Section 6

Bayesian linear models

Bayesian linear regression

- Linear regression is by far the most common statistical model
- It includes as special cases the t-test and ANOVA
- The multiple linear regression model is

$$Y_i \sim \text{Normal}(\beta_0 + X_{i1}\beta_1 + ... + X_{ip}\beta_p, \sigma^2)$$

  independently across the $i = 1, ..., n$ observations

- As we'll see, Bayesian and classical linear regression are similar if $n >> p$ and the priors are uninformative.
- However, the results can be different for challenging problems, and the interpretation is different in all cases

Outline

These notes cover Chapter 4

- Bayesian t-tests
- Bayesian linear regression
  - Gaussian priors
  - Jeffreys' priors
  - Shrinkage priors
- Generalized linear models
- Random effects
- Flexible linear models
  - Non-linear regression
  - Heteroskedastic errors
  - Non-Gaussian errors
  - Correlated errors

Bayesian one-sample (i.e., paired) t-test

- Say $Y_1, ..., Y_n \sim \text{Normal}(\mu, \sigma^2)$
- Typically Y_i is the difference of a pair of measurements, e.g., the post- minus pre-test for subject i
- Therefore the interest is to compare μ to zero
- We will consider two cases: σ^2 known and σ^2 unknown

Bayesian one-sample (i.e., paired) t-test

- Under the Jeffreys' prior $\pi(\mu) = 1$ with fixed σ,

$$\mu|\mathbf{Y}, \sigma \sim \text{Normal}\left(\bar{Y}, \frac{\sigma^2}{n}\right)$$

- Therefore the posterior mean is the sample mean,

$$\text{E}(\mu|\mathbf{Y}) = \bar{Y}$$

- The 95% credible set is the 95% confidence interval

$$\bar{Y} \pm 1.96\frac{\sigma}{\sqrt{n}}$$

- For the test of $\mathcal{H}_0 : \mu \leq 0$ versus $\mathcal{H}_1 : \mu > 0$,

$$\text{Prob}(\mathcal{H}_0|\mathbf{Y}) = \text{Prob}(\mu \leq 0|\mathbf{Y}) = \Phi(\sqrt{n}\bar{Y}/\sigma)$$

  is the frequentist p-value

Bayesian one-sample (i.e., paired) t-test

- When σ^2 is unknown, the Jeffreys' prior is

$$\pi(\mu, \sigma^2) \propto \left(\frac{1}{\sigma^2}\right)^{3/2}$$

- The marginal posterior integrating over uncertainty in σ^2 is

$$\mu|\mathbf{Y} \sim t_n\left(\bar{Y}, \frac{\hat{\sigma}^2}{n}\right)$$

  where $\hat{\sigma}^2 = \sum_{i=1}^{n}(Y_i - \bar{Y})^2/n$

- This is very similar to the frequentist t-test, except that the degrees of freedom is n rather than $n-1$

- This is the effect of the prior

Bayesian two-sample t-test

- Say the n_1 observations from group 1 are

$$Y_i \sim \text{Normal}(\mu, \sigma^2)$$

  are the n_2 observations from group 2 are

$$Y_i \sim \text{Normal}(\mu + \delta, \sigma^2)$$

- The goal is to compare δ to zero
- With σ^2 known and Jeffrey's prior $\pi(\mu, \delta) = 1$,

$$\delta | \mathbf{Y}, \sigma^2 \sim \text{Normal}\left(\bar{Y}_2 - \bar{Y}_1, \frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2}\right)$$

  and the results are identical to the two-sample z-test

Bayesian two-sample t-test

- When σ^2 is unknown, the Jeffreys' prior is

$$\pi(\mu, \delta, \sigma^2) \propto \left(\frac{1}{\sigma^2}\right)^2$$

- The marginal posterior integrating over uncertainty in σ^2 and μ is

$$\delta|\mathbf{Y} \sim t_n\left(\bar{Y}_2 - \bar{Y}_1, \frac{\hat{\sigma}^2}{n_1} + \frac{\hat{\sigma}^2}{n_2}\right)$$

where the pooled variance estimator is

$$\hat{\sigma}^2 = \left[\sum_{i=1}^{n_1}(Y_i - \bar{Y}_1)^2 + \sum_{i=n_1+1}^{n_2}(Y_i - \bar{Y}_2)^2\right]/n$$

- This is very similar to the frequentist t-test, except that the degrees of freedom is $n = n_1 + n_2$ rather than $n - 2$
- This is the effect of the prior

Review of least squares

- The least squares estimate of $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_p)^T$ is

$$\hat{\boldsymbol{\beta}}_{OLS} = \underset{\boldsymbol{\beta}}{\text{argmin}} \sum_{i=1}^{n} (Y_i - \mu_i)^2$$

  where $\mu_i = \beta_0 + X_{i1}\beta_1 + ... + X_{ip}\beta_p$

- $\hat{\boldsymbol{\beta}}_{OLS}$ is unbiased even if the errors are non-Gaussian
- If the errors are Gaussian then the likelihood is proportional to

$$\prod_{i=1}^{n} \exp\left[-\frac{(Y_i - \mu_i)^2}{2\sigma^2}\right] = \exp\left[-\frac{\sum_{i=1}^{n}(Y_i - \mu_i)^2}{2\sigma^2}\right]$$

- Therefore, if the errors are Gaussian $\hat{\boldsymbol{\beta}}_{OLS}$ is also the MLE

Review of least squares

- Linear regression is often simpler to describe using linear algebra notation
- Let $\mathbf{Y} = (Y_1, ..., Y_n)^T$ be the response vector and $\mathbf{X}$ be the $n \times (p+1)$ matrix of covariates
- Then the mean of $\mathbf{Y}$ is $\mathbf{X}\beta$ and the least squares solution is

$$\hat{\beta}_{OLS} = \underset{\beta}{\text{argmin}}(\mathbf{Y} - \mathbf{X}\beta)^T(\mathbf{Y} - \mathbf{X}\beta) = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$$

- If the errors are Gaussian then the sampling distribution is

$$\hat{\beta}_{OLS} \sim \text{Normal}\left[\beta, \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}\right]$$

- If the variance σ^2 is estimated using the mean squared residual error then the sampling distribution is multivariate t

Bayesian regression

- The likelihood remains

$$Y_i \sim \text{Normal}(\beta_0 + X_{i1}\beta_1 + ... + X_{ip}\beta_p, \sigma^2)$$

  independent for $i = 1, ..., n$ observations

- As with a least squares analysis, it is crucial to verify this is appropriate using qq-plots, added variable plots, etc.

- A Bayesian analysis also requires priors for β and σ

- We will focus on prior specification since this piece is uniquely Bayesian.

Priors

- For the purpose of setting priors, it is helpful to standardize both the response and each covariate to have mean zero and variance one.

- Many priors for β have been considered:
  1. Improper priors
  2. Gaussian priors
  3. Double exponential priors
  4. Many, many more...

Improper priors

- With σ fixed, the Jeffreys' prior is flat $p(\beta) = 1$
- This is improper, but the posterior is proper under the same conditions required by least squares
- If σ is known then

$$\beta|\mathbf{Y} \sim \text{Normal}\left[\hat{\beta}_{OLS}, \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}\right]$$

- See "Post beta" in the online derivations
- Therefore, the results should be similar to least squares
- How are they different?

Improper priors

- Of course we rarely know σ
- A conjugate uninformative prior is

$$\sigma^2 \sim \text{InvGamma}(a, b)$$

  with a and b set to be small, say $a = b = 0.01$.

- In this case the posterior of β follows a multivariate t centered on $\hat{\beta}_{OLS}$
- Again, the results are similar to OLS

Improper priors

- The objective Bayes Jeffreys prior is

$$p(\boldsymbol{\beta}, \sigma^2) = \left(\frac{1}{\sigma^2}\right)^{p/2+1}$$

  which is the inverse gamma prior with $a = p/2$ and $b \to 0$

- This gives posterior (marginal over σ^2)

$$\boldsymbol{\beta}|\mathbf{Y} \sim \mathrm{t}_n\left(\hat{\boldsymbol{\beta}}_{OLS}, \hat{\sigma}^2(\mathbf{X}^T\mathbf{X})^{-1}\right)$$

  where $\hat{\sigma}^2 = (\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}_{OLS})^T(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}_{OLS})/n$

- The posterior is proper in the same situation that the least squares solution exists

Multivariate normal prior

- Another common prior for is Zellner's g-prior

$$\beta \sim \text{Normal}\left[0, \frac{\sigma^2}{g}(\mathbf{X}^T\mathbf{X})^{-1}\right]$$

- This prior is proper assuming $\mathbf{X}$ is full rank
- The posterior mean is

$$\frac{1}{1+g}\hat{\beta}_{OLS}$$

- This shrinks the least estimate towards zero
- g controls the amount of shrinkage
- $g = 1/n$ is common, and called the unit information prior

Univariate Gaussian priors

- If there are many covariates or the covariates are collinear, then $\hat{\beta}_{OLS}$ is unstable
- Independent priors can counteract collinearity

$$\beta_j \sim \text{Normal}(0, \sigma^2/g)$$

  independent over j

- The posterior mode is

$$\underset{\beta}{\text{argmin}} \sum_{i=1}^{n} (Y_i - \mu_i)^2 + g \sum_{j=1}^{p} \beta_j^2$$

- In classical statistics, this is known as the ridge regression solution and is used to stabilize the least squares solution

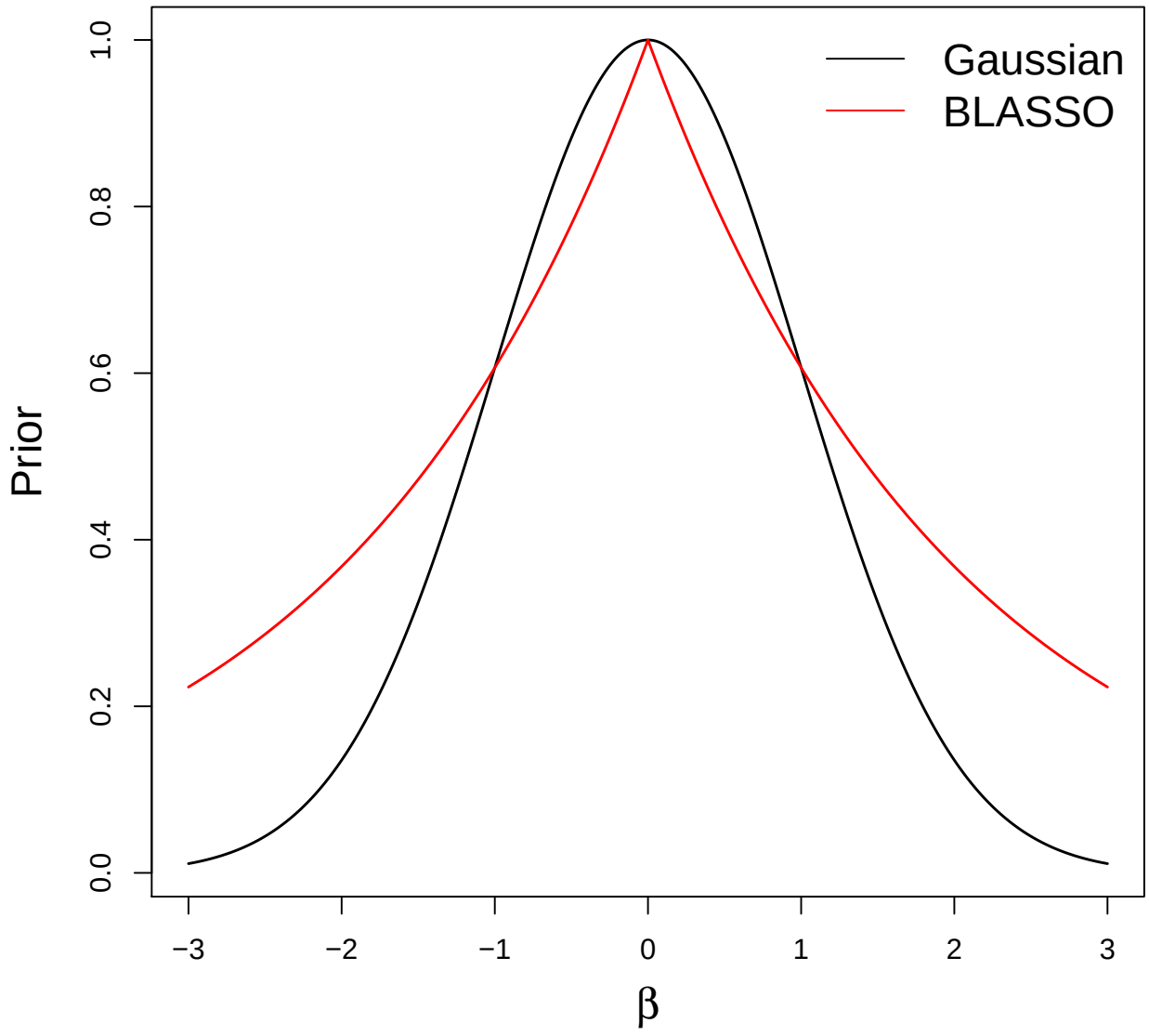
BLASSO

- An increasingly-popular prior is the double exponential or Bayesian LASSO prior
- The prior is $\beta_j \sim \text{DE}(\tau)$ which has PDF

$$f(\beta) \propto \exp\left(-\frac{|\beta|}{\tau}\right)$$

- The square in the Gaussian prior is replaced with an absolute value
- The shape of the PDF is thus more peaked at zero (next slide)
- The BLASSO prior favors settings where there are many β_j near zero and a few large β_j
- That is, p is large but most of the covariates are noise

BLASSO



BLASSO

- The posterior mode is

$$\underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{n} (Y_i - \mu_i)^2 + g \sum_{j=1}^{p} |\beta_j|$$

- In classical statistics, this is known as the LASSO solution
- It is popular because it adds stability by shrinking estimates towards zero, and also sets some coefficients to zero
- Covariates with coefficients set to zero can be removed
- Therefore, LASSO performs variables selection and estimation simultaneously

Computing

- With flat or Gaussian (with fixed prior variance) priors the posterior is available in closed-form and Monte Carlo sampling is not needed
- JAGS also works well, but there are R (and SAS and others) packages dedicated just to Bayesian linear regression that are preferred for big/hard problems
- `BLR` is probably the most common

Computing for the BLASSO

- For the BLASSO prior the full conditionals are more complicated
- There is a trick to make all full conditional conjugate so that Gibbs sampling can be used
- Metropolis sampling works fine too
- `BLR` works well for BLASSO and is super fast

Summarizing the results

- The standard summary is a table with marginal means and 95% intervals for each β_j
- This becomes unwieldy for large p
- Picking a subset of covariates is a crucial step in a linear regression analysis.
- We will discuss this later in the course.
- Common methods include cross-validation, information criteria, and stochastic search.

Predictions

- Say we have a new covariate vector $\mathbf{X}_{new}$ and we would like to predict the corresponding response Y_{new}
- A plug-in approach would fix β and σ at their posterior means $\hat{\beta}$ and $\hat{\sigma}$ to make predictions

$$Y_{new}|\hat{\beta}, \hat{\sigma} \sim \text{Normal}(\mathbf{X}_{new}\hat{\beta}, \hat{\sigma}^2)$$

- However this plug-in approach suppresses uncertainty about β and σ
- Therefore these prediction intervals will be slightly too narrow leading to undercoverage

Posterior predicitive distribution (PPD)

- We should really account for all uncertainty when making predictions, including our uncertainty about β and σ
- We really want the PPD

$$\begin{aligned} p(Y_{new}|\mathbf{Y}) &= \int f(Y_{new}, \beta, \sigma|\mathbf{Y}) d\beta d\sigma \\ &= \int f(Y_{new}|\beta, \sigma) f(\beta, \sigma|\mathbf{Y}) d\beta d\sigma \end{aligned}$$

- Marginalizing over the model parameters accounts for their uncertainty
- The concept of the PPD applies generally (e.g., logistic regression) and means the distribution of the predicted value marginally over model parameters

Posterior predicitive distribution (PPD)

- MCMC naturally gives draws from Y_{new}'s PPD
  - For MCMC iteration t we have $\beta^{(t)}$ and $\sigma^{(t)}$
  - For MCMC iteration t we sample

$$Y_{new}^{(t)} \sim \text{Normal}(\mathbf{X}\beta^{(t)}, {\sigma^{(t)}}^2)$$

  - $Y_{new}^{(1)}, ..., Y_{new}^{(S)}$ are samples from the PPD

- This is an example of the claim that "Bayesian methods naturally quantify uncertainty"

Generalized linear models

- Other forms of regression follow naturally from linear regression
- For example, for binary responses $Y_i \in \{0, 1\}$ we might use logistic regression

$$\text{logit}[\text{Prob}(Y_i = 1)] = \eta_i = \beta_0 + \beta_1 X_{i1} + ... + \beta_p X_{ip}$$

- The logit link is the log-odd $\text{logit}(x) = \log[x/(1 - x)]$
- Then β_j represents the increase in the log odds of an event corresponding to a one-unit increase in covariate j
- The expit transformation $\text{expit}(x) = \exp(x)/[1 + exp(x)]$ is the inverse, and

$$\text{Prob}(Y_i = 1) = \text{expit}(\eta_i) \in [0, 1]$$

Logistic regression

- Bayesian logistic regression requires a prior for β
- All of the prior we have discussed for linear regression (Zellner, BLASSO, etc) apply
- Computationally the full conditional distributions are no longer conjugate and so we must use Metropolis sampling
- The `R` function `MCMClogit` does this efficiently
- Other GLMs (e.g., Poisson regression, probit regression) are similar to implement using Bayesian methods

Random effects

- Linear regression assumes that the errors are independent
- This is invalid if data are grouped
- For example, n classrooms each have m students
- It might be reasonable to assume the classrooms are independent, but the students within a class may be dependent
- Random effects are a natural way to account for this dependence

One-way random effects model

- Say Y_{ij} is the score for student i in class j
- The random effects model is

$$Y_{ij} = \alpha_j + \varepsilon_{ij}$$

- The random effect for classroom j is α_j
- This is viewed as a random draw from the population,

$$\alpha_j \sim \text{Normal}(\mu, \tau^2)$$

- The population is described by μ and τ
- The random errors are $\varepsilon_{ij} \sim \text{Normal}(0, \sigma^2)$, independent over i and j

One-way random effects model

- Conditioned on the classroom mean α_j all observations are independent

- Marginalizing over the random effects gives

$$\text{Cor}(Y_{ij}, Y_{uv}) = \begin{cases} 0 & \text{for } j \neq v \\ \frac{\tau^2}{\sigma^2+\tau^2} & \text{for } j = v \end{cases}$$

- Therefore, in this model observations with the same classroom are correlated

One-way random effects model

- To complete the Bayesian model, we must specify priors for μ, σ^2 and τ
- A normal prior with large variance for μ is fine
- Improper priors must be used cautiously for complicated models
- A natural prior for the variances is

$$\tau^2, \sigma^2 \sim \text{InvGamma}(a, b)$$

- All full conditional distribution are conjugate and MCMC sampling is very fast

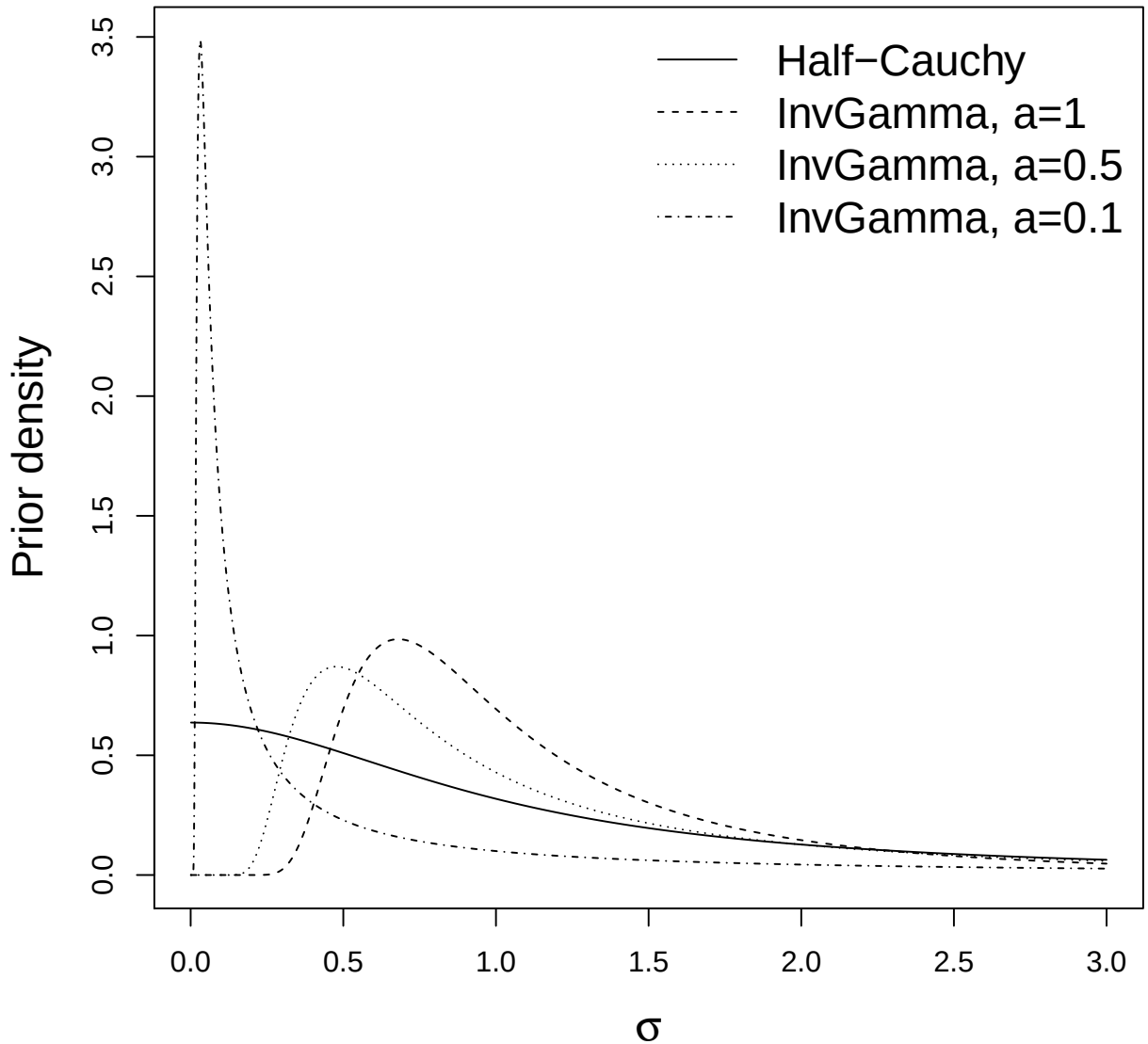
One-way random effects model

- However, under the inverse gamma prior for the variances the induced priors for σ and τ have no mass at zero
- Gelman recommends the half-Cauchy prior for the SD

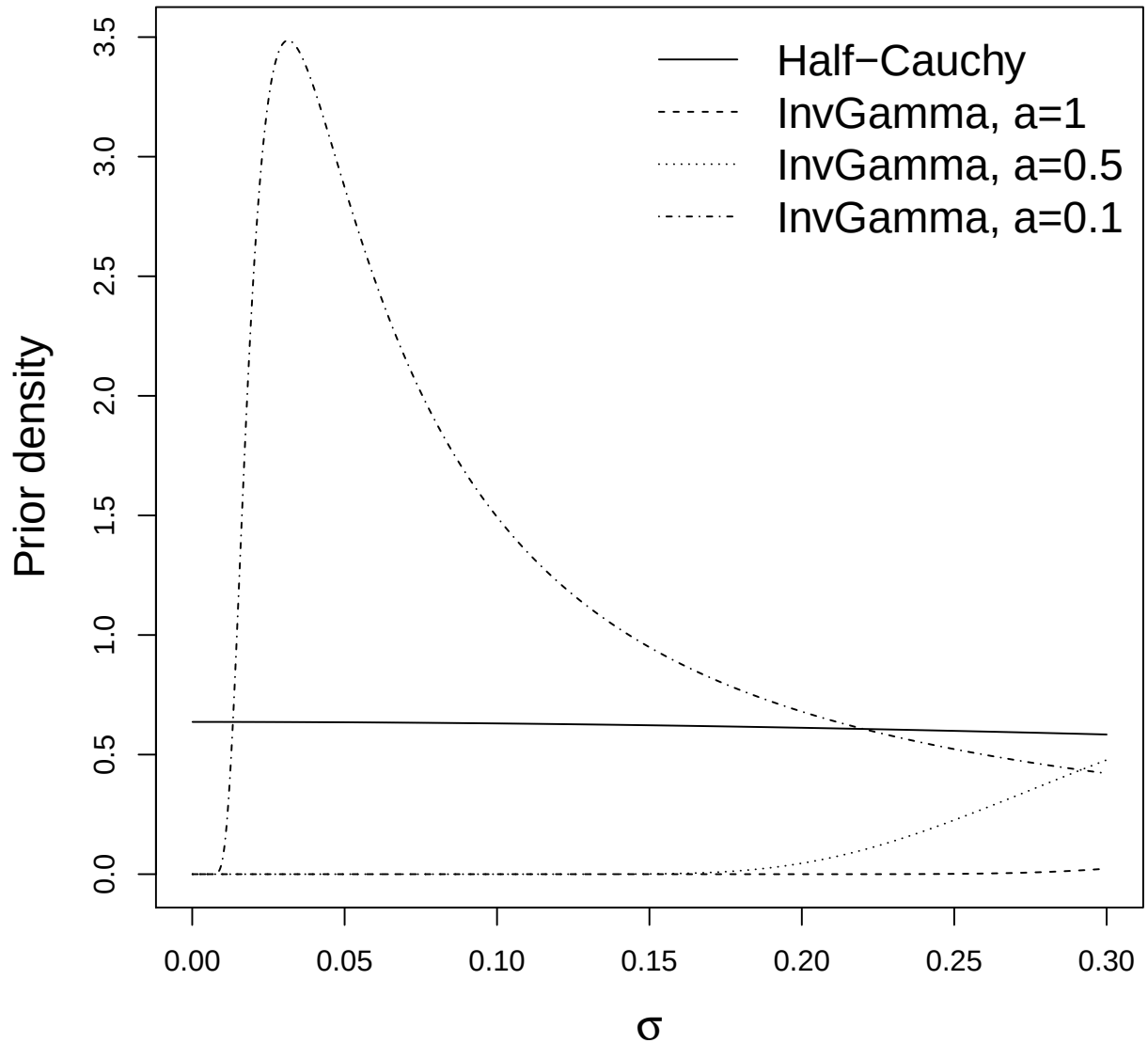
$$p(\sigma) = \frac{2}{\pi(1 + \sigma^2)},$$

  i.e., a Student-t density with 1 df restricted to be positive
- This PDF is flat around zero and has heavy tails
- This is very easy to code in JAGS
- For large sample these give similar results, but I prefer the half-Cauchy

Prior for standard deviation



Prior for standard deviation (zoomed in around 0)



Confusion about random effects

- MCMC does not distinguish between random effects and other parameters
- For example, σ, τ, μ and α_1 are all treated as random in a Bayesian analysis
- However, α_j is called a "random" effect because it represents a random draw from the fixed Normal(μ, τ^2) population of classroom means

Linear mixed models

- Consider the model

$$Y_{ij} = \beta_0 + X_{ij}\beta_1 + \alpha_j + \varepsilon_{ij}$$

  where X_{ij} is the age of student i in class j

- The regression coefficients β_0 and β_1 apply to all students are all called "fixed effects"

- The random effect is $\alpha_j \sim \text{Normal}(0, \tau^2)$

- A linear model with both fixed and random effects is called a **linear mixed model**

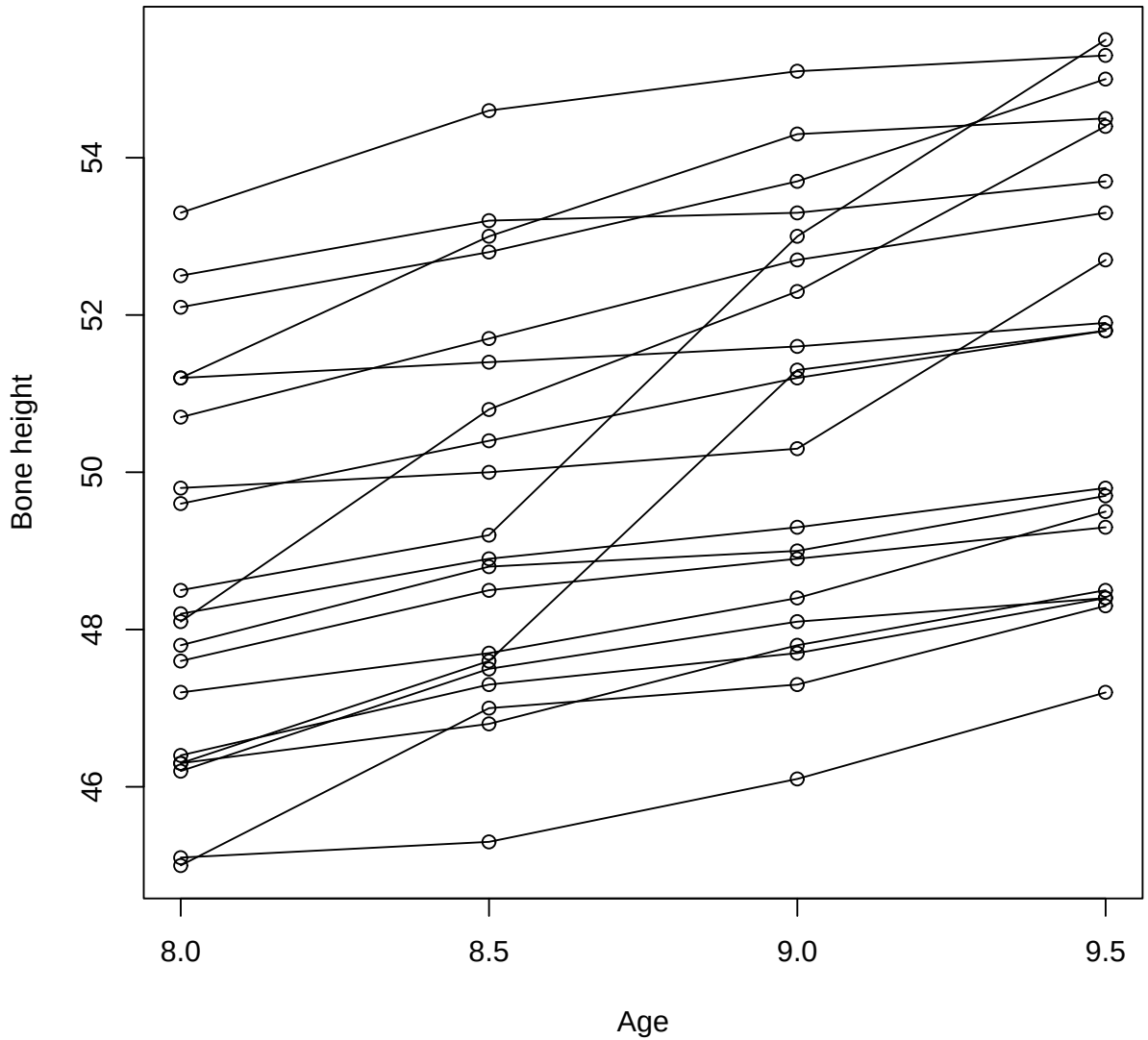
Random slopes model

- Let Y_{ij} be the j^{th} observation for subject i
- As an example, consider the data plotted on the next slide were Y_{ij} is the bone density for child i at age X_j.
- Here we might specify a different regression for each child to capture variability over the population of children:

$$Y_{ij} \sim \text{Normal}(\gamma_{0i} + X_i\gamma_{1i}, \sigma^2)$$

- $\gamma_i = (\gamma_{i0}, \gamma_{i1})^T$ controls the growth curve for child i
- These separate regression are tied together in the prior, $\gamma_i \sim \text{Normal}(\beta, \Sigma)$, which borrows strength across children
- This is a linear mixed model: γ_i are random effects specific to one child and β are fixed effects common to all children

Bone height data



Prior for a covariance matrix

- The random-effects covariance matrix is $\Sigma = \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix}$
- σ_1^2 is the variance of the intercepts across children
- σ_2^2 is the variance of the slopes across children
- σ_{12} is the covariance between the intercepts and slopes
- Prior 1: $\sigma_1^2, \sigma_2^2 \sim$ InvGamma and $\rho = \frac{\sigma_{12}}{\sigma_1 \sigma_2} \sim \text{Unif}(-1, 1)$
- Prior 2: Inverse Wishart works better in higher dimensions

Inverse Wishart distribution

- The inverse Wishart distribution is the most common prior for a $p \times p$ covariance matrix
- It reduces to the inverse gamma distribution if $p = 1$
- Say $\Sigma \sim \text{InvW}(\kappa, R)$ where $\kappa > p + 1$ and R is a $p \times p$ covariance matrix are hyperparameters
- The PDF is

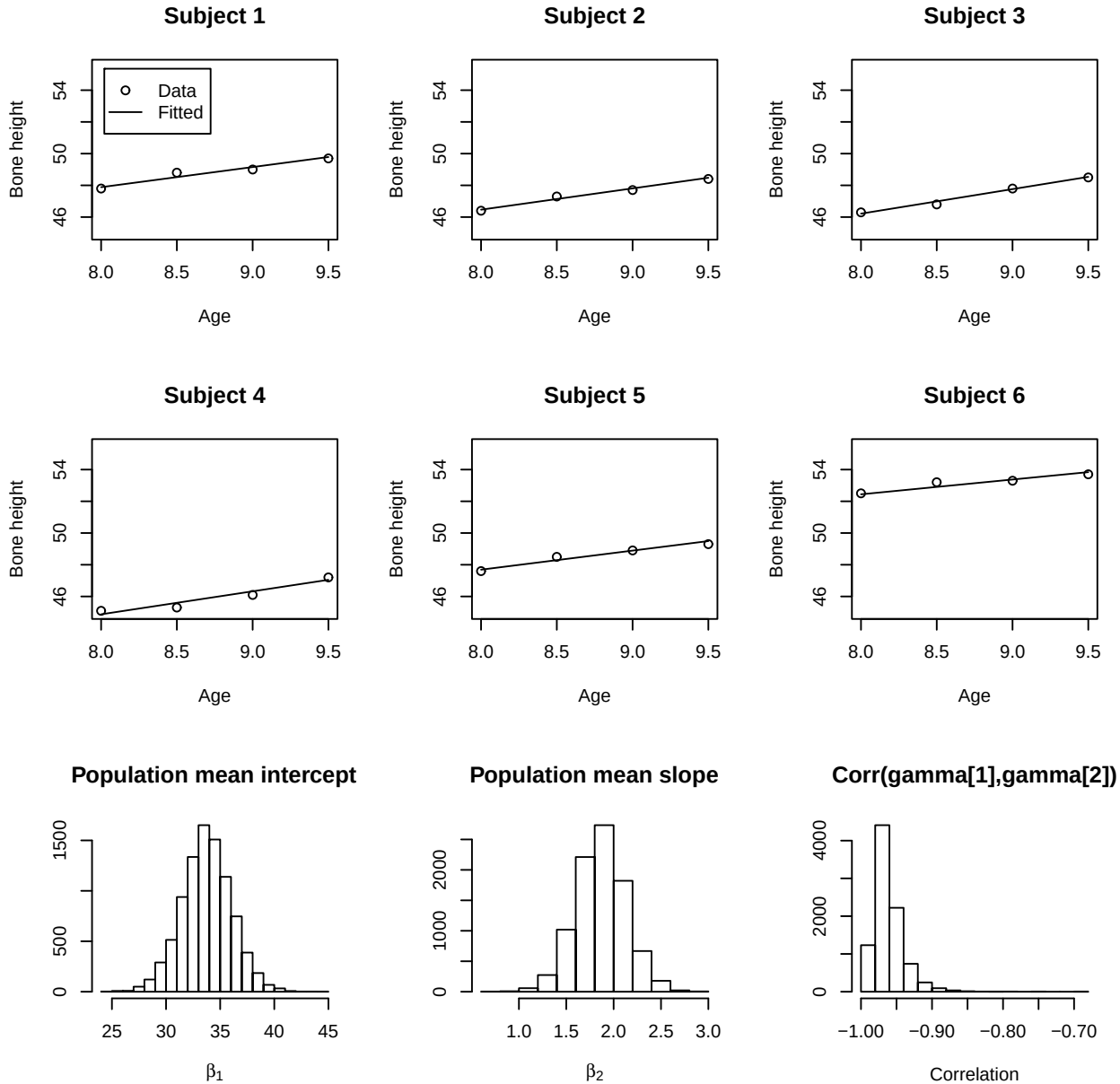
$$f(\Sigma) \propto |\Sigma|^{-(\kappa+p+1)/2} \exp\left[\frac{1}{2}\text{trace}(R\Sigma^{-1})\right]$$

- The mean is $\frac{1}{\kappa-p-1}R$

Full conditional distributions

- The hierarchical model is:
  - $Y_{ij} \sim \text{Normal}(\gamma_{0i} + X_i\gamma_{1i}, \sigma^2)$
  - $\gamma_i \sim \text{Normal}(\beta, \Sigma)$
  - $p(\beta) \propto 1$
  - $\sigma^2 \sim \text{InvGamma}(a, b)$
  - $\Sigma \sim \text{InvWishart}(\kappa, R)$
- The full conditionals are all conjugate
- JAGS code is online

Bone height data - fitted values



Linear models with correlated errors

- An alternative to using random effects to capture dependence is to model correlation directly
- For example, say the observations are collected at n different spatial locations
- Denote the measurement at lat/lon s_i as Y_i
- We might fit the model

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

  where the residual errors ε_i have spatial correlation
- A common model is

$$\text{Cov}(\varepsilon_i, \varepsilon_j) = \sigma^2 \exp(-d_{ij}/\phi)$$

- The parameter ϕ controls the exponential decay of the correlation as distance between sites, d_{ij}, increases

Linear models with correlated errors

- This is staightforward (though often slow) to fit using MCMC
- The likelihood is multivariate normal

$$\mathbf{Y}|\boldsymbol{\beta}, \sigma^2, \rho \sim \text{Normal}\left(\mathbf{X}\boldsymbol{\beta}, \sigma^2\Sigma(\phi)\right)$$

- The $n \times n$ correlation matrix $\Sigma(\phi)$ has (i, j) element $\exp(-d_{ij}/\phi)$
- This last piece is to set a prior for ϕ
- A uniform prior between 0 and the maximum distance between points is an option
- This type of modeling is also useful for time series data

Flexible regression modeling

- Nonparametric (NP) methods attempt to analyze the data by making the fewest number of assumptions as possible
- NP methods are generally more robust and flexible, but less powerful than correctly specified parametric models
- Most frequentist NP methods completely avoid specifying a model
- For example, a rank or sign test to compare two means

Non- and Semi-parametric modeling

- Bayesian methods need a likelihood in order to obtain a posterior, so you can't completely avoid specifying a model
- Bayesian NP (BNP) then attempts to specify a model that is so flexible that it almost certainly captures the true model
- One definition of the BNP model is one that has infinitely-many parameters
- In some cases, NP models are difficult conceptually and computationally, and so semiparametric models with a large but finite number of parameters are useful approximations

Parametric simple linear regression

Consider the classic parametric model:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \quad \text{where} \quad \varepsilon_i \sim \mathrm{N}(0, \sigma^2).$$

Assumptions:

1. ε_i are independent
2. ε_i are Gaussian
3. The mean of Y_i is linear in X.
4. The residual distribution does not depend on X

Alternatives:

1. Parametric alternatives such as a time series model.
2. Let $\varepsilon_i \sim F$, and place a prior on the distribution F.
3. Let $E(Y|X) = g(X)$ and put a prior on the function g.
4. Heteroskedastic regression $\mathrm{Var}(\varepsilon_i) = \exp(\alpha_0 + \alpha_1 X)$.

In 2-4 we are placing priors on functions, not parameters.

Nonparametric regression

- Let's relax the assumption of linearity in the mean.
- The mean is $g(X)$, where g is some function that relates X to $\mathrm{E}(Y|X)$.
- Parametric models include
  1. Linear: $g(X) = \beta_0 + \beta_1 X$
  2. Quadratic: $g(X) = \beta_0 + \beta_1 X + \beta_2 X^2$
  3. Logistic: $g(X) = \beta_0 + \beta_1 \frac{\exp[\beta_2 + \beta_3 X]}{1 + \exp[\beta_2 + \beta_3 X]}$.
- NP regression puts a prior on the curve $g(X)$, rather than the parameters $\beta_1, ..., \beta_p$ that determine the parametric model.

Semiparametric regression

- Semiparametric regression approximates the function g using a finite basis expansion

$$g(X) = \sum_{j=1}^{J} B_j(X)\beta_j$$

  where $B_j(X)$ are known basis functions and β_j are unknown coefficients that determine the shape of g
- Example: the cubic spline basis functions are

$$B_j(X) = (X - v_j)_+^3$$

  where v_j are fixed knots that span the range of X
- Many other expansions exist: wavelets; Fourier, etc
- Fact: A basis expansion of J terms can match the true curve g at any J points $X_1, ..., X_J$
- So increasing J gives an arbitrarily flexible model

Model fitting

- The model is $Y_i \sim \mathrm{N}\left(B_i^T \boldsymbol{\beta}, \sigma^2\right)$, where $\beta_j \sim \mathrm{N}(0, \tau^2)$ and B_i is comprised of the known basis functions $B_j(X_i)$.
- Therefore, the model is usual linear regression model and is straightforward to fit using MCMC.
- How to pick J?
- Can we have more basis functions than observations?
- What would you do if your prior was that g was probably quadratic, but you are not 100% sure about this. That is, your prior is that $g(X) \approx \beta_0 + \beta_1 X + \beta_2 X^2$.