Lec 1: March 28

- 753 Final exam notes

Q6: 2x2 table:

a.)

	Normal	cancer
smoker	55	40
non-smoker	45	60

poisson log-linear

$Z \sim$ smoking + cancer

is smoking significantly associated w/ lung cancer?

- relationship between 2x2 table and log-linear
- independence model
  - → each row has a row effect, each column has a column effect
  - → test for association between rows and columns by comparing indep. model to fully saturated model (residual deviance) 1-pchisq
    - found in R output

→ can't just look at z-values or p-values

· when can we use p-values?

b.) prospective probability that a smoker randomly sampled from the country has cancer.

→ just look at 2x2 table

→ write down log-odds from logistic regression

$= \beta_0 + \beta_1 I(\text{smoking})$

Non-smoker: $\log(40/55)$

Smoker: $\log(60/45)$

} use this to recover log odds from prospective trial (b/c coefficients are same w/ retrospective except for intercept)

can use β's in question to compute the intercept's effect

Q7: a.) residual deviance >> residual df

$\Rightarrow$ overdispersion

b.) tobacco use associated w/ disease risk?

Remember there's overdispersion!

$$\hat{\phi} \approx \frac{D}{n-p} = \frac{211.16}{85}$$

- need to use adjusted SE to construct t-statistic
- compare w/ t-dist w/ 85 df

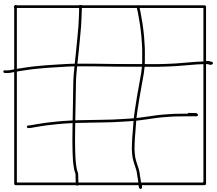
c.) CI:

d.) intercept-only model vs. current model
- use null-model
- OVERDISPERSION !!
- construct F-statistic

$$\frac{(\text{Resid} - \text{Null})/(n_1 - n_0)}{\hat{\phi}} \sim F$$

- Null deviance: deviance of intercept-only model

Fixed Effects and Random Effects



- Effects: describe the impact of a level of the factor on the outcome/response (regression coefficients) → parameters
- Fixed effects: fixed constants
- Random effects: parameters are random
- Linear Mixed Model (LMM):

$$\mu_{ij} = \mu + \alpha_i + \beta_j$$

fixed — random

$$y_{ij} = \mu_{ij} + \varepsilon_{ij}$$

- GLM:

$$g(\mu_{ij}) = \mu + \alpha_i + \beta_j$$

fixed constants

- GLMM:

$$g(\mu_{ij}) = \mu + a_i + b_j$$

random

- When to use fixed vs. random effects models?

EX:

(1) fixed effects

y_{ij} — outcome of patient j receiving drug i

$i=1$: placebo

$i=2$: new drug

Evaluate whether new drug is better

Assume fixed effects model:

$E(y_{ij}) = \mu_i = \mu + \alpha_i$ — drug-specific effect, unknown fixed constant

(2) Random effects:

y_{ij} – outcome for patient j at hospital i

$E(y_{ij}) = \mu + a_i$ — predict; random variable, follows some dist.

sometimes, main interest is to study variability across hospitals, not each individual hospital

i.e. $a_i \sim N(0, \sigma^2)$ — variability across hospitals

- also used for modelling correlated data

Inference

(1) Estimate

- Fixed effects: β
- random effects: β, σ^2 — variance for random effects

(a) Maximum likelihood approach

(b) REML - restricted maximum likelihood approach

- get rid of fixed effects parameter
  e.g. integrate them out of likelihood
- try to estimate variance parameters

$y = X\beta + Zu + \varepsilon$ — fixed; RV; noise

$$K'y = K'X\beta + K'Zu + K'\varepsilon$$

Find K s.t. $KX = 0$ — $N(0, \Sigma)$

$$\tilde{y} = K'y = \underbrace{K'Zu + K'\varepsilon}_{\text{fixed effects disappear}} \quad \text{— } N(0, \sigma^2)$$

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

ML: $\hat{\sigma}^2 = \dfrac{\sum (y_i - \bar{y})^2}{n}$

REML: $\hat{\sigma}^2 = \dfrac{\sum (y_i - \bar{y})^2}{n-1}$ → automatically accounts for df from fixed effects β

(c) Bayesian approach

$\theta = (\beta, \sigma^2)$

Assumes some prior distribution for the parameters.

$$\pi(\theta|y) = \frac{f(y|\theta)\pi(\theta)}{f(y)}$$

$f(y) \rightarrow \int f(y,\theta)\, d\theta$

↓ inference based on posterior distribution

(d) estimating equations { Quasi-likelihood; Generalized estimating eqns

(2) Prediction

$$y_{ij} = \mu + a_i + b_{ij}$$

$E(y_{ij}|a_i) = \mu + a_i$ — prediction

$a_i \sim N(0, \sigma_a^2)$

$E(y_{ij}) = 0$

μ — estimation

Best predictor:

$$\beta = E(a_i | y)$$

Fixed Effects Model → random:

- One-way classification model

1. Normal and fixed effect

Consider m groups, each has n_i obs.

y_{ij} $\quad i = 1, \cdots, m$
$\quad\quad j = 1, \cdots, n_i$

$$y_{ij} = \mu_i + \varepsilon_{ij} \overset{iid}{\sim} N(0, \sigma^2)$$

$$y_{ij} \overset{iid}{\sim} N(\mu_i, \sigma^2)$$ ← each group has a group-specific mean that we want to estimate.

(i) MLE: $\hat{\mu}_i = \bar{y}_{i\cdot}$ → unbiased

$$\hat{\sigma}^2 = \frac{1}{N} \sum_i \sum_j (y_{ij} - \bar{y}_{i\cdot})^2$$ → biased

$N = \sum_{i=1}^{m} n_i$

$$E\hat{\sigma}^2 = \frac{N-m}{N} \sigma^2 \neq \sigma^2$$

(ii) REML:

$$S^2 = \frac{1}{N-m} \sum_i \sum_j (y_{ij} - \bar{y}_{i\cdot})^2$$ → unbiased

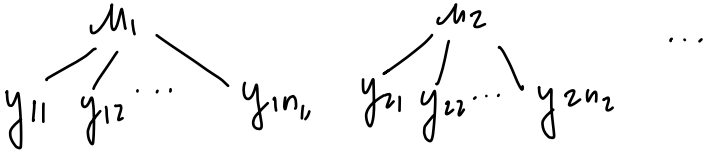
$$\frac{(N-m)S^2}{\sigma^2} \sim \chi^2_{N-m}$$ → mean of $\chi^2_{N-m} = N-m$

↳ var $= 2(N-m)$

So $Var(S^2) = \frac{\sigma^2}{(N-m)^2} \cdot 2(N-m) = \frac{2\sigma^4}{N-m}$

Lec 2: March 30

- One-way classification (Normal)



1. Normal and fixed effects

$$y_{ij} \overset{ind}{\sim} N(\mu_i, \sigma^2)$$

(i) MLE:

$$\hat{\mu}_i = \bar{y}_i$$

$$\hat{\sigma}^2 = \frac{1}{N} \sum_i \sum_j (y_{ij} - \bar{y}_i)^2$$

REML:

$$\tilde{\sigma}^2 = \frac{1}{N-m} \sum_i \sum_j (y_{ij} - \bar{y}_i)^2$$

(ii) Likelihood-ratio test

$$H_0: \mu_1 = \mu_2 = \cdots = \mu_m$$

$$LR \triangleq \Lambda = \left(\frac{\hat{\sigma}_0^2}{\hat{\sigma}^2} \right)^{-N/2}$$

$\hat{\sigma}_0^2$: MLE under null model

$\hat{\sigma}^2$: MLE for variance estimate under the alternative model

← obtained by writing the likelihood of null model and alternative model

$$\hat{\sigma}_0^2 = \frac{1}{N} \sum_i \sum_j (y_{ij} - \bar{y})^2 = \frac{1}{N} \left[\sum_i \sum_j (y_{ij} - \bar{y}_i)^2 + \sum_i \sum_j (\bar{y}_i - \bar{y})^2 \right]$$

within group sum of squares

between group sum of squares

$$\hat{\sigma}^2 = \frac{1}{N}\sum_i \sum_j (y_{ij} - \bar{y}_i)^2$$

within group sum of squares

$$-2\log\Lambda = N\log\left(\frac{\hat{\sigma}_0^2}{\hat{\sigma}^2}\right)$$

$$= N\log\left(1 + \frac{\sum_i \sum_j (\bar{y}_i - \bar{y})^2}{\sum_i \sum_j (y_{ij} - \bar{y}_i)^2}\right)$$

$$= N\log\left(1 + \frac{(m-1)\sum_i \sum_j (\bar{y}_i - \bar{y})^2/(m-1)}{(N-m)\sum_i \sum_j (\bar{y}_{ij} - \bar{y}_i)^2/(N-m)}\right)$$

$$= N\log\left(1 + \frac{m-1}{N-m}F\right)$$

F ↑ F-statistic

$\Rightarrow$ LR test equivalent to ANOVA

(iii) Confidence intervals

CI for $\mu_i - \mu_j$

$$\bar{y}_i - \bar{y}_j \pm t_{N-m, \frac{\alpha}{2}} s\sqrt{\frac{1}{n_i} + \frac{1}{n_j}}$$

CI for a linear combination of μ_i's: $\sum_i c_i \mu_i$

$$\sum_i c_i \bar{y}_i \pm t_{N-m, \alpha/2} s\sqrt{\sum_i \frac{c_i^2}{n_i}}$$

CI for σ^2

$$\frac{(N-m)s^2}{\sigma^2} \sim \chi^2_{N-m}$$

$$\Rightarrow \left(\frac{(N-m)s^2}{\chi^2_{N-m,1-\alpha/2}}, \frac{(N-m)s^2}{\chi^2_{N-m,\alpha/2}} \right)$$

(iv) Test whether a linear combination of the group means is equal to a specific value

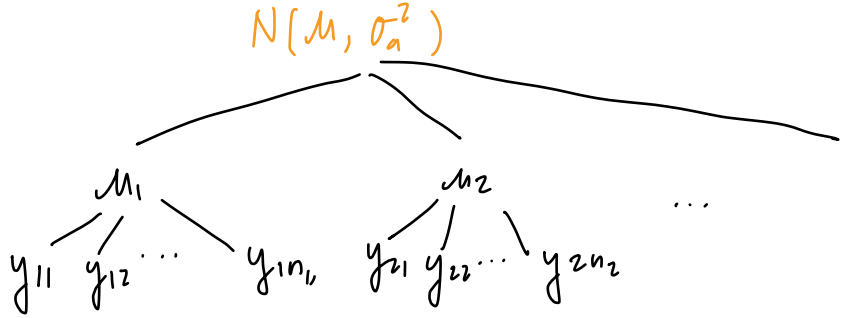
Test $H_0: \sum_i c_i \mu_i = \eta_0$

$$t = \frac{\sum_i c_i \bar{y}_{i\cdot} - \eta_0}{s\sqrt{\sum_i c_i^2/n_i}} \sim t_{N-m}$$

determined by df of sample variance

2. Normal and Random Effects

- assume μ_i's are drawn from a distribution



1. draw μ_i's
2. draw y_i's from $N(\mu_i, \sigma^2)$

$$y_{ij} \mid \mu_i \overset{ind}{\sim} N(\mu_i, \sigma^2)$$
$$\mu_i \overset{iid}{\sim} N(\mu, \sigma_a^2)$$

$$\mu_i = \mu + a_i$$
$$\begin{cases} y_{ij} \mid a_i \overset{ind}{\sim} N(\mu + a_i, \sigma^2) \\ a_i \overset{iid}{\sim} N(0, \sigma_a^2) \end{cases}$$

- this assumption introduces marginal correlation between data points in the same group

(i) covariances

$$y_{ij} = \mu + a_i + \varepsilon_{ij} \qquad a_i \sim N(0, \sigma_a^2), \quad \varepsilon_{ij} \sim N(0, \sigma^2)$$

$$a_i \perp \varepsilon_{ij}$$

$$\text{cov}(y_{ij}, y_{i'j}) = \text{cov}(\mu + a_i + \varepsilon_{ij}, \mu + a_{i'} + \varepsilon_{i'j}) \quad (\mu \text{ const})$$
$$= 0$$

$$\text{cov}(y_{ij}, y_{ik}) = \text{cov}(\mu + a_i + \varepsilon_{ij}, \mu + a_i + \varepsilon_{ik})$$
$$= \text{cov}(a_i, a_i) + \text{cov}(a_i, \varepsilon_{ik}) + \text{cov}(\varepsilon_{ij}, a_i) + \text{cov}(\varepsilon_{ij}, \varepsilon_{ik})$$
$$= \text{var}(a_i) \qquad (\text{cov}(a_i, \varepsilon_{ik}) = 0,\ \text{cov}(\varepsilon_{ij}, a_i) = 0,\ \text{cov}(\varepsilon_{ij}, \varepsilon_{ik}) = 0)$$
$$= \sigma_a^2 \neq 0$$

- don't have independent obs. in your data
  - intraclass correlations.

$$var(y_{ij}) = var(\mu + a_i + \varepsilon_{ij}) = \sigma_a^2 + \sigma^2$$

Intra-class correlation: correlation for two points in the same group

$$cor(y_{ij}, y_{ik}) = \frac{\sigma_a^2}{\sqrt{(\sigma_a^2+\sigma^2)(\sigma_a^2+\sigma^2)}} = \frac{\sigma_a^2}{\sigma_a^2+\sigma^2}$$

$f(y, a \mid \mu, \sigma_a^2, \sigma^2)$ — σ_a^2: variance for random effects; σ^2: variance for measurement error

a: unobserved, integrate it out

$$\int f(y, a \mid \mu, \sigma_a^2, \sigma^2)\, da$$

$$= f(y \mid \mu, \sigma_a^2, \sigma^2)$$

↳ this marginal distribution leads to the non-zero covariances

- σ_a^2 characterizes between-group variances

(ii) Likelihood

$$\begin{bmatrix} y_{11} \\ y_{12} \\ \vdots \\ y_{1n_1} \\ \hline y_{21} \\ y_{22} \\ \vdots \\ y_{2n_2} \end{bmatrix}$$

divide into blocks by groups.

$$y_i \triangleq \begin{bmatrix} y_{i1} \\ y_{i2} \\ \vdots \\ y_{in_i} \end{bmatrix}$$

$$y_i \sim N(\mu \mathbb{1}, V_i)$$

off diagonals are covariances of elements in the same group

$$V_i = \begin{bmatrix} \sigma_a^2+\sigma^2 & \sigma_a^2 \\ \sigma_a^2 & \ddots \\ & \sigma_a^2+\sigma^2 \end{bmatrix}$$

$$V_i = \sigma^2 I_{n_i} + \sigma_a^2 J_{n_i}$$

$$J_{n_i} \rightarrow \begin{bmatrix} 1 & 1 & \cdots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \cdots & 1 \end{bmatrix}_{n_i \times n_i} = \mathbf{1} \cdot \mathbf{1}^T$$

$$V_i^{-1} = \frac{1}{\sigma^2} I_{n_i} - \frac{\sigma_a^2}{\sigma^2(\sigma^2 + n_i\sigma_a^2)} J_{n_i}$$

$$|V_i| = (\sigma^2 + n_i\sigma_a^2)(\sigma^2)^{n_i - 1}$$

- likelihood is the product of likelihoods for different groups.

$$L = \prod_{i=1}^{m} (2\pi)^{-n_i/2} |V_i|^{-1/2} \exp\left\{ -\frac{1}{2} (y_i - \mu \mathbf{1})^T V_i^{-1} (y_i - \mu \mathbf{1}) \right\}$$

$$l = \log L = \frac{-N}{2} \log(2\pi) - \frac{1}{2} \sum_i \log(\sigma^2 + n_i\sigma_a^2) - \frac{1}{2}(N-m)\log\sigma^2$$
$$- \frac{1}{2\sigma^2} \sum_i \sum_j (y_{ij} - \mu)^2 + \frac{n\sigma_a^2}{2\sigma^2(\sigma^2 + n\sigma_a^2)} \sum_i \frac{(y_i - n_i\mu)^2}{\sigma^2 + n_i\sigma_a^2}$$

- in general, no closed form solution to score equations (need iterative algorithms)

special case: all n_i's are the same (same # obs. in each group), balanced data:

$$l = \frac{-N}{2} \log 2\pi - \frac{1}{2} m \log(\sigma^2 + n\sigma_a^2) - \frac{1}{2} m(n-1)\log\sigma^2$$
$$- \frac{1}{2\sigma^2} \sum_i \sum_j (y_{ij} - \mu)^2 + \frac{n\sigma_a^2 \sum_i (\bar{y}_i - \mu)^2}{2\sigma^2(\sigma^2 + n\sigma_a^2)}$$

$$SSA \triangleq \sum_i n(\bar{y}_{i\cdot} - \bar{y})^2$$

$$SSE \triangleq \sum_i (y_{ij} - \bar{y}_{i\cdot})^2$$

$$\lambda \triangleq \sigma^2 + n\sigma_a^2$$

$$l = \frac{-N}{2}\log 2\pi - \frac{1}{2}m\log\lambda - \frac{1}{2}m(n-1)\log\sigma^2$$

$$\frac{-SSE}{2\sigma^2} - \frac{SSA}{2\lambda} - \frac{mn(\bar{y}-\mu)^2}{2\lambda}$$

ML equations:

$$\begin{cases} l_\mu = \frac{\partial l}{\partial \mu} = \frac{mn(\bar{y}-\mu)}{\lambda} = 0 \rightarrow \hat{\mu} = \bar{y} \\ l_{\sigma^2} = \frac{\partial l}{\partial \sigma^2} = \frac{-m(n-1)}{\sigma^2} + \frac{SSE}{2\sigma^4} = 0 \rightarrow \hat{\sigma}^2 = \frac{SSE}{m(n-1)} \leftarrow df \\ l_\lambda = \frac{\partial l}{\partial \lambda} = \frac{-m}{2\lambda} + \frac{SSA}{2\lambda^2} + \frac{mn(\bar{y}-\mu)^2}{2\lambda^2} = 0 \rightarrow \hat{\lambda} = \frac{SSA}{m} \end{cases}$$

when $\mu = \bar{y}$

$$\begin{cases} \hat{\mu} = \bar{y} \\ \hat{\sigma}^2 = \frac{SSE}{m(n-1)} = MSE \\ \hat{\lambda} = \frac{SSA}{m} = \frac{m-1}{m}\frac{SSA}{m-1} = \left(1 - \frac{1}{m}\right)MSA \end{cases}$$

→ Not MLE's yet b/c $\lambda \neq \sigma_a^2$

$$\sigma_a^2 = \frac{\lambda - \sigma^2}{n}, \quad \hat{\sigma}_a^2 = \frac{\lambda - \sigma^2}{n} = \frac{\left(1 - \frac{1}{m}\right)MSA - MSE}{n}$$

↑ not necessarily MLE b/c σ_a^2 needs to be non-negative (not guaranteed by RHS)

MLE

If $(1-\frac{1}{m})MSA \geq MSE$, then

$$\begin{cases} \hat{\mu} = \bar{y} \\ \hat{\sigma}^2 = SSE \\ \hat{\sigma}_a^2 = \dfrac{(1-\frac{1}{m})MSA - MSE}{n} \end{cases}$$ are MLEs

If $(1-\frac{1}{m})MSA < MSE$, then

$\Downarrow$

$\sigma_a^2 = 0$, lies on boundary for allowed space of σ_a^2

$\Rightarrow \mu_1 = \mu_2 = \mu_3 \cdots$

Rewrite model based on this condition

$$\begin{cases} \hat{\sigma}_a^2 = 0 \\ \hat{\sigma}^2 = \dfrac{\sum_i \sum_j (y_{ij} - \bar{y})^2}{mn} = \dfrac{SST}{mn} \\ \hat{\mu} = \bar{y} \end{cases}$$

Mean

$E\hat{\mu} = \mu$

$E\sigma^2$ and $E\sigma_a^2$ don't have closed forms.

$$E\sigma^2 = E(MSE \mid \sigma_a^2 \geq 0)\, P(\sigma_a^2 \geq 0) + E(\frac{SST}{mn} \mid \sigma_a^2 < 0)\, P(\sigma_a^2 < 0)$$

generally cannot be computed

Variance

$$var(\hat{\mu}) = var(\bar{y}) = \frac{\sigma_a^2 + \frac{\sigma^2}{n}}{m} = \frac{\sigma^2 + n\sigma_a^2}{mn}$$

$\bar{y}_1, \bar{y}_2 \cdots$

$y_{ij} = \mu + a_i + \varepsilon_{ij}$

$\bar{y}_i = \mu + a_i + \bar{\varepsilon}_i$

$\text{var}\begin{bmatrix} \hat{\sigma}^2 \\ \hat{\sigma}_a^2 \end{bmatrix}$ → no closed form but can derive asymptotic variance b/c it's inverse of fisher information

$\begin{cases} \ell\sigma^2 \\ \ell\lambda \end{cases}$

$\begin{cases} -E\ell\sigma^2\sigma^2 \\ -E\ell\sigma^2\lambda \\ -E\ell\lambda\lambda \\ -E\ell\lambda\sigma^2 \end{cases}$ → Fisher info

$\text{var}\begin{pmatrix} \hat{\sigma}^2 \\ \hat{\lambda} \end{pmatrix}$ → (Fisher Info)$^{-1}$

⇒ $\text{var}\begin{pmatrix} \hat{\sigma}^2 \\ \hat{\sigma}_a^2 \end{pmatrix}$) variable transformation

Lec 3 : April 4

One-way classification

$$\mathrm{Var}(\hat{\mu}) = \mathrm{Var}(\bar{y}) = \frac{\sigma^2 + n\sigma_a^2}{mn}$$

$\hat{\sigma}^2$, $\hat{\sigma}_a^2$

$$\begin{cases} l_{\sigma^2} = \dfrac{\partial l}{\partial \sigma^2} = \dfrac{-m(n-1)}{2\sigma^2} + \dfrac{SSE}{2\sigma^4} \\ l_{\lambda} = \dfrac{\partial l}{\partial \lambda} = \dfrac{-m}{2\lambda} + \dfrac{SSA}{2\lambda^2} + \dfrac{mn(\bar{y}-\mu)^2}{2\lambda^2} \end{cases}$$

$$\begin{cases} l_{\sigma^2\sigma^2} = \dfrac{\partial^2 l}{\partial \sigma^2 \partial \sigma^2} = \dfrac{m(n-1)}{2\sigma^4} - \dfrac{2SSE}{2\sigma^6} \\ l_{\sigma^2\lambda} = \dfrac{\partial^2 l}{\partial \sigma^2 \partial \lambda} = 0 \\ l_{\lambda\lambda} = \dfrac{\partial^2 l}{\partial \lambda \partial \lambda} = \dfrac{m}{2\lambda^2} - \dfrac{2SSA}{2\lambda^3} - \dfrac{mn(\bar{y}-\mu)^2}{2\lambda^3} \end{cases}$$

$$\begin{cases} -E\,l_{\sigma^2\sigma^2} = \dfrac{m(n-1)}{2\sigma^4} \\ -E\,l_{\lambda\lambda} = \dfrac{m}{2\lambda^2} \end{cases}$$

inverse of Fisher information

$$\Rightarrow \mathrm{Var}\begin{bmatrix} \hat{\sigma}^2 \\ \hat{\lambda} \end{bmatrix} \cong \begin{bmatrix} \dfrac{2\sigma^4}{m(n-1)} & 0 \\ 0 & \dfrac{2\lambda^2}{m} \end{bmatrix}$$

$$\hat{\sigma}_a^2 = \frac{\hat{\lambda} - \hat{\sigma}^2}{n}$$

$$\Rightarrow \operatorname{var}\begin{bmatrix}\hat{\sigma}^2\\ \hat{\sigma}_a^2\end{bmatrix} \overset{n\to\infty}{\cong} 2\sigma^4 \begin{bmatrix} \frac{1}{m(n-1)} & \operatorname{cov}\left(\frac{-\sigma^2}{n}, \sigma^2\right) \\ \operatorname{cov}\left(\frac{-\sigma^2}{n}, \sigma^2\right) & \operatorname{var}(\sigma_a^2) \end{bmatrix}$$

$$= 2\sigma^4 \begin{bmatrix} \frac{1}{m(n-1)} & \frac{-1}{mn(n-1)} \\ \frac{-1}{mn(n-1)} & \frac{1}{n^2}\left(\frac{\lambda^2/\sigma^4}{m} + \frac{1}{m(n-1)}\right) \end{bmatrix}$$

$$\begin{bmatrix} y_{11}\\ y_{12}\\ \vdots\\ y_{1n}\\ y_{21}\\ \vdots\\ y_{2n}\\ \vdots\\ y_{m1}\\ \vdots\\ y_{mn} \end{bmatrix} = \begin{array}{c} 1 \ \ldots \ m \\ \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ \vdots & & & & \\ 1 & & & & \\ 0 & 1 & 0 & \cdots & 0 \\ & & & & \\ & & & & \end{bmatrix} \end{array} \begin{bmatrix} \mu_1\\ \vdots\\ \mu_m \end{bmatrix} + \begin{bmatrix} e_{11}\\ e_{12}\\ \vdots \end{bmatrix}$$

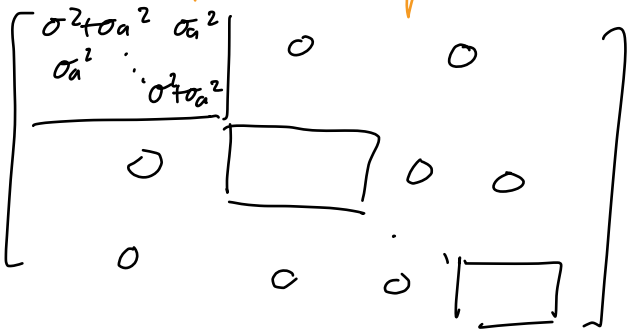
$$X \qquad \beta \qquad + \qquad \varepsilon$$

fixed effects model.

$\mu_i = \mu + a_i$ $\quad a_i \sim N(0, \sigma_a^2)$

$$\begin{bmatrix} y_{11}\\ y_{12}\\ \vdots\\ y_{1n}\\ y_{21}\\ \vdots\\ y_{2n}\\ \vdots\\ y_{m1}\\ \vdots\\ y_{mn} \end{bmatrix} = \begin{bmatrix} 1\\ 1\\ \vdots\\ 1\\ 0\\ \vdots\\ 0\\ \vdots\\ 0\\ \vdots\\ 0 \end{bmatrix} \mu + \underbrace{\begin{bmatrix} & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \\ & & \end{bmatrix}}_{e} \begin{bmatrix} a_1\\ \vdots\\ a_m \end{bmatrix} + \begin{bmatrix} e_{11}\\ e_{12}\\ \vdots \end{bmatrix}$$

$\varepsilon \sim \Sigma$, non-diagonal variance matrix

Σ: 

- have correlation within the same group.

- trying to model a set of correlated data w/ random effects model
  - ↳ common usage of ran ef models
- introduces a more parsimonious model for correlated data.

REML $\quad y = X\beta + Zu + \varepsilon$

- first try to get rid of fixed effects.
- Balanced data: $n_1 = n_2 = \dots = n_m = n$.

$$\mathcal{L}(\mu, \sigma^2, \sigma_a^2; y) \propto \exp\left(\frac{-1}{2}\left[\frac{SSE}{\sigma^2} + \frac{SSA}{\lambda} + \frac{(\bar{y}-\mu)^2}{\lambda/mn}\right]\right)$$

can treat as a normal density function

$\Rightarrow$ can integrate out

$$= L(m, y)\, L(\sigma^2, \sigma_a^2; SSA, SSE)$$

$$\frac{1}{(2\pi)^{1/2}\left(\frac{\lambda}{mn}\right)^{1/2}} \exp\left(\frac{-(\bar{y}-m)^2}{2\lambda/mn}\right)$$

$$L(\sigma^2, \sigma_a^2; y) = \int L(m, \sigma^2, \sigma_a^2; y)\, dm$$

SSA, SSE
Sufficient statistics

REML log-likelihood:

$$l_R = -\frac{1}{2}(mn-1)\log 2\pi - \frac{1}{2}\log mn$$
$$-\frac{1}{2}m(n-1)\log\sigma^2 - \frac{1}{2}(m-1)\log\lambda$$
$$-\frac{SSE}{2\sigma^2} - \frac{SSA}{2\lambda}$$

taking derivative and solving:

$$\begin{cases} \dot{\sigma}^2 = \dfrac{SSE}{m(n-1)} = MSE & \text{same as MLE} \\ \dot{\lambda} = \dfrac{SSA}{n-1} = MSA & \text{mean between group sum of squares} \end{cases}$$

MLE: $\frac{SSA}{m}$

$$\dot{\sigma}_a^2 = \frac{\dot{\lambda} - \dot{\sigma}^2}{n} = \frac{1}{n}(MSA - MSE)$$

Do not necessarily give values in allowed parameter space.

If MSA $\geq$ MSE: $\hat{\sigma}_a^2 \geq 0$

$$\begin{cases} \hat{\sigma}^2 = MSE \\ \hat{\sigma}_a^2 = \frac{1}{n}(MSA - MSE) \end{cases}$$

If MSA < MSE: $\hat{\sigma}_a^2 < 0$

$$\begin{cases} \hat{\sigma}^2 = \frac{SST}{mn-1} \\ \hat{\sigma}_a^2 = 0 \end{cases} \Rightarrow \text{no within group variance}$$

- Derivations are more complicated for unbalanced data.

Expectation - Maximization (EM)

pdf $f(X, Y \mid \theta)$ → complete data likelihood

(observed, missing) → complete data

$f(X \mid \theta)$ can be derived by integrating Y out.

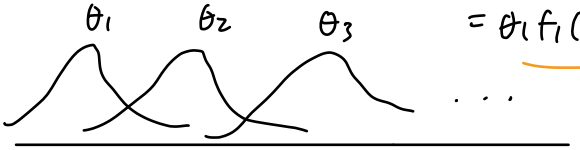
$= \int f(X, y \mid \theta) dy$ → marginal prob. for observed data.

Goal: find MLE for observed data likelihood

(find θ to maximize $\int f(X, y \mid \theta) dy$)

- problem: not easy to work w/ observed data likelihood. easier to work w/ complete data likelihood.

ex: $X_1, \dots, X_n \overset{iid}{\sim} f(x|\theta)$



$= \theta_1 f_1(x) + \theta_2 f_2(x) + \dots + \theta_k f_k(x)$ — mixture distribution

integrates to 1 ⇒ valid pdf

$\sum_i \theta_i = 1$, θ: mixing probabilities

Assume f_i's are pre-specified. only unknowns are θ's.

MLE approach: write down likelihood for obs. data

$$\ell(\theta) = \log L(\theta) = \sum_{j=1}^{n} \log\left[\theta_1 f_1(x_j) + \dots + \theta_k f_k(x_j)\right]$$

take deriv. + solve for each θ_i.

$$\begin{cases} \dfrac{\partial}{\partial \theta_k} = \displaystyle\sum \frac{f_k(\,)}{\theta_1 f_1(\,) + \dots + \theta_k f_k(\,)} = 0 \\ \text{subject to } \sum_i \theta_i = 1 \end{cases}$$

difficult to solve.

Instead, we can use the complete data likelihood. introduce a set of indicators to indicate which of the normal components was used to generate the data.

Step 1: randomly pick a category $1, \dots, K$ based on multinomial prob's $\theta_1, \dots, \theta_K$

Step 2: sample X from f_k, your selected category.

introduce indicators to augment the system:

indicators: $U_1, \dots, U_n$

$$U_j = k \Rightarrow X_i \sim f_k(\ )$$

a priori.

$$\begin{cases} P(U_j = k) = \theta_k \\ f(x_j \mid U_j = k) = f_k(x_j) \end{cases}$$

two sets of RVs: U's, X's. only X's are observed.

missing

$(U, X) \rightarrow$ complete data. can now describe complete data likelihood.

$$f(X, U \mid \theta) = \prod_{j=1}^{n} \left\{ \prod_{i=1}^{K} [\theta_i f_i(x_j)]^{I(U_j = i)} \right\}$$

multinomial likelihood

joint probability for U and X

viewing as a function of θ, it's the complete data likelihood.

- can count # points in each category if U_j's are known $\Rightarrow$ turns $f(u, x|\theta)$ into multinomial likelihood.
- But U's (membership) are not known a priori.
- But if you know θ and X, you can infer U using Bayes Thm.

iterate between these two steps (EM algorithm)

EM algorithm

- Recall the goal: Find MLE for $\mathcal{L}(\theta) = f(X|\theta)$

Define $Q(\theta|\theta_n) = E_{Y|X,\theta_n} \log f(X, Y|\theta)$ — using complete data likelihood

(Y is missing, try to get its posterior then integrate out Y from complete data likelihood.

$$= \int [\log f(X, Y|\theta)] f(Y|X, \theta_n)\, dY$$

can be computed from joint and marginal of X

(assume joint dist. is known.

Start w/ initial estimate θ_0.

↓

E-step: compute $Q(\theta|\theta_n)$

↓

M-step: try to find θ that maximizes $Q(\theta|\theta_n)$

$$\theta_{n+1} = \underset{\theta}{\operatorname{argmax}}\, Q(\theta|\theta_n)$$

iteratively run E-and M-step until convergence.

In this way, the observed data likelihood will not decrease

$$l(\theta_{n+1}) \geq l(\theta_n)$$ (hill climbing algorithm)

· algorithm will converge to a local maximum.

Lec 4: April 6

Expectation Maximization

- observed data likelihood:

$f(x|\theta)$ ← not easy to maximize

introduce: $f(x, y|\theta)$ (augmented system)

complete data likelihood

$$Q(\theta|\theta^n) = E_{y|x} \log f(x, y|\theta)$$

↑ current parameter estimate

- why does this algorithm work?

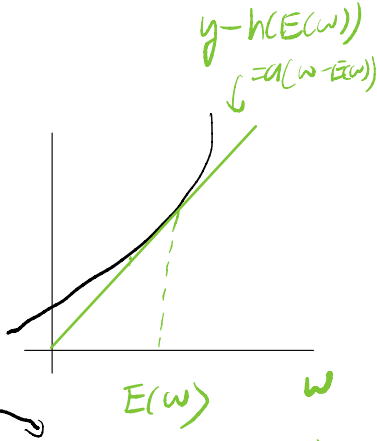
① Jensen's inequality

W : RV

$h(x)$: convex function

Then, $E(h(W)) \geq h(E(W))$

equation holds iff $W = EW$ a.s.



$h(W) \geq y = h(EW) + a(W - EW)$

$E(h(W)) \geq E(y) = h(EW)$

② Information inequality

consider: $f(x)$, $g(x)$

$$E_f \log f(x) \geq E_f \log g(x)$$

$$E_f \log f(x) - E_f \log g(x)$$

$$= E_f \log \frac{f}{g}$$

$$= E_f\left[-\log \frac{g}{f}\right]$$

$$\int \left[\log \frac{f(x)}{g(x)}\right] f(x)\,dx$$

convex

by Jensen's inequality,

$$E_f\left[-\log \frac{g}{f}\right] \geq -\log\left(E_f\left(\frac{g}{f}\right)\right)$$

$$= -\log\left(\int \frac{g(x)}{f(x)} f(x)\,dx\right)$$

b/c g is a density so it integrates to 1

$$= -\log(1) = 0$$

$$\Rightarrow E_f \log f(x) \geq E_f \log g(x).$$

③ The ascent property

$$\log f(x, y \mid \theta) = \log f(x \mid \theta) + \log f(y \mid x, \theta)$$

part we are trying to maximize

$$\log f(x \mid \theta) = \log f(x, y \mid \theta) - \log f(y \mid x, \theta)$$

taking expectation:

$$E_{Y|X,\theta^n} \log f(x|\theta) = E_{Y|X,\theta^n} \log f(x, y|\theta) - E_{Y|X,\theta^n} \log f(y|x,\theta)$$

Q

doesn't contain y

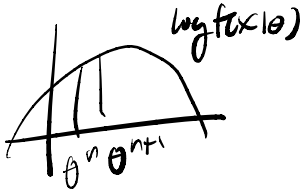
$$\log f(x|\theta) = Q(\theta|\theta^n) - E_{Y|X,\theta^n} \log f(y|x,\theta)$$

in EM, we compute Q and find θ^{n+1} that maximizes Q.

(1) $Q(\theta^{n+1}|\theta^n) \geq Q(\theta^n|\theta^n)$

(2) $-E_{Y|X,\theta^n} \log f(y|x,\theta^{n+1}) \geq -E_{Y|X,\theta^n} \log f(y|x,\theta^n)$

not the same distributions



old estimate: $\log f(x|\theta^n)$

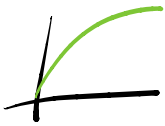
new estimate: $\log f(x|\theta^{n+1})$

$\log(x|\theta^n) \leq \log f(x|\theta^{n+1})$ by the above two inequalities

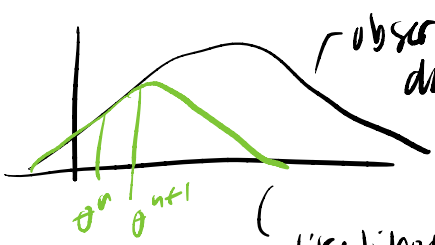
E-step: Evaluate $Q(\theta|\theta^n)$

M-step: Try to find θ to maximize Q

$\theta^{n+1} = \underset{\theta}{\operatorname{argmax}}\, Q(\theta|\theta^n)$



observed data log like



likelihood used to approximate observed

Example 1: Mixture model

$X_1, \ldots, X_n$ independently drawn from

$f(x) = \theta_1 f_1(x) + \theta_2 f_2(x) + \ldots + \theta_K f_K(x)$

introduce indicators: $u_1, \ldots, u_n$

view θ's as multinomial probabilities

$$f(X, u|\theta) = \prod_{j=1}^{n} \prod_{i=1}^{K} [\theta_i f_i(X_j)]^{I(u_j=i)}$$

$$\log f(X, u|\theta) = \sum_{j=1}^{n} \sum_{i=1}^{K} \{\log \theta_i + \log f_i(X_j)\} I(u_j=i)$$

Assume given

constant

Goal: find θ that maximizes the loglik

E-step: $u|x,\theta^t$

$$E_{u|x,\theta^t} I(u_j = i)$$

$$= P(u_j = i \mid x, \theta^n)$$

$$= \frac{\theta_i^t f_i(x_j)}{\sum_i \theta_i^t f_i(x_j)} \triangleq \tilde{u}_{j,i}$$

then

$$E \log f(x, u \mid \theta) = \sum_{j=1}^{n} \sum_{i=1}^{K} \tilde{u}_{ji} \{\log \theta_i + \log f_i(x_j)\}$$

$$Q(\theta \mid \theta^n)$$

M-step: $Q(\theta \mid \theta^t) = \sum_{j=1}^{n} \sum_{i=1}^{K} \tilde{u}_{ji} \log \theta_i + \text{const}$

($\tilde{u}_{ji}$: pseudocounts)

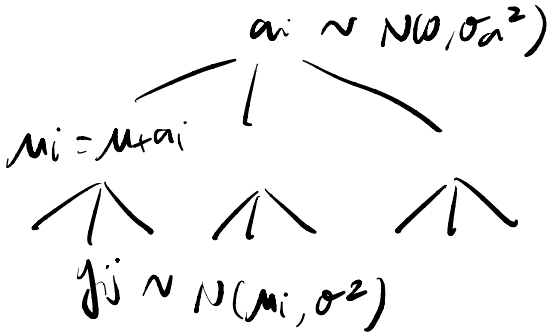
multinomial loglik

$$= \sum_{i=1}^{K} \left\{ \left[\sum_{j=1}^{n} \tilde{u}_{ji} \right] \log \theta_i \right\}$$

$$\hat{\theta}_i^{t+1} = \frac{\sum_{j=1}^{n} \tilde{u}_{ji}}{\sum_{i=1}^{K} \sum_{j=1}^{n} \tilde{u}_{ji}}$$

$$= \frac{\sum_{j=1}^{n} \tilde{u}_{ji}}{n}$$

Example 2: Mixed effects model



Assume each group has n_i datapoints (not the same across groups)

$$P(y, \mu_i \mid \mu, \sigma^2, \sigma_a^2)$$

(σ over σ^2, σ_a^2)

obs — missing

- $p(y \mid \mu, \sigma^2, \sigma_a^2)$ is too hard to work with
- use complete data likelihood instead.

$$p(y, a_i \mid \mu, \sigma^2, \sigma_a^2)$$

$$= \prod_{i=1}^{m} \left\{ \left[\prod_{j=1}^{n_i} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ \frac{-(y_{ij} - a_i)^2}{2\sigma^2} \right\} \right] \frac{1}{\sqrt{2\pi\sigma_a^2}} \exp\left\{ \frac{-(a_i - \mu)^2}{2\sigma_a^2} \right\} \right\}$$

generate a_i

$$= \prod_{i=1}^{m} (2\pi\sigma^2)^{-n_i/2} (2\pi\sigma_a^2)^{-1/2} \exp\left\{ -\frac{\sum (y_{ij} - a_i)^2}{2\sigma^2} - \frac{(a_i - \mu)^2}{2\sigma_a^2} \right\}$$

a_i follows normal dist. conditional on y and other params

Posterior dist. for a_i: prior + data means weighted by amount of info from both.

$$a_i \mid y, \mu, \sigma^2, \sigma_a^2 \sim N\left(\frac{\frac{n_i}{\sigma^2}\bar{y}_i + \frac{1}{\sigma_a^2}\mu}{\frac{n_i}{\sigma^2} + \frac{1}{\sigma_a^2}}, \frac{1}{\frac{n_i}{\sigma^2} + \frac{1}{\sigma_a^2}} \right)$$

amount of info from data

amount of info from prior

derive for HW

complete data

$$\ell_c = \log P(Y, \mu_i \mid \mu, \sigma^2, \sigma_a^2)$$

$$= -\frac{1}{2}(N+m)\log 2\pi - \frac{N}{2}\log\sigma^2 - \frac{m}{2}\log\sigma_a^2$$

$$- \sum_{i=1}^{m}\sum_{j=1}^{n_i} \frac{(y_{ij}-\mu_i)^2}{2\sigma^2} - \sum_{i=1}^{m} \frac{(\mu_i-\mu)^2}{2\sigma_a^2}$$

$$= -\frac{1}{2}(N+m)\log 2\pi - \frac{N}{2}\log\sigma^2 - \frac{m}{2}\log\sigma_a^2$$

$$-\frac{1}{2\sigma^2}\sum_{i=1}^{m}\sum_{j=1}^{n_i}(y_{ij}^2 - 2\mu_i y_{ij} + \mu_i^2)$$

$$-\frac{1}{2\sigma_a^2}\sum_{i=1}^{m}(\mu_i^2 - 2\mu_i\mu + \mu^2)$$

terms involving missing data

E-step: compute $Q(\theta \mid \theta^{(t)}) = E_{\mu_i \mid Y}\, \ell_c$

μ_i's are the missing data

Need to evaluate $E(\mu_i \mid Y, \mu, \sigma^2, \sigma_a^2)$

$E(\mu_i^2 \mid Y, \mu, \sigma^2, \sigma_a^2)$

we already know the distribution of μ_i, so these expectations are easy to evaluate.

$$E(u_i \mid y_i, \mu^t, \sigma^{2t}, \sigma_a^{2t}) = \frac{\frac{n_i}{\sigma^2}\bar{y}_{i\cdot} + \frac{1}{\sigma_a^2}\mu}{\frac{n_i}{\sigma^2} + \frac{1}{\sigma_a^2}} \triangleq \tilde{u}_i$$

$$E(u_i^2 \mid y_i, \mu^t, \sigma^{2t}, \sigma_a^{2t}) = \frac{1}{\frac{n_i}{\sigma^2} + \frac{1}{\sigma_a^2}} + \tilde{u}_i^2 \triangleq \tilde{\eta}_i$$

in l_c, replace u_i w/ $\tilde{u}_i$, u_i^2 w/ $\tilde{\eta}_i$ to get Q function.

M-Step:

$$\mu^{t+1} = \frac{1}{m}\sum_{i=1}^{m} \tilde{u}_i$$

(maximized by taking deriv. and solving)

$$\sigma^{2\,t+1} = \frac{1}{N}\sum_{i=1}^{m}\sum_{j=1}^{n}(y_{ij}^2 - 2y_{ij}\tilde{u}_i + \tilde{\eta}_i)$$

similar to sum of squared residuals.

$$\sigma_a^{2\,t+1} = \frac{1}{m}\left\{\sum_{i=1}^{m}\tilde{\eta}_i - \left(\frac{\sum_{i=1}^{m}\tilde{u}_i}{m}\right)^2 m\right\}$$

$$= \frac{1}{m} \sum_{i=1}^{m} \tilde{\eta}_i - \left[\mu^{t+1} \right]^2$$

Lec 5: April 11

Random effects model:

$\mu, \sigma^2, \sigma_a^2$ $\quad \sigma_a^2 = 0 \Rightarrow$ all means are equal

$N(\mu, \sigma_a^2)$

m = # groups

Balanced data

$\hat{\mu}: \bar{y} \sim N\left(\mu, \frac{\sigma_a^2 + \frac{\sigma^2}{n}}{m}\right)$ $\quad \frac{s^2}{\sigma^2} \sim \frac{\chi^2_{N-m}}{N-m}$

$\hat{\sigma}^2$: $\text{MSE} = \frac{\sum\sum (y_{ij} - \bar{y}_i)^2}{N-m} \triangleq s^2$

$\hat{\sigma}_a^2$: $\quad H_0: \sigma_a^2 = 0$ $\quad \lambda = \sigma^2 + n\sigma_a^2$

$\Rightarrow \lambda = \sigma^2$ $\quad$ used to estimate σ_a^2

$\uparrow$

estimated using MSA (between group variance divided by corresponding df)

$$\frac{MSA}{MSE} = \frac{SSA/(m-1)}{SSE/(N-m)} \sim F_{m-1,\,N-m}$$

used to evaluate whether $\sigma_a^2 = 0$

unbalanced data:

- take deriv. of obs. data likelihood, can also take 2nd deriv.

$$\frac{\partial \log f(x|\theta)}{\partial \theta}$$ — plug in $\hat{\theta}$ and evaluate (=0)

$$\frac{\partial^2 \log f(x|\theta)}{\partial \theta^2}\Big|_{\hat{\theta}} \Rightarrow$$ can evaluate Fisher information → take inverse to get asymptotic variance.

- since these are MLEs, we can use asymptotic properties of MLEs to do inference.

- In mixed effect models, sometimes we are interested in estimating μ_i's (group specific means)

Predicting random effects:

X, Y are RVs

Y is unobserved, X is observed.

find $g(x)$ that can predict Y.

The best predictor is the one with minimum squared error.

$E(Y - g(x))^2$ } $g(x)$ should minimize this.

$g(x) = E(Y|X)$ minimizes this.

so best predictor of u_i's is

$E(u_i | y, \underbrace{\mu, \sigma^2, \sigma_u^2}_{\text{estimated using EM}})$ last time: showed u_i has a normal dist.

$$= \frac{\frac{n_i}{\sigma^2}}{\frac{n_i}{\sigma^2} + \frac{1}{\sigma_u^2}} \bar{y}_i + \frac{\frac{1}{\sigma_u^2}}{\frac{n_i}{\sigma^2} + \frac{1}{\sigma_u^2}} \mu$$

shrinkage factor

combination of sample group mean and prior mean.

- plug in EM MLE's of $\mu, \sigma^2, \sigma_a^2$

$$\mu_i = \mu + a_i \qquad a_i \sim N(0, \sigma_a^2)$$

$$E(a_i \mid y_i, \mu, \sigma^2, \sigma_a^2)$$

b/c $a_i = \mu_i - \mu$

$$= \frac{\frac{n_i}{\sigma^2}}{\frac{n_i}{\sigma^2} + \frac{1}{\sigma_a^2}} (\bar{y}_i - \mu) + \frac{\frac{1}{\sigma_a^2}}{\frac{n_i}{\sigma^2} + \frac{1}{\sigma_a^2}} 0$$

Note: we are trying to pull the sample mean ($\bar{y}_i$) towards its prior (μ) in $E(\mu_i \mid y, \sigma^2, \sigma_a^2, \mu)$

And trying to pull a_i estimate to 0.

Both cases: trying to shrink the data driven estimate towards the prior.

Posterior mean also called shrinkage estimator

Let B = shrinkage factor

$$= \frac{\frac{1}{\sigma_a^2}}{\frac{n_i}{\sigma^2} + \frac{1}{\sigma_a^2}}$$

σ_a^2 large → info contributed by prior is small, rely on data more.

$$E(\mu_i | y, \mu, \sigma^2, \sigma_\mu^2) = (1-B) y_i + B\mu$$

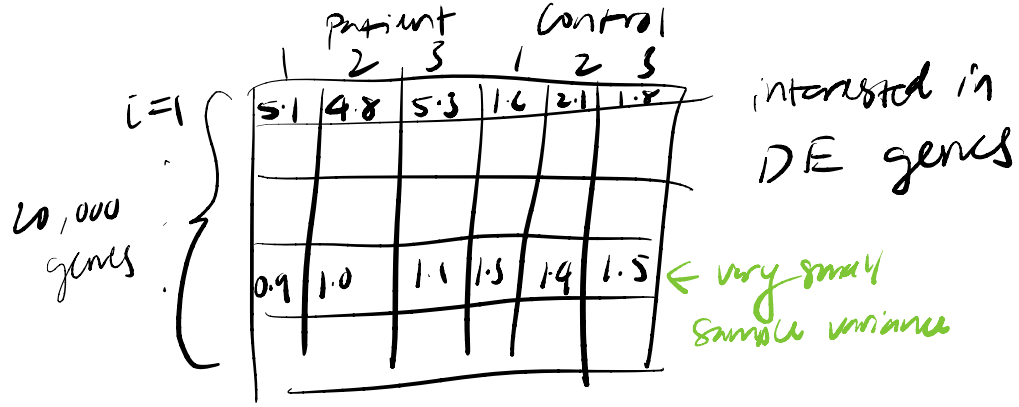
σ_μ^2 small → more info contributed by prior.

- groups very different → mostly use data
- groups similar → mostly use prior

Empirical Bayes

- used data from all groups to estimate overall mean
- for group specific mean: combined overall mean w/ group-specific empirical mean (borrowed info from other groups)

Genomics example:



- can run a t-test for each row.

H_0: patient mean gene expr = control mean gene expr

- lots of false positives (variance can't be stably estimated w/ small sample size)

$$\frac{\bar{x} - \bar{y}}{\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right) s^2}}$$

assume the two groups have same variance

$$s^2 = \frac{\sum (x_i - \bar{x})^2 + \sum (y_i - \bar{y})^2}{n_1 + n_2 - 2}$$ } pooled sample variance

Suppose data is normally distributed:

$X \sim N(\mu_X, \sigma^2)$, $Y \sim N(\mu_Y, \sigma^2)$

$$\frac{df\, s^2}{\sigma^2} \sim \chi^2_{df}$$

probability of getting a very small sample variance is low, but we are considering ~20,000 genes.

Scenario: No DE genes.

But, mean difference is non-negligible, and s^2 is small. A few hundred out of 20,000 will have this → huge t-statistic (false positive)

- very low statistical power to detect true DE genes.

using shrinkage to solve this problem:

$$x_{ijk} \mid \mu_{ij}, \sigma_i^2 \sim N(\mu_{ij}, \sigma_i^2)$$

- i: gene
- j: group 1: patient, 2: control
- k: sample replicate 1, 2, 3

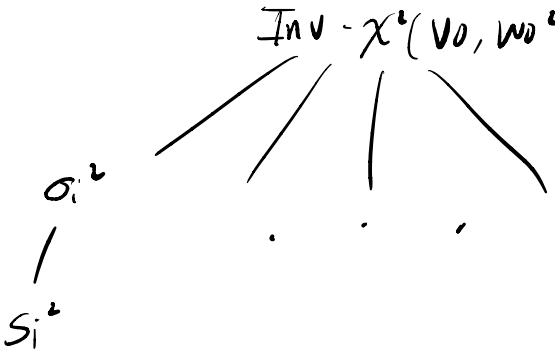
$$s_i^2 = \frac{\sum\sum (x_{ijk} - \bar{x}_{ij})^2}{n_1 + n_2 - 2}$$

$\nu = n_1 + n_2 - 2$

$$s_i^2 \mid \sigma_i^2 \sim \frac{\sigma_i^2 \chi_\nu^2}{\nu}$$ under normality assumption

If we assume σ_i^2 are random effects:

$$\sigma_i^2 \sim \text{Inv-}\chi^2(\nu_0, w_0^2)$$



$$X \sim \chi_\nu^2, \quad Y = \frac{\nu_0 \cdot w_0^2}{X} \sim \text{Inv-}\chi^2(\nu_0, w_0^2)$$

conjugate prior for χ^2 ⇒ posterior will be Inv-χ^2

PDF of Inv-$\chi^2(\nu, w^2)$:

$$f(\theta) = \frac{\left(\frac{\nu}{2}\right)^{\nu/2}}{\Gamma\left(\frac{\nu}{2}\right)} (w^2)^{\frac{\nu}{2}} \theta^{-\left(\frac{\nu}{2}+1\right)} e^{\frac{-\nu w^2}{2\theta}}$$

$$E(\theta) = \frac{v}{v-2} w^2$$

$$var(\theta) = \frac{2v^2}{(v-2)^2(v-4)} w^4$$

$$mode(\theta) = \frac{v}{v+2} w^2$$

$$E s_i^2 = E \sigma_i^2 = \frac{v_0}{v_0 - 2} w_0^2$$

$$\hat{\sigma}_i^2 = E(\sigma_i^2 | s_i^2, v_0, w_0^2)$$

$$= (1 - B) s_i^2 + B\left(\frac{v_0}{v_0 - 2} w_0^2\right)$$

(data-based estimate; prior mean for σ_i^2)

Shrinkage estimator

$$B = \frac{E(var(s_i | \sigma_i^2))}{var(s_i^2)}$$

ratio between within group var and between group var.

integrate missing data out to get marginal dist of s_i^2

- borrowing information from other genes to estimate gene i's true variance

In the normal model:

$$B = \frac{\frac{1}{\sigma_a^2}}{\frac{n_i}{\sigma_i^2} + \frac{1}{\sigma_a^2}} \cdot \frac{\frac{\sigma_a^2 \sigma^2}{n_i}}{\frac{\sigma_a^2 \cdot \sigma^2}{n_i}}$$

$$= \frac{\frac{\sigma^2}{n_i}}{\frac{\sigma^2}{n_i} + \sigma_a^2}$$

$\frac{\sigma^2}{n_i}$ — within group variance

$\frac{\sigma^2}{n_i} + \sigma_a^2$ — total variance

Lec 6: April 13

HW 2: due April 27

Proj. 3: random + mixed effects
districts of Baltimore

Shrinkage:

$(1-B)$ Data + B Prior

$$B = \frac{\text{within-group variance}}{\text{total variance}}$$

$$= \frac{E[\text{var}(s_i^2 | \sigma_i^2)]}{\text{var } s_i^2} = \frac{v_0 - 2}{v_0 + v_i - 2}$$

prior has higher df ($\Rightarrow$ contains more info, give it more weight)

$$s_i^2 | \sigma_i^2 \sim \frac{\sigma_i^2 \chi_{v_i}^2}{v_i}$$

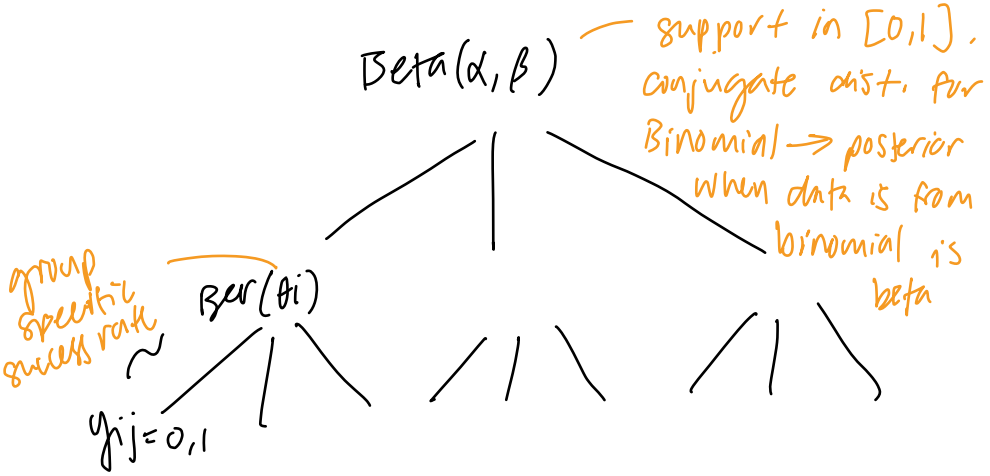
df contributed by the prior / df contributed by gene i

$$\sigma_i^2 \sim \frac{v_0 w_0^2}{\chi_{v_0}^2}$$

- small within-group variance → more difference between groups, B should be close to 0.

How to handle non-normal models?

- Fixed effects model: compute p_i from data.
- Mixed effects: p_i are random



$$p(y, \theta \mid \alpha, \beta)$$

m groups, indep.

$$= \prod_{i=1}^{m} \left[\frac{\theta_i^{\alpha-1}(1-\theta_i)^{\beta-1}}{\text{Beta}(\alpha,\beta)} \prod_{j=1}^{n_i} \left[\theta_i^{y_{ij}} (1-\theta_i)^{1-y_{ij}} \right] \right]$$

generate θ

$$\propto \prod_{i=1}^{m} \theta_i^{\alpha + y_{i+} - 1} (1-\theta_i)^{\beta + n_i - y_{i+} - 1}$$

y_{i+}: sum of all y_{ij}'s in the ith group (total # success)

$n_i - y_{i+}$: total # failures

joint dist. of θ, y.

has form of Beta dist.

if we treat y as known,

$$p(\theta \mid y, \alpha, \beta) \sim \text{Beta}(\alpha + y_{i+}, \beta + m_i - y_{i+})$$

best predictor of θ_i is posterior mean

$$E(y_{ij}) = E[E(y_{ij} \mid \theta_i)]$$

$$= E\theta_i$$

$$= \frac{\alpha}{\alpha+\beta}$$

view α, β as pseudocounts
- α: prior count of first category
- β: prior count of second category

$$\text{var}(y_{ij}) = \text{var}[E(y_{ij} \mid \theta_i)] + E[\text{var}(y_{ij} \mid \theta_i)]$$

$$= \frac{\alpha\beta}{(\alpha+\beta)^2} \quad = E(y_{ij})[1 - E(y_{ij})]$$

$\text{cov}(y_{ij}, y_{kl}) = 0$ if $i \neq k$ b/c their θ's are independently generated.

$$\text{cov}(y_{ij}, y_{il}) = \text{cov}[E(y_{ij} \mid \theta_i), E(y_{il} \mid \theta_i)]$$

$$+ E[\text{cov}(y_{ij}, y_{il} \mid \theta_i)]$$

independently generated given θ_i

$$= \text{cov}(\theta_i, \theta_i) + E(0)$$

$$= \text{var}(\theta_i) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$$

How to estimate α, β, and predict θ_i?

$$p(y \mid \alpha, \beta) = \int p(y, \theta \mid \alpha, \beta)\, d\theta$$

$$= \prod_{i=1}^{m} \frac{\text{Beta}(\alpha + y_{it},\ \beta + n_i - y_{it})}{\text{Beta}(\alpha, \beta)}$$

$$\log p(y \mid \alpha, \beta) = \sum_{i=1}^{m} \log\left(\frac{\text{Beta}(\alpha + y_{it},\ \beta + n_i - y_{it})}{\text{Beta}(\alpha, \beta)}\right)$$

Beta functions can be expressed as gamma functions, log-gamma can be computed in R.

posterior mean is the best predictor for θ_i

$$E(\theta_i \mid y, \alpha, \beta) = \frac{\alpha + y_{it}}{\alpha + y_{it} + \beta + n_i - y_{it}}$$

$$= \frac{\alpha + y_{it}}{\alpha + \beta + n_i}$$

$$= \frac{\alpha+\beta}{\alpha+\beta+n_i}\left(\frac{\alpha}{\alpha+\beta}\right) + \frac{n_i}{\alpha+\beta+n_i}\left(\frac{y_{it}}{n_i}\right)$$

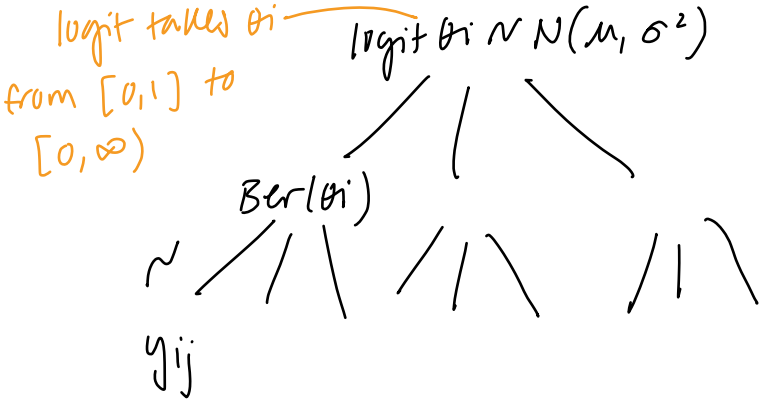
prior mean

observed success rate

only use this term in the fixed effects model

reduces MSE b/c variance is reduced.

If we use a different prior:



$$p(y, \theta \mid \mu, \sigma^2) = \prod_{i=1}^{m} \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(\text{logit}\theta_i - \mu)^2}{2\sigma^2}} \prod_{j=1}^{n_i} \theta_i^{y_{ij}} (1-\theta_i)^{1-y_{ij}}$$

generate θ_i from Normal

- $p(y, \theta \mid \mu, \sigma^2)$ is hard to integrate wrt θ
- EM also does not help b/c can't compute expectation
- use Markov chain Monte Carlo (MCMC)

Markov Chains

1. A MC is a stochastic process $\{X_i\}$, such that

$$P(X_i \mid X_0, \dots X_{i-1}) = P(X_i \mid X_{i-1})$$

the current state of the chain depends only on the immediate previous state

State space: the collection of values of X_i

★ $i = 0, 1, 2, \dots$ discrete time

t – continuous time

- finite state space
- ★ countable state space

2. One-step transition probability:

$$P_{ij}^{n, n+1} = \Pr\{X_{n+1} = j \mid X_n = i\}$$

(time)

If $P_{ij}^{n, n+1} = P_{ij}$ (doesn't depend on time), it is a

Stationary transition probability.

$$Pr(X_{n+1}=j \mid X_n=i) = Pr(X_i=j \mid X_0=i)$$

$$\sum_j P_{ij} = 1$$

3. n-step transition probability

$$P_{ij}^{(n)} = Pr\{X_n=j \mid X_0=i\}$$

$$P^{(n)} = \begin{array}{c} \\ 1 \\ \vdots \\ S \end{array} \begin{array}{c} \begin{array}{ccc} 1 & \cdots & S \end{array} \\ \begin{bmatrix} & P_{ij}^{(n)} & \\ & & \end{bmatrix}_{S \times S} \end{array}$$

If the MC has a stationary one-step transition probability, then

$$P^{(n)} = P^n, \quad P = \begin{bmatrix} P_{ij} \end{bmatrix}_{S \times S}$$

4. Right Multiplication

$$E[f(X_n) \mid X_{n-1}=i]$$

$$= \sum_j P_{ij} f(j)$$

5. Left multiplication

$$\pi = (\pi_1, \cdots, \pi_S), \quad \sum_{i=1}^{S} \pi_i = 1$$

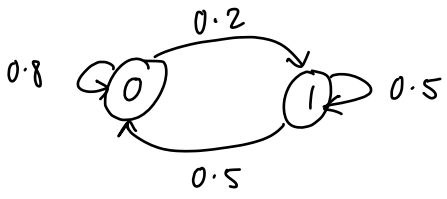
$$P(X_i = j \mid X_0 = i) = P_{ij}$$

$$P(X_i = j) = \sum_i P(X_i = j \mid X_0 = i)\, P(X_0 = i)$$

$$= \sum_i \pi_i P_{ij}$$

Lec 7: April 18

· Markov chains



$$P = \begin{array}{c} \\ 0 \\ 1 \end{array} \begin{array}{c} \begin{array}{cc} 0 & 1 \end{array} \\ \begin{bmatrix} 0.8 & 0.2 \\ 0.5 & 0.5 \end{bmatrix} \end{array}$$

} one-step transition probability matrix

Assuming stationary transition probabilities:

$$P^{(n)} = P^n$$

Suppose $Pr(X_0 = i) = \pi_i$, then

$$Pr(X_1 = j \mid X_0 = i) = p_{ij}$$

$$Pr(X_1 = j) = \sum_i \pi_i p_{ij}$$

} marginalize over X_0

If $Pr(X_1 = j) = \pi_j$, $\pi P = \pi$ then π is called the stationary distribution of the Markov chain.

NOT the same as stationary one-step transition probabilities

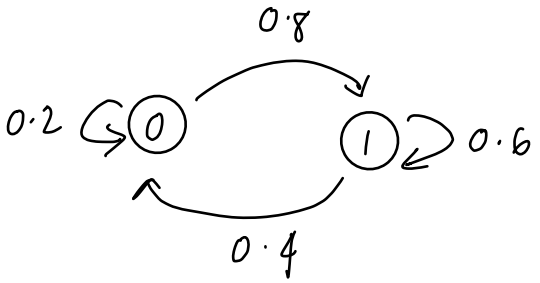
Stationary distribution:

$$\pi = \pi P = \pi P \cdot P = \pi P^n$$

So $Pr(X_n = j) = \pi_j$

Limiting distribution:

$$P = \begin{bmatrix} 0.2 & 0.8 \\ 0.4 & 0.6 \end{bmatrix}$$



$$P^2 = \begin{bmatrix} 0.36 & 0.64 \\ 0.32 & 0.68 \end{bmatrix}$$

$$P^4 = \begin{bmatrix} 0.33 & 0.67 \\ 0.33 & 0.67 \end{bmatrix}$$

each row becomes the same

Suppose you start your chain with state 0, $X_0 = 0$

$$[1 \quad 0] P^n = [0.33 \quad 0.67]$$

probability that $X_n = 0$

probability that $X_n = 1$

Starting from state 1 instead:

$$[0 \quad 1]\, P^n = [0.33 \quad 0.67]$$

regardless of where you start your chain, the probability of visiting the two states at step n will be the same.

For a finite state space MC:

- P is regular if $\exists\, k$ such that P^k has all its elements > 0.

- A MC with regular P has limiting distribution $\pi = (\pi_1, \ldots, \pi_n)$ that satisfy

$$\lim_{n\to\infty} P_{ij}^{(n)} = \pi_j$$ ← does not depend on initial state i

$$\sum_j \pi_j = 1, \quad \pi_j > 0$$

$$\pi \lim_{n\to\infty} P^{(n)} = \pi$$

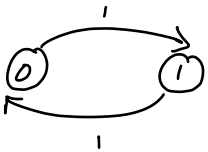
$$\lim_{n\to\infty} \pi P^{n+1} = \pi P^n$$

$$= (\pi P) P^n = \pi P^n \Rightarrow \pi P = \pi$$

Solve this equation to get the limiting distribution (soln is unique)

- Stationary distribution ≠ limiting distribution
- limiting ⇒ stationary, stationary ⇏ limiting
- ex:

$$P = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$$



$$\pi = [0.5 \quad 0.5]$$

$$\pi P = [0.5 \quad 0.5] = \pi \text{ (stationary)}$$

$$X_0 = 0 \quad P_{01}^{(1)} = 1, \quad P_{01}^{(2)} = 0 \quad P_{01}^{(3)} = 1$$

$$P_{01}^{(n)} = 1, 0, 1, 0, 1, 0 \cdots$$

$$P_{00}^{(n)} = 0, 1, 0, 1, 0, 1 \cdots$$

$$P_{10}^{(n)} = 1, 0, 1, 0, 1, 0 \cdots$$

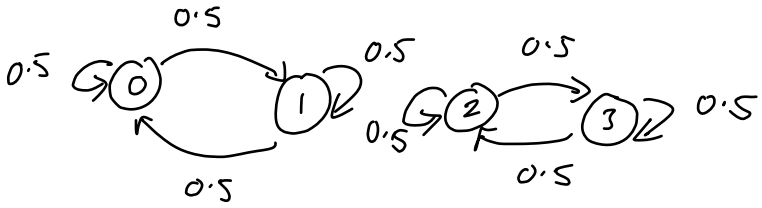
$$P_{11}^{(n)} = 0, 1, 0, 1, 0, 1 \cdots$$

don't have limits as $n \to \infty$ ⇒ no limiting distribution

- can't find a k such that P^k has all entries > 0

$$P^2 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad P^3 = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \cdots$$

- ex:



$$P = \begin{array}{c} \begin{matrix} 0 & 1 & 2 & 3 \end{matrix} \\ \begin{bmatrix} 0.5 & 0.5 & 0 & 0 \\ 0.5 & 0.5 & 0 & 0 \\ 0 & 0 & 0.5 & 0.5 \\ 0 & 0 & 0.5 & 0.5 \end{bmatrix} \end{array} \begin{matrix} 0 \\ 1 \\ 2 \\ 3 \end{matrix}$$

- this MC has a stationary distribution.

$$\pi = [0.25 \quad 0.25 \quad 0.25 \quad 0.25]$$

$$\pi P = \pi$$

$X_0 = 0$, $P_{00}^{(1)} = 0.5$ $P_{00}^{(2)} = 0.5$

$X_0 = 2$ $P_{20}^{(1)} = 0$ $P_{20}^{(2)} = 0$

have different limits (depends on initial state i)

thus, we don't have a limiting distribution.

- If you have a stationary distribution, your limiting distribution will be unique.

- Reasons for not having a limit:
  - periodic behavior (first example)
  - two different groups of states (second example)

Classification of states of a MC

① accessible: $i \to j$ if $\exists\, n$ such that $P_{ij}^{(n)} > 0$

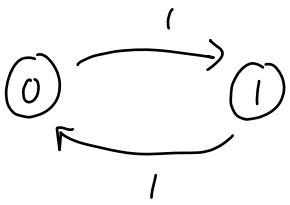
- initial state is i, after n-step transitions, there is non-zero probability of transitioning to n.

② communicate: $i \leftrightarrow j \iff \begin{matrix} i \to j \\ j \to i \end{matrix}$

③ Irreducible: $\forall\, i, j$, $i \leftrightarrow j$ or $P_{ij}^{(n)} > 0$ for some n.

④ Period of a state i, $d(i)$:

the greatest common divisor (GCD) of all integers $n \geq 1$ for which $P_{ii}^{(n)} > 0$



$d(0) = ?$

$P_{00}^{(1)} = 0$

$P_{00}^{(2)} = 1 \longrightarrow n = 2$

$P_{00}^{(3)} = 0$

$P_{00}^{(4)} = 1 \longrightarrow n = 4$

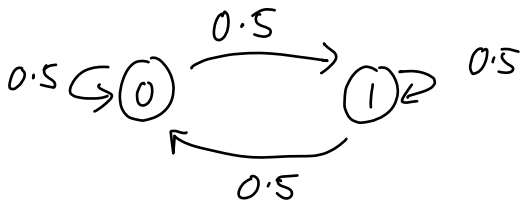
$\vdots$

> 0

sequence of n's : 2, 4, 6, 8, ...

$GCD = 2 = d(0)$

$d(1) = 2$ as well.



$$P = \begin{bmatrix} 0.5 & 0.5 \\ 0.5 & 0.5 \end{bmatrix}$$

$P_{00}^{(1)} = 0.5$ (1) $\quad P_{00}^{(2)} > 0$ (2) $\quad > 0$ (3 ...)

$d(0) = 1 \Rightarrow 0$ is an aperiodic state

⑤ Aperiodic: period = 1

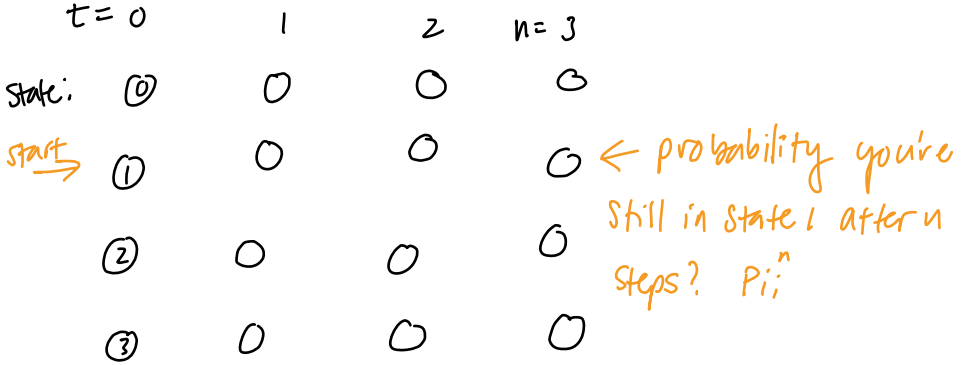
- If $i \leftrightarrow j$, then $d(i) = d(j)$
- In an irreducible MC, all states have the same period.
  - Period is a class property for irreducible MCs.

⑥ $f_{ii}^{(n)} = Pr\{X_n = i,\ X_v \neq i \text{ for } v = 1, \dots, n-1 \mid X_0 = i\}$

The probability that starting from i, the first return to i occurs at the n^{th} transition.

$f_{ii}^{0} = 0$, $f_{ii}^{1} = P_{ii}$.

$$P_{ii}^{(n)} = \sum_{k=0}^{n} f_{ii}^{k} P_{ii}^{(n-k)}$$



$\{X_0 = i,\ X_n = i\} = \{\text{1st return at } n=1\} \cup \{2\} \cup \{3\} \cdots$

disjoint events

Adding them up shows

$$P_{ii}^{n} = \sum_{k=0}^{n} f_{ii}^{k} P_{ii}^{(n-k)}$$

⑦ recurrent:

State i is recurrent $\Leftrightarrow \sum_{n=1}^{\infty} f_{ii}^{n} = 1$

i.e. the probability to return to state i after some finite length of time is equal to 1.

· class property

⑧ transient: a non-recurrent state

i.e. $\sum_{n=1}^{\infty} f_{ii}^{n} < 1$

i is recurrent if and only if $\sum_{n=1}^{\infty} P_{ii}^{(n)} = \infty$

$$\sum_{n=1}^{\infty} E\left[I(X_n = i) \mid X_0 = i \right] = \sum_{n=1}^{\infty} P_{ii}^{(n)} = \infty$$

count # times you've visited state i

② Positive recurrent (strongly ergodic):

$$\sum_{n=0}^{\infty} n f_{ii}^{n} < \infty$$

Null recurrent (weakly ergodic):

$$\sum_{n=0}^{\infty} n f_{ii}^{n} = \infty$$

} expected length of time for first occurrence

· class properties $i \leftrightarrow j$, i positive recurrent $\Rightarrow$ j positive recurrent.

Lec 8: April 20

Markov chains

- An irreducible MC with finite state space is positive recurrent.

Basic Limit Theorem of MCs:

① A recurrent, irreducible, aperiodic MC:

$$\lim_{n\to\infty} P_{ii}^{(n)} = \frac{1}{\sum_{n=0}^{\infty} n f_{ii}^{n}}$$

} probability that first return occurs at time n

$$\triangleq \frac{1}{m_i} = 0 \quad \text{if MC is null-recurrent}$$

As well,

$$\lim_{n\to\infty} P_{ij}^{(n)} = \lim_{n\to\infty} P_{ii}^{(n)}$$

converge to a number that doesn't depend on i

② For positive recurrent, irreducible, aperiodic MCs:

$$\lim_{n\to\infty} P_{jj}^{(n)} = \pi_j = \sum_{i=0}^{\infty} \pi_i P_{ij}$$

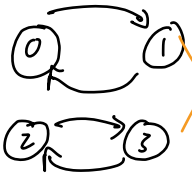
$$\sum_{i=0}^{\infty} \pi_i = 1$$

i.e. limiting distribution exists and it is equal to the stationary distribution. π is unique and

(limiting distribution

uniquely determined by:

$$\begin{cases} \pi = \pi P \\ \sum \pi_i = 1 \end{cases}$$



irreducible subchains

MC

- irreducible
  - recurrent
    - aperiodic
    - periodic
  - transient
- reducible

$\lim_{n\to\infty} P_{ji}^{(n)} = 0$ transient

reducible → irreducible

$M = M_1 \cup M_2 \cup T$

transient state

can now study the limiting behavior.

aperiodic: $\lim_{n\to\infty} P_{ij}^{(n)} = \frac{1}{m_i}$

$= \frac{1}{\sum_{n=0}^{\infty} n f_{ii}^{(n)}} = \frac{d}{m_i}$

periodic: $\lim_{n\to\infty} P_{ii}^{(nd)}$ — period of state i; only consider integer multiples of the period

$= \frac{d}{\sum_{n=0}^{\infty} n f_{ii}^{(n)}}$ } subsequence has a limit but the whole sequence doesn't.

transient irreducible

$$\lim_{n\to\infty} a_n = C \Rightarrow \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} = C$$

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} P_{ji}^{(k)} = \lim_{n\to\infty} P_{ji}^{(n)} = \frac{1}{m_i}$$

$$\frac{1}{n}\sum_{k=1}^{n} P_{ji}^{(k)} = \frac{1}{n}\sum_{k=1}^{n} E\left[I(X_n = i) \mid X_0 = j\right]$$

count # times state i has been visited from first transition to n^{th}

$$= \frac{1}{n} E\left[\sum_{k=1}^{n} I(X_n = i) \mid X_0 = j\right]$$

empirical frequency of visiting state i has some limit as transition probability

longitudinal behavior of the MC

- empirical longitudinal behavior can be used to study the cross-sectional behavior because they converge to the same limiting distribution.

Reversible Markov chains

- A reversible MC is a MC such that:

$$\exists\ \pi,\ \pi_i P_{ij} = \pi_j P_{ji} \quad (*)$$

detailed balance condition

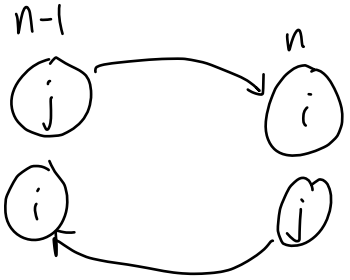
$$\pi_i = \sum_j \pi_i P_{ij} = \sum_j \pi_j P_{ji}$$

$$\Rightarrow \pi = \pi P$$

· Why "reversible"?

$$\underbrace{Pr(X_{n-1}=i \mid X_n=j)}_{\text{reverse transition probability}} = \frac{Pr(X_{n-1}=i,\ X_n=j)}{Pr(X_n=j)}$$

$$= \frac{Pr(X_{n-1}=i)\, Pr(X_n=j \mid X_{n-1}=i)}{Pr(X_n=j)}$$



$$= \frac{\pi_i P_{ij}}{\pi_j}$$

$$= \frac{\pi_j P_{ji}}{\pi_j} = P_{ji} = Pr(X_n=i \mid X_{n-1}=j)$$

Markov chain Monte Carlo

Metropolis-Hastings

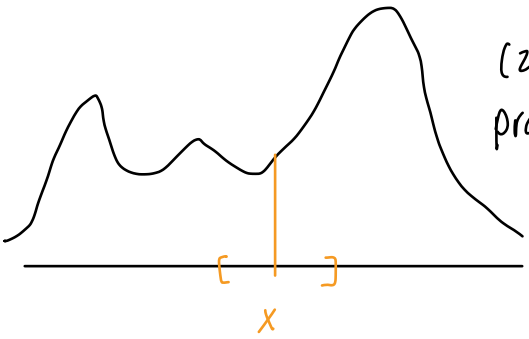
Goal: Draw samples from a target distribution $\pi(\cdot)$

initial value: X_0

$$X_{n-1} = x$$

(1) simulate a candidate from the jumping distribution

$$y \sim q(x, \cdot)$$



(2) Accept the candidate with probability

$$\alpha(x,y) = \min\left[\frac{\pi(y)\,q(y,x)}{\pi(x)\,q(x,y)}, 1\right]$$

conditional that current state is x, what's the probability of transitioning to y

- In other words, draw u from unif$(0,1)$ and accept y if $u \le \alpha(x,y)$. If y is accepted, $X_n = y$. Otherwise, $X_n = X_{n-1} = x$.

$$\text{i.e. } X_n = \begin{cases} y & \text{if } u \le \alpha(x,y) \\ x & \text{if } u > \alpha(x,y) \end{cases}$$

- Sequence of X's is a Markov chain.

Why does MH work?

(1) Creates a MC

(2) Chain is reversible wrt $\pi(\cdot)$.

Pf:/ Let $k(x,y)$ denote the transition density of the MC.

$$k(x,y) = \alpha(x,y)\, q(x,y) \quad \} x \neq y$$

$$+ \delta_x(y)\Big[\sum_t [1-\alpha(x,t)]\, q(x,t)\Big] \quad \} x = y$$

$\delta_x(y)$: indicator for $x=y$

$[1-\alpha(x,t)]$: propose a new t, reject it and keep x

WTS: $\pi(x)\, k(x,y) = \pi(y)\, k(y,x)$

$$\pi(x)\, k(x,y) = \pi(x)\alpha(x,y)\, q(x,y)$$

$$+ \pi(x)\, \delta_x(y)\Big[\sum_t [1-\alpha(x,t)]\, q(x,t)\Big]$$

$$\pi(x)\alpha(x,y)\, q(x,y) = \pi(x)\, q(x,y)\min\left(1, \frac{\pi(y)\, q(y,x)}{\pi(x)\, q(x,y)}\right)$$

$$= \min\left(q(y,x)\,\pi(y),\ \pi(x)\, q(x,y)\right)$$

$$\pi(x)\, k(x,y) = \min\left(q(y,x)\,\pi(y),\ \pi(x)\, q(x,y)\right)$$

$$+ \pi(x)\, \delta_x(y)\Big[\sum_t [1-\alpha(x,t)]\, q(x,t)\Big]$$

$$\pi(y)\, k(y,x) = \pi(y)\alpha(y,x)\, q(y,x)$$

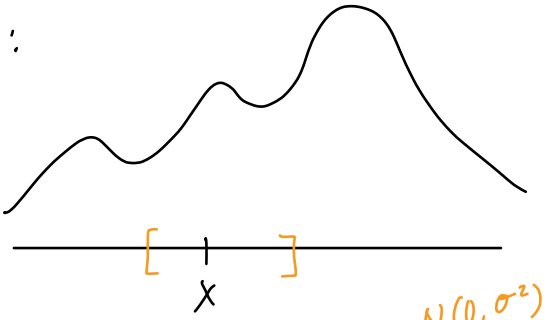
$$+ \pi(y)\, \delta_y(x)\Big[\sum_t [1-\alpha(y,t)]\, q(y,t)\Big]$$

$$= \min\left(\pi(x)q(x,y),\ \pi(y)q(y,x)\right)$$

$$+ \pi(y)\, \delta_y(x)\Big[\sum_t [1-\alpha(y,t)]\, q(y,t)\Big]$$

So $\pi(x)\,k(x,y) = \pi(y)\,k(y,x)$

Ex:



propose: $y = x + \epsilon$ $\sim N(0, \sigma^2)$ $\sim$ unif $[-a, a]$

$q(x,y) = N(x, \sigma^2)$ if we use $N(0, \sigma^2)$