Chapters 4.3–4.5

Advanced modeling

Outline of 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

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-odds $\text{logit}(x) = \log[x/(1 - x)]$
- 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

Steps to selecting a Bayesian GLM

1. Identify the support of the response distribution
2. Select the likelihood by picking a parametric family of distributions with this support
3. Choose a link function g that transforms the range of parameters to the whole real line
4. Specify a linear model on the transformed parameters
5. Select priors for the regression coefficients

Example of selecting a Bayesian GLM

1. Support: $Y_i \in \{0, 1, 2, ...\}$
2. Likelihood family: $Y_i \sim \text{Poisson}(\lambda_i)$
3. Link: $g(\lambda_i) = \log(\lambda_i) \in (-\infty, \infty)$
4. Regression model: $\log(\lambda_i) = \beta_0 + \sum_{j=1}^{p} X_{ij}\beta_j$
5. Priors: $\beta_j \sim \text{Normal}(0, 10^2)$

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 are likely 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

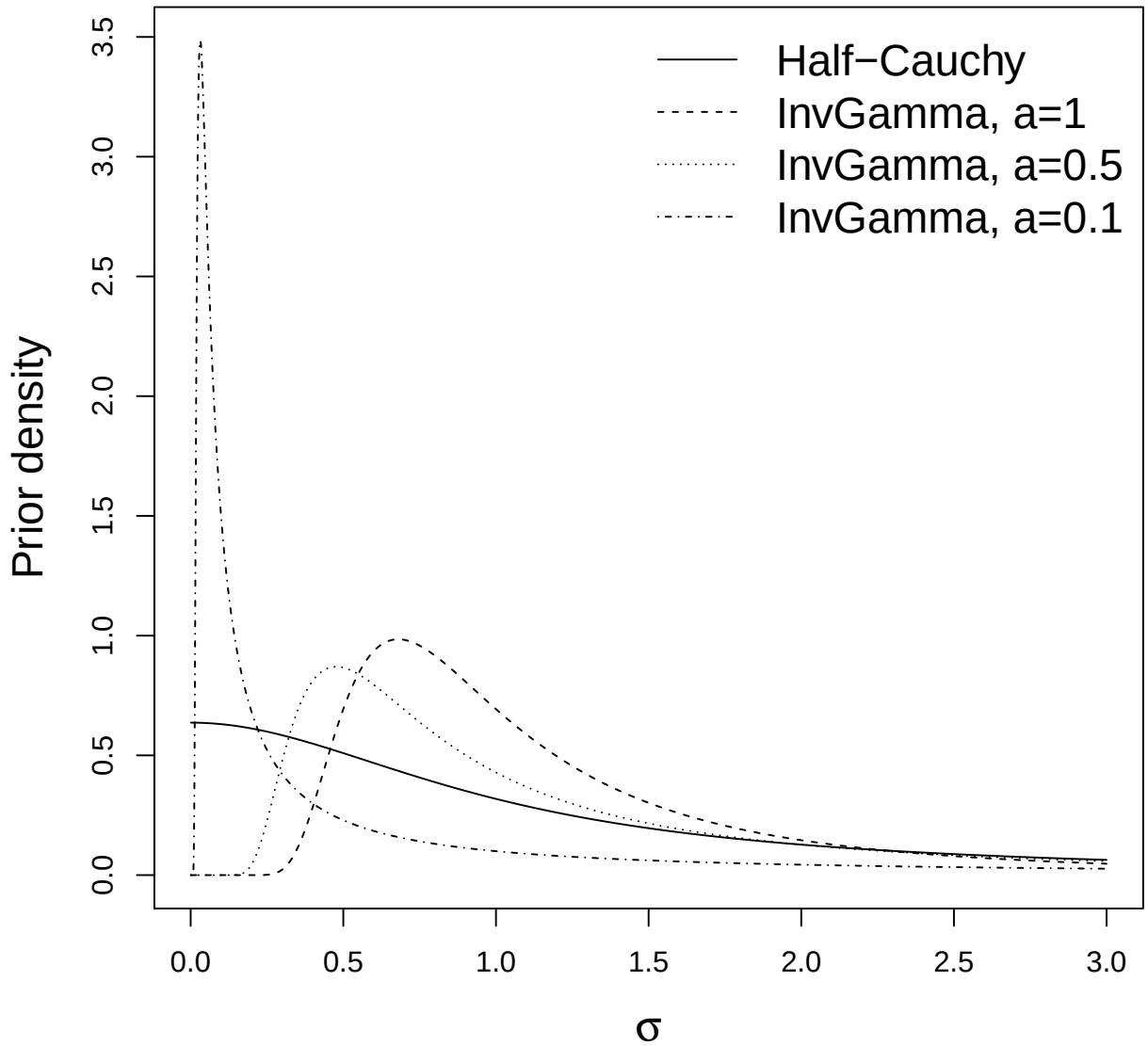
One-way random effects model

- 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

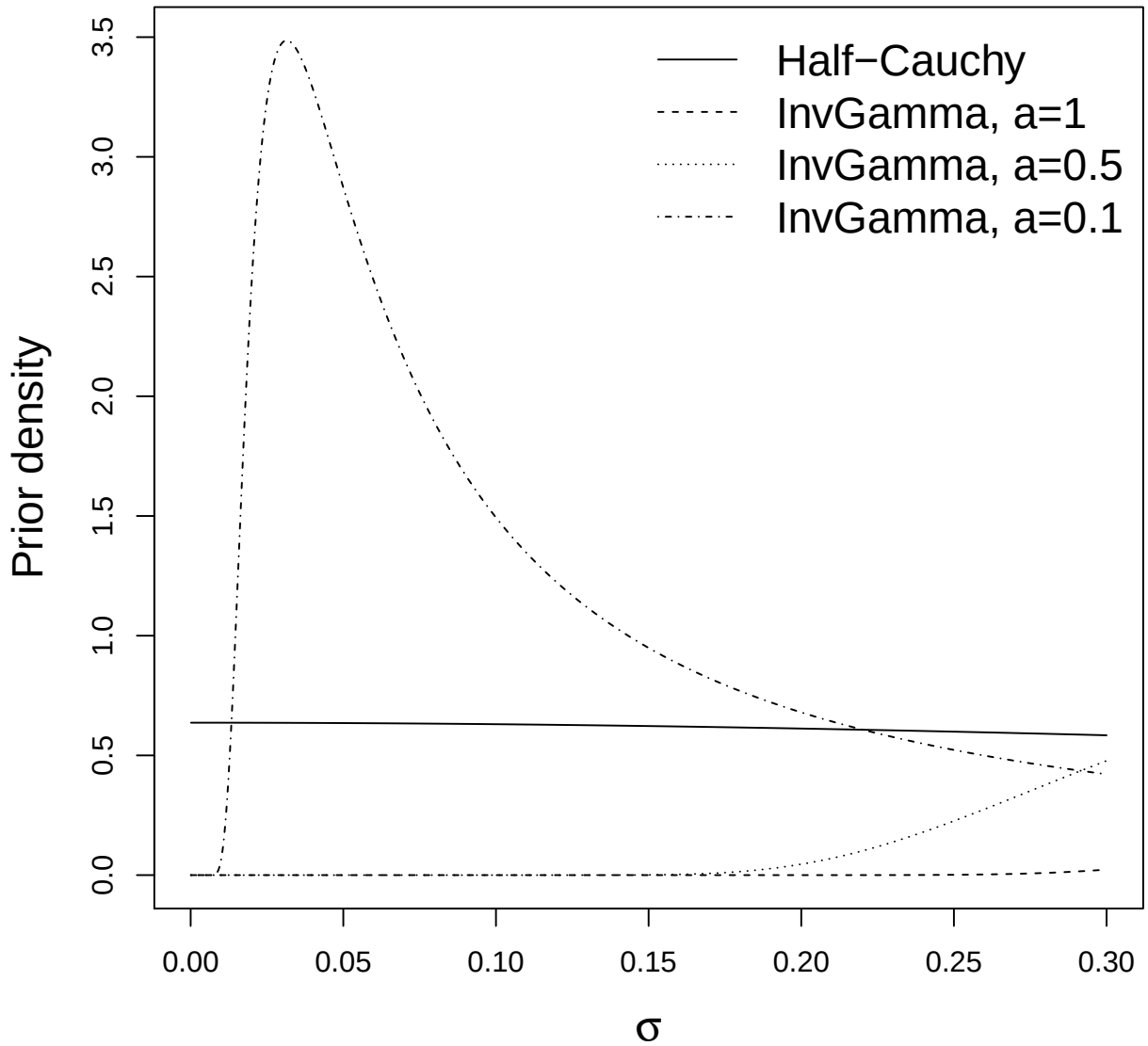
$$p(\sigma) = \frac{2}{\pi(1+\sigma^2)},$$

  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**

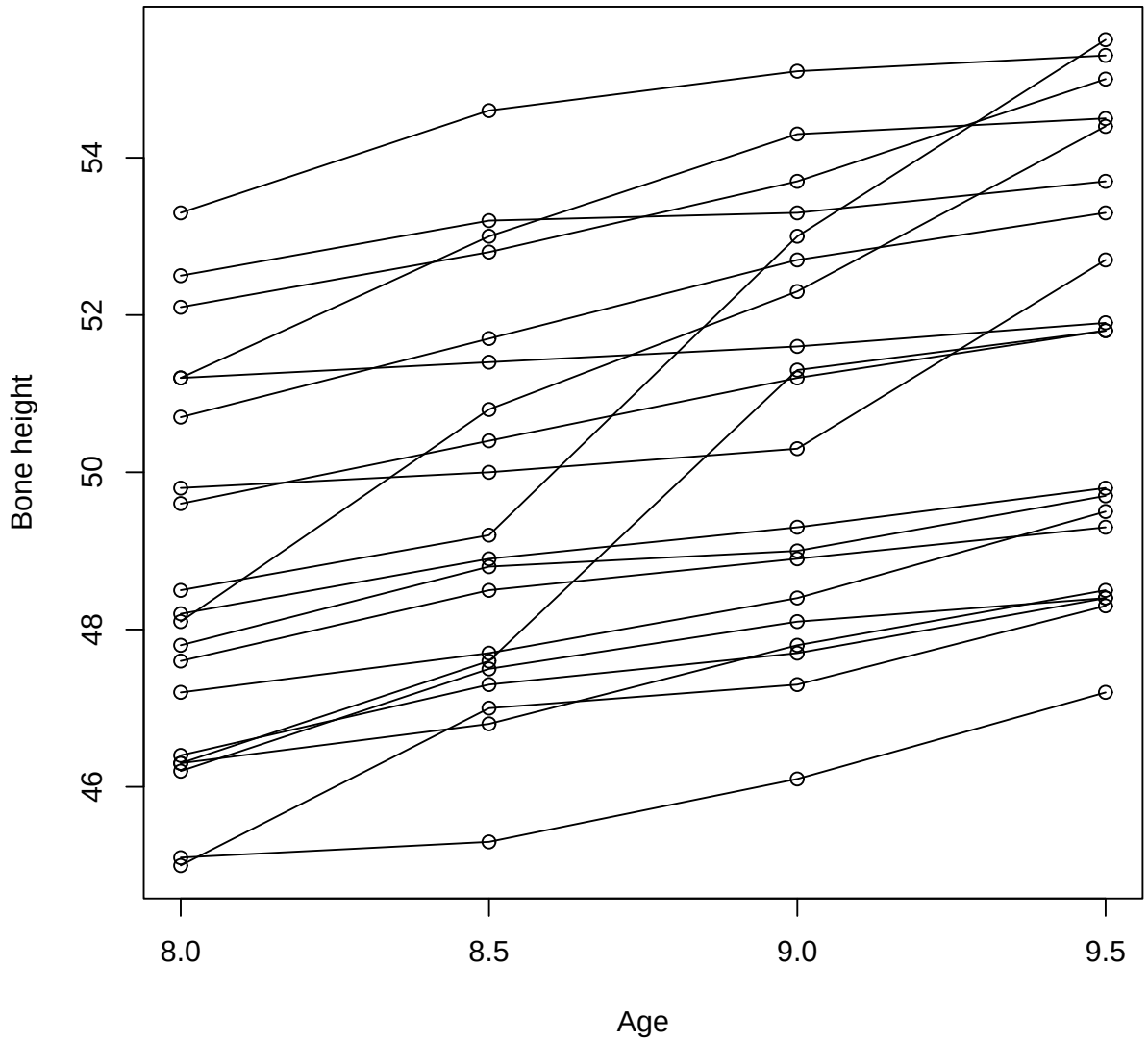
Random slopes model

- 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

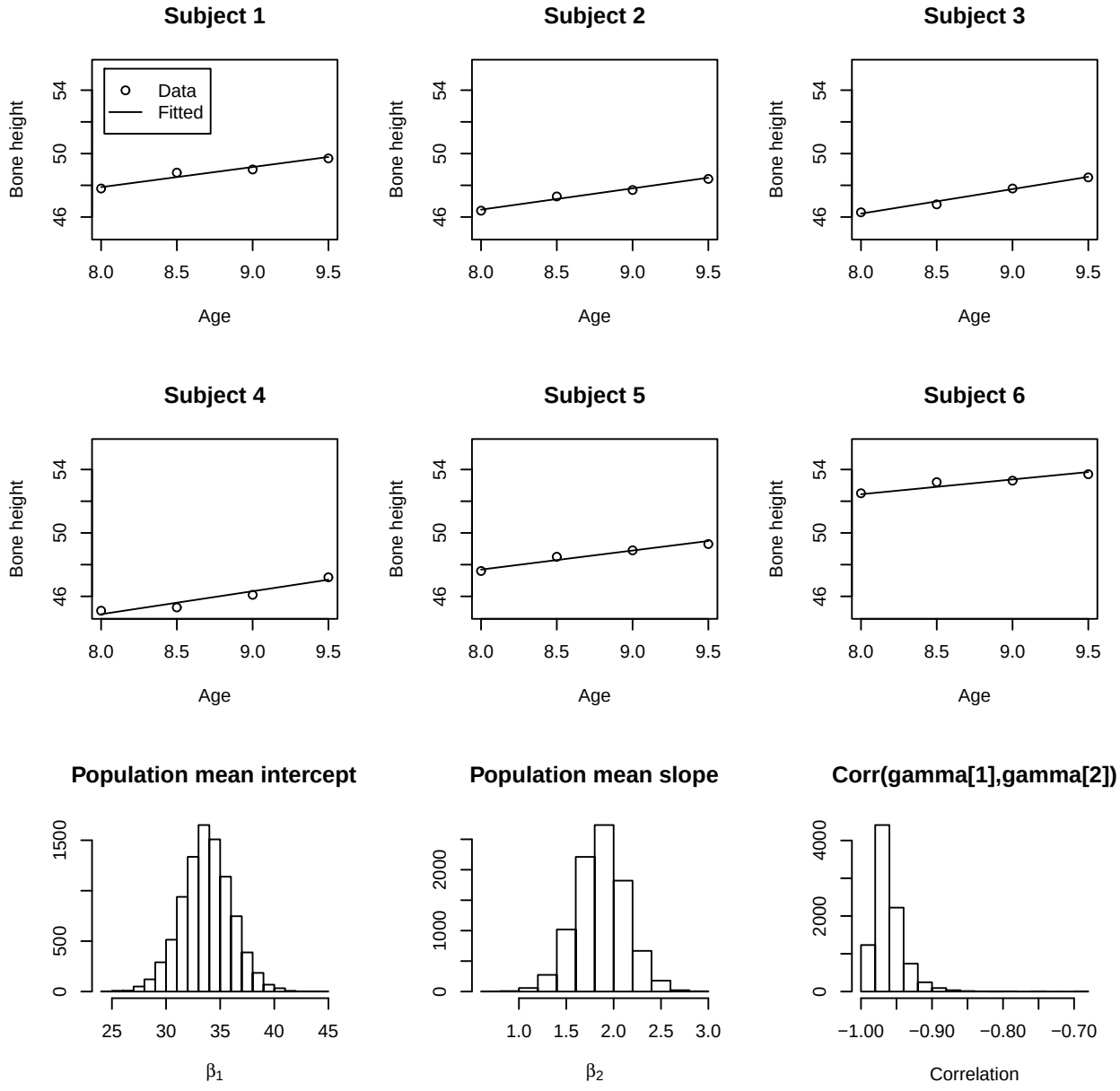
$$f(\Sigma) \propto |\Sigma|^{-(\kappa+p+1)/2} \exp\left[\frac{1}{2}\text{trace}(R\Sigma^{-1})\right]$$

- 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
- NP regression methods are also popular in machine learning because it removes the need to specify a valid model

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 non-linear regressions models include:
  1. Quadratic: $g(X) = \beta_0 + \beta_1 X + \beta_2 X^2$
  2. Exponential: $g(X) = \exp(\beta_0 + \beta_1 X)$
  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: polynomial regression takes $B_j(X) = X^j$
- 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

Semiparametric regression

- Many other expansions exist: wavelets; Fourier, neural networks, regression trees, 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
- This is allows the machine to learn patterns in the data without prior knowledge
- It also makes interpreting the results very difficult

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
- Bayesian methods are excellent for quantifying uncertainty in the fitted model and predictions
- How to pick J? Can we $J > n$?