A Short Introduction to Bayesian Statistics

Daniel F. Schmidt

Faculty of Information Technology, Monash University

Fritz Haber Institute
April 26, 2024

Outline

Outline

Thomas Bayes

Reverend Thomas Bayes (1701 - 1761). Born in England. Studied logic and theology at University of Edinburgh, and became a Presbyterian minister. Became interested in problems of chance, and is most famous for the theorem on conditional probability that bears his name.

Schools of Statistical Inference

- Since statistics became a discipline, there have been two major schools of inference
  1. Frequentist statistics, pioneered by Ronald Fisher
  2. Bayesian statistics, named after Reverend Thomas Bayes
- More recently, a third paradigm – empirical risk minimisation – has become popular; I would consider it frequentist-adjacent
- Fisher disliked Bayesian statistics, and his personality dominated
  - Frequentist approach largely ruled until the 90s
- This is largely due to the increase in computing power
  - Bayesian approaches influenced much of modern machine learning

Why Bayesian Statistics?

- There are many strong reasons to be a Bayesian
  1. A unified framework for inference
     - Point/interval estimation and testing using one idea
  2. Extremely flexible model specification
     - Complex hierarchical models
     - “Random” parameters
     - Hidden/latent variables
  3. Marries well with computational advances
  4. Directly incorporates uncertainty
     - Takes into account uncertainty/variability in estimation
  5. Allows natural incorporation of prior information

Bayes' Rule (1)

- The primary tool we will use is Bayes' Rule
  - Named after Rev. Thomas Bayes
- Let X, Y be two R.V.s
  - Let $\mathbb{P}(X = x)$ be the marginal distribution of X
  - Let $\mathbb{P}(Y = y \,|\, X = x)$ be the conditional distribution of Y
  - Then, if we observe Y, Bayes' rule tells us

$$\mathbb{P}(X = x \,|\, Y = y) = \frac{\mathbb{P}(Y = y \,|\, X = x)\mathbb{P}(X = x)}{\mathbb{P}(Y = y)}$$

  where

$$\mathbb{P}(Y = y) = \sum_{X \in x} \mathbb{P}(Y = y \,|\, X = x)\mathbb{P}(X = x)$$

  is the marginal distribution of Y
- Bayes' rule gives us conditional probability of X given Y

Bayes' Rule Example

- A woman attends a GP clinic regarding a breast lump
  - The population frequency of breast cancer $(C = 1)$ 0.0066 (our prior probability)
  - The probability of developing a breast lump $(L = 1)$ if :
    - a woman has breast cancer $(C = 1)$ is 60%
    - if a woman does not have breast cancer $(C = 0)$ is 5%
- What is the probability the woman has breast cancer?

$$\begin{aligned}\mathbb{P}(C = 1 \,|\, L = 1) &= \frac{\mathbb{P}(L = 1 \,|\, C = 1)\mathbb{P}(C = 1)}{\sum_{c=0}^{1} \mathbb{P}(L = 1 \,|\, C = c)\mathbb{P}(C = c)} \\ &= \frac{0.6 \cdot 0.0066}{0.05 \cdot (1 - 0.0066) + 0.6 \cdot 0.0066} \\ &= 0.0738\end{aligned}$$

- So before seeing lump, $\mathbb{P}(C = 1)$ was 0.0066; after seeing lump the revised probability is 0.0738

Bayes' Rule Example

- A woman attends a GP clinic regarding a breast lump
  - The population frequency of breast cancer $(C = 1)$ 0.0066 (our prior probability)
  - The probability of developing a breast lump $(L = 1)$ if :
    - a woman has breast cancer $(C = 1)$ is 60%
    - if a woman does not have breast cancer $(C = 0)$ is 5%
- What is the probability the woman has breast cancer?

$$\begin{aligned}\mathbb{P}(C = 1 \,|\, L = 1) &= \frac{\mathbb{P}(L = 1 \,|\, C = 1)\mathbb{P}(C = 1)}{\sum_{c=0}^{1} \mathbb{P}(L = 1 \,|\, C = c)\mathbb{P}(C = c)} \\ &= \frac{0.6 \cdot 0.0066}{0.05 \cdot (1 - 0.0066) + 0.6 \cdot 0.0066} \\ &= 0.0738\end{aligned}$$

- So before seeing lump, $\mathbb{P}(C = 1)$ was 0.0066; after seeing lump the revised probability is 0.0738

Outline

Bayesian Inference – Setting

- How is this related to statistical inference?
- In Bayesian inference, we have the following ingredients:
  1. An observed sample $\mathbf{y} = (y_1, \ldots, y_n)$ from our population
  2. A model of our population

$$p(\mathbf{y} \,|\, \theta), \;\; \mathbf{y} \in \mathcal{Y}^n, \; \theta \in \Theta,$$

     parameterised by an unknown θ
     $\Rightarrow$ describes probability of $\mathbf{y}$ given true parameter is θ
  3. A prior probability distribution for our unknown parameter

$$\pi(\theta), \;\; \theta \in \Theta$$

     $\Rightarrow$ describes probability that θ is the true parameter before seeing data

- We now treat the unknown parameter as a random variable
  $\Longrightarrow$ Allows us to make probabilistic statements about θ

Bayesian Inference – The Posterior Distribution (1)

- We have seen $\mathbf{y}$; we know $p(\mathbf{y} \mid \theta)$ and $\pi(\theta)$
  - We then apply Bayes' rule to find $p(\theta \mid \mathbf{y})$:

$$p(\theta \mid \mathbf{y}) = \frac{p(\mathbf{y} \mid \theta)\pi(\theta)}{p(\mathbf{y})} \propto p(\mathbf{y} \mid \theta)\pi(\theta)$$

  where

$$p(\mathbf{y}) = \int_{\Theta} p(\mathbf{y} \mid \theta)\pi(\theta)d\theta$$

  is the marginal distribution of the data
  $\Longrightarrow$ This quantity is called the posterior distribution
- In this framework
  - $\pi(\theta)$ is the prior probability of model θ generating the data
  - $p(\mathbf{y} \mid \theta)$ is the probability of data $\mathbf{y}$ if the true model is θ
  - $p(\theta \mid \mathbf{y})$ is the posterior probability of model θ being true after observing data $\mathbf{y}$

Bayesian Inference – The Posterior Distribution (1)

- We have seen $\mathbf{y}$; we know $p(\mathbf{y} \,|\, \theta)$ and $\pi(\theta)$
  - We then apply Bayes' rule to find $p(\theta \,|\, \mathbf{y})$:

$$p(\theta \,|\, \mathbf{y}) = \frac{p(\mathbf{y} \,|\, \theta)\pi(\theta)}{p(\mathbf{y})} \propto p(\mathbf{y} \,|\, \theta)\pi(\theta)$$

where

$$p(\mathbf{y}) = \int_{\Theta} p(\mathbf{y} \,|\, \theta)\pi(\theta)d\theta$$

is the marginal distribution of the data

$\Longrightarrow$ This quantity is called the posterior distribution

- In this framework
  - $\pi(\theta)$ is the prior probability of model θ generating the data
  - $p(\mathbf{y} \,|\, \theta)$ is the probability of data $\mathbf{y}$ if the true model is θ
  - $p(\theta \,|\, \mathbf{y})$ is the posterior probability of model θ being true after observing data $\mathbf{y}$

Bayesian Inference – The Posterior Distribution (2)

- How to interpret the posterior distribution?
- If our prior distribution, $\pi(\theta)$, accurately describes the probability that different values of θ are the truth (i.e., the population value), then

$$\mathbb{P}(\theta \in A \,|\, \mathbf{y}) = \int_A p(\theta \,|\, \mathbf{y}) d\theta$$

  is the probability the population value of θ is in the set A, given that we observed the data $\mathbf{y} = (y_1, \ldots, y_n)$
- The posterior takes the data we have observed, and uses it to update our beliefs about how likely different values of θ are to be the population value

Bayesian Inference – The Prior Distribution (1)

- The prior distribution is the most controversial element of Bayesian inference
- How to interpret the prior distribution?
  - As a subjective description of prior beliefs about θ
    - E.g., probability of rat being dead after leaving out bait
    - It either is or isn't, but we don't know for sure until observed – has no frequency interpretation
  - As a model of a truly random process
    - Probability of failure of a component made from a manufacturing line
    - Yield of a corn-plant of a particular species
- Frequentists attack Bayesianism by targeting the prior
  - Claim is that frequentist stats is free of "personal priors"

Bayesian Inference – The Prior Distribution (2)

- Where do prior distributions come from?
  1. Chosen to reflect prior information/beliefs about problem
     - Prior information can be specific or general, depending on how we choose $\pi(\cdot)$
  2. Chosen for mathematical convenience
     - The choice of prior $\pi(\cdot)$ leads to simple posterior distributions
  3. Created to express prior ignorance
     - Sometimes called uninformative priors
     - Created by defining a mathematical concept of ignorance
  4. Chosen to match classical procedures (e.g., LASSO or ridge prior)
- Can combine different approaches, i.e., convenient prior distribution that (partially) reflects real prior information

Bayesian Inference – Summary

- The likelihood $p(\mathbf{y} \mid \theta)$ describes the probability of seeing data $\mathbf{y}$, if the population parameter was θ
- The prior distribution $\pi(\theta)$ describes the probability that the population parameter is θ, if we have not seen any data
- These form a joint distribution

$$p(\mathbf{y}, \theta) = p(\mathbf{y} \mid \theta)\pi(\theta)$$

- The posterior distribution $p(\theta \mid \mathbf{y})$ describes the probability θ is the population parameter, given we have observed $\mathbf{y}$
- The marginal distribution $p(\mathbf{y})$ describes the probability of observing data $\mathbf{y}$ if all we know about the population parameter is that it follows $\pi(\theta)$

Bayesian Inference – Summary

- The likelihood $p(\mathbf{y} \mid \theta)$ describes the probability of seeing data $\mathbf{y}$, if the population parameter was θ
- The prior distribution $\pi(\theta)$ describes the probability that the population parameter is θ, if we have not seen any data
- These form a joint distribution

$$p(\mathbf{y}, \theta) = p(\mathbf{y} \mid \theta)\pi(\theta)$$

- The posterior distribution $p(\theta \mid \mathbf{y})$ describes the probability θ is the population parameter, given we have observed $\mathbf{y}$
- The marginal distribution $p(\mathbf{y})$ describes the probability of observing data $\mathbf{y}$ if all we know about the population parameter is that it follows $\pi(\theta)$

Outline

Bayesian Point Estimation

- How do we actually use the posterior distribution to make inferences?
- Point estimates are statistics of the posterior
  - Posterior maximum (MAP) – choose θ that maximises posterior

$$\hat{\theta}_{\mathrm{MAP}} = \arg\max_{\theta} \{p(\theta \,|\, \mathbf{y})\}$$

    Tries to select the "most likely" estimate
  - Posterior mean

$$\hat{\theta}_{\mathrm{PM}} = \int \theta \, p(\theta \,|\, \mathbf{y}) d\theta = \mathbb{E}\left[\theta \,|\, \mathbf{y}\right]$$

    Uses the posterior average value of θ as the estimate
- Bayesian estimates combine information in the prior with information in the likelihood (i.e., from the observed data)

Uncertainty of Bayesian Point Estimates

- Point estimates give a best "guess" at the parameter values
  - They do not capture variability/uncertainty
- These aspects can be naturally measured using the posterior distribution
- One way to measure the uncertainty about the estimate is posterior standard deviation:

$$\sqrt{\mathbb{V}\left[\theta \mid \mathbf{y}\right]}$$

- The more informative is your prior distribution, the smaller (less uncertainty) the posterior standard deviation will be
- What about interval estimates to capture uncertainty?

Bayesian Credible Sets

- Bayesian equivalent of confidence intervals called credible intervals
- A $100\alpha\%$ credible interval is any interval (θ_-, θ_+) such that

$$\mathbb{P}(\theta_- < \theta < \theta_+ \,|\, \mathbf{y}) = \int_{\theta_-}^{\theta_+} p(\theta \,|\, \mathbf{y}) d\theta = \alpha$$

  where $\alpha \in (0, 1)$ is the level of the set
  - Generally we use centred intervals (e.g., from 2.5% to 97.5%)
- Different interpretation from confidence interval:
  - A $100\alpha\%$ confidence interval is an interval such that for $100\alpha\%$ of possible datasets, the interval will contain the (fixed) unknown true θ
  - A $100\alpha\%$ credible interval says that if our prior is accurate, then the probability that $\theta \in (\hat{\theta}_-, \hat{\theta}_+)$ is α, given we have observed the data $\mathbf{y}$

Bayesian Credible Sets

- Bayesian equivalent of confidence intervals called credible intervals
- A $100\alpha\%$ credible interval is any interval (θ_-, θ_+) such that

$$\mathbb{P}(\theta_- < \theta < \theta_+ \,|\, \mathbf{y}) = \int_{\theta_-}^{\theta_+} p(\theta \,|\, \mathbf{y}) d\theta = \alpha$$

  where $\alpha \in (0, 1)$ is the level of the set
  - Generally we use centred intervals (e.g., from 2.5% to 97.5%)
- Different interpretation from confidence interval:
  - A $100\alpha\%$ confidence interval is an interval such that for $100\alpha\%$ of possible datasets, the interval will contain the (fixed) unknown true θ
  - A $100\alpha\%$ credible interval says that if our prior is accurate, then the probability that $\theta \in (\hat{\theta}_-, \hat{\theta}_+)$ is α, given we have observed the data $\mathbf{y}$

The elephant in the room

- As we know, the key formula in the Bayesian approach is

$$p(\theta \,|\, \mathbf{y}) = \frac{p(\mathbf{y} \,|\, \theta)\pi(\theta)}{\int p(\mathbf{y} \,|\, \theta)\pi(\theta) d\theta}$$

  which gives us the posterior (after data) distribution describing how likely different values of θ are to be the value of the population parameter, given our prior beliefs
- This formula depends crucially on evaluating the denominator
- Yet for almost all real problems, it cannot be evaluated
  - Even numerical approaches tend to fail – it is a nasty integral!
- Even if we could, we still need to somehow manipulate multidimensional densities

  $\Longrightarrow$ instead we usual approximate the posterior

Monte Carlo Markov Chain (MCMC)

- MCMC is very popular for Bayesian inference
- Here we approximate the posterior by a set of m samples

$$\theta^{(1)}, \ldots, \theta^{(m)}$$

  randomly draw from the posterior
- We can then approximate posterior statistics using empirical quantities, e.g.,

$$\mathbb{E}\left[\theta \,|\, \mathbf{y}\right] \approx \frac{1}{m} \sum_{i=1}^{m} \theta^{(i)}$$

- Similarly for medians, quantiles, etc.
- MCMC algorithms are general and are simulation consistent but can be slow, especially if you need many samples
- General purpose tools (i.e., JAGS, Stan) available

Variational Bayes

- An alternative to MCMC is variational Bayes
- We replace the posterior $p(\theta \,|\, \mathbf{y})$ with an approximation
  - We choose some parametric distributions to model the posterior
- We adjust parameters of approximating distributions to minimise approximation error
  - Based on the KL divergence from approximators to true posterior
- This formulation avoids the need to compute $p(\mathbf{y})$, i.e., we can use unnormalised posteriors
- In comparison to MCMC, can be much faster and more scalable
- There are drawbacks though:
  - we never know how close our approximation actually is
  - no matter how long we run the VB search, we are limited in quality of approximation by the choice of approximating distributions

Bayesian Prediction (1)

- Consider a model $p(y \,|\, \boldsymbol{\theta})$ that we want to use for prediction
- A prediction is some function of the model, and therefore, a function of the model parameters, i.e., $f(\boldsymbol{\theta})$
- Examples of predictions
  - The average value of future realisations of Y from our population would be predicted by the mean of the fitted distribution:

$$f(\boldsymbol{\theta}) \equiv \mathbb{E}\left[Y \,|\, \hat{\boldsymbol{\theta}}\right] = \int_{-\infty}^{\infty} y \, p(y \,|\, \hat{\boldsymbol{\theta}}) dy$$

  - Or the probability that a random individual from our population has a value greater than c would be predicted by

$$f(\boldsymbol{\theta}) \equiv \mathbb{P}(Y > c \,|\, \hat{\boldsymbol{\theta}}) = \int_{c}^{\infty} p(y \,|\, \hat{\boldsymbol{\theta}}) dy$$

  and so on

Bayesian Prediction (1)

- Consider a model $p(y \,|\, \boldsymbol{\theta})$ that we want to use for prediction
- A prediction is some function of the model, and therefore, a function of the model parameters, i.e., $f(\boldsymbol{\theta})$
- Examples of predictions
  - The average value of future realisations of Y from our population would be predicted by the mean of the fitted distribution:

$$f(\boldsymbol{\theta}) \equiv \mathbb{E}\left[Y \,|\, \hat{\boldsymbol{\theta}}\right] = \int_{-\infty}^{\infty} y \, p(y \,|\, \hat{\boldsymbol{\theta}}) dy$$

  - Or the probability that a random individual from our population has a value greater than c would be predicted by

$$f(\boldsymbol{\theta}) \equiv \mathbb{P}(Y > c \,|\, \hat{\boldsymbol{\theta}}) = \int_{c}^{\infty} p(y \,|\, \hat{\boldsymbol{\theta}}) dy$$

and so on

Bayesian Prediction (1)

- Consider a model $p(y \,|\, \boldsymbol{\theta})$ that we want to use for prediction
- A prediction is some function of the model, and therefore, a function of the model parameters, i.e., $f(\boldsymbol{\theta})$
- Examples of predictions
  - The average value of future realisations of Y from our population would be predicted by the mean of the fitted distribution:

$$f(\boldsymbol{\theta}) \equiv \mathbb{E}\left[Y \,|\, \hat{\boldsymbol{\theta}}\right] = \int_{-\infty}^{\infty} y\, p(y \,|\, \hat{\boldsymbol{\theta}}) dy$$

  - Or the probability that a random individual from our population has a value greater than c would be predicted by

$$f(\boldsymbol{\theta}) \equiv \mathbb{P}(Y > c \,|\, \hat{\boldsymbol{\theta}}) = \int_{c}^{\infty} p(y \,|\, \hat{\boldsymbol{\theta}}) dy$$

  and so on

Bayesian Prediction (2)

- How to do Bayesian prediction?
- One way is to use a Bayesian estimate of $\boldsymbol{\theta}$, such as the posterior mean $\mathbb{E}[\boldsymbol{\theta} \mid \mathbf{y}]$ and plug it in to our model as usual
  $\Longrightarrow$ but this ignores the variability in our estimates
- Alternatively, use the posterior $p(\boldsymbol{\theta} \mid \mathbf{y})$ to incorporate the uncertainty
- As a prediction $f(\boldsymbol{\theta})$ is just a function of a $\boldsymbol{\theta}$, and $\boldsymbol{\theta}$ is a random variable distributed as per the posterior distribution, it follows that $f(\boldsymbol{\theta})$ is a random variable as well with distribution $p(f(\boldsymbol{\theta}) \mid \mathbf{y})$, i.e., there exists a posterior distribution over the predictions
- In general this is difficult, but it is easy if we have posterior samples $\boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(m)}$; we just evaluate $f(\boldsymbol{\theta})$ for every sample:

$$f(\boldsymbol{\theta}^{(1)}), \ldots, f(\boldsymbol{\theta}^{(m)})$$

$\Longrightarrow$ these samples now approximate the posterior of $f(\boldsymbol{\theta})$

Bayesian Prediction (2)

- How to do Bayesian prediction?
- One way is to use a Bayesian estimate of $\boldsymbol{\theta}$, such as the posterior mean $\mathbb{E}\left[\boldsymbol{\theta} \,|\, \mathbf{y}\right]$ and plug it in to our model as usual
  $\Longrightarrow$ but this ignores the variability in our estimates
- Alternatively, use the posterior $p(\boldsymbol{\theta} \,|\, \mathbf{y})$ to incorporate the uncertainty
- As a prediction $f(\boldsymbol{\theta})$ is just a function of a $\boldsymbol{\theta}$, and $\boldsymbol{\theta}$ is a random variable distributed as per the posterior distribution, it follows that $f(\boldsymbol{\theta})$ is a random variable as well with distribution $p(f(\boldsymbol{\theta}) \,|\, \mathbf{y})$, i.e., there exists a posterior distribution over the predictions
- In general this is difficult, but it is easy if we have posterior samples $\boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(m)}$; we just evaluate $f(\boldsymbol{\theta})$ for every sample:

$$f(\boldsymbol{\theta}^{(1)}), \ldots, f(\boldsymbol{\theta}^{(m)})$$

$\Longrightarrow$ these samples now approximate the posterior of $f(\boldsymbol{\theta})$

Outline

The Linear Regression Model (1)

- In this session we will examine the linear regression model
- We have a target (outcome) variable, Y, that we wish to predict
- We say that Y is modelled as a linear combination of p explanatory variables, plus an intercept and a random error:

$$Y = \beta_0 + \sum_{j=1}^{p} \beta_j X_j + \varepsilon$$

where

- β_0 is the intercept
- $X_1, \ldots, X_p$ are explanatory variables
- $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)$ are the coefficients
- ε is the random error

The Linear Regression Model (2)

- If we assume that the error is normally distributed, i.e.

$$\varepsilon \sim \mathrm{N}(0, \sigma^2)$$

  then we can say that

$$Y \sim \mathrm{N}\left(\beta_0 + \sum_{j=1}^{p} \beta_j X_j,\ \sigma^2\right)$$

  and
  - β_0 sets the average value of Y when all the predictors are zero
  - β_j is the increase in mean of Y per unit increase in predictor X_j, *above and beyond* the effect of β_0
  - σ sets the *scale* of our errors
- For simplicity of exposition, let us assume β_0 and σ are known

Prior Distributions for β_j (1)

- How to choose a prior for the coefficients β_j?
- Coefficient β_j expresses the effect of expanatory variable X_j on the mean of Y, above and beyond the average value β_0
  - We might expect, *a priori*, it is just as likely to be a negative effect as a positive effect
  - We might expect, *a priori*, that any given explanatory variable is likely to be unassociated with Y
- We use a symmetric, bell-shaped distribution centered at $\beta_j = 0$
  - Prior "guess" is that X_j is unassociated with Y
  - Prior probability that $\mathbb{P}(\beta_j < 0)$ is same as that $\mathbb{P}(\beta_j > 0)$

Bayesian Ridge Regression (1)

- Let us choose to use a normal prior on β_j centered on $\beta_j = 0$
- We have the Bayesian hierarchy

$$\begin{aligned} \mathbf{y} \mid \boldsymbol{\beta}, \mathbf{X} &\sim \mathrm{N}(\mathbf{X}\boldsymbol{\beta}, \sigma^2 \mathbf{I}_n) \\ \boldsymbol{\beta} \mid \tau &\sim \mathrm{N}(\mathbf{0}_p, \tau^2 \sigma^2 \mathbf{I}_p) \end{aligned}$$

  where τ is a hyperparameter controlling the prior variance (i.e., how tightly our prior is concentrated around $\beta_j = 0$)
- Apply Bayes rule (multiplying likelihood and prior and normalizing) yields the posterior distribution for $\boldsymbol{\beta}$

$$\boldsymbol{\beta} \mid \mathbf{y} \sim \mathrm{N}(\mathbf{A}\mathbf{X}'\mathbf{y}, \sigma^2 \mathbf{A})$$

  where

$$\mathbf{A} = \left(\mathbf{X}'\mathbf{X} + \tau^{-2}\mathbf{I}_p\right)^{-1}$$

- The fact the posterior is also normal is because the prior is conjugate to the likelihood

Bayesian Ridge Regression (2)

- The posterior mean estimate of $\boldsymbol{\beta}$ is

$$\mathbb{E}\left[\boldsymbol{\beta} \,|\, \mathbf{y}\right] = \left(\mathbf{X}'\mathbf{X} + \tau^{-2}\mathbf{I}_p\right)^{-1}\mathbf{X}'\mathbf{y}$$

  which is also the solution to the ridge regression

$$\arg\min_{\boldsymbol{\beta}} \left\{ ||\mathbf{y} - \mathbf{X}\boldsymbol{\beta}||^2 + \tau^{-2}||\boldsymbol{\beta}||^2 \right\}$$

- This is why it is called the Bayesian ridge
- As the hyperparameter $\tau \to 0$, the estimates shrink towards $\boldsymbol{\beta} \to \mathbf{0}_p$ $\implies$ we become more confident in our prior guess that $\beta_j = 0$

Bayesian Ridge Regression (3)

- The posterior covariance of $\boldsymbol{\beta}$ is

$$\text{Cov}\left[\boldsymbol{\beta} \,|\, \mathbf{y}\right] = \sigma^2 \left(\mathbf{X}'\mathbf{X} + \tau^{-2}\mathbf{I}_p\right)^{-1}$$

- As the hyperparameter $\tau \to 0$, the variances become smaller $\implies$ our prior becomes more informative about $\boldsymbol{\beta}$ relative to the data

- Recall that squared-prediction error is composed of bias and variance
- If we choose τ carefully, we can reduce variance a lot while only introducing a small amount of bias, and obtain improved prediction performance over least-squares
- In regular ridge regression we would use cross-validation to choose τ

Bayesian Hyperpriors (1)

- How to select the prior hyperparameter τ?
  - This controls how much prior probability is concentrated around $\beta_j = 0$
- This is where the beauty of Bayes comes to the fore
  - We don't use heuristic methods like cross-validation.
- Instead, we treat it as an another unknown parameter, put a prior on it, and estimate it along with everything else!
- We use the same machinery to estimate hyperparameters and parameters.
- In contrast to methods like CV, the final posterior incorporates uncertainty about τ into our estimates of $\boldsymbol{\beta}$
- A good default prior for scale-type hyperparameters is the half-Cauchy distribution

$$\pi(\tau) = \frac{2}{\pi(1 + \tau^2)}$$

Bayesian Hyperpriors (2)

- Why can we put a prior on our hyperparameter?
- Consider a prior distribution $\pi(\theta \mid \alpha)$ where α is a hyperparameter
  - Place a hyperprior on α, say $\pi(\alpha)$
- We can write the joint prior distribution as

$$\pi(\theta, \alpha) = \pi(\theta \mid \alpha)\pi(\alpha)$$

- We could then remove α from the problem by integrating (marginalising) it out

$$\pi(\theta) = \int \pi(\theta \mid \alpha)\pi(\alpha) d\alpha$$

to get a marginal prior distribution free of α
$\implies$ so priors on hyperparameters really just lead to new priors on θ

Thank you!

- An implementation of Bayesian ridge regression in python that outperforms leave-one-out cross-validation
    - S. Tew, M. Boley and D.F.Schmidt, "*Bayes beats Cross Validation: Fast and Accurate Ridge Regression via Expectation Maximization*", NeuRIPS, 2023
- `pip install fastridge`
- Code is available at
    - `https://github.com/marioboley/fastridge.git`
- "bayesreg" R package for Bayesian penalized linear and logistic regression
    - Available on CRAN
- Thank you for your attention!

Bayes Inference - A Recap (2)

Quantity	Frequentist	Bayesian
Model of population	$p(\mathbf{y} \mid \theta)$, true population parameter θ unknown	
Population Parameter	True θ unknown, but fixed	True θ is a random variable i.e., $\theta \sim \theta(\pi) d\theta$
Point Estimates	Maximum Likelihood $\hat{\theta}_{\mathrm{ML}}$ Penalized Maximum Likelihood, etc.	Posterior mean, posterior mode General Bayes estimator
Measures of Uncertainty	Standard error $\sqrt{\mathbb{V}\left[\hat{\theta}_{\mathrm{ML}}\right]}$	Posterior standard deviation $\sqrt{\mathbb{V}\left[\theta \mid \mathbf{y}\right]}$
Interval Estimates	$100\alpha\%$ Confidence Intervals $A(\mathbf{y})$ such that $\mathbb{P}(\theta \in A(\mathbf{y})) = \alpha$ if $\mathbf{y} \sim p(\mathbf{y} \mid \theta)$, θ unknown but fixed	$100\alpha\%$ Credible Intervals A such that $\mathbb{P}(\theta \in A \mid \mathbf{y}) = \alpha$ conditional on seeing $\mathbf{y}$

Frequentist vs Bayesian Inference