I(\theta) = \mathcal{I}(\theta)$ for all θ!

So, observed information = expected information for canonical families so uncertainty about θ in this sample ~ uncertainty of long-run frequency quantity

---

Cramér-Rao lower bound

For 'nice' distributions (like exponential pd) and 'nice' estimators $T(x) = \hat{\theta}(x)$ ($E_\theta|T(x)| < \infty$ etc)

If $T(X)$ is an unbiased estimator of $g(\theta)$ with finite variance

$$\boxed{Var_\theta T(X) \geq g'(\theta)' FI(\theta)^{-1} g'(\theta)}$$

The proof was done by using conditional variance $\Rightarrow$ reduction in variance

$\Rightarrow$ bound can only be met by $T(x)$ that are linearly related to the score $S(\theta)$

$\Rightarrow$ exponential family

$\Rightarrow$ (loss of) UMVUEs $\Rightarrow$ alternative to MLE

(compare in - bias

General estimation

- parameter of interest $g(\theta)$
- estimator $T(X^n)$

Is $T(X^n)$ any good? What is "good"

bias: $E_\theta T(X^n) - g(\theta)$

variance: $V_\theta T(X^n)$

$MSE: E_\theta (T(X^n) - g(\theta))^2$ (common combo Criterion)

$MSE(T, \theta)$

usually depends on true value of θ

so there is no single best estimator for all θ $\Rightarrow$ but some restrictions on form of $T(X^n)$

(common restriction $\Rightarrow$ Unbiasedness $\rightarrow$ then look for minimum variance

$\Rightarrow$ worst-case (over θ) $\rightarrow$ minimax $\min_T \max_\theta MSE(T, \theta)$

Y stay with unbiased

→ Rao-Blackwell

If $T(X^n)$ sufficient for θ and $S(X^n)$ is an estimator of $g(\theta)$,

the new estimator $E(S(X^n)|T(X^n)) = T^*(X^n)$

has property

$$\forall \theta \quad MSE(T^*, \theta) \leq MSE(S, \theta)$$

→ Lehmann-Scheffé

If $T(X^n)$ is a complete sufficient stat for θ, and

$S(X^n)$ is an unbiased estimator of $g(\theta)$,

then $T^*(X^n)$ is an UMVUE of $g(\theta)$

(Complete: $E_\theta h(T) = 0 \ \forall \theta$ implies $h(T) = 0$)

Construction of UMVUE

If have complete sufficient T and any unbiased S

$\Rightarrow T^* = E(S|T)$ ✓

That part is saying T is complete — but usually case for exponential families!

For exponential families

→ (can under one parametrization find a $T^*(x^n)$

that is UMVUE and meets the CR bound!

$$\left(p_\theta(x) = \exp\{\eta(\theta)T^*(x) - A(\theta) + c(x)\} \right)$$

→ doesn't necc work for all parametrizations since

we may have $\hat{\theta}$ unbiased for θ but $g(\hat{\theta})$ not

unbiased for $g(\theta)$

MLE v. UMVUE

- MLE - easy to work with
- UMVUE - requires sometimes work to find

- MLE - invariant to parameterizations
- UMVUE - may find unbiased $\hat{\theta}$ but not $g(\hat{\theta})$

- For large n, MLE and UMVUE tend to agree

- for exponential families, both MLE & UMVUEs tend to be functions of the sufficient statistics $T(x)$