Bayesian Statistics: Techniques & Models

Wk 1: Statistical Modeling & Mc

↳ Real models are complicated

- MCMC: Allows sampling of post. w/ no analytical solutions
- Expands models possible for bayesians

→ Statistical Modeling

Stats - process of planning, collecting & analyzing data to draw conclusions

↳ Many applied fields

Model ∈ imitate & approx data gen. process

given an estimate

← How will number change if measured again?

← What patterns are in the noise?

Objectives

1. Quantify uncertainty
2. Inference
3. Measure support for hypo
4. Prediction

· Modeling Process

↳ Stat. model process.

① Understand the problem
② Plan & collect data
③ Explore data
④ Postulate Model
⑤ fit model (This class: bayesian)
⑥ Check model
⑦ Iterative
⑧ Use model

→ Bayesian Modeling

data
↳ heights of Men, $n = 15$

Models can have increasing complexity

↳ Model... $y_i = \mu + \varepsilon_i$

ε_i error, iid norm

ie $\varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$ $i = 1 \dots N$

$\therefore y_i \overset{iid}{\sim} N(\mu, \sigma^2)$

↑ frequentist model

How do we turn freq. model into bayesian model?

↳ add priors to hyperparameters

ie, μ & σ^2 are RVS

Likelihood ← describes how data is generated

$$p(y|\theta)$$

prior ← describes probability of parameters

$$p(\theta)$$

posterior ← beliefs of parameters given y

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

$$= \int p(y,\theta)\,d\theta$$

↑ get joint dist & marginalise over θ

→ Model specification

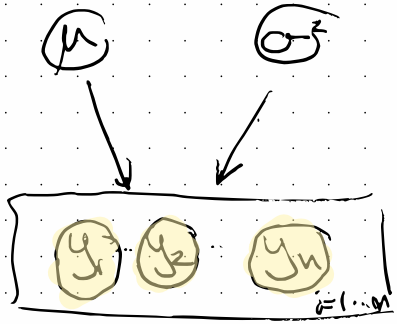
$y_i \mid \mu, \sigma^2 \overset{iid}{\sim} N(\mu, \sigma^2)$ ← priors

if μ or σ^2 are unknown

$\text{prior}(\mu) = N(\mu_0, \sigma_0^2)$

$\text{prior}(\sigma^2) = IG(\nu_0, \beta_0)$ ← inverse gamma

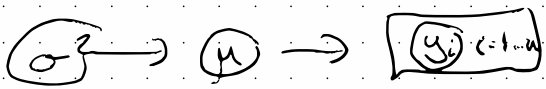
↓ convert to graphic representation



⊗ = RV

○ = Observed RV

↳ you can even make μ dependant on σ



S.t. $y_i \mid \mu, \sigma^2 \overset{iid}{\sim} N(\mu, \sigma^2)$

$\mu \mid \sigma^2 \sim N(\mu_0, \frac{\sigma^2}{\omega_0})$

$\sigma^2 \sim IG(\nu_0, \beta_0)$

Hence, joint dist. for posterior =

$$p(y_1, \dots, y_n, \mu, \sigma^2) = p(y_1, \dots, y_n \mid \mu, \sigma^2) * p(\mu \mid \sigma^2) * p(\sigma^2)$$

$$= \prod_{i=1}^{n} \left[N(y_i \mid \mu, \sigma^2) \right] \times N\left(\mu \mid \mu_0, \frac{\sigma^2}{\omega_0}\right) * IG(\sigma^2 \mid v_0, \beta_0)$$

likelihood of data ← priors

thus, equal to numerator in bayes theorem

↳ posterior is proportional to this.

If this value can be integrated, then work is done.

↳ problem, often non-analytical solutions

→ Non conjugate models

$n = 10$, measuring % change in employee count for a company

$$y_i | \mu \overset{iid}{\sim} N(\mu, 1)$$

$$\mu \sim t(0, 1, 1)$$

$$p(\mu | y_1 \dots y_n) \propto \prod_{i=1}^{n} \left[\frac{1}{\sqrt{2\pi}} e^{\frac{-(y_i - \mu)^2}{2}} \right] \frac{1}{\pi(1+\mu^2)}$$

$$\propto e^{-\frac{1}{2}\sum (y_i - \mu)^2} * \frac{1}{1+\mu^2}$$

↳ expand & simplify

$$\propto \boxed{\frac{e^{n(\bar{y}\mu - \mu^2/2)}}{1+\mu^2}}$$

(we can find posterior dist $g(\theta)$ but cannot be $\int d\theta$ easily.

→ Prevented bayesian methods from being mainstream for a long time

→ Monte carlo Estimation

↳ calculating parameters via rvs

Say $\Theta \sim \text{Gamma}(a, b)$ st $a = 2$, $b = \frac{1}{3}$

LOTUS

$$E(\Theta) = \int_0^\infty \Theta P(\Theta) = \int_0^\infty \Theta \frac{b^a}{\Gamma(a)} \Theta^{a-1} e^{-b\Theta} d\Theta$$

$$= \frac{a}{b}$$

or Simulate Θ_i^*, $i = 1 \dots n$

$$\text{Avg}(\Theta_i^*) \to E(\Theta) \text{ as } n \to \infty$$

$$= \frac{1}{m} \sum_{i=1}^{m} \Theta_i^* = \bar{\Theta}^*$$

$$\text{Var}(\Theta_i^*) = \int_0^\infty (\Theta - E(\Theta))^2 p\, \Theta\, d\Theta$$ ✓

continued...

Say we have $h(\Theta)$ st $\int h(\Theta) p(\Theta) d\Theta$

$$= E(h(\Theta)) \simeq h(\bar{\Theta}_i^*)$$

ie. $E[h(\theta)] = \int_0^\infty I_{0<\theta<5}(\theta) p(\theta)\, d\theta$

$h(\theta) = I_{0<\theta<5}(\theta)$ $= \int_0^5 1 \cdot p(\theta)\, d\theta$

$= Pr[0<\theta<5]$

$\simeq \frac{1}{m} \sum_{i=1}^{M} I_{0<\theta<5}(\theta_i^*)$ ← people do this all the time

↳ by CLT, $\bar{\theta}^* \sim N(E(\theta), \frac{var(\theta)}{m})$

$\widehat{var}(\theta) = \frac{1}{m} \sum_{i=1}^{m} (\theta_i^* - \bar{\theta}_i^*)^2$

thus

$\sqrt{\frac{\widehat{var}(\theta)}{m}}$ = standard error

↳ example

$y|\phi \sim Binom(10, \phi)$

$\phi \sim Beta(2,2)$

↳ thus $p(y, \phi) = p(\phi) p(y|\phi)$

How do we simulate from this posterior?

repeat:

1. ϕ_i^* from Beta(2,2)
2. w/ given ϕ_i^*, draw $y_i^* \sim Bin(10, \phi_i^*)$
3. generates (y_i^*, ϕ_i^*)

ϕ_i^*: draws from posterior distribution

→ Markov Chains

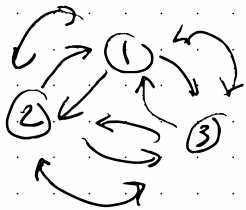
↳ Markov chains have sequences of observations w/ the property that $P(X_t | X_{t-1})$

ie $P(X_1, X_2, \ldots X_t) = P(X_1) P(X_2|X_1) \ldots P(X_t | X_{t-1})$

↳ ie you have a coin & as you flip it you increment/decrement a number

Where this number goes is called a random walk

If we have multiple hidden states, we can calculate go the prob of one to another w/ a transition matrix



$\Rightarrow$

$$\begin{bmatrix} 1\to1 & 1\to2 & 1\to3 \\ 2\to1 & 2\to2 & 2\to3 \\ 3\to1 & 3\to2 & 3\to3 \end{bmatrix}$$

A or Q

← rows of matrix add to 1

Applying A repeatedly will sometimes make it converge to steady state ↳ found w/ Em Alg

→ Metropolis Hastings

↳ Allows us to sample from prob dist using MC w/ stationary dist = target dist
ie our posterior

ie $p(\theta) \propto g(\theta)$ ... $p(\theta) = c\, g(\theta)$
↑ c is unknown

Metropolis Hastings

1. Select θ_0
2. for $i \dots n$:
   a) draw candidate $\theta^* \sim q(\theta^* | \theta_{i-1})$
   b) $\alpha = \dfrac{g(\theta^*)/q(\theta^* | \theta_{i-1})}{g(\theta_{i-1})/q(\theta_{i-1} | \theta^*)}$

      $= \dfrac{g(\theta^*)\, q(\theta_{i-1} | \theta^*)}{g(\theta_{i-1})\, q(\theta^* | \theta_{i-1})}$

   c) if $\alpha \geq 1$, accept θ^*, $\theta_i := \theta^*$
      if $0 < \alpha < 1$, " , $\theta_i := \theta^*$ w/ prob α
      else reject θ^* & set $\theta_i := \theta_{i-1}$ w/ prob $1-\alpha$

is a markov chain

Notes on candidate generating distribution

$q(\theta^* | \theta_{i-1})$

↳ generates candidates to step toward

we can use different distributions.

→ Random Walk Met-Hastings

uses normal distribution

$\therefore q(\theta_{i-1} | \theta^*) = q(\theta^* | \theta_{i-1})$,

$\alpha = \frac{g(\theta^*)}{g(\theta_{i-1})}$ ← evaluating here makes sure q evaluates to p

This should only be done if $q \sim N$ approximates p well.

Consider this problem,

$\Theta \in \{$ fair, loaded $\}$ (coin example)

let's also say
0 is fair
1 is loaded

Prior, $P(\Theta = \text{loaded}) = .6$

likelihood, $f(x|\theta) = \binom{5}{x}\left(\frac{1}{2}\right)^5 I_{\{\theta = \text{fair}\}} + \binom{5}{x}(.7)^x(.3)^{5-x} I_{\{\theta = \text{loaded}\}}$

post, $P(\Theta = \text{loaded} \mid X = 2) = .388$

↑ can we simulate this w/ mcmc?

1) start @ either $\Theta_0 = \text{fair}$ or $\Theta_0 = \text{loaded}$

2) for $i = 1, \ldots, m$

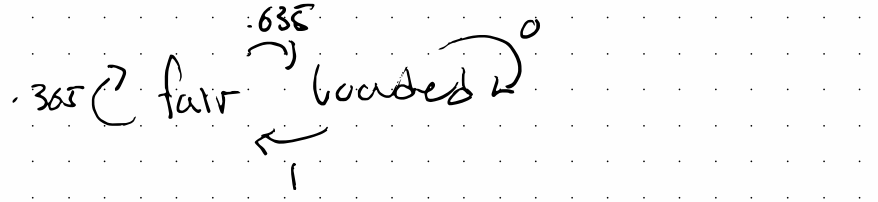
a) Propose Θ^* to be the other state as Θ_{i-1}

b) $\alpha = \dfrac{g(\Theta^*)/q(\Theta^*|\Theta_{i-1})}{g(\Theta_{i-1})/q(\Theta_{i-1}|\Theta^*)}$

$= \dfrac{f(x=2|\Theta^*)/1}{f(x=2|\Theta_{i-1})/1}$ ← simple case prob 1 propose Θ as other state.

If $\theta^* = \text{loaded}$, $\alpha = \frac{0.0796}{0.125} = .635$ ← accept w/ prob = .635

$\theta^* = \text{fair}$, $\alpha = \frac{0.125}{0.0796} = 1.57$ ← always accept



↳ $Q = \begin{bmatrix} .365 & .635 \\ 1 & 0 \end{bmatrix}$, $\pi = [.612 \quad .388]$

↳ Stationary dist = $[.612 \quad .388]$

So, it works

↳ ∴ we only need to maximise post via est w/ prior & likelihood

no need for norm. const.

Side bar: Chain rule of probability

↳ for the next § we need Ch. rule of PS,
so, what is it?

$P(A, B) = P(A|B)P(B)$ by definition

↳ gives bayes thrm: $P(A|B) = \frac{P(B|A)P(A)}{P(B)}$ ✓

Chain rule of prob.

If $A_1, \ldots A_n$ & $P(A_1 \cap \ldots \cap A_{n-1}) > 0$

$$P(A_1 \cap \ldots \cap A_n) = P(A_1)\, P(A_2|A_1)\, P(A_3|A_1 \cap A_2) \cdots P(A_n|A_1 \cap \ldots \cap A_{n-1})$$

proof by induction:

if $n = 2$ then chain rule holds

$P(A_1 \cap A_2) = P(A_1)P(A_2|A_1)$

let $B = A_1 \cap \ldots \cap A_{n-1}$

$P(B) = P(A_1)P(A_2|A_1) \ldots P(A_{n-1}|A_1 \cap \ldots \cap A_{n-2})$

$P(B \cap A_n) = P(A_n|B)P(B)$ ▨

→ Gibbs Sampling

↳ We saw how to draw post w/ 1 ~~variable~~ parameter

What about multiple parameters

↳ say $p(\theta, \phi | y) \propto g(\theta, \phi | y)$

↳ update values 1 at a time using new values for each draw

JAGS → Just Another Gibbs Sampler

Note that

$$p(\theta, \phi | y) = p(\phi | y)\, p(\theta | \phi, y)$$

(no θ)

by chain rule of probability

↳ $p(\theta | \phi, y) \propto p(\theta, \phi | y) \propto g(\theta, \phi)$

$p(\phi | \theta, y) \propto p(\theta, \phi | y) \propto g(\theta, \phi)$

Basically: for each unknown param, start w/ each one & update one at a time

→ Algorithm

① Init. θ_0, ϕ_0

② for i in $1 \dots m$

a) using ϕ_{i-1}, draw $\theta_i \sim P(\theta \mid \phi_{i-1}, y)$

b) using θ_i, draw $\phi_i \sim P(\phi \mid \theta_i, y)$

(θ_i, ϕ_i)

Example

↳ Similar non-conj dist from before

$$y_i \mid \mu, \sigma^2 \overset{iid}{\sim} N(\mu, \sigma^2), \quad i = 1, \dots, n$$

$$\mu \sim N(\mu_0, \sigma_0^2)$$

$$\sigma^2 \sim IG(\nu_0, \beta_0)$$

← we have multiple parameters, which is why we need Gibbs

$$P(\mu, \sigma^2 \mid y_1, \dots, y_n)$$

$$\propto p(y_1, \dots, y_n \mid \mu, \sigma^2) p(\mu) p(\sigma^2)$$

↑ what does this density look like?

$$\text{post} \propto \prod_{i=1}^{n} [N(y_i|\mu,\sigma^2)]\, N(\mu|\mu_0,\sigma_0^2)\, IG(\sigma^2|\nu_0,\beta_0)$$

↳ insert actual densities & drop proportionality consts.

$$\propto (\sigma^2)^{\frac{-n}{2}} e^{\frac{-\sum_{i=1}^{n}(y_i-\mu)^2}{2\sigma^2}} e^{\frac{-(\mu-\mu_0)^2}{2\sigma^2}} (\sigma^2)^{-(\nu_0+1)} e^{\frac{-\beta_0}{\sigma^2}}$$

full posterior up to proportionality

Now that we have post to prop. we can change $\propto$ as necessary for Gibbs sampling

ie

$$P(\mu|\sigma^2, y_1 \dots y_n) \propto e^{\frac{-\sum_{i=1}^{n}(y_i-\mu)^2}{2\sigma^2}} e^{\frac{-(\mu-\mu_0)^2}{2\sigma^2}}$$ ✓

&

$$P(\sigma^2|\mu, y_1 \dots y_n) \propto (\sigma^2)^{\frac{-n}{2}} e^{\frac{-\sum_{i=1}^{n}(y_i-\mu)^2}{2\sigma^2}} (\sigma^2)^{-(\nu_0+1)} e^{\frac{-\beta_0}{\sigma^2}}$$ ✓

↳ Both can be simplified further but not so important. ↳ they have recognizable distributions to sample from

Note: What happens if we can't draw from dist directly (like we did for the previous example?

↳ Place mh algorithm inside of Gibbs sampling.

→ Assessing Convergence

↳ trace plot → Shows where sampler is spending its time

- Shouldn't show long term trends
- No wandering
  ↳ though, you can still get convergence even when wandering

↳ this can be assessed using autocorrelation

AC will be high when chain is searching the space

↳ When good value is found, autocorr will be low. (ie Randomness dictates search)

Effective Sample Size

↳ How many indep samples from stationary dist you would need to draw to have equiv info in your chain.

coda has function effectiveSize

↳ Check my effective sample size gives a lot of info, so it is worth doing.

you can also use raftery diagnostic

↳ Shows how much more samples you'd need for some accuracy w/ some prob for a quantile est

↳ vs how many from completely indep chain.

↳ Make sure to check documentation

What happens if it takes long time to sample from stationary dist?

↳ remove samples as "burn in" period

we can find this visually on a trace plot

Gelman-Rubin diagnostic

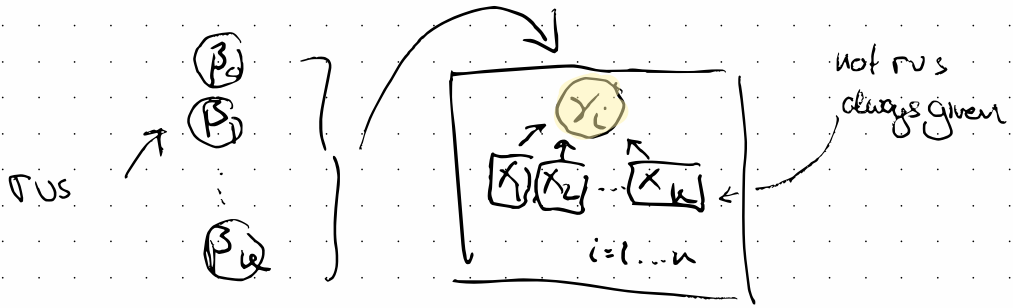
↳ checks if multiple chains have converged. Should give value close to 1.

↳ function gelman.diag

→ Bayesian Linear Regression

$$y_i = \beta_0 + \beta_1 x_1 + \dots + \beta_k x_k + \varepsilon_i, \quad \varepsilon_i \overset{iid}{\sim} N(0, \sigma^2), \; i = 1 \dots n$$

$$y_i \mid \underset{\sim}{x}_i, \underset{\sim}{\beta}, \sigma^2 \overset{ind}{\sim} N(\beta X, \sigma^2)$$



$\beta \leftarrow$ while holding all other x const. a 1-fold rise in x will change the mean of y_i by β

Note that all β & σ have their own priors

Note that adding Laplace prior to β is Lasso regression

↳ $p(\beta) = \frac{1}{2} e^{-|\beta|}$

→ ANOVA (Analysis of Variance)

"Factors"

categorical vars, obs belong to groups

↳ comparing responses inside vs below groups

↳ if var below > ins, conclude grouping effect

ex Sound (music, ? music) Food size (small, med., large) (for website)

2 × 3 = 6 total combos

→ model: $y_i \mid g_i, \mu, \sigma^2 \overset{iid}{\sim} N(\mu_g, \sigma^2)$

$g_i \in \{1, \dots, G\}$

$i = 1, \dots, n$

Alternative

$$E(y_i) = \beta_0 + \beta_1 X_1 + \dots + \beta_{a-1} X_{a-1,i}$$

↑ dummy vars (indicator that group$_i$ = true)

Note: Alternative formulation shows how ANOVA is related to linear regression

Consider {sound: A, font: b}

		B		
		1	2	3
A	1	μ_{11}	μ_{12}	μ_{13}
	2	μ_{21}	μ_{122}	μ_{23}

Cell means model

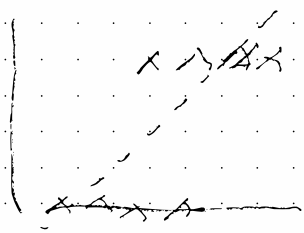
$E(y_i) = \mu + \alpha_2 I_{(a_i=2)} + \beta_2 I_{(b_i=2)}$ Additive Model

$+ \beta_3 I_{(b_i=3)}$ ← Appropriate when no interaction between independent variables

→ Logistic Regression

What about classification tasks

↳ Linear regression doesn't quite work



Likelihood here is bernoulli

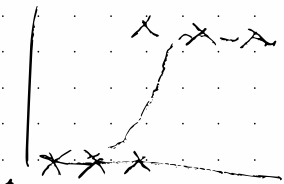
$y_i | \phi_i \overset{ind}{\sim} Bern(\phi_i)$ $i = 1, \dots, n$

Lin reg: $E(y_i) = \beta_0 + \beta_1 x_1$ ← use link function to bound $E(y_i)$ between 0 & 1

$$\frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1)}}$$

Note that this is the result of using

$\log\left(\frac{\phi}{1-\phi}\right)$ as the link (Logit)



↑ Now we can see probs of conversion

→ Poisson Regression

What about count data?

↳ Linear reg? Problem, counts cannot -ve

variance might not be constant

$y_i \overset{ind}{\sim} Pois(\lambda_i) \quad i = 1, \ldots, n$

Linear reg: $E(y_i) = \beta_0 + \beta_1 x_{1i}$ ← Same problem w/ Logistic regression

use Log Link

$$\log(\lambda_i) = \beta_0 + \beta_1 x_1$$

$$\Rightarrow E(y_i) = \lambda_i = e^{\beta_0 + \beta_1 x_{1i}}$$ ← Note $E(\log(y)) \neq \log(E(y))$

Prior Sensitivity Analysis

↳ Detect dependence of result on prior

- Why did you choose the prior you did
- How do different priors change things?

One thing you can try is to create a skeptical prior that disfavours your hypothesis.

↳ If you can still support your own hypothesis then you have good justifications to claim so.

Hierarchical Models

So far all observations have been independent

Not always true

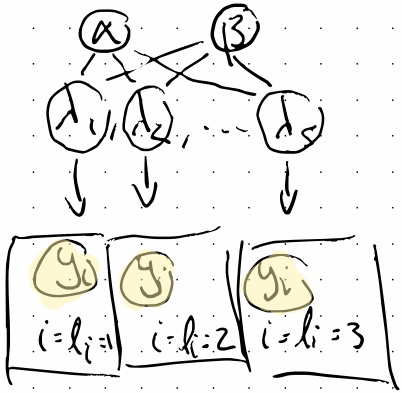
Example: 150 cookies 30 from 5 bags.

fully indep. $y_i | \lambda \overset{iid}{\sim} Pois(\lambda)$ $i = 1, \dots, 50$

loc dep. $y_i | t_i, \lambda_{t_i} \overset{ind}{\sim} Pois(\lambda_{t_i})$ $t_i \in \{1, \dots, 5\}$ $i = 1, \dots, 150$

$\lambda_t | \alpha, \beta \overset{iid}{\sim} Ga(\alpha, \beta)$ $t = 1, \dots, 5$

Graphical Representation



Alternatives

- Make all vectors independent
  - ignore differences between vectors
- fit 5 models for all.
  - ignores data from other vectors

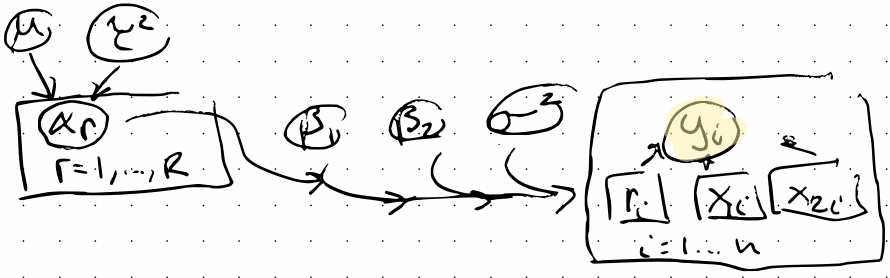
Linear Regression as Hierarchical Model

Linear model: $y_i \mid \underset{\sim}{x}_i, \beta, \sigma^2 \overset{ind}{\sim} N(\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i}, \sigma^2)$ $\quad i = 1, \dots, n$

random intercept model, change β_0

$y_i \mid r_i, \underset{\sim}{x}_i, \beta, \underset{\sim}{\alpha}, \sigma^2 \overset{ind}{\sim} N(\alpha_{r_i} + \beta_1 x_{1i} + \beta_2 x_{2i}, \sigma^2)$ $\quad r \in \{1, \dots, R\}$, $i = 1, \dots, n$

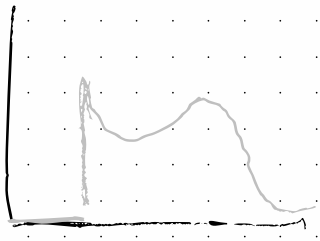
$\alpha_r \mid \mu, \tau^2 \overset{iid}{\sim} N(\mu, \tau^2)$ $\quad r = 1, \dots, R$



Mixture Models

Problem, what if data doesn't follow a standard distribution?

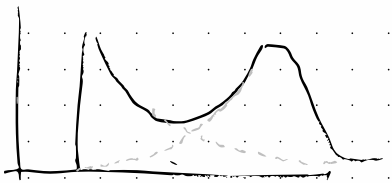
↳ Mixture dists are a combo of multiple distributions



ie a pdf for the graph above might look like this:

$$p(y) = .4\, e^{-y} I_{y \geq 0} + .6 \frac{1}{2\pi} e^{\frac{-(y-3)^2}{2}}$$

↳ we can see how these dists combine here →



the general form for an MM is

↳ $$p(y) = \sum_{j=1}^{J} \omega_j \cdot f_j(y) \quad j = 1, \ldots J \text{ pdfs}$$

$$\text{s.t. } \sum_{j=1}^{J} \omega_j = 1, \quad f_j(y) \in \{\text{PDF}\}$$

we can simulate from this ~~the~~ MM by selecting from either $f_j(y)$ ~~to~~ w/ ω_j probability

↳ ie $[j=1, j=2]$ w/ prob $[.4, .6]$

when j is selected, take rv from $f_j(y)$

issue: when we have real data, we can't see j, latent variable

↳ we can use bayesian hierarchical models to infer them