MSA220 - Statistical Learning for Big Data

Lecture 16

Rebecka Jörnsten

Mathematical Sciences
University of Gothenburg and Chalmers University of Technology

Bayesian vs Frequentist

Brad Efron: "A 250-year argument"

Frequentist:

- Data are a random sample and the data generating process can be repeated
- Parameters are fixed
- Asymptotic frequencies over repeated sampling
- P-values: Prob(Reject null given null is true) (a frequency over repeated sampling)
- We can never accept the null, only reject it.

Bayesian:

- Data are observed and fixed
- Parameters are unknown and described probabilistically (describing subjective beliefs as probabilities)
- Probabilities interpreted as subjective beliefs (Prob(model is true))

Bayesian vs Frequentist

Frequentist:

- Point estimates, SE and CI: $\hat{\theta}(X)$, $CI(X)$ are random quantities through the sample X
- Deduction from $P(data|H0)$, $H0$ null hypothesis
  - Reject H0 if $P(data|H0) < \alpha$.
  - Fail to reject H0 if $\mathrm{P}(data|H0) \geq \alpha$.

Bayesian:

- Induction from posterior $P(\theta|data)$, starting with prior belief $\pi(\theta)$.
- That is, data is used to update our prior beliefs
- posterior density intervals - credible region

Bayesian vs Frequentist

Frequentist:

- A 95% confidence interval *covers the true, unknown parameter θ for 90% of CIs generated from repeated sampling*

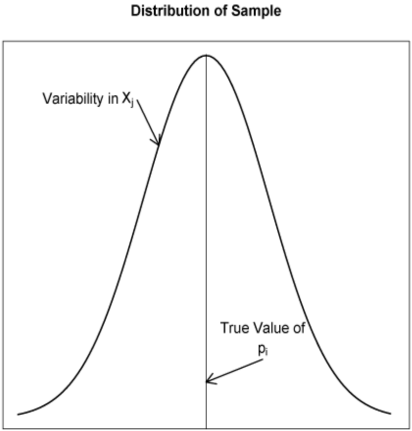
Bayesian:

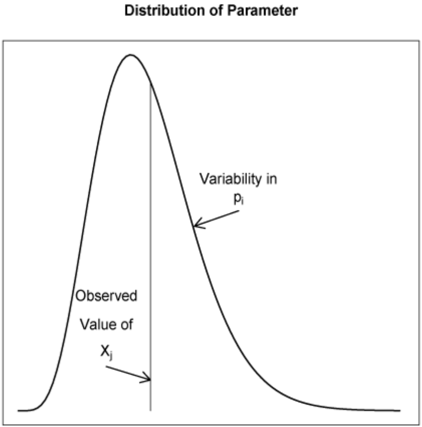
- For this data, a 95% credible region has probability 95% of including the parameter in the interval

Bayesian vs Frequentist

Frequentist: Describe variability in X given fixed parameter

Bayesian: Describe variability of the parameter for fixed X.





Bayesian vs Frequentist

Frequentist:

- Repeatable experiments in a controlled setting
- Parameters are fixed throughout the experiments

Bayesian:

- View the world as probabilistic
- Utilize subjective beliefs and translate to probabilities on parameters

Frequentist

- Key to analysis is the data likelihood

$$L(\theta|x_1^n) = \prod_{i=1}^{n} f_\theta(x_i)$$

- θ is fixed
- We view x_1^n as just one sample drawn from the data distribution and repeated sampling is possible
- We draw inference about θ from *statistics* $T(x_1^n)$
- T is random through the randomness of the sample
- p-value: $Pr(T(x^{rep}) > T(x^{obs})|H_0)$
- Probability of a repeated-sample statistic larger than observed statistic if null is true, i.e. just by chance alone
- NOT Probability that null is true or Probability that alternative is true
- It's a frequency statement over repeated sampling!

Bayesian

- The data x_1^n is fixed
- We have subjective beliefs about parameter that we express as a prior $\pi(\theta)$
- We update the belief to a posterior probability using Bayes rule
- $\pi(\theta|X) \propto \pi(\theta)L(\theta|X)$
- Credible region $Pr(\theta \in CR|X) = 95\%$
- Instead of p-value: Bayes Factor, $BF = \frac{Pr(M_1|X)}{Pr(M_0|X)}$ used to quantify relative evidence for candidate models.

- All about the prior!
- Subjective prior: we use knowledge of the world, prior experiments etc to formulate $\pi(\theta)$ (Frequentists are usually on board with this one)
- Objective prior: When we don't have much to go on, use an *uninformative prior* (a prior that says very little about the parameters, high variance).
- Frequentists don't like this one as much.
- Problem? Prior can have a big effect on marginal probabilities (one parameter of interest say) even though they're vague enough to not influence the fit much overall. We'll see an example later.

P-VALUES

- Frequentists: hypothesis testing
- Type I error: Prob(reject null — null is true) - we want to control this at some level α
- Type II error: Prob(fail to reject null — null is false) - this relates to the power of the test, can we detect a real effect?
- p-value depends on both the sample size and the effect size
- effect size: e.g. correlation, r-squared, group-mean differences,...

P-VALUES

- What happens when n is very large?
- Uncertainties of estimates become tiny
- "just by chance" variation becomes tiny
- All models are approximations and when n is large the approximations dominate over estimation uncertainty
- p-values become small! reflecting the imperfection or lack-of-fit of the model

- Does that mean p-values are meaningless?
- No, they do what they're designed to do - assess uncertainty due to sampling
- BUT, significance is not the same thing as important
- You should check the R^2 also (or some other measure of effect size).
- Small p-value $+$ big effect size to select

- Example (from Sullivan and Feinn, 2012)
- Study of 22000 subjects over 5 years
- Found that aspirin associated with a reduction in myocardial infarction
- p-value less than 10^{-5}!!!
- BUT... effect size $R^2 = 0.001$ or a reduction in risk for infarction 0.77%

- Example from Gelman, 2013
- Consider two sample with mean(SE): 25(10) and 10(10)
- The first sample results in a small p-value for testing $H_0 : \mu = 0$ and the second is not significant
- BUT the difference (two-sample t): 15(14) is NOT significant...
- What happened here? Myopic view but also we forgot that the p-value is ALSO a statistic and subject to random error

Bayes and big n

- Does being Bayesian fix the problem with big n?
- Not really - well, the focus is not on a p-value
- However, when n is large the prior has very little influence on the estimation and then how you compare models with BF is almost like doing likelihood-ratio testing only
- It boils down again to choosing a cutoff
- Divide and Conquer methods for Bayesian analysis looks very similar to the methods we talked about, just Bayesian estimation in each chunk instead of MLE or LS.

Bayes and the choice of prior

- We can all agree that subjective priors make sense
- What about the uninformative priors?
- Another example from Gelman, 2012
- Study found that 56% of children born to attractive parents are girls, whereas it's only 48% to less attractive parents (Kid you not: published study in J. Theor. Bio).
- Null hypothesis: sex-ratio difference $\theta = 0$: p-value 0.2 (original study 0.02 but didn't correct for multiple testing).
- OK - let's be Bayesian. No clear prior we can use so let's use an uninformative one Uniform on -1 to 1.
- 90% posterior probability that $\theta > 0$

Bayes and the choice of prior

- What happened?
- p-value: if we sampled attractive and unattractive parent sets repeatedly there's a 20% chance that we would see a sex-ratio difference as large as 56-48% just by chance.
- BUT, Bayesian analysis says the probability of more girls born to attractive parents is 90%
- Danger of flat or uninformative priors, especially in small samples.
- Can have weird effects on marginal posterior probabilities.

Bayes and the choice of prior

- More reasonable prior
- $N(0, v)$, believe that sex-ratio difference is 0 a priori
- The posterior probability that sex-ratio difference is bigger than 0 drops to 0.6.

- What's the trick in Bayesian analysis?
- In simple examples like above, we can compute posterior relatively easily
- In more complex models we use Monte-Carlo simulations, Gibbs sampling, or MCMC
- This is about *sampling* the model space to compute the posterior

- Example Raftery, Madigan and Hoeting, 1999
- Want to run a big regression model $Y = X\beta + \epsilon$
- Identify important predictors (model selection) and come up with a good final prediction scheme via model averaging

Bayes and linear modeling

- Frequentist version
- Subset selection
- Average top-models (based on AIC or BIC or Cp)
- Check which variables are in top models.

Bayes and linear modeling

- Here, set of candidate models $M_k, k = 1, \cdots, K$
- Posterior probability for model

$$Prob(M_k|D) = \frac{Pr(D|M_k)Pr(M_k)}{\sum_l Pr(D|M_l)Pr(M_l)}$$

- Each model involves parameters β_k with prior $Pr(\beta_k|M_k)$
- Data likelihood $Pr(D|\beta_k, M_k)$ is $Y \sim N(X\beta_k, \sigma^2 I)$

Bayes and linear modeling

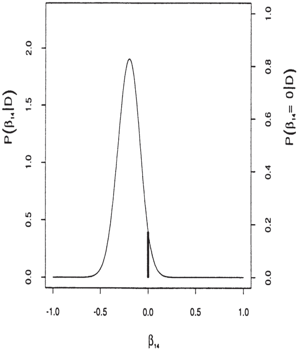
- Prior $\beta \sim N(0, \sigma^2 V)$
- where $V_{ii} \propto (X_i' X_i)^{-1}$, i.e. related to the information content in the i-th variable.
- Prior $\frac{\nu\lambda}{\sigma^2} \sim \chi^2_\nu$

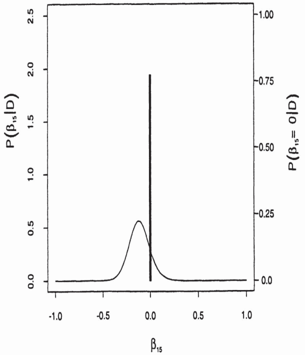
- Define a neighborhood for all models (like only one variable difference)
- Travel in model space (MCMC) exploring model neighborhoods and accept a new model if the BF(new vs old) is bigger than 1.
- You can approximate the posterior of any quantity of interest by taking averages over all states visited in the MCMC.

Bayes and linear modeling

Table 2. *Crime Data: Occam's Window Posterior Model Probabilities*

Model											Posterior model probability (%)
1		3	4			9	11		13	14	12.6
1		3	4				11		13	14	9.0
1		3	4			9			13	14	8.4
1		3		5		9	11		13	14	8.0
		3	4		8	9			13	14	7.6
1		3	4						13	14	6.3
1		3	4				11		13		5.8
1		3		5			11		13	14	5.7
1		3	4						13		4.9
1		3		5		9			13	14	4.8
		3		5	8	9			13	14	4.4
		3	4			9			13	14	4.1
		3		5		9			13	14	3.6
1		3		5					13	14	3.5
	2	3	4						13	14	2.0
1		3		5			11		13		1.9
		3	4						13	14	1.6
		3		5					13	14	1.6
		3	4						13		1.4
1		3		5					13		1.4
		3		5					13		.7
1			4					12	13		.7


P(β14|D)
P(β14= 0|D)
β14
-1.0
-0.5
0.0
0.5
1.0
2.0
1.5
0.8
0.6
0.4
0.2


P(β15|D)
P(β15= 0|D)
β15
-1.0
-0.5
0.0
0.5
1.0
1.5
2.0
2.5
0.25
0.50
0.75
1.00

Empirical Bayes

- We are Bayesian but we use the data to estimate they hyperparameters in the prior
- E.g. Let's say we have a prior $N(0, v)$ on each regression coefficient
- We can compute the *marginal distribution*

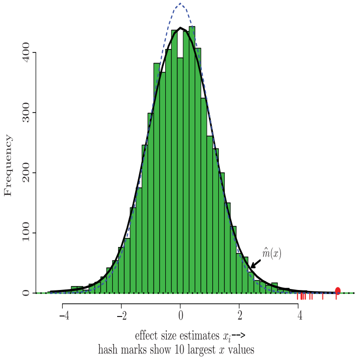
$$m(y|v) = \int_{\beta} f(y|\beta)\pi(\beta|v)d\beta$$

- Maximize the marginal distribution with respect to v to get $\hat{v}$
- Plug in to get posteriors $Prob(\beta_k|D, \hat{v})$

Empirical Bayes

- The point is that we use the fact that through the hyperparameters there is shared information
- An example from Efron, 2012
- Gene expression data (like the TCGA demo data), 6033 genes
- We want to identify the genes with expression levels different from 0
- $x_i \sim N(\delta_i, 1)$
- marginal $m(x) = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-.5(x-\delta)^2} \pi(\delta) d\delta$
- We don't know the prior BUT we can use ALL THE DATA to come up with an estimate for $m(x)$ without it!
- Natural estimate: the density of observed expression levels across all genes $\hat{m}(x)$.
- Posterior estimate $E(\delta_i | x_i) = x_i + \frac{d}{dx} \log \hat{m}(x)|_{x_i}$

Empirical Bayes



144 BRADLEY EFRON

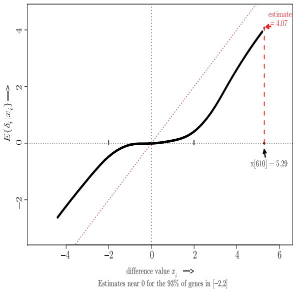


FIGURE 11. Empirical Bayes estimates of $E\{\delta|x\}$, the expected true difference δ_i given the observed difference x_i.

- Another example: Bayesian Lasso
- $\pi(\beta) = \prod_{j=1}^{p} \frac{\lambda}{2\sigma} e^{-\lambda|\beta_j|/\sigma}$
- Notice how all the prior components share hyperparameter λ (and σ)
- Yuan and Lin use this prior mixed with a "spike" at 0
- Park and Casella (Blasso) use the fact that the double-exponential prior can be written as a mixture of normals

$$\frac{a}{2} e^{-a|s|} = \int_0^\infty \frac{1}{\sqrt{2\pi s}} e^{-s^2/(2s)} \frac{a^2}{2} e^{-a^2 s/2} ds$$

Empirical Bayes

- Write prior for β $\pi(\beta|\tau_j, j = 1, \cdots, p) = N(0, \sigma^2 D_\tau)$ where D_τ is $diag(\tau_1, \cdots, \tau_p)$
- $\pi(\tau) = \prod_{j=1}^{p} \frac{\lambda^2}{2} e^{-\lambda^2 \tau_j^2}$
- Notice the shared hyperparameter λ!

- For current λ
- Gibbs sampling from posterior $p(\beta, \sigma, \tau | D, \lambda)$
- Approximate likelihood with respect to λ with average Gibbs plug-in for expected values β and τ
- Maximize with respect to λ
- Repeat

Empirical Bayes

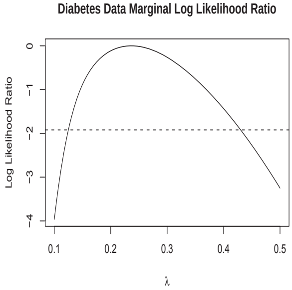


Fig. 5. The log likelihood ratio $\log\{L(\lambda|\tilde{y})/L(\lambda_{\text{MLE}}|\tilde{y})\}$ for the diabetes data, as approximated by a Monte Carlo method described in the text. The horizontal reference line at $-\chi^2_{1,0.95}/2$ suggests the approximate 95% confidence interval $(0.125, 0.430)$.

- What we get?
- Credible intervals for each β
- posterior distributions for β
- Empirical Bayes estimate for λ

Bayesian of Frequentist

- Why not both?
- Depends on situation at hand.
- Controlled experiments - frequentist approach natural
- Observational studies where much is known a priori - Bayesian setting is natural, especially if the notion of repeated samples make no sense
- BF or p-values: different perspective on modeling
- Empirical Bayes: really useful in high-dimensional modeling. Borrow information across multiple studies.