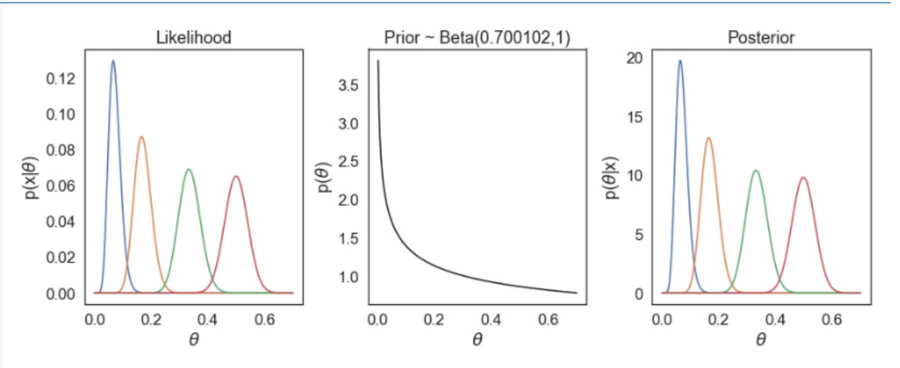
Mathematical Statistics II

STA2212H S LEC9101



Week 8

March 10 2021

Start recording!

More on vaccines

Boyang Zhao's blog post

$1 - \frac{8 = n_v}{162 = n_c}$

link

Boyang Zhao

Senior data scientist at ING Bank
Computational biology, genomics, systems biology, finance, banking
@MIT PhD

Netherlands
Twitter
LinkedIn
GitHub

Vaccine efficacy

The way we measure vaccine efficacy is defined as follows,

$$\text{vaccine efficacy (VE)} = 1 - R$$

The R can be ratio of risks (RR, risk ratio); rates (IRR, incidence rate ratio); or hazards (HR, hazard ratio). Because of the ratios, we see that vaccine efficacy is a relative measure - in how much relative reduction in infection or disease in the vaccinated group compared to the unvaccinated group. A VE of 90% means there are 90% fewer cases in the vaccinated group compared to the placebo group. We will use the subscripts $_v$ and $_p$ to denote vaccinated and placebo groups, respectively; but obviously the discussion is applicable for comparing between any two treatment arms - not necessarily have to be a placebo.

With RR

$$\mathrm{VE} = 1 - \mathrm{RR} = 1 - \frac{c_v/N_v}{c_p/N_p}$$

where N_v and N_p are the total number of participants in the vaccinated and placebo group, respectively.

With IRR

$$\mathrm{VE} = 1 - \mathrm{IRR} = 1 - \frac{c_v/T_v}{c_p/T_p}$$

where T_v and T_p are the time-person years for the vaccinated and placebo group, respectively.

With HR

$$\mathrm{VE} = 1 - \mathrm{HR} = 1 - \frac{\lambda_v}{\lambda_p}$$

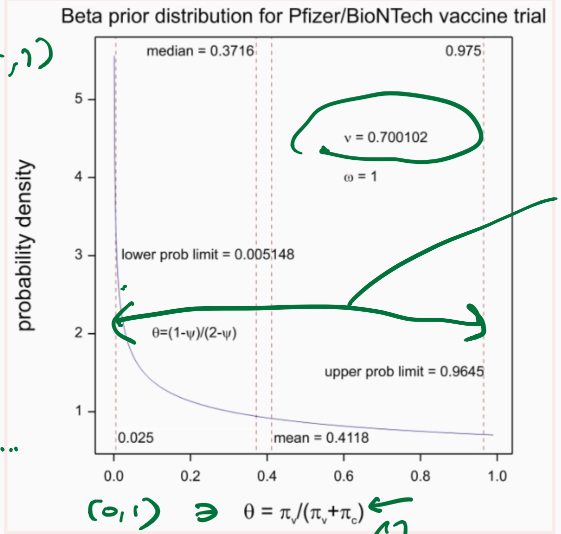
where λ_v and λ_p are the hazard rates for the vaccinated and placebo group, respectively. This measures the relative reduction in the hazard of infection. The hazard ratio can be

More on vaccines

Stephen Senn's blog post

link

$Be(.7, 1)$



.70000...

$(0,1) \ni \theta = \pi_v/(\pi_v+\pi_c)$

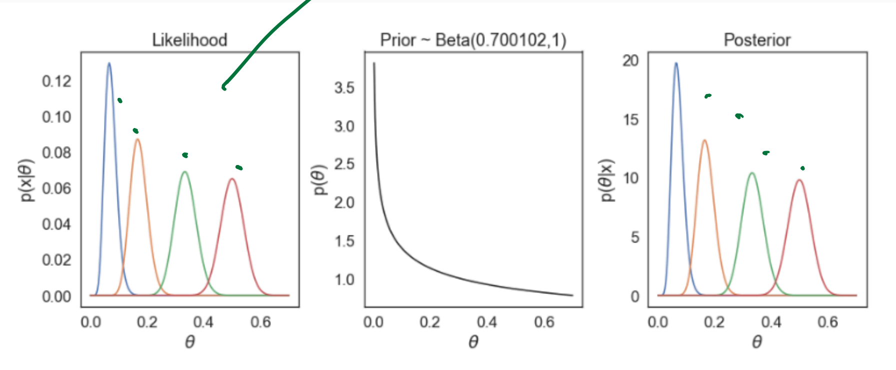
Figure 4 Prior distribution for θ

95% prior

fake data

$n_x = 10, 25, 50, 75$

$n_v + n_c = 150$



$$\theta = \frac{\pi_v}{\pi_v + \pi_c} \Rightarrow \pi(VE)$$

$\theta \in (0,1)$

$VE \in (-\infty, 1)$

$$\boxed{\theta = (1 - VE)/(2 - VE)}$$

POLLING

US election polls: a quick postmortem

How did the 2020 US presidential election polls really do? **Ole J. Forsberg** gives his assessment

The American Association for Public Opinion Research (AAPOR) is expected to produce a report early this year that explores the strengths and weaknesses of the polls in the 2020 US election cycle. The polls were criticised in some quarters immediately after the election, when it became clear that Donald Trump had done better than expected and that Joseph R. Biden Jr's margin of victory in the popular vote was not as large as anticipated.[1]

In preparation for this report, I wanted to provide some insight into the polls and some suggestions of my own for moving forward. Specifically, I hope to convince polling houses to use some type of model averaging – or even Bayesian methods – to closing weeks of the campaign.

The first source of error, faulty weighting, is extremely important for polling houses to take seriously. While the number of US polling houses taking education level into consideration increased in 2020, the education characteristics of the voting population remain uncertain.

"Shy voters" – the second source of error – may be more myth than reality (53eig.ht/30NEb6R). But whether shy or not, there are some voters who either choose not to respond to polls, or who choose not to answer honestly when surveyed. Pollsters need to address this, either by asking additional questions to model respondent preference for those who choose not to say how they will vote, or by finding new ways to encourage the public to interpretation, not of polling. The mistake happens in how we interpret a poll result such as "48% Biden, 44% Trump". Do we focus on the two-party vote and claim that Biden is ahead, or do we acknowledge that there is a sizeable portion of voters – 8% – who may only decide how to vote once in the polling booth? Clearly, the latter interpretation is more appropriate, but it makes for a less straightforward story, so these undecided voters tend to be overlooked in media reports.

Missing data

The majority of polls in the 2020 election cycle contained just three response options for those asked about their intended vote: "Biden", "Trump", and "undecided". The implied fourth mentioned earlier constitute a huge amount of missing data about voting intention. Ignoring these missing data leads to false precision in the polls' assessment of the state of the election.

While some undecided voters ultimately will not vote, many will eventually decide between the two candidates. This increases the uncertainty in polling estimates beyond what is reported in terms of confidence intervals and margins of error. As a result, when those late-deciding voters finally vote, polls may look very wrong.

To illustrate this point, compare the polls in the final two weeks of the 2020 election to the final election result (Table 1). In this sample of 174 polls, the actual Biden vote was within the polls' margins of error 85% of the time, while the actual Trump vote was within the polls' margins of error only 43% of the time. For the 57% of confidence intervals that missed Trump's actual vote, they were always too low, never too high – meaning that the polls consistently underestimated Trump's final vote. The 15% of confidence intervals

Polling analysis

Table 1: Results from comparing candidate support levels in polls from the last two weeks of the US presidential election with the actual outcome of the election (vote share). Polls are a mix of state-level and national polls from a variety of polling houses, using a variety of methods.

		Confidence interval hits		Average miss ~~(standard error)~~	
Source	***n***	**Biden**	**Trump**	**Biden**	**Trump**
All polls	174	85% (79% to 90%)	43% (35% to 50%)	−0.09	+2.41
Online only	23	96% (78% to 99%)	30% (13% to 53%)	−0.79	+2.21
Online + telephone	26	92% (75% to 99%)	54% (33% to 73%)	−0.78	+2.24
Telephone only	125	82% (74% to 88%)	42% (34% to 52%)	−0.18	+2.48
University	60	92% (82% to 97%)	27% (16% to 40%)	−0.10	+2.99
Non-university	114	82% (73% to 88%)	51% (41% to 60%)	−0.09	+2.10
Partisan	52	79% (65% to 89%)	75% (61% to 86%)	+0.62	+1.33
Non-partisan	122	88% (81% to 93%)	29% (21% to 38%)	−0.40	+2.87

"Personally, I favour the Bayesian solution because it provides a solid statistical structure for estimation and communication of results." $\pm \sqrt{\tilde{p}(1-\tilde{p})} / \sqrt{n} \times 2$ link

$\sqrt{\frac{1}{2} \times \frac{1}{2} \times n^{-1}} \times 2 \doteq \sqrt{n}^{-1}$ $\quad 1/\sqrt{n} \quad \tilde{p} = \frac{1}{2}$

Latest issue of Applied Statistics (JRSS C)

Journal of the Royal Statistical Society: Series C (Applied Statistics)
Volume 70, Issue 2

Pages: 249-506

March 2021

ISSUE INFORMATION

Free Access

Issue Information

Pages: 249-250 | First Published: 08 March 2021

ORIGINAL ARTICLES

Finite mixtures of semiparametric Bayesian survival kernel machine regressions: Application to breast cancer gene pathway subgroup analysis

Lin Zhang, Inyoung Kim
Pages: 251-269 | First Published: 01 December 2020

Quantile-frequency analysis and spectral measures for diagnostic checks of time series with nonlinear dynamics

Ta-Hsin Li
Pages: 270-290 | First Published: 22 November 2020

Latest issue of Applied Statistics (JRSS C)

Journal of the Royal Statistical Society: Series C (Applied Statistics)
Volume 70, Issue 2

Pages: 249-506

March 2021

ISSUE INFORMATION

Free Access

Issue Information

Pages: 249-250 I First Published: 08 March 2021

ORIGINAL ARTICLES

Finite mixtures of semiparametric Bayesian survival kernel Application to breast cancer gene pathway subgroup analy

Lin Zhang, Inyoung Kim
Pages: 251-269 I First Published: 01 December 2020

Quantile-frequency analysis and spectral measures for dia with nonlinear dynamics

Ta-Hsin Li
Pages: 270-290 I First Published: 22 November 2020

Open Access

Future proofing a building design using history matching inspired level-set techniques

Evan Baker, Peter Challenor, Matt Eames
Pages: 335-350 I First Published: 19 December 2020

Recurrent events modelling of haemophilia bleeding events

Andrew C. Titman, Martin J. Wolfsegger, Thomas F. Jaki
Pages: 351-371 I First Published: 07 January 2021

Multiscale null hypothesis testing for network-valued data: Analysis of brain networks of patients with autism

Ilenia Lovato, Alessia Pini, Aymeric Stamm, Maxime Taquet, Simone Vantini
Pages: 372-397 I First Published: 22 January 2021

Open Access

Bayesian semi-parametric G-computation for causal inference in a cohort study with MNAR dropout and death

Maria Josefsson, Michael J. Daniels

Annals of Applied Statistics 2019

STATISTICS

AN OFFICIAL JOURNAL OF THE
INSTITUTE OF MATHEMATICAL STATISTICS

Articles

9/24 B. 14/24 collab.

Vol. 13, No. 3—September 2019

of

APPLIED STATISTICS

AN OFFICIAL JOURNAL OF THE
INSTITUTE OF MATHEMATICAL STATISTICS

Articles—*Continued from front cover*

- approximate posterior normality

  if n not large, need better → $\tilde{\theta}_B \pm z_{\alpha/2}\, \tilde{se}_B$ $\sim$ $\hat{\theta} \pm z_{\alpha/2}\, \hat{se}$ if n is large

- choosing a prior: subjective, conjugate, flat, convenience
- matching priors; Jeffreys' prior $\quad \pi_J(\theta) \propto |i(\theta)|^{1/2} \quad i(\theta) = E\{-\frac{\partial^2 \ell(\theta)}{\partial \theta}\}$

  $O(\frac{1}{n})$ (not m.)

- multiple parameters, marginal posterior $\quad \pi(\theta \mid x^n) : \pi_m(\psi \mid x^n) = \int_{\{\psi(\theta)=\psi\}} \cdot\, d\theta$

  $\pi(p_1, p_2 \mid x^n) \to \pi(p_1 - p_2 \mid x^n)$

- Bayesian and frequentist philosophy $\quad$ AoS 11.1 (*) $\quad$ AoS §11.1
- empirical and epistemic probability

  freq. $\quad$ uncertainty of knowledge

Today

Start Recording

1. Friday: Jeffreys-Lindley paradox (HW 6 (c)); DF re χ^2; Pf. of B-H (?)
2. Bayesian inference overview ✓
3. Two weird examples — AoS 11.8, 11.10
4. ~~4.~~ 3. empirical Bayes
5. ~~5.~~ 4. hierarchical Bayes

- Mar 15 5.15 – 6.15 pm EDT
  Data Science Speaker Series
  Jesse Cresswell
  Machine Learning Scientist, Layer 6 AI at TD
  "Evaluating Model Performance on
  Highly Imbalanced Datasets"

Overview

- all information about θ contained in posterior density $\pi(\theta \mid x^n) = f(x^n \mid \theta)\pi(\theta)/f_{X^n}(x^n)$

Overview

- all information about θ contained in posterior density $\pi(\theta \mid x^n) = f(x^n \mid \theta)\pi(\theta)/f_{X^n}(x^n)$
- inference about $\psi(\theta)$ based on marginal posterior

Overview

- all information about θ contained in posterior density $\pi(\theta \mid x^n) = f(x^n \mid \theta)\pi(\theta)/f_{X^n}(x^n)$
- inference about $\psi(\theta)$ based on marginal posterior
- for comparing two (or more) points θ in $\pi(\theta \mid x^n)$, don't need marginal distribution of X^n

Overview

- all information about θ contained in posterior density $\pi(\theta \mid x^n) = f(x^n \mid \theta)\pi(\theta)/f_{X^n}(x^n)$
- inference about $\psi(\theta)$ based on marginal posterior
- for comparing two (or more) points θ in $\pi(\theta \mid x^n)$, don't need marginal distribution of X^n
- for choosing between models, do need marginal distribution of X^n, as in HW 6 (c)

Overview

- all information about θ contained in posterior density $\pi(\theta \mid x^n) = f(x^n \mid \theta)\pi(\theta)/f_{X^n}(x^n)$
- inference about $\psi(\theta)$ based on marginal posterior
- for comparing two (or more) points θ in $\pi(\theta \mid x^n)$, don't need marginal distribution of X^n
- for choosing between models, do need marginal distribution of X^n, as in HW 6 (c)
- Bayesian predictions of future values: posterior predictive

$$\pi(x_{new} \mid x^n) = \int f(x_{new} \mid x^n, \theta)\pi(\theta \mid x^n)d\theta,$$

Overview

- all information about θ contained in posterior density $\pi(\theta \mid x^n) = f(x^n \mid \theta)\pi(\theta)/f_{X^n}(x^n)$
- inference about $\psi(\theta)$ based on marginal posterior
- for comparing two (or more) points θ in $\pi(\theta \mid x^n)$, don't need marginal distribution of X^n
- for choosing between models, do need marginal distribution of X^n, as in HW 6 (c)
- Bayesian predictions of future values: posterior predictive

$$\pi(x_{new} \mid x^n) = \int f(x_{new} \mid x^n, \theta)\pi(\theta \mid x^n)d\theta,$$

- probability statements refer to uncertainty of knowledge

Overview

- all information about θ contained in posterior density $\pi(\theta \mid x^n) = f(x^n \mid \theta)\pi(\theta)/f_{X^n}(x^n)$
- inference about $\psi(\theta)$ based on marginal posterior
- for comparing two (or more) points θ in $\pi(\theta \mid x^n)$, don't need marginal distribution of X^n
- for choosing between models, do need marginal distribution of X^n, as in HW 6 (c)
- Bayesian predictions of future values: posterior predictive

$$\pi(x_{new} \mid x^n) = \int f(x_{new} \mid x^n, \theta)\pi(\theta \mid x^n)d\theta,$$

- probability statements refer to uncertainty of knowledge
- choosing priors can be difficult, and can have large impact in high-dimensional settings

Overview

- all information about θ contained in posterior density $\pi(\theta \mid x^n) = f(x^n \mid \theta)\pi(\theta)/f_{X^n}(x^n)$
- inference about $\psi(\theta)$ based on marginal posterior
- for comparing two (or more) points θ in $\pi(\theta \mid x^n)$, don't need marginal distribution of X^n
- for choosing between models, do need marginal distribution of X^n, as in HW 6 (c)
- Bayesian predictions of future values: posterior predictive

$$\pi(x_{new} \mid x^n) = \int f(x_{new} \mid x^n, \theta)\pi(\theta \mid x^n)d\theta,$$

- probability statements refer to uncertainty of knowledge
- choosing priors can be difficult, and can have large impact in high-dimensional settings
- most applications of Bayesian inference involve sampling from the posterior density

Overview

$\int \text{num}\, d\theta$

$\pi(\theta) = f(\theta)$ $\qquad$ $\pi(\theta | x) \propto L(\theta)\pi(\theta)$

- all information about θ contained in posterior density $\pi(\theta \mid x^n) = f(x^n \mid \theta)\pi(\theta)/f_{X^n}(x^n)$
- inference about $\psi(\theta)$ based on marginal posterior
- for comparing two (or more) points θ in $\pi(\theta \mid x^n)$, don't need marginal distribution of X^n
- for choosing between models, do need marginal distribution of X^n, as in HW 6 (c)
- Bayesian predictions of future values: posterior predictive

$$\pi(x_{new} \mid x^n) = \int f(x_{new} \mid x^n, \theta)\pi(\theta \mid x^n)d\theta,$$

$\Pr(H_1 | X) = \int \pi(\mu | x^n) \cdot d\mu$

freq.: $f(x_{new} | x^n; \hat{\theta})$

- probability statements refer to uncertainty of knowledge
- choosing priors can be difficult, and can have large impact in high-dimensional settings
- most applications of Bayesian inference involve sampling from the posterior density
- or approximating the posterior density $\qquad$ normal, Laplace

- excellent overview of Bayesian computational methods (in SM book)
- Laplace approximation of integrals ← analytic approx^n using Normal + higher order terms
- importance sampling ←
- Markov chain Monte Carlo sampling ← $\theta_1^*, \theta_2^*, \ldots, \theta_B^* \sim \pi(\theta \mid \underline{x}^n)$
  - Gibbs sampling ←
  - Metropolis-Hastings algorithm ← (improvement on)

$\underbrace{\theta_1^*, \theta_2^*, \ldots, \theta_B^*}$ → hist, or cdf, or $\hat{\pi}(\theta \mid \underline{x})$ kernel

- excellent overview of Bayesian computational methods
- Laplace approximation of integrals
- importance sampling
- Markov chain Monte Carlo sampling
  - Gibbs sampling
  - Metropolis-Hastings algorithm

616 11 · Bayesian Models

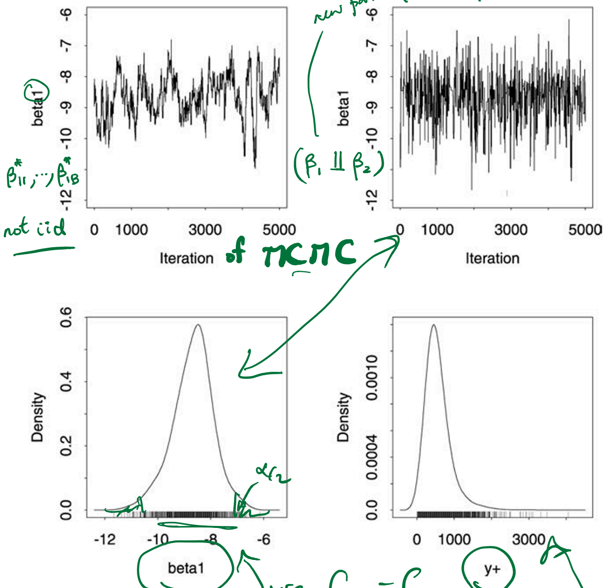


Figure 11.10 Bayesian analysis of motorette data using Metropolis–Hastings algorithm. Upper panels: sample paths for β_1 using two parametrizations, the right one more nearly orthogonal. Lower left: kernel density estimates of $\pi(\beta_1 \mid y)$ and of $\pi(Y_+ \mid y)$, where Y_+ is failure time predicted for 130°F.

new param. of nuisance par.

$(\beta_1 \perp\!\!\!\perp \beta_2)$

$\beta^*_{11}, \dots, \beta^*_{1B}$

not iid

of MCMC

$\alpha/2$

use for inf.

$\pi_m(\beta_1 \mid x^n) \to$ post. pred.

$(X_1, R_1, Y_1), \ldots, (X_n, R_n, Y_n)$ i.i.d.; parameter $\theta = (\theta_1, \ldots, \theta_B)$, B very large

parameter $\theta = (\theta_1, \ldots, \theta_B)$ B v. large $\gg n$ $0 \le \theta_j \le 1, \forall j$

known v. $\xi = (\xi_1, \ldots, \xi_B)$ $0 \le \delta < \xi_j < 1-\delta \le 1$

Sampling: $X_i \sim U\{1, \ldots, B\}$ rand. integer

$R_i \sim Ber(\xi_{X_i}) \left\{ \begin{array}{ll} 1 & \text{w. p. } \xi_{X_i} \\ 0 & 1-\xi_{X_i} \end{array} \right.$

if $R_i = 1$, obs. $Y_i \sim Ber(\theta_{X_i}) \longleftarrow$

if $R_i = 0$ don't obs Y_i

missing data

θ par. of interest = $P_r(Y_i = 1) = \psi = \left.\sum_{j=1}^{B} \theta_j / B \right\} \sigma$

Example 11.9 AoS

$$L\{\theta \mid (X, Y, R)\} = \prod_{i=1}^{n} f(x_i)\, \underbrace{f(r_i \mid x_i)}\, f(y_i \mid x_i, r_i)$$

$$= \prod_{i=1}^{n} \frac{1}{B} \cdot \xi_{x_i}^{r_i} (1-\xi_{x_i})^{1-r_i}\, \theta_{x_i}^{y_i r_i} (1-\theta_{x_i})^{(1-y_i) r_i}$$

$$\propto \prod \theta_{x_i}^{y_i r_i} (1-\theta_{x_i})^{(1-y_i) r_i}$$

$$\ell(\theta \mid (\ \ )\} = \sum_{i=1}^{n} \{ r_i y_i \log \theta_{x_i} + (1-y_i) r_i \log (1-\theta_{x_i}) \}$$

$$\textcircled{}\ \ell\{\theta \mid (X, Y, R)\} = \sum_{j=1}^{B} n_j \log \theta_j + \sum_{j=1}^{B} m_j \log (1-\theta_j)$$

for most θ_j

$m_j = \#\{Y_i = 0, R_i = 1, X_i = j\}$ $\quad n_j = \#\{Y_i = 1, R_i = 1, X_i = j\}$

To use this for B. inf.

$$\pi(\underline{\theta} \mid \underline{X}, \underline{Y}, \underline{R}) \propto \underbrace{L(\underset{\sim}{\theta} \mid (\underline{X}, \underline{Y}, \underline{R}))}_{\text{many are } 1}\,\pi(\underline{\theta})$$

on $\mathbb{R}^B$

$$\prod_{j \in \text{small set}} L(\theta_j)\pi(\theta_j) \times \underbrace{\prod}_{\text{most } j \in \{1, \dots, B\}} 1 \cdot \pi(\theta_j)$$

$$\boxed{\psi = \frac{1}{B}\sum_{j=1}^{B} \theta_j}$$

$$\pi(\psi \mid (\underline{X}, \underline{Y}, \underline{R})) = \sum_{\theta_j} \pi(\theta_j \mid \underline{X}, \underline{Y}, \underline{R})$$

$$\approx \pi(\psi) \quad \text{only}$$

Bayesian inf. fails

$$\hat{\psi} = \frac{1}{n}\sum_{i=1}^{n} \frac{R_i Y_i}{\xi_{x_i}}$$ — weighting by $pr(R_i = 1)$

$$E(\hat{\psi} \mid \psi) = \psi$$ in this weird model

perfectly good estimate of ψ

$$\mathrm{var}(\hat{\psi} \mid \psi) \le \frac{1}{n\delta^2}$$

"if the likelihood is weird, then Bayesian methods fail"

if your model parametrization ... could be in / is in trouble model $\boxed{x_i \mid \theta_i \sim N(\underline{\theta_i}, \underline{\underline{v_i}})}$ $i=1,\dots,n$ ind't

prior $\theta_i \mid \mu \sim N(\mu, \sigma^2)$ iid $i=1,\dots,n$

hyperprior $\underline{\underline{\mu}} \sim N(\mu_0, \tau^2)$ f μ_0, τ^2 are $\underline{\text{known}}$

$$\pi(\underset{\uparrow}{\underset{\sim}{\theta}}, \underset{\uparrow}{\mu} \mid \underset{\sim}{x}) \propto f(\underset{\sim}{x} \mid \underset{\sim}{\theta}) f(\underset{\sim}{\theta} \mid \mu) \pi(\mu)$$

$$= \prod_{i=1}^{n} \downarrow \quad \prod_{i=1}^{n} \downarrow \quad \frac{1}{\sqrt{2\pi}\,\tau} e^{-\frac{1}{2\tau^2}(\mu-\mu_0)^2}$$

Multivariate n on $\mathbb{R}^{n+1}$

$$\tilde{\mu}_B = E(\mu \mid \underset{\sim}{x}) = \frac{\mu_0/\tau^2 + \sum x_i/(\sigma^2+v_i)}{1/\tau^2 + \sum 1/(\sigma^2+v_i)}$$

Bayes est. of μ

hyper. par μ

"pop^{n} mean"

$$\operatorname{var}(\mu \mid \underline{x}) = \frac{1}{1/\tau^2 + \sum 1/(\sigma^2+v_i)}$$

$$\tilde{\theta}_{Bi} = E(\theta_i \mid \underline{x}) = E(\mu \mid \underline{x}) + \frac{\sigma^2}{\sigma^2+v_i}\left(x_i - E(\mu \mid \underline{x})\right)$$

Bayes est. of θ_i $= x_i \dfrac{\sigma^2}{\sigma^2+v_i} + E(\mu \mid \underline{x})\left(1 - \dfrac{\sigma^2}{\sigma^2+v_i}\right)$

Fig 11.11

$$v_i = \mathrm{var}(x_i \mid \theta_i) \qquad \sigma^2 = \mathrm{var}(\theta_i \mid \mu)$$

$$\tilde{\theta}_{i,B} = w_i x_i + (1 - w_i)\tilde{\mu}_B$$

$$\tilde{\mu}_B = \cancel{\bar{w}\mu_0 + (1-\bar{w})}$$ see previous slide

$$\mathrm{var}(\tilde{\theta}_{iB}) = \;? \; = \mathrm{var}(\tilde{\theta}_{iB} \mid \underset{\sim}{x})$$

— not available in closed form —

Empirical Bayes

SM 11.4. Eg. 11.28

Empirical Bayes

SM 11.4. Eg. 11.28

Empirical Bayes