Part 1

Stochastic simulation

Part 2

Simulation { MCMC } Bayes, aims to compute $\pi(\theta|y)$

Laplace approx. { INLA, RTMB } Frequentist, aim is MLE $\hat{\theta} = \underset{\theta}{\operatorname{argmax}}\ \pi(y|\theta)$

Part 3

Estimate sampling distribution of $\hat{\theta}$ given θ (incl. bias and SD)

using a non-parametric or parametric model

(bootstrapping). Frequentist

E-M alg. (after easter)

Non-parametric bootstrap

- Simplest case: iid sample $x_1, x_2, \dots, x_n \sim F$

where F is an unknown cdf.

Find estimator of

$$\theta = t(F)$$

e.g.

$$\theta = E(X) = \underbrace{\int_{-\infty}^{\infty} x \, dF(x)}_{\text{Lebesgue-Stieltjes integral}} = \begin{cases} \int_{-\infty}^{\infty} x f(x)\, dx & \text{if } X \text{ is cont.} \\ \sum_x x f(x) & \text{if } X \text{ is discr.} \end{cases}$$

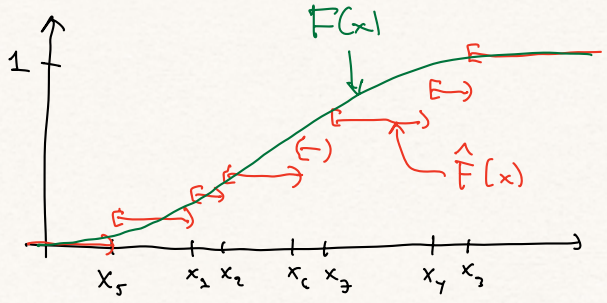
or

$$\sigma = \operatorname{Var}(X) = \int (x - EX)^2 \, dF$$

– F can be estimated non-parametrically by the estimator

$$\hat{F}(\underline{x}) = \frac{1}{n}\sum_{i=1}^{n} I(\underline{x}_i \le \underline{x}),$$

the empirical cdf. Note that x can be a vector



– The corresponding (plug-in) estimator of θ is

$$\hat{\theta} = t(\hat{F})$$

Thus, if $\theta = EX = \int x\, dF$

$$\hat{\theta} = \int_{-a}^{a} x\, d\hat{F} = \sum_{x} x\, \hat{f}(x) = \sum_{i=1}^{n} x_i \frac{1}{n} = \bar{x} = E_{\hat{F}}(X)$$

and if $\theta = \mathrm{Var}\, X = \int (x - EX)^2\, dF$

$$\hat{\theta} = \widehat{\mathrm{Var}}(X) = \sum_{x} (x - EX)\, \hat{f}(x)$$

$$= \sum_{i=1}^{n} (x_i - \bar{x})^2 \frac{1}{n}$$

$$= \frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x})^2 = \mathrm{Var}_{\hat{F}}(X)$$

If $\theta = SD(X)$,

$$\hat{\theta} = \widehat{SD(X)} = SD_{\hat{F}}(X)$$

If $\theta = Skew(X) : E\left(\left(\frac{X-\mu}{\sigma}\right)^3\right)$ the plug-in estimator of θ is

$$\hat{\theta} = Skew_{\hat{F}}(X)$$

What is the sampling distribution of these plug-in estimators if we replace F by $\hat{F}$?

- Approximate method

1. Generate B bootstrap samples

$X^{1*}, X^{2*}, \ldots, X^{B*}$ where $X^{b*} = (X_1^{b*}, X_2^{b*}, \ldots, X_n^{b*})$

and $X_i^{b*} \overset{iid}{\sim} \hat{F}$ for $i = 1, 2, \ldots, n$, $b = 1, 2, \ldots, B$.

2. Compute corresponding bootstrap replicates of plug-in estimator $\hat{\theta}$ of θ,

$$\hat{\theta}^{b*} = t(\hat{F}^{b*})$$

3. Estimate e.g. $E_{\hat{F}}(\hat{\theta})$ and $SD_{\hat{F}}(\hat{\theta})$ or, in general $E(h(\hat{\theta}))$, by Monte-Carlo integration, i.e.

$$\widehat{E(\hat{\theta})} = \frac{1}{B}\sum_{b=1}^{B}\hat{\theta}^{b*}$$

and

$$\widehat{SD(\hat{\theta})} = \sqrt{\frac{1}{B-1}\sum_{b=1}^{B}\left(\hat{\theta}^{b*} - \widehat{E(\hat{\theta})}\right)^2}$$

- Exact method (applicable for small n)

If the original sample $x_1, x_2, \dots, x_n$ consist of distinct values the number of unordered outcomes ((x_4, x_4, x_1, x_2) the same as (x_1, x_2, x_4, x_4)) is

$$\binom{2n-1}{n-1} \approx (n\pi)^{-\frac{1}{2}} 2^{2n-1}$$

$$\underset{n=10}{=} 92378$$

and the probability of observing $m_1, m_2, \dots, m_n$ of each value $x_1, x_2, \dots, x_n$ is

$$\frac{n!}{m_1! m_2! m_3!} \cdot \underbrace{\left(\frac{1}{n}\right)^{m_1}\left(\frac{1}{n}\right)^{m_2}\cdots\left(\frac{1}{n}\right)^{m_n}}_{n^{-n}}$$

since $m_1, \dots, m_n \sim \text{multinom}\left(n, \left(\frac{1}{n}, \frac{1}{n}, \dots, \frac{1}{n}\right)\right)$

Ex: $x = (5, 7, 10)$ $n = 3$

#5's	#7's	#10's	Ball-bar repr.	Prob
3	0	0	0 0 0 | |	3^{-3}
2	1	0	0 0 | 0 |	3^{-2}
2	0	1	0 0 | | 0	3^{-2}
1	2	0	0 | 0 0 |	⋮
1	1	1	0 | 0 | 0	⋮
1	0	2	0 | | 0 0	
0	3	0	| 0 0 0 |	
0	2	1	| 0 0 | 0	
0	1	2	| 0 | 0 0	
0	0	3	| | 0 0 0	

n balls

$n-1$ bars

has

$\binom{2n-1}{n-1}$ combinations

$= \binom{5}{2} = 10$

Original sample with $k < n$ distinct values can be treated similarly,

$\binom{n+k-1}{k-1}$ distinct outcomes

Parametric bootstrapping

Suppose original data $\underline{x} = (x_1, x_2, \ldots, x_n) \sim F(\underline{x}; \underline{\theta})$ (no iid assumption)
where F is known and $\underline{\theta}$ is unknown.
Estimate $\underline{\theta}$ by $\hat{\underline{\theta}}$ (by given method, e.g. ML) and
$F(\underline{x}; \underline{\theta})$ by $F(\underline{x}; \hat{\underline{\theta}})$. Algorithm

1. Simulate $\underline{x}_1^{1*}, \underline{x}_2^{2*}, \ldots, \underline{x}^{B*} \overset{iid}{\sim} F(\underline{x}; \hat{\theta})$
2. Compute corresponding bootstrap replicates of $\hat{\underline{\theta}}$, $\hat{\underline{\theta}}^{b*}$, $b = 1, 2, \ldots, B$ (by same given method)
3. Estimate $E(h(\hat{\theta}))$ by Monte-Carlo integration

Estimating and correcting for bias

Having estimated $E(\hat{\theta})$ by $\widehat{E(\hat{\theta})} = \frac{1}{B}\sum_{b=1}^{B} \hat{\theta}^{b*}$

an estimator of

$$\text{Bias}(\hat{\theta}) = E(\hat{\theta}) - \theta$$

is

$$\widehat{\text{Bias}(\hat{\theta})} = \widehat{E(\hat{\theta})} - \hat{\theta}$$

Subtracting the estimated bias from $\hat{\theta}$ we obtain a bias-corrected estimator

$$\hat{\theta}_c = \hat{\theta} - \widehat{\text{Bias}(\hat{\theta})}$$

Typically has less but not zero bias and higher variance.

Bootstrap confidence intervals

→ Percentile method

1. Generate $x^{1*}, x^{2*}, \ldots, x^{B*} \overset{iid}{\sim} \hat{F}$
2. Compute $\hat{\theta}^{1*}, \hat{\theta}^{2*}, \ldots, \hat{\theta}^{B*}$
3. Compute empirical $\alpha/2$ and $1-\alpha/2$ quantiles of $\hat{\theta}^{b*}$

Justification: Valid if a strictly increasing transformation $\phi(\theta)$ exist such that $\phi(\hat{\theta}) - \phi(\theta)$ has cdf $H(z) = 1 - H(-z)$.

Then

$$P\left(h_{\alpha/2} \leq \phi(\hat{\theta}) - \phi(\theta) \leq h_{1-\alpha/2} \right) = 1 - \alpha \qquad (1)$$

where $h_\alpha = H^{-1}(\alpha)$.

Given an existing ϕ bootstrap replicates of $\phi(\hat{\theta}) - \phi(\theta)$ are given by $\phi(\hat{\theta}^*) - \phi(\hat{\theta})$ from which $h_{\alpha/2}$ and $h_{1-\alpha/2}$ can be estimated by empirical quantiles, i.e.

$$P\left(h_{\alpha/2} \leq \phi(\hat{\theta}^*) - \phi(\hat{\theta}) \leq h_{1-\alpha/2} \right) \approx 1-\alpha$$

and so

$$P\left(\phi^{-1}\left(h_{\alpha/2} + \phi(\hat{\theta}) \right) \leq \hat{\theta}^* \leq \phi^{-1}\left(h_{1-\alpha/2} + \phi(\hat{\theta}) \right) \right) \approx 1-\alpha$$

$\phi^{-1}\left(h_{\alpha/2} + \phi(\hat{\theta}) \right) = \hat{\theta}^*_{\alpha/2}$

$\phi^{-1}\left(h_{1-\alpha/2} + \phi(\hat{\theta}) \right) = \hat{\theta}^*_{1-\alpha/2}$

The empirical quantiles $\hat{\theta}^*_{\alpha/2}$ and $\hat{\theta}^*_{1-\alpha/2}$ of $\hat{\theta}^*$ are explicitly known.

From (1) we have

$$P\left(-h_{\alpha/2} \geq \phi(\theta) - \phi(\hat{\theta}) \geq -h_{1-\alpha/2} \right) = 1-\alpha$$

Since $H(z) = 1 - H(-z)$, $h_{\alpha/2} = -h_{1-\alpha/2}$ and so

$$P\left(h_{\alpha/2} \leq \phi(\theta) - \phi(\hat{\theta}) \leq h_{1-\alpha/2} \right) = 1-\alpha$$

and

$$P\left(\underbrace{\phi^{-1}(\phi(\theta) - h_{\alpha/2})}_{= \hat{\theta}^*_{\alpha/2}} \leq \theta \leq \underbrace{\phi^{-1}(\phi(\hat{\theta}) - h_{1-\alpha/2})}_{= \hat{\theta}^*_{1-\alpha/2}} \right) = 1-\alpha$$

Thus, $(\hat{\theta}^*_{\alpha/2}, \hat{\theta}^*_{1-\alpha/2})$ is a conf. int. for θ. Note that ϕ and $h_{\alpha/2}, h_{1-\alpha/2}$ do not need to be known explicitly!

Ex. of failure of percentile method:

Suppose $X_1, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$.

It follows that parametric bootstrap replicates of $\frac{\hat{\sigma}^2(n-1)}{\sigma^2}$, i.e. $\frac{\hat{\sigma}^{2*}(n-1)}{\hat{\sigma}^2} \sim \chi^2_{n-1}$ and bootstrap replicates of $\hat{\sigma}^2$, i.e. $\hat{\sigma}^{2*} \sim \frac{\hat{\sigma}^2}{n-1}\chi^2_{n-1}$ so the percentil-method interval is (up to Monte-Carlo error)

$$\left(\frac{\hat{\sigma}^2 \chi^2_{\alpha/2, n-1}}{n-1}, \frac{\hat{\sigma}^2 \chi^2_{1-\alpha/2, n-1}}{n-1} \right)$$

But the classical exact interval is

$$\left(\frac{\hat{\sigma}^2(n-1)}{\chi^2_{1-\alpha/2, n-1}}, \frac{\hat{\sigma}^2(n-1)}{\chi^2_{\alpha/2, n-1}} \right)$$

Bootstrapping regression

Model

$$y_i = x_i^T \beta + \varepsilon_i$$

where the errors ε_i are iid zero mean constant variance.
In matrix notation

$$y = X\beta + \varepsilon$$

Least-square estimate of β is

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \| y - X\beta \|_2^2 = (X^T X)^{-1} X^T y$$

Residuals defined as

$$\hat{\varepsilon}_i = y_i - x_i^T \hat{\beta}$$

- Bootstrapping residuals

1. Generate B bootstrap replicates $\hat{\varepsilon}_1^*, \hat{\varepsilon}_2^*, \ldots, \hat{\varepsilon}_n^*$ by resampling with replacement from $\hat{\varepsilon}_1, \ldots, \hat{\varepsilon}_n$.
2. Compute $y_i^* = x_i^T \hat{\beta} + \hat{\varepsilon}_i^*$
3. Compute bootstrap replicates $\hat{\beta}^*$ by regressing each y^* on X

More appropriate for experimental data

R demo

- Bootstrapping pairs

1. Generate B bootstrap samples (y_i^*, x_i^*), $i = 1, 2, \ldots, n$ by sampling (y_i, x_i), $i = 1, \ldots, n$ with replacement.
2. Compute bootstrap replicates $\hat{\beta}^*$ by regressing each y^* on each X^*

More appropriate for observational data

Accelerated Bias-corrected Percentile Method BC_a

Assume strictly increasing $\phi(\,)$ and constants a and b exist such that

$$U = \frac{\phi(\hat{\theta}) - \phi(\theta)}{1 + a\phi(\theta)} + b \sim N(0,1) \qquad (1)$$

Bootstrap replicates of U from $\hat{F}$,

$$U^* = \frac{\phi(\hat{\theta}^*) - \phi(\hat{\theta})}{1 + a\phi(\hat{\theta})} + b \sim N(0,1) \qquad (2)$$

Thus, from (2),

$$\beta = P(U^* \le z_\beta)$$

$$= P\left(\frac{\phi(\hat{\theta}^*) - \phi(\hat{\theta})}{1 + a\phi(\hat{\theta})} + b \le z_\beta \right)$$

$$= P\Big(\hat{\theta}^* \le \underbrace{\phi^{-1}\Big(\phi(\hat{\theta}) + (z_\beta - b)\big[1 + a\phi(\hat{\theta})\big] \Big)}_{= \hat{\theta}^*_\beta = \text{observable empirical } \beta\text{-quantile of } \hat{\theta}^*} \Big) \qquad (3)$$

Similarly, from (1), isolating θ rather than $\hat{\theta}$

$$1 - \alpha = P(U > z_\alpha)$$

$$= P\left(\frac{\phi(\hat{\theta}) - \phi(\theta)}{1 + a\phi(\theta)} + b > z_\alpha \right)$$

$$= P\Big(\phi(\hat{\theta}) - \phi(\theta) > (z_\alpha - b)\big(1 + a\phi(\theta)\big) \Big)$$

$$= P\Big(\phi(\hat{\theta}) + b - z_\alpha > \big[1 - a(b - z_\alpha)\big]\phi(\theta) \Big)$$

$$= P\left(\theta < \phi^{-1}\left(\frac{\phi(\hat{\theta}) + b - z_\alpha}{1 - a(b - z_\alpha)}\right)\right)$$

$$= P\left(\theta < \underbrace{\phi^{-1}\left(\phi(\hat{\theta}) + \frac{b - z_\alpha}{1 - a(b - z_\alpha)}\left(1 + a\phi(\hat{\theta})\right)\right)}_{*}\right)$$

$$\phi(\hat{\theta}) + \frac{b - z_\alpha}{1 - a(b - z_\alpha)}\left(1 + a\phi(\hat{\theta})\right)$$

$$= \frac{\left[1 - a(b - z_\alpha)\right]\phi(\hat{\theta}) + (b - z_\alpha)\left(1 + a\phi(\hat{\theta})\right)}{1 - a(b - z_\alpha)}$$

$$= \frac{\phi(\hat{\theta}) + b - z_\alpha}{1 + b - z_\alpha}$$

i.e. $(*)$ is the upper limit of a $(1-\alpha)$ one sided conf. int. for θ, equal to $\hat{\theta}^*_\beta$ for

$$\frac{b - z_\alpha}{1 - a(b - z_\alpha)} = z_\beta - b$$

$$\beta = \Phi\left(b + \frac{b - z_\alpha}{1 - a(b - z_\alpha)}\right)$$

Two-sided confidence limits given by $(\hat{\theta}^*_{\beta_1}, \hat{\theta}^*_{\beta_2})$ with α replaced by $\alpha/2$ and $1-\alpha/2$, respectively.

Choosing a and b (Shao & Tu, 1996):

$b = \Phi^{-1}\left(F_{\hat{\theta}^*}(\hat{\theta})\right)$, where $F_{\hat{\theta}^*}$ is the cdf of $\hat{\theta}^*$

$$a = \frac{\frac{1}{6}\sum_{i=1}^{n} \psi_i^3}{\left(\sum \psi_i^2\right)^{\frac{2}{3}}}, \text{ where } \psi_i = \hat{\theta}_{(\cdot)} - \hat{\theta}_{(-i)},$$

$\hat{\theta}_{(-i)}$ is estimate based on ommiting i'th observation and

$$\hat{\theta}_{(\cdot)} = \frac{1}{n}\sum_{i=1}^{n} \theta_{(-i)}$$

R demo

Bootstrap t confidence interval

Suppose an estimate $\hat{V}$ of the variance of $\hat{\theta}$ is "directly" available.

Ex.: Regression, $\hat{\theta} = f(\hat{\beta}_0, \hat{\beta}_1) = \frac{\hat{\beta}_1}{\hat{\beta}_0}$

$$\hat{V} = \widehat{Var}(\hat{\theta}) \approx \left(\frac{\partial f}{\partial \beta_0}\right)^2 \widehat{Var}(\hat{\beta}_0) + 2\left(\frac{\partial f}{\partial \beta_0}\right)\left(\frac{\partial f}{\partial \beta_1}\right) \widehat{Cov}(\hat{\beta}_0, \hat{\beta}_1) + \left(\frac{\partial f}{\partial \beta_1}\right)^2 \widehat{Var}(\hat{\beta}_1)$$

Then

$$R = \frac{\hat{\theta} - \theta}{\sqrt{\hat{V}}}$$

is approximately pivotal. The bootstrap replicates

$$R^* = \frac{\hat{\theta}^* - \hat{\theta}}{\sqrt{\hat{V}^*}}$$

should have approximately same distribution and observable quantiles $R^*_{\alpha/2}$ and $R^*_{1-\alpha/2}$. Thus

$$P\left(R^*_{\alpha/2} < \frac{\hat{\theta} - \theta}{\sqrt{\hat{V}}} < R^*_{1-\alpha/2}\right) \approx 1 - \alpha$$

and so

$$\left(\hat{\theta} - \sqrt{\hat{V}}\, R^*_{1-\alpha/2} < \theta < \hat{\theta} - \sqrt{\hat{V}}\, R^*_{\alpha/2}\right)$$

is an approximate $1-\alpha$ conf. int for θ.

R demo

Bootstrapping time-series models

Ex. $Y_t - \mu = \phi(Y_{t-1} - \mu) + \varepsilon_t, \quad \varepsilon_t \sim WN(\sigma^2)$ (1)

For $|\phi| < 1$, (1) has a causal stationary solution with stationary mean μ. Taking variance of (1), the stationary variance $\gamma_0 = Var(Y_t) = Var(Y_{t-1})$ satisfies

$$\gamma_0 = \phi^2 \gamma_0 + \sigma^2$$

$$\Rightarrow \gamma_0 = \frac{\sigma^2}{1-\phi^2}$$

MLEs of μ, ϕ and σ^2 (assuming $\varepsilon_t \overset{iid}{\sim} N(0, \sigma^2)$) obtained by maximising $\pi(y_1, y_2, \ldots, y_n \mid \theta, \phi, \sigma^2)$.

Let $\hat{\varepsilon}_t = Y_t - \widehat{E(Y_t \mid Y_{t-1})}$, i.e. $\hat{\varepsilon}_1 = Y_1 - \mu$, $\hat{\varepsilon}_2 = Y_2 - \hat{\phi} Y_1, \ldots$

Bootstraping residuals: For $b = 1, 2, \ldots, B$

1. Create $n+m+1$ bootstrap innovations $\varepsilon^*_{-m}, \varepsilon^*_{-m+1}, \ldots, \varepsilon^*_n$ by resampling $\hat{\varepsilon}_2, \hat{\varepsilon}_3, \ldots, \hat{\varepsilon}_n$ with replacement.

2. Compute $Y^*_{-m} = \varepsilon^m_{-m} + \hat{\mu}$ and for $t = -m+1, \ldots, n$

$$Y^*_t = \hat{\mu} + \hat{\phi}(\hat{Y}^*_{t-1} - \hat{\mu}) + \hat{\varepsilon}^*_t$$

3. Refit model to $Y^*_1, Y^*_2, \ldots, Y^*_n$ to obtain bootstrap replicates $\hat{\mu}^{b*}, \hat{\phi}^{b*}, \hat{\sigma}^{2b*}$

Scale parameter bootstrap CI

Approximate pivotal

$$V = \frac{\hat{\theta}}{\theta}$$

Bootstrap replicates

$$V^* = \frac{\hat{\theta}^*}{\hat{\theta}}$$

has known quantiles $V^*_{\alpha/2}$, $V^*_{1-\alpha/2}$. Thus

$$P\left(V^*_{\alpha/2} < \frac{\hat{\theta}}{\theta} < V^*_{1-\alpha/2} \right) = 1 - \alpha$$

$$P\left(\frac{\hat{\theta}}{V^*_{1-\alpha/2}} < \theta < \frac{\hat{\theta}}{V^*_{\alpha/2}} \right) = 1 - \alpha$$

EM-algorithm

Motivation:

Likelihood $f_Y(y|\theta)$ of "complete" data Y has simple form but only $X = M(Y)$ is observed where M is a many-to-fewer mapping.

Aim: Maximise "difficult" likelihood $f_X(x|\theta)$ of X.

Algorithm: Let

$$Q(\theta|\theta^{(t)}) = E\left(\ln f_Y(Y|\theta) \,\middle|\, X = x, \Theta = \theta^{(t)}\right)$$

$$= \int_{\{y: M(y)=x\}} \ln f_Y(y|\theta)\, f_{Y|X}(y|x,\theta^{(t)})\, dy$$

1. E-step: Find $Q(\theta|\theta^{(t)})$
2. M-step: Set $\theta^{(t+1)} = \underset{\theta}{\operatorname{argmax}}\, Q(\theta|\theta^{(t)})$

Repeat until convergence.

- Example: ABO-blood types (Ceppellini et al. 1955)

Genotypes	Observable phenotypes	Y	X
AA	A	n_{AA}	$n_A = n_{AA} + n_{AO}$
AO		n_{AO}	
BB	B	n_{BB}	$n_B = n_{BB} + n_{BO}$
BO		n_{BO}	
AB	AB	n_{AB}	n_{AB}
OO	O	n_{OO}	n_O

Model: $Y = (n_{AA}, n_{AO}, \ldots, n_{OO}) \sim \text{multinom}(n, \underline{p})$

where

$$\underline{p} = \begin{bmatrix} p_A^2 \\ 2p_Ap_O \\ p_B^2 \\ 2p_Bp_O \\ 2p_Ap_B \\ p_O^2 \end{bmatrix} \left.\vphantom{\begin{bmatrix} a \\ b \\ c \\ d \\ e \\ f \end{bmatrix}}\right\} \text{Hardy-Weinberg proportions}$$

and $p_O = 1 - p_A - p_B$

E-step: Given $\theta^{(t)} = (p_A^{(t)}, p_B^{(t)}, p_O^{(t)})$ and $X = (n_A, n_B, n_{AB}, n_O)$

$$Q(\theta \mid \theta^{(t)}) = E \ln f_Y(Y \mid \theta) \mid X = x, \Theta = \theta^{(t)}$$

$$\stackrel{c}{=} E\left(n_{AA} \ln(p_A^2) + n_{AO} \cdot \ln(2 p_A p_O) + \dots + n_{OO} \ln(p_O^2) \mid n_A, n_B, n_{AB}, n_O \right)$$

$$= E\Big((2n_{AA} + n_{AO} + n_{AB}) \ln p_A + (2n_{BB} + n_{AB} + n_{BO}) \ln p_B + (2n_{OO} + n_{AO} + n_{BO}) \ln p_O \mid n_A, n_B, n_{AB}, n_O \Big)$$

$$= (2n_{AA}^* + n_{AO}^* + n_{AB}) \ln p_A + (2n_{BB}^* + n_{AB} + n_{BO}^*) \ln p_B + (2n_O + n_{AO}^* + n_{BO}^*) \ln p_O$$

where

$$n_{AA}^* = E(n_{AA} \mid n_A, p_A^{(t)}, p_B^{(t)}, p_O^{(t)})$$

$$= n_A \cdot \frac{p_A^{(t)2}}{p_A^{(t)2} + 2 p_A^{(t)} p_O^{(t)}}$$

$$n_{AO}^* = n_A \frac{2 p_A^{(t)} p_B^{(t)}}{p_A^{(t)2} + 2 p_A^{(t)} p_O^{(t)}}$$

etc. since $n_{AA} \mid n_A \sim \text{bin}(n_A, \quad)$

M-step: Maximising $Q(\theta \mid \theta^{(t)})$ w.r.t. θ yields

$$\theta^{(t+1)} = (p_A, p_B, p_O)^{(t+1)}$$

$$= \frac{1}{2n} \left(2n_{AA}^* + n_{AO}^* + n_{AB}, \; 2n_{BB}^* + n_{BO}^* + n_{AB}, \; 2n_O \right)$$

= "gene counting"

- Ex.: Gaussian mixture:

$$x_1, x_2, \dots, x_n \sim f(x|\theta), \quad \theta = (\pi_1, \mu_1, \mu_2)$$

where

$$f(x|\theta) = \sum_{k=1}^{2} \pi_k \phi(x - \mu_k), \quad \pi_2 = 1 - \pi_1$$

Full likelihood is

$$L(\theta) = \prod_{i=1}^{n} f(x_i|\theta)$$

$$= \prod_{i=1}^{n} \sum_{k=1}^{2} \pi_k \phi(x_i - \mu_k)$$

and log-likelihood

$$l(\theta) = \sum_{i=1}^{n} \ln \sum_{k=1}^{2} \pi_k \phi(x_i - \mu_k)$$

Can be maximised numerically to find MLEs $\hat{\pi}_1, \hat{\mu}_1, \hat{\mu}_2$

Via EM-alg.: Introduce "missing" "component memberships"

$$z_1, z_2, \dots, z_n \overset{iid}{\sim} f(z|\pi_1) = \pi_1 I(z=1) + (1-\pi_1) I(z=2)$$

Letting

$$f(x_i|z_i=k) = \phi(x_i - \mu_k)$$

$x_1, \dots, x_n$ has the correct marginal distribution.

Conditional on x_i, and $\theta^{(t)} = (\mu_1^{(t)}, \mu_2^{(t)}, \pi_1^{(t)})$

$$w_{i,k}^{(t)} = f(z_i = k \mid x_i, \theta^{(t)}) = \frac{f(x_i|z_i=k, \theta^{(t)}) f(z_i=k|\theta^{(t)})}{f(x_i|\theta^{(t)})}$$

i.e

$$w_{i,1}^{(t)} = \frac{\phi(x_i - \mu_1)\pi_1}{\phi(x_i - \mu_1)\pi_1 + \phi(x_i - \mu_2)(1-\pi_1)}$$

and

$$w_{i,2}^{(t)} = 1 - w_{i,1}^{(t)}$$

The likelihood based on "complete" data is

$$L(\theta) = \prod_{i=1}^{n} f(x_i | z_i, \theta) f(z_i | \theta)$$

and

$$\ell(\theta) = \sum_{i=1}^{n} \ln f(x_i | z_i, \theta) + \ln f(z_i | \theta)$$

E-step:

$$Q(\theta | \theta^{(t)}) = \sum_{i=1}^{n} E\left(\ln f(x_i | z_i, \theta) + \ln f(z_i | \theta) \,\middle|\, x_i, \theta^{(t)} \right)$$

$$= \sum_{i=1}^{n} \sum_{k=1}^{2} w_{i,k}^{(t)} \left(\ln f(x_i | z_i = k, \theta) + \ln f(z_i = k | \theta) \right)$$

$$= \sum_{i=1}^{n} \sum_{k=1}^{2} w_{i,k}^{(t)} \left(-\frac{1}{2}(x_i - \mu_k)^2 + \ln \pi_k \right)$$

M-step:

$$\frac{\partial Q}{\partial \mu_k} = \sum_{i=1}^{n} w_{i,k}^{(t)} (x_i - \mu_k) = 0 \Rightarrow \mu_k^{(t+1)} = \frac{\sum_{i=1}^{n} x_i w_{i,k}^{(t)}}{\sum_{i=1}^{n} w_{i,k}^{(t)}}$$

for $k = 1, 2$.

$$\frac{\partial Q}{\partial \pi_1} = \sum_{i=1}^{n} \left(\frac{w_{i,1}^{(t)}}{\pi_1} - \frac{1 - w_{i,1}^{(t)}}{1 - \pi_1} \right)$$

$$= \frac{1}{\pi_1 (1 - \pi_1)} \sum_{i=1}^{n} \left[(1 - \pi_1) w_{i,1}^{(t)} - \pi_1 (1 - w_{i,1}^{(t)}) \right] = 0$$

$$\sum w_{i,1} - n\pi_1 = 0$$

$$\pi_1^{(t+1)} = \frac{1}{n} \sum_{i=1}^{n} w_{i,1}^{(t)}$$

Convergence: EM-alg. only guaranteed to find local optima or saddle point of $f_X(x|\theta)$.

Proof that $f_X(x|\theta^{(t+1)}) \geq f_X(x|\theta^{(t)})$.

Let $(X, Z) = M(Y)$ be a one-to-one mapping between X, Z and Y

Since

$$f_{Z|X}(z|x,\theta) = \frac{f_{X,Z}(x,z|\theta)}{f_X(x|\theta)}$$

$$\ln f_X(x|\theta) = \ln f_{X,Z}(x,z|\theta) - \ln f_{Z|X}(z|x,\theta)$$

and

$$E\left(\ln f_X(x|\theta) \middle| x, \theta^{(t)}\right) = E\left(\ln f_{X,Z}(x,Z|\theta) \middle| x, \theta^{(t)}\right) - E\left(\ln f_{Z|X}(Z|x,\theta) \middle| x, \theta^{(t)}\right)$$

i.e.

$$\ln f_X(x|\theta) \overset{c}{=} Q(\theta|\theta^{(t)}) - H(\theta|\theta^{(t)})$$

It follows that

$$-H(\theta|\theta^{(t)}) + H(\theta^{(t)}|\theta^{(t)})$$

$$= E\left(-\ln f_{Z|X}(Z|x,\theta) + \ln f_{Z|X}(Z|x,\theta^{(t)}) \middle| x, \theta^{(t)}\right)$$

$$= E\left(-\ln \frac{f_{Z|X}(Z|x,\theta)}{f_{Z|X}(Z|x,\theta^{(t)})} \middle| x, \theta^{(t)}\right)$$

$$\geq -\ln\left[E\left(\frac{f_{Z|X}(Z|x,\theta)}{f_{Z|X}(Z|x,\theta^{(t)})} \middle| x, \theta^{(t)}\right)\right] \qquad \text{(Jensen)}$$

$$= -\ln \underbrace{\int \frac{f_{Z|X}(z|x,\theta)}{f_{Z|X}(z|x,\theta^{(t)})} f_{Z|X}(z|x,\theta^{(t)})\, dz}_{=1}$$

$$= 0$$

for any θ.

Thus, since $Q(\theta^{(t+1)} | \theta^{(t)}) \geq 0$,

$$\ln f_X(x | \theta^{(t+1)}) = Q(\theta^{(t+1)} | \theta^{(t)}) - H(\theta^{(t+1)} | \theta^{(t)})$$

$$\geq Q(\theta^{(t)} | \theta^{(t)}) - H(\theta^{(t)} | \theta^{(t)})$$

$$= \ln f_X(x | \theta^{(t)})$$

Summary, part 1

Example: Suppose X has density

$$f(x) \propto \cosh(x)\, e^{-\frac{x^2}{2}}$$

Find alg. to simulate X.

$$f(x) \propto \frac{e^x + e^{-x}}{2} e^{-\frac{x^2}{2}}$$

$$\propto e^{-\frac{1}{2}(x^2 - 2x)} + e^{-\frac{1}{2}(x^2 + 2x)}$$

$$\propto e^{-\frac{1}{2}\left[(x-1)^2 - 1\right]} + e^{-\frac{1}{2}\left[(x+1)^2 - 1\right]}$$

$$= e + e$$

$$\propto \sum_{y \in \{-1, 1\}} f(x|y) f(y)$$

where $f(x|y) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-y)^2}$

and

$$f(y) = \frac{1}{2} \quad \text{for } y = \pm 1$$

Alg.: Generate $u \sim \text{unif}(0, 1)$

If $u < \frac{1}{2}$

$y \leftarrow -1$

else

$y \leftarrow 1$

Generate $x \sim N(y, 1)$

Exercise: Find alg. to simulate from

$$f(x) \propto (\cosh(x))^k e^{-\frac{x^2}{2}}$$

where $k \in \mathbb{N}$

Example: Estimate $\int_0^3 \frac{e^{-x}}{1+1/x}\,dx$

Monte-Carlo:

$$\int_0^3 \frac{e^{-x}}{1+1/x}\,dx = E(h(X))$$

where $X \sim \text{exp}(1)$ and $h(x) = \frac{1}{1+1/x} I(x<3)$

Monte-Carlo integration:

$$\widehat{E(h(X))} = \frac{1}{n}\sum_{i=1}^{n} h(x_i) = \frac{1}{n}\sum \frac{1}{1+1/x_i} I(x_i<3)$$

Improvement using importance sampling. Use proposal

$$g(x) = \frac{e^{-x}}{1-e^{-3}} \quad \text{for } 0<x<3$$

since $g(x)$ is approx. prop. to $f(x)h(x)$.

The cdf of X is then

$$G(x) = \int_0^x \frac{e^{-t}}{1-e^{-3}}\,dt = \frac{1-e^{-x}}{1-e^{-3}}$$

and X can be generated using the inversion method:

$$U = G(X)$$

$$U(1-e^{-3}) = 1-e^{-X}$$

$$e^{-X} = 1-U(1-e^{-3})$$

$$X = -\ln\left(1-U(1-e^{-3})\right)$$

Importance sampling estimate

$$\widehat{\int_0^3 \frac{e^{-x}}{1+1/x}\,dx} = \widehat{E(h(X))} = \frac{1}{n}\sum_{i=1}^{n} h(x_i)\frac{f(x_i)}{g(x_i)} = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{1+1/x_i}(1-e^{-3})$$

Further improvement using antithetic sampling.

$$\overbrace{\int_0^3 \frac{e^{-x}}{1+1/x}\,dx}^{} = \frac{1}{2n}\left[\sum_{i=1}^{n} \frac{1}{1+1/x_i}\left(1-e^{-3}\right) + \sum_{i=1}^{n} \frac{1}{1+1/x_i^*}\left(1-e^{-3}\right)\right]$$

where $x_i = -\ln\left(1-U\left(1-e^{-3}\right)\right)$

and $x_i^* = -\ln\left(1-(1-U)\left(1-e^{-3}\right)\right)$

Variance reduced by a factor of $1+\rho$

where

$$\rho = \text{corr}\left(h(X_i)\frac{f(X_i)}{g(X_i)}, h(X_i^*)\frac{f(X_i^*)}{g(X_i^*)}\right)$$

$$= \text{corr}\left(\frac{1}{1+1/X_i}, \frac{1}{1+1/X_i^*}\right)$$

R demo: $\hat{\rho} = -0.9978$

Summary part 2

Example: Consider a sample of arrival times $y_1, y_2, \dots, y_n$ from a renewal process on $[0, \tau]$ where the iid interarrival times have density

$$f(x) = \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x}$$

Prior:

$$\lambda \sim \text{Gamma}(a_0, b_0)$$

$$\alpha \sim \exp(1)$$

Construct a MCMC-alg. to sample from $\pi(\lambda, \alpha \mid y_1, \dots, y_n)$

The density of $y_1, y_2, \dots, y_n$ is given by

$$\phi(y_1) f(y_2 \mid y_1) \cdots f(y_n \mid y_{n-1}) P(y_{n+1} > \tau)$$

where $\phi(y_1)$ is the pdf of the residual time (see wikipedia) equal to

$$\phi(y_1) = \frac{1 - F(y_1)}{\mu}$$

where $\mu = E(X) = \frac{\alpha}{\lambda}$ and F is the expectation and cdf of the interarrival times.

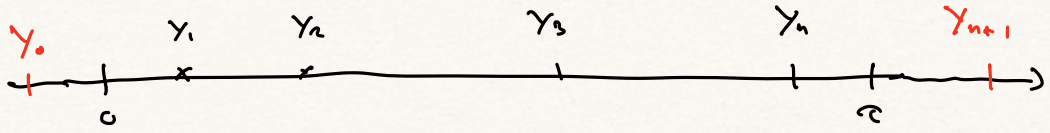
An alternative approach is to introduce latent arrival times y_0 and y_{n+1} and sample from

$$\pi(\lambda, \alpha, y_0, y_{n+1} \mid y_1, \dots, y_n)$$

Given y_0 and y_{n+1} we observe the arrival process instead on the interval $[y_0, y_{n+1}]$ and the density of $y_0, \dots, y_{n+1}$ simplifies to

$$\phi(y_0) \prod_{i=1}^{n+1} f(y_i \mid y_{i-1}) P(y_{n+2} > y_{n+1})$$

$$\propto \frac{\lambda}{\alpha} \prod_{i=1}^{n+1} \frac{\lambda^\alpha}{\Gamma(\alpha)} (y_i - y_{i-1})^{\alpha-1} e^{-\lambda(y_i - y_{i-1})} \cdot (1 - F(0))$$



Gibbs sampler:

- Sample y_0' from $\pi(y_0 \mid \underline{y}_{-0}, \alpha, \lambda)$
  1. Generate $s_1 \sim \text{Gamma}(\alpha, \lambda)$
  2. Accept s_1 if $s_1 > y_1$
  3. Set $y_0 = y_1 - s_1$
- Sample y_{n+1}' from $\pi(y_{n+1} \mid y_n, \alpha, \lambda)$ using similar method
- Block Metropolis-within Gibbs for α, λ:

$$\alpha' \mid \alpha \sim N(\alpha, \sigma^2)$$ Random-walk proposal

Conditional on α' and $\underline{y}$, sample λ' from

$$\pi(\lambda \mid \alpha', \underline{y}) \propto \pi(\lambda) f(\underline{y} \mid \alpha', \lambda)$$

$$\propto \lambda^{a_0 - 1} e^{-b_0 \lambda} \frac{\lambda}{\alpha'} \prod_{i=1}^{n+1} \frac{\lambda^{\alpha'}}{\Gamma(\alpha')} (y_i - y_{i-1})^{\alpha' - 1} e^{-\lambda(y_i - y_{i-1})}$$

$$\propto \lambda^{a_0 - 1 + (n+2)\alpha'} e^{-\lambda(b_0 + y_{n+1} - y_0)},$$

that is,

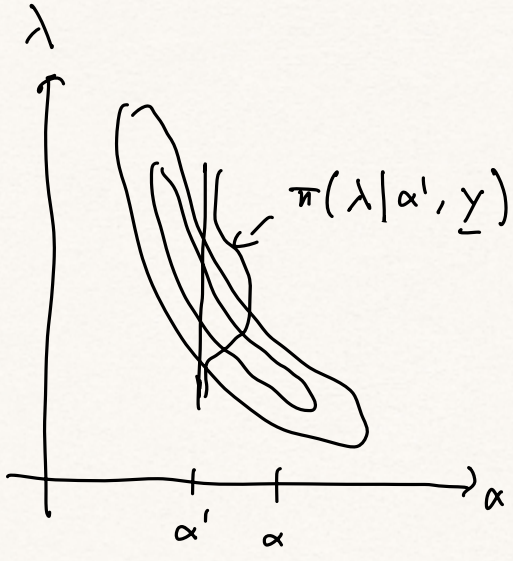
$$\lambda' \mid \alpha', \underline{y} \sim \text{Gamma}\left(a_0 + (n+2)\alpha', \; b_0 + y_{n+1} - y_0\right)$$

so the joint proposal is

$$Q(\alpha', \lambda' \mid \alpha, \lambda, \underline{y}) = \frac{1}{\sigma} \phi\left(\frac{\alpha' - \alpha}{\sigma}\right) \pi(\lambda' \mid \alpha', \underline{y})$$

Accept α', λ' with prob.

$$\alpha = \min\left(1, \frac{\pi(\alpha', \lambda' \mid \underline{y})\, G(\alpha, \lambda \mid \alpha', \lambda', \underline{y})}{\pi(\alpha, \lambda' \mid \underline{y})\, Q(\alpha', \lambda' \mid \alpha, \lambda, \underline{y})}\right)$$



Similarly, in project 2 (and for latent Gaussian models in general), a single-block proposal

$$Q(\underline{x}', \underline{\theta}' \mid \underline{x}, \underline{\theta}, \underline{y}) \propto \underbrace{\pi(\underline{x}' \mid \underline{\theta}', \underline{y})}_{\text{via Gaussian approximation around mode}} \underbrace{\pi(\theta' \mid \theta)}_{\text{random-walk}}$$

$\hat{x} = \underset{x}{\operatorname{argmax}}\, \pi(\underline{x} \mid \underline{\theta}', \underline{y})$

is feasible (see discussion section to Rue et. al. 2009, p. 356)

Summary, part 3

Example EM-alg.: Suppose $X_1, \ldots, X_n, Y_1, \ldots, Y_n$ are indep. and $Y_i \sim \text{Pois}(\beta\tau_i)$, $X_i \sim \text{Pois}(\tau_i)$. (Casella & Berger)

Complete data likelihood

$$L(\beta, \underline{\tau}; \underline{x}, \underline{y}) = \prod_{i=1}^{n} \frac{e^{-\beta\tau_i}(\beta\tau_i)^{y_i}}{y_i!} \frac{e^{-\tau_i}\tau_i^{x_i}}{x_i!}$$

$$\ell(\beta, \underline{\tau}; \underline{x}, \underline{y}) = -(\beta+1)\sum_{i=1}^{n}\tau_i + \sum(x_i+y_i)\ln\tau_i + (\sum y_i)\ln\beta$$

MLEs

$$\frac{\partial \ell}{\partial \beta} = -\sum\tau_i + \frac{\sum y_i}{\beta} = 0 \Rightarrow \beta = \frac{\sum y_i}{\sum \tau_i} = \frac{(\beta+1)\sum y_i}{\sum(x_i+y_i)}$$

$$\frac{\partial \ell}{\partial \tau_i} = -(\beta+1) + \frac{x_i+y_i}{\tau_i} = 0 \Rightarrow \hat{\tau}_i = \frac{x_i+y_i}{\beta+1}$$

$$\beta\sum x_i + \beta\sum y_i = \beta\sum y_i + \sum y_i$$

$$\hat{\beta} = \frac{\sum y_i}{\sum x_i}$$

Suppose x_1 is missing.

E-step: Given $\theta^{(t)} = (\beta^{(t)}, \underline{\tau}^{(t)})$,

$$Q(\beta, \underline{\tau} \mid \beta^{(t)}, \underline{\tau}^{(t)}) = E\left(\ell(\beta, \underline{\tau}; \underline{X}, \underline{Y}) \,\middle|\, \underline{X}_{-1} = \underline{x}_{-1}, \underline{Y} = \underline{y}, \beta^{(t)}, \underline{\tau}^{(t)}\right)$$

$$= -(\beta+1)\sum_{i=1}^{n}\tau_i + \left(x_1^* + \sum_{i=2}^{n} x_i + \sum_{i=1}^{n} y_i\right)\ln\tau_i + (\sum y_i)\ln\beta$$

where $x_1^* = E\left(X_1 \mid Y_1 = y_1, \beta^{(t)}, \tau_1^{(t)}\right)$

$$= \beta^{(t)}\tau_1^{(t)}$$

M-step:

Maximising $Q(\ )$ w.r.t. β and $\underline{\tau}$ yields

$$\beta^{(t+1)} = \frac{\sum_{i=1}^{n} y_i}{x_1^* + \sum_{i=2}^{n} y_i}$$

$$\tau_i^{(t+1)} = \begin{cases} \dfrac{x_i + y_i}{\beta^{(t+1)} + 1} & \text{for } i \geq 2 \\ \dfrac{x_i^* + y_i}{\beta^{(t+1)} + 1} & \text{for } i = 1 \end{cases}$$

Exercise:

Can we find the MLE via the observed data likelihood?

$$L(\beta, \underline{\tau}) = \frac{e^{-\beta\tau_1}(\beta\tau_1)^{y_1}}{y_1!} \prod_{i=2}^{n} \frac{e^{-\beta\tau_i}(\beta\tau_i)^{y_i}}{y_i!} \frac{e^{-\tau_i}\tau_i^{x_i}}{x_i!}$$

Show that the MLEs are

$$\hat{\beta} = \frac{\sum_{i=2}^{n} y_i}{\sum_{i=2}^{n} x_i}, \quad \hat{\tau}_i = \frac{x_i + y_i}{\beta + 1} \quad \text{for } i \geq 2$$

$$\tau_1 = \frac{y_1}{\hat{\beta}}$$

Proof: Profile likelihood for $(\beta, \underline{\tau}_{-1})$? Log of first term (only term containg τ_1),

$$\ell_1(\beta, \tau_1) = -\beta\tau_1 + y_1(\ln\beta + \ln\tau_1),$$

is maximised when

$$\frac{\partial \ell_1}{\partial \alpha_1} = 0$$

$$-\beta + \frac{Y_1}{\alpha_1} = 0$$

$$\hat{\alpha}_1 = \frac{Y_1}{\beta}.$$

Thus, the profile likelihood

$$L(\beta, \underline{\alpha}_{-1}) = \frac{e^{-Y_1} Y_1}{Y_1!} \prod_{i=2}^{n} \left(\quad \right) \left(\quad \right)$$

$\propto$ complete data likelihood for observations $i = 2, 3, \ldots, n$.