Introduction:
- Bayes theorem,
- Predictive distribution
- Sequential learning

Inference:
- Inference about functions of parameters
- Point & interval estimates
- Summary for hypothesis
- Statistical decision theory; optimal decision, when mean/mode/median
- Likelihood principle

Priors:
- Conjugate priors; exponential family
- Non-informative; uniform & Jeffreys
- Hierarchical priors

Graphs:
- DAG, moralising, $X_1 \perp\!\!\!\perp X_2 \mid S$
- Factorisation theorem, Markov blanket, full conditional distribution

Hierarchical models:
- Marginal prior
- exchangeability

MCMC:
- Gibbs sampling; sampling from full cond distributions
- Burn-in; determine M, Gelman-Rubin diagnostic
- Determining N, MCSE, $SE(\bar{\xi}_{MN})$ (batching)

1: Introduction

Bayes Theorem

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

posterior — prior — likelihood

θ is parameters, y is data
(can also use on RVs, $P(A|B) = \frac{P(B|A)P(A)}{P(B)}$)

Interpretation of Probability

- Frequentist: $P(A) = \lim_{n\to\infty} \frac{m}{n}$ ← occurrences of A in n repeats
  - concrete, but what about unobservable quantities?
- Bayesian: $P(A)$ = subjective belief that A will occur

Bayesian Inference

- Classical inference: ← what does y tell us about unknown θ
  - y random, θ fixed but unknown
- Bayesian inference: ← how does y change our belief about θ
  - y and θ random
  - $p(\theta|y) \propto p(y|\theta)\,p(\theta)$

Predictive Distributions

$$p(\tilde{Y} = \tilde{y} | Y = y) = \int p(\tilde{y}|\theta)\,p(\theta|y)\,d\theta$$

(predict new $\tilde{Y}$ given already seen y)

- Sequential learning: ← 'todays posterior is tomorrows prior'

$$p(\theta|y_1, y_2) \propto p(y_2|\theta) \times p(\theta|y_1)$$

allows us to 'update' priors as we observe data

2: Inference

Bayesian inference for the Normal distribution

$Y_i | \theta, \tau \sim N(\theta, \tau^{-1}), i = 1, \ldots, n$ (τ = precision, $\sigma^2 = \tau^{-1}$)

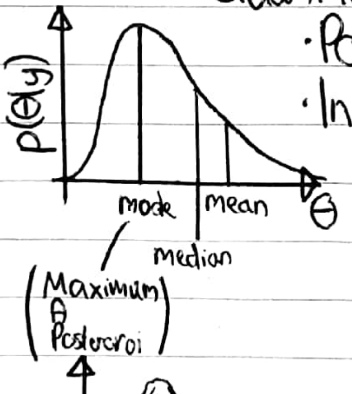
- θ and/or τ unknown, with priors
  eg $p(\theta, \tau) \propto \tau^{-1}$
  $\Rightarrow p(\theta, \tau | \underline{y}) \propto p(\theta, \tau) \cdot p(\underline{y} | \theta, \tau)$
  $p(\tau | \underline{y}) = \int_{-\infty}^{\infty} p(\theta, \tau | \underline{y}) \, d\theta$ (similar for θ)

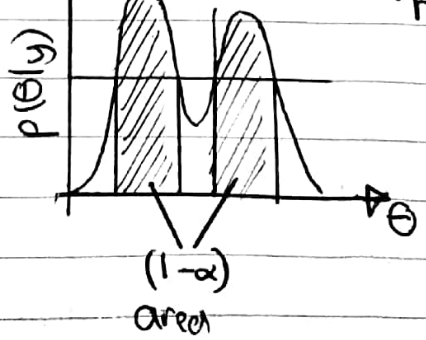
- Inference about functions of parameters:
  $\phi = g(\theta)$
  $\Rightarrow p_{\phi|y}(\phi | \underline{y}) = p_{\theta|y}(g^{-1}(\phi) | \underline{y}) \cdot \left| \frac{d\theta}{d\phi} \right|$

Summarisation of posterior distributions

- 3 ways to summarise $p(\theta | \underline{y})$:
  - Graphically:
    - good for 1-2D; need to marginalise for higher dimensions (integrals often intractable; use MCMC)
  - Quantitative:
    - Point estimates: mean, mode, median
    - Interval estimates:
      - Credible interval:
        $p(\theta \in [c_1, c_2] | \underline{y}) = 1 - \alpha$,
      - Credible region:
        $p(\theta \in C | \underline{y}) = 1 - \alpha$

        (both) 100(1-α)% confidence
    - Highest posterior density:
      - interval s.t. any point inside has density > any point outside
      - 100(1-α)% HPD interval is the shortest width 100(1-α)% credible interval





- Summaries for specific hypothesis:
  - eg $H_0: \Theta < \Theta_0$ $(H_1: \Theta \geq \Theta_0)$
    reject H_0 if $P(\Theta < \Theta_0 | y) < P(\Theta > \Theta_0 | y)$
    (can scale by asymmetric losses too)
    $\left[\underset{\substack{\uparrow \\ \text{(Type I loss)}}}{C_\alpha} P(\Theta < \Theta_0 | y) < \underset{\substack{\uparrow \\ \text{(Type II loss)}}}{C_\beta} P(\Theta > \Theta_0 | y) \right]$

Statistical Decision Theory

- parameter space Θ, set of actions A, loss function $L(\Theta, a) \geq 0$
- decision function $d(y)$ maps observed data y to action
- posterior expected loss:
$$E_{\Theta|y}[L(\Theta, d(y))] = \int L(\Theta, d(y))\, p(\Theta|y)\, d\Theta$$
- optimal decision minimises this
  - Mean under square loss
  - median under absolute loss
  - Mode under zero-one loss

Comparison of Bayesian & classical inferences

- Point estimation:
  - classical criteria: $\underset{\substack{\text{expected} \\ = \text{true}}}{\text{bias}}$, $\underset{\text{MSE}}{\text{efficiency}}$, $\underset{\substack{n \to \infty \\ \text{converge to true}}}{\text{consistency}}$, $\underset{\substack{\text{not uniformly higher} \\ \text{MSE than another} \\ \text{estimator}}}{\text{admissability}}$
  - based around $p(y|\Theta)$, as Θ fixed
  - uses asymptotic arguments, not used in Bayesian so much
  - MAP = ML under uninformative prior / large samples
- Interval estimation:
  - Bayesian: $100(1-\alpha)\%$ chance Θ lies in interval
  - Classical: Random interval containing $\hat{\Theta}_{\text{true}}$ w/ probability $100(1-\alpha)$ ← cannot be probability statement, as Θ not random
  - In general can be quite different, in practise Bayesian credible interval similar to classical confidence interval
- Likelihood principle:
  - $p(y|\Theta)$ summarises all information about Θ provided by y
  - obeyed by Bayesian, violated by classical (as sampling distribution of Y can influence inference) ← (ML estimator obeys likelihood principle)

Pros / Cons of Bayesian Approaches

+ Theoretically sound, uses all available information
− where does prior come from (hard to justify), computational difficulties

3: Priors

Basic Considerations

θ support	Suitable priors
$(-\infty, +\infty)$	Normal, t-dist
$(0, \infty)$	Gamma, log-normal
$(0, 1)$	Beta

- functional form (support of θ must match) & parameter values

Conjugate Priors

Prior & Posterior in same class

- class P of prior distributions $p(\theta)$ is conjugate family for $p(y|\theta)$ if $p(\theta|y)$ is also in P for all y
- eg:

likelihood	conjugate prior
Binomial	Beta
Normal (known τ)	Normal
$f(y)g(\theta)e^{h(\theta)t(y)}$	$g(\theta)^{\nu} e^{h(\theta)\delta}$

exponential family

$t(y)$ is sufficient statistic

(ν = prior 'sample size'
δ = prior sufficient statistic)

- Exponential family: $p(y|\theta) = f(y)g(\theta)e^{h(\theta)t(y)}$

Distribution	$g(\theta)$	$h(\theta)$	$t(y)$
Normal $(\theta \mid \tau^{-1})$	$\exp\left[\frac{-\tau\theta^2}{2}\right]$	$\tau\theta$	y_i
Bin (m, θ)	$(1-\theta)^m$	$\log\frac{\theta}{1-\theta}$	y

- Choosing parameters:
  - choose ν, δ ~~by~~ roughly, or by knowing mean/variance of θ, or confidence interval for θ etc...

Non-informative Priors

- Uniform: gives $p(\theta|y) \propto p(y|\theta)$
  - if support is infinite, $p(\theta)$ is improper; $\int p(\theta)d\theta = \infty$
  - $\Rightarrow$ may give improper posterior

eg large variance gaussian

- Locally uniform:
  - roughly uniform where likelihood is non-negligible; dominated by likelihood
- Beta(1,1) vs Beta(0,0):
  - Beta(1,1) uniform & proper, but influences posterior more than Beta(0,0) (which is improper)

maybe do Beta(0.001, 0.001) instead!

- Beta(0.001, 0.001) best of both, but if likelihood non-negligable at $\theta = 0$ or 1, then highly informative
- Jeffreys prior:
  - goal: $p(\theta)$ remain invariant under transformation $\phi = g(\theta)$

Fisher information $\rightarrow$ $I(\theta) = -E_{Y|\theta}\left[\frac{\partial^2}{\partial\theta^2}\log p(Y|\theta)\right] = E_{Y|\theta}\left[\left(\frac{\partial}{\partial\theta}\log p(Y|\theta)\right)^2\right]$

$p(\theta) \propto \sqrt{I(\theta)}$ $\quad (\Rightarrow p(\phi) = \sqrt{I(\phi)}$ for $\phi = g(\theta))$

proper

Normal(θ, τ^{-1}), known $\theta \Rightarrow p(\tau) \propto \frac{1}{\tau}$, $\tau \sim$ Gamma(0,0)

Bin(n, θ), $\quad \theta \sim$ Beta$(\frac{1}{2}, \frac{1}{2})$

improper

- often improper, violates likelihood principle, can lead to inconsistencies in multiparameter case (θ_1 can influence prior on θ_2 etc)

Hierarchical priors

- divide model into stages; hyperpriors on priors
- eg $\quad Y \sim$ Bin$(10, \theta)$

  $\theta \sim$ Beta(α, β)

  $\alpha \sim$ Gamma(4,4), $\beta \sim$ Gamma(5,10)
- often useful when $\underline{\theta} = (\theta_1, \theta_2, \dots, \theta_k)$ & θs are exchangeable

Summary

- Conjugate computationally convenient, but limiting
- Non-informative priors must be used carefully
- Sensitivity analysis of priors is important

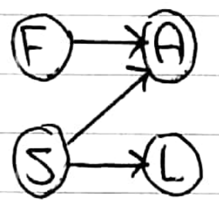
4: Graphs

Introduction

- Graphical models visual conditional independence structure

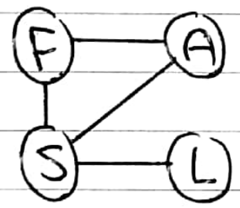
$$X \perp\!\!\!\perp Y \text{ if } p(x,y) = p(x)p(y) \quad \forall x,y$$
$$X \perp\!\!\!\perp Y|Z \text{ if } p(x,y|z) = p(x|z)p(y|z)$$

Directed Acyclic Graphs



- Lines ~~are~~ indicated 'causation'; direct influence
- Node is conditionally independent of ancestors given parents
- Founders (no parents) are marginally independent
- Parent's are conditionally dependent given shared child

Moralising a DAG



- 'marry' parents, drop directions
- to determine if $X_1 \perp\!\!\!\perp X_2 | X_3$,
  - keep X_1, X_2, X_3 & their ancestors
  - Moralise this subgraph
  - true if X_3 blocks all paths from X_1 to X_2 (else false)

ie X_3 separates X_1 & X_2 $\longrightarrow$

(This is global Markov property)

Factorisation theorem; Markov blankets & full conditional distributions

- Factorisation theorem:

$$p(X) = \prod_k p(X_k | \text{parents}[X_k])$$

(eg, above $P(F,A,S,L) = P(F)p(S)p(A|F,S)p(L|S)$)

- Markov blanket:

$$X_k \perp\!\!\!\perp X_{\setminus(X_k,\, bl[X_k])} | bl[X_k]$$

where $bl[X_k] = \text{parents}[X_k] \cup \text{children}[X_k] \cup \text{partners}[X_k]$

(ie all nodes connected to X_k in Moralised graph)

useful when using MCMC methods to fit Bayesian models

Scanned by CamScanner

- Full-conditional distributions:

$$p(X_k \mid X_{\setminus X_k}) \propto p(X_k \mid \text{parents}[X_k]) \times \prod_{w \in ch[X_k]} p(W \mid \text{parents}[W])$$

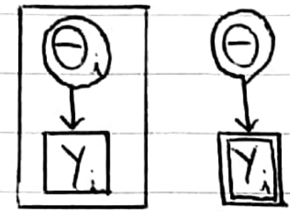
- ie full-conditional dist of X_k depends only on variables in X_k's Markov blanket

Summary

- DAG's summarise statistical models as factorisation of simple relationships
- Moralising shows conditional independences

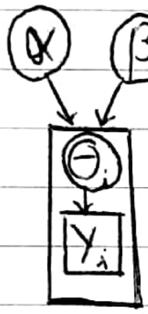
5: Hierarchical Models

Non-hierarchical models



- must fully share ~~a have a copy of~~ parameter, or have all θ_i independent
- want to encode some dependence between θ_i's

Hierarchical models



$$p(\underline{\theta}, \alpha, \beta) = \left[\prod_i p(\theta_i | \alpha, \beta)\right] p(\alpha) p(\beta)$$

- Can give weak priors on α, β (eg Exp(0.01))
- ensures θ_i's weakly influence each other; ie assume θ_is similar but not identical
- shifts θ_i's towards overall mean, and improve reliability when n_i is small

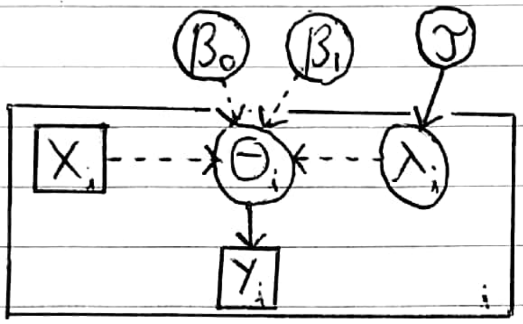
- in general, marginal prior
$$p(\underline{\theta}) = \int p(\underline{\theta} | \phi_2) p(\phi_2) d\phi_2$$

- Exchangeability:
  - $\theta_1, \dots, \theta_n$ exchangeable if any permutation of these RVs has the same joint distribution
  - Marginal independence & same marginal distribution $\Rightarrow$ exchangeability (but not other way)
  - Exchangeability implies a hierarchical model with some prior ϕ, such that $\theta_i \perp\!\!\!\perp \theta_j | \phi$

Scanned by CamScanner

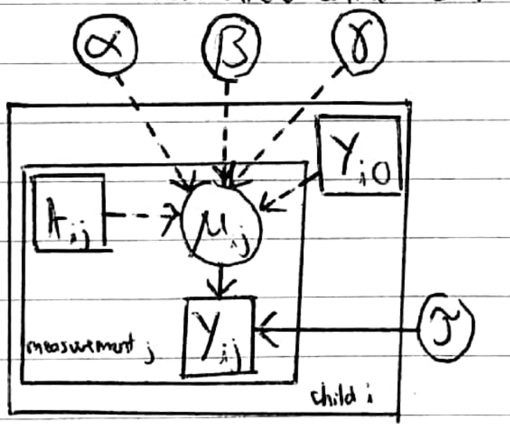
DAGs for hierarchical models

- circle nodes for unknowns
- Square nodes for observed RVs
- rectangular boxes for repetitive structure
- GLMM: Generalized Linear Mixed model

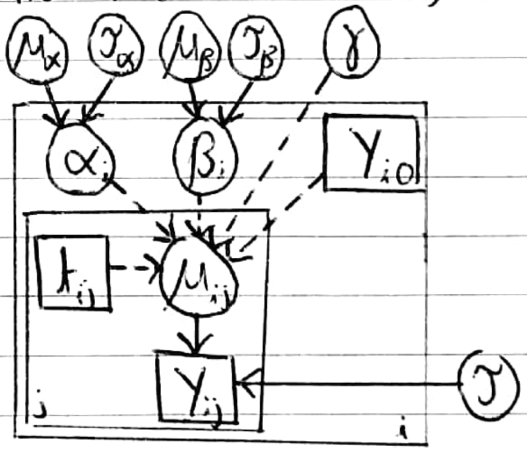


$Y_i | \theta_i \sim \text{Poisson}(C_i \theta_i)$
$\log \theta_i = \beta_0 + \beta_1 X_i + \lambda_i$
$\lambda_i | \tau \sim N(0, \tau^{-1})$
$\beta_0, \beta_1, \tau \sim$ non-informative

- Non-hierarchical LM:



Hierarchical LM (LMM):



$Y_{ij} | \mu_{ij}, \tau \sim N(\mu_{ij}, \tau^{-1})$
$\mu_{ij} = \alpha + \beta(t_{ij} - \bar{t}) + \gamma(Y_{i0} - \bar{Y}_0)$
$\alpha \sim N(0, 10000)$
$\beta \sim N(0, 10000)$
$\gamma \sim N(0, 10000)$
$\tau \sim \text{Gamma}(0.001, 0.001)$

$Y_{ij} | \mu_{ij}, \tau \sim N(\mu_{ij}, \tau^{-1})$
$\mu_{ij} = \alpha_i + \beta_i(t_{ij} - \bar{t}) + \gamma(Y_{i0} - \bar{Y}_0)$
$\alpha_i | \mu_\alpha, \tau_\alpha \sim N(\mu_\alpha, \tau_\alpha^{-1})$
$\beta_i | \mu_\beta, \tau_\beta \sim N(\mu_\beta, \tau_\beta^{-1})$
$\mu_\beta, \mu_\alpha \sim N(0, 10000)$
$\tau_\alpha, \tau_\beta \sim \text{Gamma}(0.001, 0.001)$

Summary

hard to specify informative hyperpriors

- Breaking model into layers; parameters & hyper-parameters
- useful when data from similar but not identical 'units' which are exchangeable

6: Markov chain Monte Carlo

Motivation

- Monte Carlo integration:

$$E[f(\Theta)|x] = \int f(\Theta)\, p(\Theta|x)\, d\Theta \approx \frac{1}{N}\sum_{i=1}^{N} f(\Theta^{(i)}), \quad \cancel{\Theta^{(i)} \sim p}$$

  - $\Theta^{(i)}$ sampled from $p(\Theta|x)$
- Markov chain:
  - used to sample from $p(\Theta|x)$; sample from an MC w/ $p(\Theta|x)$ as it's equilibrium distribution

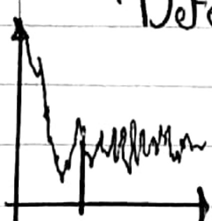
MCMC

- Metropolis-Hastings algorithm:
  - general framework for MCMC; focus on special case; Gibbs sampling:
    - split $\underline{\Theta}$ into k components
    - sample $\Theta_1^{(k+1)}$ from $p(\Theta_1|\Theta_{2:k}^{(i)})$ ← full conditional distribution
    - repeat for Θ_2 to Θ_k (always using most up to date Θ_j)
    - now have $\underline{\Theta}^{(i+1)}$, repeat
  - Sampling from full conditional distributions:
    - may have closed form, if not use, eg, rejection sampling or ratio-of-uniforms method

Convergence & Monte Carlo standard errors

$$E[f(\Theta)|x] \approx \frac{1}{N-M}\sum_{i=M+1}^{N} f(\Theta^{(i)}) = \bar{f}_{MN}$$

- early iterations, 1 to M, are burn-in; ~~a~~ chain hasn't converged well yet, so don't use these
- Determining M:
  - only truly converge at $M=\infty$



  - use trace plots; burnt in once samples look like random scatter about some value

- Convergence diagnostics:
  - Gelman-Rubin diagnostic:
    - run R chains, measure difference in behaviour ~~of~~ (when converge, should behave similarly)
    - $R = \sqrt{\frac{V}{W}}$ V = under-estimate of $\sigma_R^2 = Var(\Theta_R | x)$
      W = over-estimate of σ_R^2
      ($W, V \to \sigma_R^2$ as $n \to \infty$)
    - $R < 1.05 \implies$ 'practical' convergence
      (calculate R for all or several Θ)
- Determining N:
  - run chain until Monte carlo standard error (MCSE), less than 5% of parameters posterior standard deviation, for all parameters
  - estimating $SE(\bar{f}_{MN})$:
    - Batching:
      - divide sequence into Q uncorrelated batches of length L

$$\left(\begin{array}{l} b_q = \frac{1}{L} \sum_{i \in batch(q)} f(\Theta^{(i)}) \\ \bar{b} = \frac{1}{Q} \sum_{i=1}^{Q} b_i \end{array} \right)$$

$$\widehat{SE}(\bar{f}_{MN}) = \sqrt{\frac{1}{Q(Q-1)} \sum_{q=1}^{Q} (b_q - \bar{b})^2}$$

Pros & Cons of MCMC

+ freedom in modelling & inference
+ only available method for complex problems
- slow & computationally expensive
- hard to diagnose (lack of) convergence