Lecture 2: Mixed-effects models

April 5, 2016

How do we estimate things?

1. Specify a model
   – Function generating predictions
2. Identify plausible values for any unknown parameters
   – Maximize probability of observations given function
3. Assess uncertainty
   – Explore function around plausible values

Laws of probability

1. Axiom of conditional probability

$$\Pr(X, Y) = \Pr(Y|X)\Pr(X)$$

   – Often easier to specify conditional probabilities than joint probabilities

2. Law of total probability

$$\Pr(X) = \int \Pr(X, Y)\mathrm{d}Y$$

   – Used when justifying hierarchical models

Why use maximum likelihood estimation?

$$\widehat{\boldsymbol{\theta}} = \operatorname{argmax}_{\boldsymbol{\theta}}\big(L(\boldsymbol{\theta};\mathbf{y})\big)$$

Where $p(\mathbf{y}|\boldsymbol{\theta}) = L(\boldsymbol{\theta};\mathbf{y})$ is your specified probability distribution

1. Consistency (correct model)
2. Consistency (incorrect model)
3. Asymptotic normality

Likelihood statistics

Problem:

We often can't write the probability of data given parameters

Examples:

1. Tag-recapture
   - What's the probability of tagging an animal in 2008, seeing it again in 2010 and 2011, and then never seeing it again?
2. Time-series
   - What's the probability distribution for escapement of chinook salmon in the snake river in 2011, given that you've sampled escapement from 1980-2010?
3. Occupancy
   - Three volunteers look for an endangered butterfly at a site, and only two find it. These volunteers sample at a new site, and none see the butterfly. What is the probability that is present but wasn't detected?

Likelihood statistics

Solution:

- Introduce "latent" variables

$$\Pr(y, \varepsilon|\theta) = \Pr(y|\theta, \varepsilon) \Pr(\varepsilon)$$

  - where ε is a unobserved random variable
  - $\Pr(\varepsilon)$ is a "prior" or "hyper-distribution" for latent variables
  - ε is sometimes called "augmented data"
    - Left side of the joint-likelihood
- Calculate the marginal likelihood of parameters when integrating across random effects

$$\Pr(y|\theta) = \int \Pr(y|\theta_1, \varepsilon) \Pr(\varepsilon|\theta_2) \, \mathrm{d}\varepsilon$$

  - *Marginalize – take a weighted average of likelihoods, where weights are given according to the probability of random effects,* $\Pr(\varepsilon|\theta_2)$

Definitions

Term	Definition
Random effect	Coefficient that is “exchangeable” with one or more other coefficients
Hyperdistribution	Distribution for “exchangeable” random effects
Exchangeable	No information is available to distinguish between residual variability in random effects
Fixed effect	Coefficient that is not exchangeable with others, and which hence is estimated without a hyperdistribution
Mixed-effect model	Model with both fixed and random effects

Why would you make a hierarchy of parameters

1. Biological intuition – Formulate models based on knowledge of constituent parts (Burnham and Anderson 2008)
2. Variance partitioning – Separate different sources of variability (e.g., measurement errors!)
3. Shrinkage – Often improve precision from assuming parameters arise from a distribution

Stein's paradox

- Pooling parameters towards a mean will be more accurate on average (Efron and Morris 1977)
  - Say we have a batter with 100 at bats, and 35 hits
    - x: Batting average (x=0.35)
    - z: Best prediction of future probability of hit (z=0.35)
  - Say we have three batters
    - $\mathbf{x}$: Batting average ($\mathbf{x} = (0.3, 0.35, 0.4)^{\mathrm{T}}$)
    - $\mathbf{z}$: Best prediction of future probability of hit

$$\mathbf{z} = c\bar{x} + (1 - c)\mathbf{x}$$

  - Where c is the magnitude of shrinkage, $0 < c < 1$

Stein's paradox

- Why is this a paradox?
  - No reference to things being pooled!
  - Say we have three batters, and the proportion of Japanese-made cars
    - $\mathbf{x}$: Batting and car-sales averages ($\mathbf{x} = (0.3,0.35,0.4,0.2)^{\mathrm{T}}$)
    - $\mathbf{z}$: Best prediction of future probability of hit

$$\mathbf{z} = c\bar{x} + (1-c)\mathbf{x}$$

      - Where c is the magnitude of shrinkage, $0 < c < 1$
  - Works regardless of definition of $\mathbf{x}$
    - Contamination leads to lower shrinkage on average, $c \to 0$

Likelihood statistics

Predicting random variables

- *Empirical Bayes* – Predict random variables ε via fixed values for θ

$$\hat{\varepsilon} = \text{argmax}_{\varepsilon}\left(\Pr\left(y \middle| \hat{\theta}_1, \varepsilon\right) \Pr\left(\varepsilon \middle| \hat{\theta}_2\right)\right)$$

  - Where $\hat{\theta}$ is the maximum likelihood estimate of fixed effects θ

- Fisheries has historically used "penalized likelihood" (Ludwig and Walters 1981)

$$(\hat{\theta}, \hat{\varepsilon}) = \text{argmax}_{\theta,\varepsilon}(\Pr(y|\theta_1, \varepsilon) \Pr(\varepsilon|\theta_2))$$

- … but this precludes estimating θ_2

Likelihood statistics

Estimation

$$L(\theta; y) = \Pr(y|\theta) = \int \Pr(y|\theta_1, \varepsilon) \Pr(\varepsilon|\theta_2)\, \mathrm{d}\varepsilon$$

where

- $L(\theta|y)$ is the likelihood
- $\Pr(\varepsilon|\theta_2)$ is the hyper-distribution
- $\Pr(y|\theta_1, \varepsilon) \Pr(\varepsilon|\theta_2)$is the "penalized likelihood"

How do we estimate the marginal likelihood?

1. "Hierarchical Bayes"
   - Generally involves MCMC
   - Already integrating across parameters, so integrates across latent variables automatically
2. "Maximum marginal likelihood"
   - Use the "Laplace approximation" to approximate integral
   - Use alternating estimation of fixed and random effects
     - "Inner optimization" – Optimize random effects given fixed effects
     - "Outer optimization" – Optimize fixed effects given random effects

Likelihood statistics

Laplace approximation

- Define joint log-likelihood:

$$f(\theta,\varepsilon) = \log(\Pr(y|\theta_1,\varepsilon)\Pr(\varepsilon|\theta_2))$$

- Taylor series expansion of joint log-likelihood

$$f(\varepsilon|\theta) \approx f(\hat{\varepsilon}|\theta) + f'(\hat{\varepsilon}|\theta)(\hat{\varepsilon}-\varepsilon) + \frac{1}{2}f''(\hat{\varepsilon}|\theta)(\hat{\varepsilon}-\varepsilon)^2$$

- Evaluate Taylor series around "inner maximum"

$$\hat{\varepsilon} = \mathrm{argmax}_\varepsilon\big(f(\theta,\varepsilon)\big)$$

- Approximate joint likelihood via Taylor series expansion

$$\Pr(y|\theta_1,\varepsilon)\Pr(\varepsilon|\theta_2) = e^{f(\varepsilon|\theta)} \approx e^{f(\hat{\varepsilon}|\theta)-\frac{1}{2}|f''(\hat{\varepsilon})|(\hat{\varepsilon}-\varepsilon)^2}$$

Likelihood statistics

Laplace approximation

- Approximate joint likelihood via Taylor series expansion

$$\Pr(y|\theta_1,\varepsilon)\Pr(\varepsilon|\theta_2) = e^{f(\varepsilon|\theta)} \approx e^{f(\hat{\varepsilon}|\theta)-\frac{1}{2}|f''(\hat{\varepsilon})|(\hat{\varepsilon}-\varepsilon)^2}$$

- Integrate both sides

$$\int \Pr(y|\theta_1,\varepsilon)\Pr(\varepsilon|\theta_2)\,\mathrm{d}\varepsilon = \int e^{f(\varepsilon|\theta)}\mathrm{d}\varepsilon$$

$$\int \Pr(y|\theta_1,\varepsilon)\Pr(\varepsilon|\theta_2)\,\mathrm{d}\varepsilon \approx e^{f(\hat{\varepsilon}|\theta)} \int e^{-\frac{1}{2}|f''(\hat{\varepsilon})|(\hat{\varepsilon}-\varepsilon)^2}\mathrm{d}\varepsilon$$

- Looks like a normal distribution
  - $\hat{\varepsilon}$ is the mean of the normal distribution
  - $f''(\hat{\varepsilon})$ is the hessian of the normal distribution ($f''(\hat{\varepsilon}) = \Sigma^{-1}$)

$$\mathit{Normal\ PDF}\colon \Pr(\varepsilon|\mu,\Sigma) = \frac{1}{(2\pi)^{k/2}|\Sigma|^{1/2}}\exp\left(\frac{-(\varepsilon-\mu)^{\mathrm{T}}\Sigma^{-1}(\varepsilon-\mu)}{2}\right)$$

Chi-squared example

$$\Pr(x) = \frac{x^{\frac{k}{2}-1}e^{\frac{-x}{2}}}{c}$$

Defining the log-likelihood

$$\log(\Pr(x)) \equiv f(x)$$

Taking derivatives:

$$f(x) \propto \left(\frac{k}{2} - 1\right)\log(x) - \frac{x}{2}$$

$$f'(x) \propto \left(\frac{k}{2} - 1\right)x^{-1} - \frac{1}{2}$$

$$f''(x) \propto -\left(\frac{k}{2} - 1\right)x^{-2}$$

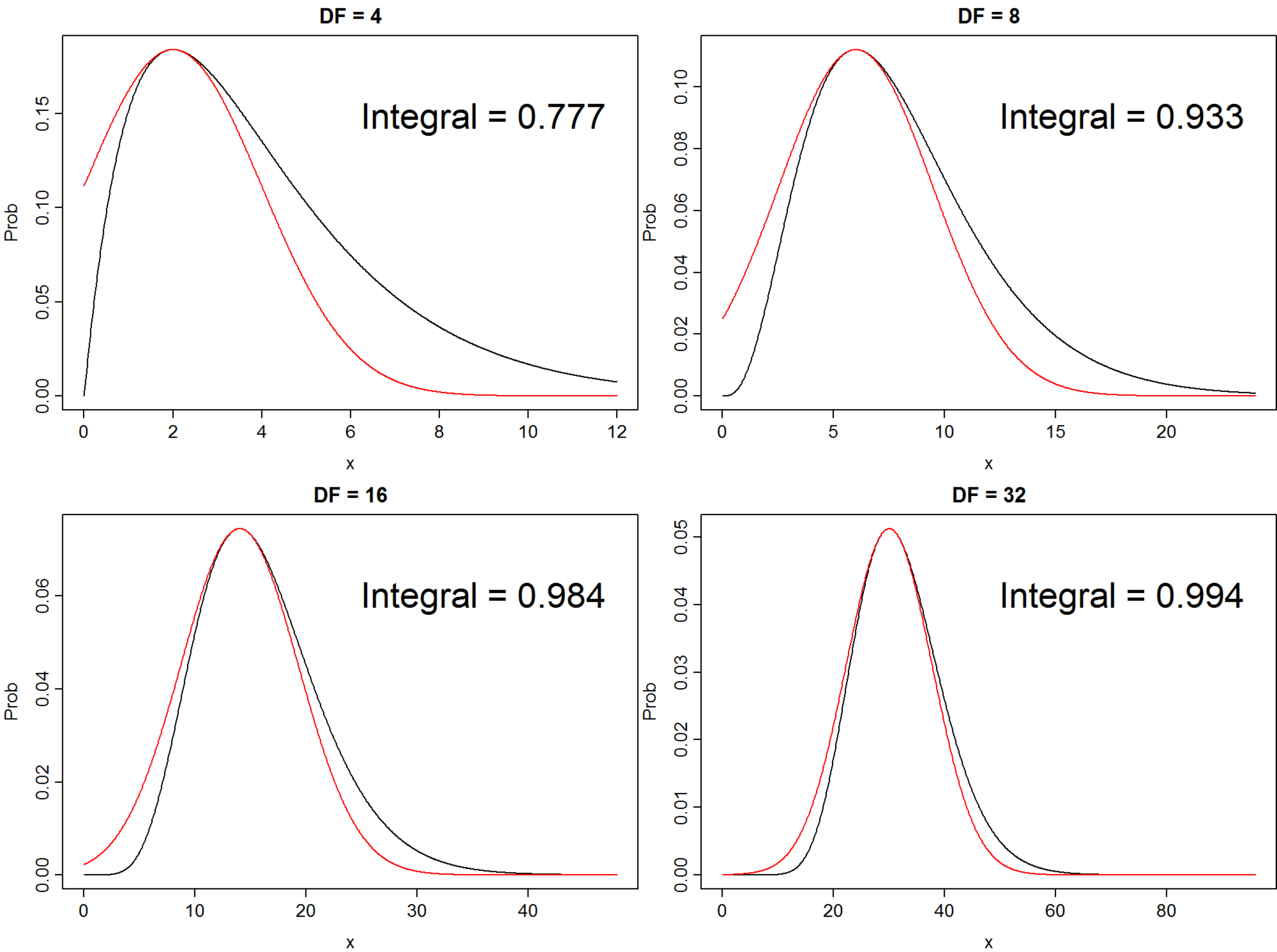
Solving for mode and Hessian:

$$f'(x) = 0 \quad \rightarrow \quad \hat{x} = k - 2$$

$$f''(\hat{x}) = -\left(\frac{1}{2(k-2)}\right)$$

Hence:

$$\Pr(x) \propto Normal(k-2, 2(k-2))$$



DF = 4
Integral = 0.777
DF = 8
Integral = 0.933
DF = 16
Integral = 0.984
DF = 32
Integral = 0.994
Prob
x

Likelihood statistics

Bottom line

$$\ln L(\theta; y) \cong \log(\Pr(y, \varepsilon|\theta)) - \frac{1}{2}\log(|\mathbf{H}|)$$

– Where

$$\Pr(y, \varepsilon|\theta) = \Pr(y|\theta_1, \varepsilon)\Pr(\varepsilon|\theta_2)$$

– And

$$\mathbf{H} = \frac{\partial^2}{\partial \varepsilon^2}(\log(\Pr(y, \varepsilon|\theta)))$$

- Definitions
  - $\log(L(\theta; y))$ is the marginal log-likelihood
  - $\Pr(y, \varepsilon|\theta)$ is the joint likelihood
  - $|\mathbf{H}|$ is the determinant of the Hessian matrix

Steps during optimization

1. Write joint log-likelihood $\Pr(y, \varepsilon|\theta)$ in CPP file

$$f(\theta, \varepsilon) = \log(\Pr(y|\theta_1, \varepsilon)\Pr(\varepsilon|\theta_2))$$

2. Choose initial values for fixed θ_0 and random ε_0

3. "Inner optimization" – Optimize random effects with $\theta_0$held constant

$$\hat{\varepsilon} = \mathrm{argmax}_{\varepsilon}\big(f(\theta_0, \varepsilon)\big)$$

4. Calculate Laplace approx. for marginal likelihood of fixed effects

$$\ln L(\theta_0; y) \cong f(\theta_0, \hat{\varepsilon}) - \frac{1}{2}\log(|\mathbf{H}|)$$

   – TMB also provides the gradient of the penalized likelihood with respect to fixed effects

5. "Outer optimization" – Repeat steps 2-3

   – Outer optimization is done in R using the function value and gradient provided by TMB

Generalized linear mixed model

1. Specify distribution for response variable

$$c_i \sim \mathrm{Poisson}(\lambda_i)$$

2. Specify function for expected value

$$\lambda_i = \exp(\beta_0 + \boldsymbol{\beta}\mathbf{x}_i + \boldsymbol{\varepsilon}\mathbf{z}_i)$$

3. Specify distribution for random effects

$$\varepsilon_i \sim Normal(0, \sigma_u^2)$$

= General linear model + mixed effect(s)

Shrinkage

- Suppose you have density samples $d_{i,j}$ for site *j*
  - You assume the following model:

$$d_{i,j} \sim Normal\left(0, \sigma^2_{within}\right)$$

$$d_j \sim Normal\left(\mu, \sigma^2_{among}\right)$$

  - Three fixed effects (σ^2_{within}, σ^2_{among}, and μ)
  - n_j random effects (d_j)

Shrinkage

- Estimated random effects are weighted average of:
  - Optimal predictor

$$\hat{d}_j = c\bar{d} + (1 - c)\bar{d}_j$$

  - Where

$$c = \frac{\widehat{w}_1}{\widehat{w}_1 + \widehat{w}_{2,j}}$$

$$\widehat{w}_1 = \frac{1}{\sigma^2_{among}}$$

$$\widehat{w}_2 = \frac{n_j}{\sigma^2_{within}}$$

  - And where
    - σ^2_{among} is the variance in d_j among groups
    - σ^2_{within} is the variance of density samples within a given group
    - $\bar{d}_j$ is the sample mean for group j
    - $\bar{d}$ is the sample mean for $\bar{d}_j$ for all groups

[Look at code]

Likelihood statistics

Separability

- What if different components of the model are statistically independent?

$$\Pr(y|\theta_1, \varepsilon)\Pr(\varepsilon|\theta_2) = \prod_{i=1}^{N} \Pr(y|\theta_1, \varepsilon_i)\Pr(\varepsilon_i|\theta_2)$$

- Examples:
  - Overdispersed samples

$$C_i \sim Poisson(\lambda_i)$$
$$\log(\lambda_i) \sim Normal(\mu, \sigma^2)$$

  - Each λ_i is independent conditional on μ, σ^2

$$\Pr(C|\lambda)\Pr(\lambda|\mu, \sigma^2) = \prod_{i=1}^{N} \Pr(C_i|\lambda_i)\Pr(\lambda_i|\mu, \sigma^2)$$

Likelihood statistics

Separability

- Then we can factor the integral

$$\int \Pr(y|\theta_1, \varepsilon) \Pr(\varepsilon|\theta_2)\, \mathrm{d}\varepsilon = \prod_{i=1}^{N} \int \Pr(y|\theta_1, \varepsilon_i) \Pr(\varepsilon_i|\theta_2)\, \mathrm{d}\varepsilon_i$$

  – Where we replace a *N*-dimensional integral with *N* 1-dimenstional integrals

Uses

1. Meta-analysis: species are often independent
2. Time series: years are often “conditionally” independent