BAYESIAN INFERENCE

We start knowing
- θ (or $\underline{\theta}$): the parameters about the population characteristics
- $\underline{y}$: the numerical description of the sample we get (a dataset)

So we have $\theta(\underline{\theta}) \in \Theta$, the parameter space, and $\underline{y} \in \mathcal{Y}$, the sample space

We start as follows:

(1) we assign a prior distribution about how we believe initially that the population characteristics behave

$$\theta \sim \pi(\theta)$$

(2) we assign the likelihood $f(y|\theta)$, ie the conditional law of a sample obs y given the true parameter θ

$$y_1, \dots, y_n | \theta \overset{iid}{\sim} f_{Y|\theta}(y|\theta)$$

$$\Rightarrow f_{\underline{Y}|\theta}(\underline{y}|\theta) = f_{\underline{Y}|\theta}(y_1, \dots, y_n|\theta) = \prod_{i=1}^{n} f_{Y_i|\theta}(y_i|\theta)$$

(3) once we collect the data $\underline{y}$ we can get the posterior distribution which updates the belief about θ. We can compute it by the Bayes thm:

$$\pi(\theta|\underline{y}) = \frac{f(\underline{y}|\theta) \cdot \pi(\theta)}{\underbrace{f(\underline{y})}_{\text{marginal density of } \underline{y} \text{ which we also call } m(\underline{y})}} =\Big|_{\text{argument trick}} \frac{f(\underline{y}|\theta) \cdot \pi(\theta)}{\int_\Theta f(\underline{y}|\theta) \cdot \pi(\theta)\, d\theta} =$$

$$=\Big|_{iid} \frac{\prod_i f(y_i|\theta) \cdot \pi(\theta)}{\int_\Theta \prod_i f(y_i|\theta) \pi(\theta)\, d\theta} \propto \underbrace{(\text{LIKELIHOOD}) \cdot (\text{PRIOR})}_{\text{for the computation we can just leave out the den, to leave the functional part, the num}}$$

A quite interesting example was about the posterior computation of θ when we have/observe two samples $\underline{y}_1$ and $\underline{y}_2$. But morally however we proceed we get the same distribution in the end.

Case of using both $\underline{y}_1$ and $\underline{y}_2$ together

$$\pi(\theta|\underline{y}_1, \underline{y}_2) = \frac{f(\underline{y}_1, \underline{y}_2|\theta) \cdot \pi(\theta)}{f(\underline{y}_1, \underline{y}_2)} =\Big|_{\text{cond} \perp} \frac{f(\underline{y}_1|\theta) \cdot f(\underline{y}_2|\theta) \cdot \pi(\theta)}{f(\underline{y}_1, \underline{y}_2)} =$$

$$= \frac{f(\underline{y}_2|\theta)}{f(\underline{y}_2|\underline{y}_1)} \cdot \underbrace{\frac{f(\underline{y}_1|\theta) \cdot \pi(\theta)}{f(\underline{y}_1)}}_{\pi(\theta|\underline{y}_1)} = \frac{f(\underline{y}_2|\theta) \cdot \pi(\theta|\underline{y}_1)}{f(\underline{y}_2|\underline{y}_1)} =$$

$$= \pi(\theta|\underline{y}_1, \underline{y}_2)$$ Case of first using $\underline{y}_1$ and then $\underline{y}_2$

INFERENTIAL PROBLEMS

(1) Bayesian point estimation: we could take the posterior distr (we always refer to the posterior when doing inference) and we could take the posterior mean or median, for example.

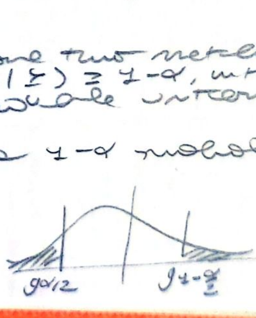
$$\hat{\theta} = \arg\min_a E[\ell(\theta, a)|\underline{y}] \qquad \ell(\theta, a) = \begin{cases} (\theta - a)^2 \\ |\theta - a| \\ 1 - \delta_\theta(a) \end{cases} \Rightarrow \hat{\theta} = \begin{cases} E[\theta|\underline{y}] \\ \text{median}(\theta|\underline{y}) \\ \text{MAP}(\theta|\underline{y}) \end{cases}$$

(2) Bayesian interval estimation: here we have two methods, where the target was to build $CR \subset \Theta$ st $P(\theta \in CR|\underline{y}) \geq 1-\alpha$, with α low. We call it CR as credible region (or CI as credible interval).

- we could take the quantiles to delimit a $1-\alpha$ probability

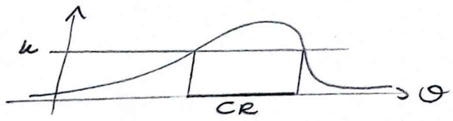
$$CR_{\theta|\underline{y}} = \left(q_{\frac{\alpha}{2}}^{\pi(\theta|\underline{y})}, q_{1-\frac{\alpha}{2}}^{\pi(\theta|\underline{y})}\right) \text{ if } \Theta \subset \mathbb{R}$$

$$CR_{\theta_j|\underline{y}} = \left(q_{\frac{\alpha}{2}}^{\pi_j(\theta|\underline{y})}, q_{1-\frac{\alpha}{2}}^{\pi_j(\theta|\underline{y})}\right) \text{ if } \Theta \subset \mathbb{R}^p$$



- we could take the HPD region (highest probability density region) which would be the narrowest region/interval that meets that request. For a uni-modal distr the idea makes sense:

$$CR = \{ \theta \in \Theta : \pi(\theta | \underline{y}) \geq k \}$$
$$k : \mathbb{P}(\theta \in CR | \underline{y}) = 1-\alpha$$



(3) Hypothesis testing. We could have two cases, the conventional set comparison or the punctual test.

- $\begin{cases} H_0 \,|\, \theta \in \Theta_0 \\ H_1 \,|\, \theta \in \Theta_1 \end{cases}$ prior: $\theta \sim \pi(\theta)$ or the more general $\theta \sim \begin{cases} g_0(\theta) \,|\, \text{if } \theta \in \Theta_0 \\ g_1(\theta) \,|\, \text{if } \theta \in \Theta_1 \end{cases}$

Let's call $\pi_0 = \pi(\Theta_0)$ and $\pi_1 = \pi(\Theta_1) = 1 - \pi_0$ since $\Theta_1 = \Theta_0^c$. These will give the prior odds. For the posterior one we firstly need the likelihood/marginal

$$f(\underline{y}) = \int_\Theta f(\underline{y}|\theta)\pi(\theta)\,d\theta =$$
$$= \pi_0 \int_{\Theta_0} f(\underline{y}|\theta) g_0(\theta)\,d\theta + \pi_1 \int_{\Theta_1} f(\underline{y}|\theta) g_1(\theta)\,d\theta$$

$$\Rightarrow \pi(\theta|\underline{y}) = \begin{cases} \dfrac{\pi_0 f(\underline{y}|\theta) g_0(\theta)}{f(\underline{y})} & \text{if } \theta \in \Theta_0 \\ \dfrac{\pi_1 f(\underline{y}|\theta) g_1(\theta)}{f(\underline{y})} & \text{if } \theta \in \Theta_1 \end{cases}$$

Now we could just use for example if $\mathbb{P}(\theta \in \Theta_0 | \underline{y}) \overset{?}{\geq} 0.5$ or similar, but the idea is to compare also with the prior ones.

$$\left(\begin{matrix}\text{POST} \\ \text{ODDS}\end{matrix}\right) = \frac{\mathbb{P}(\theta \in \Theta_0|\underline{y})}{\mathbb{P}(\theta \in \Theta_1|\underline{y})} = \frac{\pi_0}{\pi_1} \cdot \frac{\int_{\Theta_0} f(\underline{y}|\theta) g_0(\theta)\,d\theta}{\int_{\Theta_1} f(\underline{y}|\theta) g_1(\theta)\,d\theta} =$$
$$= \frac{\pi_0}{\pi_1} \cdot BF_{01} = \left(\begin{matrix}\text{PRIOR} \\ \text{ODDS}\end{matrix}\right) \cdot BF_{01}$$

- $\begin{cases} H_0 \,|\, \theta = \theta_0 \\ H_1 \,|\, \theta \neq \theta_0 \end{cases}$ For this case the idea is the same as the above where we can set as g the following

$$\theta \sim \begin{cases} g_0(\theta) = \delta_{\theta_0}(\theta)\, \mathbb{1}_{\{\theta = \theta_0\}} \\ g_1(\theta) = \pi(\theta)\, \mathbb{1}_{\{\theta \neq \theta_0\}} \end{cases} \Rightarrow BF_{01} = \frac{f(\underline{y}|\theta_0)}{f(\underline{y})}$$

Then once we have BF_{01} we can interpret it according to this table from literature:

$$\log_{10}(BF_{01}) \in \begin{cases} (0, 1/2) & | \text{ barely mentionable} \\ (1/2, 1) & | \text{ substantial} \\ (1, 2) & | \text{ strong} \\ (2, +\infty) & | \text{ decisive} \end{cases} \quad \left(\begin{matrix}\text{evidence} \\ \text{for } H_0\end{matrix}\right)$$

(4) Posterior predictive distribution. (while the prior pred distr would just be the marginal, we $f(y) = \int_\Theta f(y|\theta)\pi(\theta)\,d\theta$). Here we want to get the law of a new obs given all the others. So

$$f_{Y_{n+1}|Y_1,\dots,Y_n}(y|\underline{y}) = \frac{f(y,\underline{y})}{f(\underline{y})} = \Big|_{\substack{\text{augment} \\ \text{trick}}} \frac{\int_\Theta f(y,\underline{y}|\theta)\pi(\theta)\,d\theta}{\int_\Theta f(\underline{y}|\theta)\pi(\theta)\,d\theta} =$$
$$= \Big|_{\text{cond} \perp} \int_\Theta f(y|\theta) \left(\frac{f(\underline{y}|\theta)\pi(\theta)}{\int_\Theta f(\underline{y}|\theta)\pi(\theta)\,d\theta} \right) d\theta = \int_\Theta f(y|\theta)\pi(\theta|\underline{y})\,d\theta$$

Or another way, quicker:

$$f_{Y_{n+1}|\underline{Y}}(y|\underline{y}) = \Big|_{\substack{\text{augment} \\ \text{trick}}} \int_\Theta f_{Y_{n+1}|\underline{Y}}(y,\theta|\underline{y})\,d\theta =$$
$$= \int_\Theta f(y|\theta,\cancel{\underline{y}}) \cdot \pi(\theta|\underline{y}) = \Big|_{\text{cond} \perp} \int_\Theta f(y|\theta) \cdot \pi(\theta|\underline{y})\,d\theta$$

PRIORS STUFF

We say that the r.v.s $(Y_1, \dots, Y_n)$ is exchangeable if for any permutation π of $(1, \dots, n)$, we have that $\mathcal{L}(Y_1, \dots, Y_n) = \mathcal{L}(Y_{\pi(1)}, \dots, Y_{\pi(n)})$. This condition implies that the order in which data are recorded is irrelevant for inferential purposes.

This implies a symmetry about the role of the individuals in the sampling, like an "analogous conditions". It is relevant for Bayesian stuff through a

Thm (De Finetti's representation)

a (an) sequence $(Y_n)_{n \geq 1}$ of binary r.v.s is exchangeable $\iff$ $\exists$ a prob measure F on $([0,1], \mathcal{B}([0,1]))$ st

$$p_{\underline{Y}}(\underline{y}) = \int_0^1 \left(\theta^{\sum y_i} (1-\theta)^{n - \sum y_i} \right) F(\theta) \, d\theta$$

$\forall n \geq 1$ and $\forall y_i \in \{0,1\}$

$\iff$ $\exists$ some r.v. $\tilde{\theta}$ st we can model

$$Y_1, \dots, Y_n | \tilde{\theta} \overset{iid}{\sim} Ber(\tilde{\theta})$$
$$\tilde{\theta} \sim F(\theta)$$

About actually defining priors we can have different choices.

(1) Reference (or convenience) priors. We use them when we want to have a minimal impact on the Bayesian analysis. They are also called "non-informative", but this definition is misleading, they are more a reliable/common choice to start modelling with.

An example is a flat prior, $\pi(\theta) = c$ $\forall \theta \in \Theta$. This makes it a $\mathcal{U}(\Theta)$ if Θ is bounded, otherwise becomes improper. But a flat prior in a representation may not be flat in another reparametrization. $\smile$ $\theta \sim Beta(0,0)$ and $\gamma = \ln(\theta / 1-\theta)$

(2) Jeffreys priors. They are often improper, but are invariant to monotone transformations. They are defined through the Fisher information

$$I(\theta) = E\left[\left(\frac{\partial}{\partial \theta} \ln f(y|\theta) \right)^2 \right] = E\left[-\frac{\partial^2}{\partial \theta^2} \ln f(y|\theta) \right]$$

E wrt y, not θ

$\pi(\theta) \propto \sqrt{I(\theta)}$ $\smile$ Given the proportional we take just the functional part of θ, discard constants $\Rightarrow$ we can use both $f(y|\theta)$ and $f(\underline{y}|\theta)$ to compute it

(3) Informative (or scientifically informed) priors. The best choice to use, especially for delicate/relevant params of the model.

Anyway, we could also be more free in this choice as eventually, having more data, all the priors will merge anyway (converge) to the right posterior. The

Thm (AN of the posterior distr). Let $Y_1, \dots, Y_n | \theta \overset{iid}{\sim} f(y|\theta)$, and $\pi(\theta)$ the prior of $\theta \in \Theta \subseteq \mathbb{R}^p$. Under suitable regularity conditions we have that

$$\pi(\theta | \underline{y}) \underset{n \to +\infty}{\approx} \mathcal{N}(\tilde{\theta}_n, V_n)$$

$\tilde{\theta}_n$: posterior mean
V_n: posterior cov matrix

POPULAR DISTRIBUTIONS

$X \sim Pois(\lambda) \iff p_X(k) = e^{-\lambda} \frac{\lambda^k}{k!} \mathbb{1}_{\mathbb{N}_0}(k)$ $E(X) = \lambda$, $var(X) = \lambda$

$X \sim Exp(\lambda) \iff f_X(x) = \lambda e^{-\lambda x} \mathbb{1}_{[0,+\infty)}(x)$ $E(X) = 1/\lambda$, $var(X) = 1/\lambda^2$

$X \sim \text{Beta}(a,b) \iff f_X(x) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1} (1-x)^{b-1} \mathbb{1}_{[0,1]}(x)$

Comments:

- $X_i \overset{\perp}{\sim} \Gamma(a_i, b)$ for $i=1,2$ $\Rightarrow$ $\frac{X_i}{X_1+X_2} \sim \text{Beta}(a_i, a_j)$
- It's also an extension of the $U([0,1])$ law to also model different masses rather than the first flat one

$E(X) = \frac{a}{a+b}$

$\text{var}(X) = \frac{ab}{(a+b+1)(a+b)^2}$

$S^{2-1} = S^1 = \{x \in \mathbb{R} : \begin{smallmatrix} x\in[0,1] \\ 0 \le x \le 1 \end{smallmatrix}\}$

$$\boxed{\left(\prod_{i=1}^{n-1} x_i^{\alpha_i - 1}\right)\left(1 - \sum_{i=1}^{n-1} x_i\right)^{\alpha_n - 1}}$$

$\begin{pmatrix} X_1 \\ \vdots \\ X_n \end{pmatrix} \sim \text{Dir}\left(\underline{\alpha} = \begin{pmatrix} \alpha_1 \\ \vdots \\ \alpha_n \end{pmatrix}\right) \iff f_{\underline{X}}(\underline{x}) = \frac{\Gamma(\sum \alpha_i)}{\prod \Gamma(\alpha_i)} \prod_{i=1}^{n} x_i^{\alpha_i - 1} \mathbb{1}_{S^{n-1}}\left(\underline{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_{n-1} \end{pmatrix}\right)$

the last component is determined by $x_1, \dots, x_{n-1}$

Comments:

- The support is $S^{n-1}(\underline{x}) = \{(x_1, \dots, x_{n-1}) \in \mathbb{R}^{n-1} : x_i \in \{0,1\}, 0 \le \sum x_i \le 1\}$ and then x_n will be $1 - \sum_{i=1}^{n-1} x_i$
- It's the extension of the Beta law
- $U_i \overset{\perp}{\sim} \Gamma(\alpha_i, \beta)$ for $i = 1, \dots, n$ $\iff$ $X_i = \frac{U_i}{\sum U_i}$, $\underline{X} \sim \text{Dir}(\underline{\alpha})$
- $\begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix} \sim \text{Dir}\begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{pmatrix} \Rightarrow \begin{pmatrix} X_1 + X_3 \\ X_2 \end{pmatrix} \sim \text{Dir}\begin{pmatrix} \alpha_1 + \alpha_3 \\ \alpha_2 \end{pmatrix}$ and so on
- $\underline{X} \sim \text{Dir}(\underline{\alpha}) \Rightarrow \underbrace{X_1}_{\sim \text{Beta}(\alpha_1, \alpha_0 - \alpha_1)} \perp \underbrace{\left(\frac{X_2}{1-X_1}, \dots, \frac{X_n}{1-X_1}\right)}_{\sim \text{Dir}(\alpha_2, \dots, \alpha_n)}$

$E(X_i) = \frac{\alpha_i}{\alpha_0}$

$\text{var}(X_i) = \frac{\alpha_i(\alpha_0 - \alpha_i)}{(\alpha_0 + 1)\alpha_0^2}$

$(\alpha_0 = \sum \alpha_i)$

$X \sim \text{Gamma}(\alpha, \lambda) \iff f_X(x) = \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x} \mathbb{1}_{[0,+\infty)}(x)$

Comments:

- $\Gamma(\alpha+1) = \alpha\Gamma(\alpha)$ so $\Gamma(n+1) = n!$
- $X_i \overset{\perp}{\sim} \Gamma(\alpha_i, \lambda)$ for $i = 1, \dots, n$ $\Rightarrow$ $\sum X_i \sim \Gamma(\sum \alpha_i, \lambda)$
- $\Gamma(1/2) = \sqrt{\pi}$
- $c \cdot \Gamma(\alpha, \lambda) \sim \Gamma(\alpha, \lambda/c)$ for $c>0$

$E(X) = \alpha/\lambda$

$\text{var}(X) = \alpha/\lambda^2$

$Y = \frac{1}{X} \sim \text{inv Gamma}(\alpha, \lambda)$

$E(Y) = \frac{\lambda}{\alpha - 1}$

$\frac{\lambda^\alpha}{\Gamma(\alpha)}\left(\frac{1}{x}\right)^{\alpha+1} e^{-\lambda\left(\frac{1}{x}\right)}$

$X \sim \text{Categorical}(\underline{p}) \iff p_X(i) = p_i \mathbb{1}_{\{1,\dots,k\}}(i)$
$= \prod_{j=1}^{k} p_j \mathbb{1}_{\{i=j\}}$

$\Gamma(a) = \int_0^{+\infty} x^{a-1} e^{-x} \, dx$

Comments:

- The supp$(X) = \{1, \dots, k\}$, so X gets a label from 1 to k
- The vector $\underline{p}$ is $\underline{p} = (p_1, \dots, p_k) \in S^k$, so $\sum p_i = 1$ and $p_i \in [0,1]$ $\forall i$

$\begin{pmatrix} X_1 \\ \vdots \\ X_k \end{pmatrix} \sim \text{Multinomial}\left(n, \begin{pmatrix} p_1 \\ \vdots \\ p_k \end{pmatrix}\right) \iff p_{\underline{X}}(\underline{x}) = \frac{n!}{x_1! \cdots x_k!} \prod_{i=1}^{k} p_i^{x_i}$

Comments:

- It is the extension of the Binomial, since each of x_i counts the # of occurrence (successes) of extracting label i (out of labels 1 to k) when done n experiments
- So the supp$(\underline{X})$ is the set of values $x_i \in \mathbb{N}_0$ that sum up to n, $\sum x_i = n$ (so we get n outcomes)

$E(X_i) = np_i$

$\text{var}(X_i) = np_i(1-p_i)$

FAMOUS MODELS

Beta-Binomial model. Useful when we work on data which are obtained through a series of Bernoulli experiments.

> prior: $\theta \sim Beta(a,b)$
> likelih: $y_1,\dots,y_n \mid \theta \sim Ber(\theta) \iff \underline{y} \mid \theta \sim Bin(n,\theta)$
> $\Rightarrow$ post: $\theta \mid \underline{y} \sim Beta\left(a+\sum y_i,\ b+(n-\sum y_i)\right)$

$$f(\theta\mid\underline{y}) = \frac{f(\theta,\underline{y})}{f(\underline{y})} \propto f(\theta,\underline{y}) = f(\underline{y},\theta) = f(\underline{y}\mid\theta)\cdot f(\theta) =$$

$$= \left[\theta^{\sum y_i}(1-\theta)^{n-\sum y_i}\right]\left[\frac{1}{B(a,b)}\theta^{a-1}(1-\theta)^{b-1}\mathbb{1}_{[0,1]}(\theta)\right] \propto$$

$$\propto \theta^{(a+\sum y_i)-1}(1-\theta)^{(b+n-\sum y_i)-1}\mathbb{1}_{[0,1]}(\theta) \Rightarrow \text{kernel of the updated Beta}$$

Normal-Normal model. Useful to model data which follows a normal law of known variance, and we want to model μ also with a normal law.

> prior: $\mu \sim N(\mu_0, \tau^2 = var)$
> likelih: $y_1,\dots,y_n \mid \mu \sim N(\mu, \sigma^2 = var)$
> $\Rightarrow$ post: $\mu \mid \underline{y} \sim N(\mu_n, \tau_n^2)$
>
> $\mu_n = \frac{\sigma^2\mu_0 + n\tau^2\bar{y}}{\sigma^2+n\tau^2}$, $\tau_n^2 = \frac{\sigma^2\tau^2}{\sigma^2+n\tau^2}$

all the likelihood are actually with $\dots \overset{iid}{\sim} \dots$, as data are cond $\perp$

$$f(\mu\mid\underline{y}) \propto f(\underline{y}\mid\mu)\cdot f(\mu) = \left[\prod_{i=1}^{n}\cancel{\frac{1}{\sqrt{2\pi\sigma^2}}}e^{-\frac{1}{2\sigma^2}(y_i-\mu)^2}\right]\cdot\left[\cancel{\frac{1}{\sqrt{2\pi\tau^2}}}e^{-\frac{1}{2\tau^2}(\mu-\mu_0)^2}\right] \propto$$

$$\propto e^{-\frac{1}{2\sigma^2}\sum(y_i-\mu)^2}\, e^{-\frac{1}{2\tau^2}(\mu-\mu_0)^2} = \dots =$$

$$= e^{-\frac{1}{2}\left[\frac{n}{\sigma^2}(\mu-\bar{y})^2+\frac{1}{\tau^2}(\mu-\mu_0)^2\right]} = \dots$$

Gamma-Poisson model. Useful when our data describes rare events that we want to count, so a Poisson law is well suited for them.

> prior: $\theta \sim Gamma(\alpha,\beta)$
> likelih: $y_1,\dots,y_n \mid \theta \sim Poi(\theta)$
> $\Rightarrow$ post: $\theta \mid \underline{y} \sim Gamma(\alpha+\sum y_i, \beta+n)$

$$f(\theta\mid\underline{y}) \propto f(\underline{y}\mid\theta)\cdot f(\theta) = \left[\prod_{i=1}^{n} e^{-\theta}\frac{\theta^{y_i}}{y_i!}\right]\left[\frac{\beta^{\alpha}}{\Gamma(\alpha)}\theta^{\alpha-1}e^{-\beta\theta}\mathbb{1}_{[0,+\infty)}(\theta)\right] \propto$$

$$\propto e^{-n\theta}\,\theta^{\sum y_i}\,\theta^{\alpha-1}e^{-\beta\theta}\mathbb{1}_{[0,+\infty)}(\theta) =$$

$$= \theta^{(\alpha+\sum y_i)-1}e^{-(\beta+n)\theta}\mathbb{1}_{[0,+\infty)}(\theta) \Rightarrow \text{kernel of the updated Gamma}$$

Dirichlet-multinomial model. Useful when we have data values that belong to different categories/classes (in fact as data we get a multinomial law), of which the probabilities are modeled through a Dirichlet law.

> prior: $\underline{p} \sim Dir(\underline{\alpha})$, $\underline{p} \in \mathbb{R}^k$, $k = \#$classes
> likelih: $\underline{y} = (y_1,\dots,y_k) \mid \underline{p} \sim Mult(n,\underline{p})$
> $\Rightarrow$ post: $\underline{p} \sim Dir(\underline{\alpha} + \underline{y})$

remember that $\underline{y}\mid\underline{p} \sim Mult(n,\underline{p})$ means that $\underline{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_k \end{pmatrix}$, $y_i = \#$ events that belong to category i ($n = \sum y_i = \#$ total observations)

$$f(p \mid y) \propto f(y \mid p) \cdot f(p) = \left[\frac{n!}{y_1! \cdots y_k!} \prod_{i=1}^{k} p_i^{y_i} \right] \cdot \left[\frac{1}{B(\alpha)} \prod_{j=1}^{k} p_j^{\alpha_j - 1} \mathbb{1}_{S_{k-1}}(p) \right]$$

$$\propto \prod_{j=1}^{k} p_j^{(\alpha_j + y_j) - 1} \mathbb{1}_{S_{k-1}}(p) \rightarrow \text{kernel of the updated Dir}$$

normal, inverse Gamma - normal model. Useful to model normal data, but now with both parameters unknown.

prior: $\mu \mid \sigma^2 \sim N(\mu_0, \sigma^2/\lambda)$ $\;\;\sigma^2 \sim \text{inv Gamma}(\alpha, \beta)$ $\Rightarrow \mu \not\perp \sigma^2$

data: $Y_1, \dots, Y_n \mid (\mu, \sigma^2) \sim N(\mu, \sigma^2)$

$\Rightarrow$ post: $\mu \mid \sigma^2 \sim N\left(\frac{n\bar{y} + \lambda \mu_0}{n + \lambda}, \frac{\sigma^2}{n + \lambda}\right) \rightarrow \mu \not\perp \sigma^2$

$$\sigma^2 \sim \text{inv Gamma}\left(\alpha + \frac{n}{2}, \beta + \frac{1}{2}\sum (y_i - \bar{y})^2 + \frac{1}{2} \frac{n\lambda(\bar{y} - \mu_0)^2}{n + \lambda}\right)$$

SIMULATION METHODS (MCMC)

We now want to get relevant features about our target distr $\pi(\cdot \mid \underline{z})$, the posterior, through simulations, as often we can't directly sample and from it (we sample means get values that follow a certain distribution).

Idea: resort on two thms about simulations and long-runs properties.

<u>Thm</u> (SLLN). Let $\theta^{(1)}, \theta^{(2)}, \dots$ be an iid sequence of samples from $\pi(\theta)$, and let $h: \Theta \to \mathbb{R}$ st $E[|h(\theta)|] < +\infty$. Our target is computing

$$\bar{h} = E_\pi(h(\theta)) = \int_\Theta h(\theta)\,\pi(\theta)\,d\theta$$

Then we have that

(1) $$\bar{h}^{(T)} = \frac{1}{T}\sum_{t=1}^{T} h(\theta^{(t)}) \xrightarrow[T\to+\infty]{a.s.} \bar{h}$$

(2) $$g_\alpha^{(T)} = \left(\substack{\alpha\text{-quantile} \\ \text{of } h(\theta^{(t)})}\right) \xrightarrow[T\to+\infty]{a.s.} g_\alpha = \left(\substack{\alpha\text{-quantile} \\ \text{of } f(\theta)}\right)$$

<u>Thm</u> (CLT). If we have the above assumptions plus $var[h(\theta)] < +\infty$, then we get also info about the velocity of convergence:

set $\sigma^2 = var[h(\theta)]$
$\quad\sigma^{2(T)} = var[h(\theta^{(t)})]$ over $t = 1, \dots, T$

(1) $$\sqrt{T}\,(\bar{h}^{(T)} - \bar{h}) \xrightarrow[T\to+\infty]{\mathcal{L}} \mathcal{N}(0, \sigma^2) \iff \bar{h}^{(T)} \underset{T\to+\infty}{\approx} \mathcal{N}\left(\bar{h}, \frac{\sigma^2}{T}\right)$$

(2) $$\sigma^{2(T)} \xrightarrow[T\to+\infty]{a.s.} \sigma^2$$

MARKOV CHAINS MONTE CARLO

The idea is to get samples from the posterior $\pi(\theta \mid \underline{z})$ by building a MC whose limiting distr is that posterior, and so the samples will be given by the states we are visiting.

Let $E \subset \mathbb{R}^k$ be the state space. A time-homogeneous MC $(X_n)_{n\geq 0}$ with values in E is a sequence of rvs E-valued st

$$\mathbb{P}(X_{n+1} \in A \mid X_n = x_n, \dots, X_0 = x_0) = \mathbb{P}(X_{n+1} \in A \mid X_n = x_n)$$

we call
$$\left.\begin{aligned} P(x_n, A) &= \mathbb{P}(X_{n+1} \in A \mid X_n = x_n) \\ P^m(x, A) &= \mathbb{P}(X_m \in A \mid X_0 = x) \\ P_x(\cdot) &= \mathbb{P}(\cdot \mid X_0 = x) \end{aligned}\right\}$$
$P(\cdot,\cdot)$ is the transition probability kernel

We now see some terminology and properties, and then focus on the needed requirements for the convergence algos.

(1) <u>Invariant</u> (or stationary) distribution (as we need also a distr for X_0 to define a MC). A prob meas π on $(E, \mathcal{E})$ st

$$\int_E \underset{\Longrightarrow}{\pi(x)}\ \underset{\boxed{d}}{P(x,A)}\ dx = \pi(A) \quad \forall A \in \mathcal{E} \quad (\pi P = \pi)$$

(2) <u>Irreducibility</u>: a MC is irreducible if $i \to j$ (j can be reached from i) $\forall i, \forall j \in E$ in a finite # of steps. This means that eventually (sooner or later) we will visit all the reachable states/sets of states.

$$\exists\, \psi \text{ a prob distr}: \ \forall A \text{ st } \psi(A) > 0 \ \ \exists\, m_{x,A} \geq 1 : P^m(x, A) > 0 \quad \forall x \in E$$

This is good as it means that the MCMC will be able to visit the whole support of the post distr. This property is guaranteed:
- for GS if $\exists\, n \geq 1$ st P^n has a (strictly) positive density f (wrt ψ)
- for MH the one of GS plus if P has discrete and absolutely continuous components

(3) <u>Recurrence</u>. An irreducible (assumption!) MC is recurrent if we will visit infinitely often (io) the sets of reachable states a infinite number of times. This, only at least from almost every initial point.

$$P_x(X_n \in A \text{ io}) > 0 \quad \forall x$$
$$P_x(X_n \in A \text{ io}) = 1 \quad \psi\text{-ae wrt } x$$

Thm Let $(X_n)_{n\geq 0}$ an irreducible MC, and π a stationary distr for it.
$\Rightarrow$ • the MC is π-irreducible
• π is the unique stat distr $\left(\begin{array}{l}\text{irreducible} \\ +\exists \pi \text{ invariant}\end{array} \Rightarrow \text{recurrent}\right)$
• the MC is also recurrent

ergodic
Thm (SLLN) Let $(X_n)_{n\geq 0}$ an irreducible MC, with π its (unique) invariant distr. Let $f: E \to \mathbb{R}$ st $E_\pi[|f|] < +\infty$. Then

$$P_x\left(\frac{1}{n+1}\sum_{i=0}^{n} f(X_i) \xrightarrow[n\to+\infty]{} \int_E f(x)\pi(x)\,dx\right) = 1 \qquad \begin{array}{c}\pi\text{-ae} \\ \text{wrt } x\end{array}$$

So starting from the "correct" points that estimation is good. But we need to solve this initialization problem, as that limit may not hold for $x \in C$ st $\pi(C) = 0$.

Anyway the theorem is really useful as to compute $\pi(A)$ we can do

$$\pi(A) = \int_A \pi(x)\,dx = \int_E [\mathbb{1}_A(x)]\pi(x)\,dx \Rightarrow f(x) = \mathbb{1}_A(x)$$

$$\Rightarrow \pi(A) \underset{n\text{ large}}{\approx} \frac{1}{n+1}\sum_{i=0}^{n} f(X_i) = \frac{\#(X_i \in A)}{n+1}$$

To solve that issue of a.e wrt x we move to (a) Harris recurrence. a MC is Harris recurrent if (it is irreducible firstly and then if)

$$\forall A \text{ st } \psi(A) > 0,\ P_x(X_n \in A \text{ i.o.}) = 1 \quad \forall x \in E$$

And to be sure that we reach the invariant distr we need (b) Aperiodicity. a MC is aperiodic if all states have period 1. Then we get this important

Thm Let $(X_n)_{n\geq 0}$ an irreducible, aperiodic MC, with P its transition matrix and π its invariant distr (recall $\exists$, so the MC is also recurrent). Then

$$\| P^n(x,\cdot) - \pi(\cdot) \| \xrightarrow[n\to+\infty]{} 0 \qquad \pi\text{-ae wrt } x$$

Again, to get it be true $\forall x \in E$ we need to ask Harris recurrent instead of just recurrent.

So, final idea of the method: the goal is to approximate $E_\pi[h(\theta)]$, with the posterior distr π, and $h: \Theta \to \mathbb{R}$ good. Then we can

- build a MC $(\theta_n)_{n\geq 0}$ of state space Θ st it is irreducible and Harris recurrent, and has $\pi(\theta|x)$ as invariant distr
- set a sensible initialization θ_0
- simulate the MC, taking just the values after a Burn In (BI) period
- estimate

$$E_\pi[h(\theta)] = \int_E h(\theta)\pi(\theta|x)\,d\theta \approx \frac{1}{T+1}\sum_{i=0}^{T} h(\theta_i)$$

Last problem: how to set π to be the invariant distr?
(c) Reversibility. a MC $(X_n)_{n\geq 0}$ of matrix P is π-reversible if

$$\pi(x)P(x,y) = \pi(y)P(y,x) \quad \forall x \forall y \in E$$

Thm π reversible $\Rightarrow$ π invariant
Proof: we have to show $\pi P = \pi$. In the discrete case we have

$$(\pi P)_j = \sum_i \pi_i P_{ij} = \sum_i \pi_j P_{ji} = \pi_j\Big(\sum_i P_{ji}\Big) = \pi_j \cdot 1 = \pi_j$$

In the continuous case

$$\int_E \pi(x)P(x,y)\,dx = \int_E \pi(y)P(y,x)\,dx = \pi(y)\left(\int_E P(y,x)\,dx\right) = \pi(y)$$

METROPOLIS-HASTINGS ALGORITHM

Suppose that the target distribution π has a density (wrt a measure ν). Consider a transition probability $q(x,y) = g(x,y)$, with g being the proposal density, and st $Q(x,E^+) = 1 \ \forall x \notin E^+$ where $E^+ = \text{supp}(\pi) = \{x \in E : \pi(x) > 0\}$.

Then the MH alg does this:

(1) set $X_n = x$
(2) generate a candidate point y sampled from $Q(x,\cdot)$) so we need to be able to sample from this
(3) define

$$\alpha(x,y) = \begin{cases} \min\left(\frac{\pi(y)\,g(y,x)}{\pi(x)\,g(x,y)}, 1\right) & \text{if den} \neq 0 \\ 1 & \text{if den} = 0 \end{cases}$$

(4) accept with probability $\alpha(x,y)$ (so generate $u \sim \mathcal{U}([0,1])$ and decide) or reject. So we get $X_{n+1} = y$ or $X_{n+1} = x$
(5) advance to the next iteration $n+1$ and repeat

So we built a MC which has as transition kernel

$$P(x,y) = [g(x,y)\,\alpha(x,y)]\,\mathbb{1}_{\{x \neq y\}} + \underbrace{[r(x)]}\,\mathbb{1}_{\{x=y\}}$$

we $\delta_x(y)$

$$r(x) = \mathbb{P}\left(\begin{matrix}\text{remaining in}\\ x \text{ from } n \to n+1\end{matrix}\right) = 1 - \int_E g(x,y)\,\alpha(x,y)\,dy$$

We have to check the requirements about the MC we got, which were irreducibility, H recurrent, and π invariant (reversible)

(1) We check that $\pi(x)$ is reversible, ie

$$\pi(x)\,p(x,y) = \pi(y)\,p(y,x)$$
$$\pi(x)[g(x,y)\,\alpha(x,y)] = \pi(y)[g(y,x)\,\alpha(y,x)]$$

Suppose we are in the case of α interesting, defined, ie the minimum of ● and 1 is ●. So ● ≤ 1 and we get

$$LHS = \pi(x)[g(x,y)\,\alpha(x,y)] = \cancel{\pi(x)}\,\cancel{g(x,y)}\,\frac{\pi(y)\,g(y,x)}{\cancel{\pi(x)}\,\cancel{g(x,y)}} =$$
$$= \pi(y)\,g(y,x) = \pi(y)[g(y,x) \cdot \underset{\sim}{1}] = RHS$$

since here we have $\alpha(y,x)$, but if $\alpha(x,y)$ was ≤ 1 then $\alpha(y,x)$ which is $1/\alpha(x,y)$ will be ≥ 1 so gets clipped

(2) For irreducible and H recurrent we need stronger assumptions on the choice of $g(x,y)$.

- Random walk MH chain: we let $g(x,y) = f(y-x)$ where f is a density, like $\mathcal{N}$. That is equivalent to setting $y = x + z$ with $z \sim f$.
And we have the requirements if $f(x) > 0 \ \forall x \in E$.
- Independence MH chain: we let $g(x,y) = f(y)$, so we are setting $y \sim f$ in an $\perp$ way from x.
And we have the requirements if $f(x) > 0$ ν-ae in E^+.

GIBBS SAMPLER

The MH alg becomes very inefficient when θ is multidimensional, and since it uses a joint proposal density this can give problems when the components of θ are on different scales or multimodal or skewed.

So GS resorts to a divide & conquer approach. Let $\theta = (X,Y)$ and then $\pi(x,y)$ the target distribution. Assume we know the full conditional distributions, $\boxed{f_{X|Y}}$ and $\boxed{f_{Y|X}}$. Then the idea is:

(1) at iteration n we have (X_n, Y_n)
(2) perform sequentially the updates:

sample X_{n+1} from $f_{X|Y}(\cdot \mid Y_n)$
sample Y_{n+1} from $f_{Y|X}(\cdot \mid X_{n+1})$

(3) advance to next iteration and repeat

In this way we get a bivariate MC simulation, for $(X_n, Y_n)_{n\geq 0}$. Again we need to check that we have the desired/required properties on this MC. Typically they are met when:

- $spt(\pi_X) \times spt(\pi_Y) = spt(\pi)$
- $f_{X|Y}$ and $f_{Y|X}$ are >0 on the respective supports of the marginals π_X and π_Y
- the marginals π_X and π_Y exists (ie π is not improper)

The power of this method is even clearer in higher dimensional cases, as we do

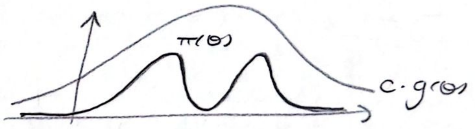
$$\text{sample } X_1^{(n+1)} \text{ from } f_{X_1|X_2,\dots,X_p}(\cdot \mid x_2^{(n)}, \dots, x_p^{(n)})$$
$$\text{sample } X_2^{(n+1)} \text{ from } f_{X_2|X_1,X_3,\dots}(\cdot \mid x_1^{(n+1)}, x_3^{(n)}, \dots)$$
$$\vdots$$
$$\text{sample } X_p^{(n+1)} \text{ from } f_{X_p|X_1,\dots,X_{p-1}}(\cdot \mid x_1^{(n+1)}, \dots, x_p^{(n+1)})$$

So anyway this methods generate a MC where the transition kernel is given by the product of the full conditionals and we sample θ sequentially and always accept it. This allows to see GS as a particular case of MH.

SAMPLING METHODS

(1) Rejection sampling. Let $\pi(\theta)$ the target distr. from which we can't directly sample, but assume we can evaluate $\pi(\theta)$ ∀θ. To apply this method we need

- a proposal density $g(\theta)$, from which we are able to sample
- a constant c st $\pi(\theta) \leq c \cdot g(\theta)$ $\forall \theta \in \Theta$ (like an envelope)

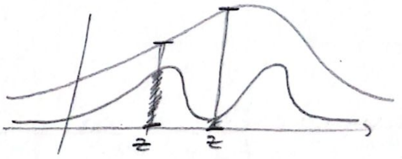


Then the method is

- draw a sample z from $g(\cdot)$, and compute $r(z) = \frac{\pi(z)}{c \cdot g(z)}$
- generate $u \sim \mathcal{U}([0,1])$
- accept z (as sample of $\pi(\cdot)$) if $u \leq r(z)$

Which is equivalent to this:

- sample z from $g(\cdot)$, on the x axis
- draw the vertical line till $c \cdot g(z)$
- sample $u \sim \mathcal{U}([0, c\cdot g(z)])$
- accept if $u \leq \pi(z)$



(2) Inverse CDF. The idea is that given X a r.v. of cdf $F_X(x)$, then there is a result that $U = F_X(X) \sim \mathcal{U}([0,1])$. So

$$X \sim F_X(x) \Rightarrow F_X(X) \sim \mathcal{U}([0,1]) \Rightarrow \left(\begin{matrix} U \sim \mathcal{U}([0,1]) \\ \text{"}U = F_X(x)\text{"} \end{matrix} \Rightarrow F_X^{-1}(U) \sim \mathcal{L}(x) \right)$$

So the idea is, to sample from a r.v. X of law $\mathcal{L}(x)$, of doing this:

- set $u = F_X(x)$ and invert it, to get $x = F_X^{-1}(u)$
- sample $u_1, \dots, u_n$ from $\mathcal{U}([0,1])$
- the samples x_i will be $x_i = F_X^{-1}(u_i)$

(3) Importance sampling. a method to just compute an integral, not to really build samples from a distribution. The idea is: we are computing a $E[h(\theta)]$ wrt to a density $f(\theta)$ which is difficult to sample from. So we convert to another law:

$$E_f[h(\theta)] = \int_\Theta (h(\theta)) f(\theta)\, d\theta = \int_\Theta h(\theta) f(\theta) \left(\frac{g(\theta)}{g(\theta)}\right) d\theta =$$
$$= \int_\Theta \left[h(\theta) \frac{f(\theta)}{g(\theta)}\right] \cdot g(\theta)\, d\theta =$$
$$= \int_\Theta [h(\theta) w(\theta)]\, g(\theta)\, d\theta = E_g[h(\theta) w(\theta)]$$

$g(\cdot)$ the importance distr, easier to sample from, but having the same spt of $f(\theta)$

weight to correct the similarity or not among $f(\cdot)$ and $g(\cdot)$

Small comment about the "full conditional": it is the law/distr of a param given all the rest, i.e. data and other params. When computing it we just leave the functional interesting part.

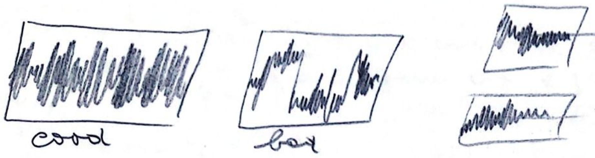
$$f(\theta|\tau,y) = \frac{f(\theta,\tau,y)}{f(\tau,y)} \propto f(\theta,\tau,y) = f(\underline{y}|\theta,\tau)\cdot f(\theta,\tau) =$$

$$= \dots =\Big|_{\text{if } \theta\perp\tau} f(\underline{y}|\theta,\tau)\cdot f(\theta)\cdot \cancel{f(\tau)} \propto \dots$$

CONVERGENCE DIAGNOSTICS

To be checked after we do our MCMC simulations (like GS or MH) to generate samples of the posterior distr.

(1) Trace plots. They show the history of the generated iterates for each parameter/variable, as a time series. Is good when they look like WN and if parallel chains settle in the end to the same value



same level for // chains, good, we converged to the "right" one

(2) Markov chain standard error. It measures the variance of the estimator $\hat{g}^{(T)}$ when we wanted to approximate the real integral $\bar{g}$.

$$\bar{g} = \int_\Theta g(\theta)\pi(\theta|\underline{y})\,d\theta \qquad \hat{g}^{(T)} = \frac{1}{T}\sum_{i=1}^{T} g(\theta^{(i)})$$

$$\widehat{\left(\frac{\sigma^2_g}{T}\right)} = \frac{\hat{\sigma}^2_g}{T}\left(1+2\sum_{j=1}^{M}\hat{\rho}_j\right)$$ good convergence if this value is small

(3) Effective sample size. It is the # of iid iterations that we would have to run (if it were possible) to get the same MC standard error that we actually obtained.

So the higher is this value (the closer to our # of simulated draws) the better, as it means that our chain is "reliable".

(4) Autocorrelation plots: the barplot of $\hat{\rho}_j = \text{cor}(g(\theta^{(i)}), g(\theta^{(i+j)}))$. We expect the corr to decrease as the lag increases. And the faster it decreases the better, as in theory from MC theory we would have correlation only to the previous value, all the rest are forgotten.

BAYESIAN LINEAR MODELS

a linear model describes a functional relation between the mean of a response variable y and some covariates $\underline{x}$. We have n observations, $y_1,\dots,y_n$, and the corresponding covariates $\underline{x}_1,\dots,\underline{x}_n \in \mathbb{R}^k$. We model as

$$\boxed{y_i|\underline{x}_i,\beta,\sigma^2 \overset{ind}{\sim} N(\underline{x}_i^T\beta,\sigma^2) \qquad i=1,\dots,n}$$

or in matrix form:

$$\underline{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \qquad \underline{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix} \in \mathbb{R}^{p=k+1} \qquad \underline{y}|X,\underline{\beta},\sigma^2 \overset{\perp}{\sim} N_n(X\underline{\beta},\sigma^2 I_n)$$

$$X = \begin{pmatrix} - \underline{x}_1 - \\ \vdots \\ - \underline{x}_n - \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1k} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n1} & \cdots & x_{nk} \end{pmatrix}_{n\times p}$$

In theory also the covariates X could be random, but if we set X and $(\underline{\beta},\sigma^2)$ to be $\perp$ (a priori), then there will not be any functional part involving X in the likelihood, so we can discard it.

$$f(\underline{y},X|\psi,\underline{\beta},\sigma^2) = f(\underline{y}|X,\psi,\underline{\beta},\sigma^2)\cdot f(X|\psi,\underline{\beta},\sigma^2) \sim \text{const factor wrt } \beta \text{ and } \sigma^2$$

So in this context a fixed (deterministic) covariate is the same of a random covariate, if we assume a priori that $X \perp (\underline{\beta},\sigma^2)$

The likelihood of the model is

$$f(\underline{y}|\beta,\sigma^2) = \frac{1}{(\sigma^2)^{n/2}} e^{-\frac{1}{2\sigma^2}[(\underline{y}-X\beta)^T(\underline{y}-X\beta)]}$$
$$= \frac{1}{(\sigma^2)^{n/2}} e^{-\frac{1}{2\sigma^2}[s^2+(\beta-\hat{\beta})^T X^T X(\beta-\hat{\beta})]}$$

$$\left.\begin{array}{l}\hat{\beta}_{MLE}=(X^TX)^{-1}X^T\underline{y}\\ s^2=(\underline{y}-X\hat{\beta})^T(\underline{y}-X\hat{\beta})\end{array}\right\}$$ MLE estimators for β and σ^2 from the frequentist approach

$$B_n=(X^TX+B_0^{-1})^{-1}$$
$$\underline{b}_n = B_n(X^TX\hat{\beta}_{MLE}+B_0^{-1}\underline{b}_0)$$

PRIORS AND CONJUGATE MODELS

(1) If σ^2 is known we have a conjugate prior for β.

$$\beta \sim N(\underline{b}_0, B_0) \Rightarrow \beta|\underline{y} \sim N(\underbrace{\underline{b}_n}, B_n)$$

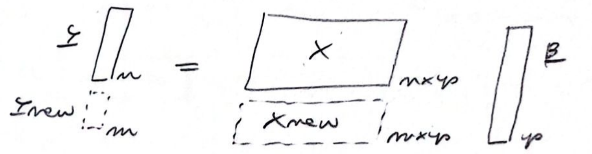
weighted average of the prior mean b_0 and the MLE $\hat{\beta}$

(2) Conjugate prior for β and σ^2.

$$\pi(\beta,\sigma^2)=\pi(\beta|\sigma^2)\cdot\pi(\sigma^2)$$

$$\begin{array}{l}\beta|\sigma^2 \sim N_p(\underline{b}_0,\sigma^2B_0)\\ \sigma^2\sim \text{inv-Gamma}(\frac{\nu_0}{2},\frac{\nu_0\sigma_0^2}{2})\end{array} \Rightarrow \begin{array}{l}\beta|\sigma,\underline{y},X\sim N_p(\underline{b}_n,\sigma^2B_n)\\ \sigma^2|\underline{y},X\sim\text{inv-Gamma}(\frac{\nu_n}{2},\frac{\nu_n\sigma_n^2}{2})\end{array}$$

Also here we can recover the posterior marginal of $\beta|\underline{y},X$, which turns out to be a multivariate (in $\mathbb{R}^p$) t Student. And if we want to get predictions for new subjects, like $\underline{y}_{new}$ the new set of m samples, we get



$$\underline{y}_{new}|\underline{y},X,X_{new}\sim$$
$$\sim t_m(X_{new}\underline{b}_n,\dots)$$

(the marginal for β was
$\beta|\underline{y},X\sim t_p(\underline{b}_n,\dots)$)

(3) Zellner's g prior. This one gets a way to fix the hyperparameters of the previous model. The idea is

$$\begin{array}{l}\beta|\sigma^2\sim N_p(\underline{b}_0,\sigma^2B_0)\\ \sigma^2\sim\text{inv-Gamma}(\frac{\nu_0}{2},\frac{\nu_0\sigma_0^2}{2})\end{array} \quad \Bigg| \quad \boxed{\begin{array}{l}B_0=c(X^TX)^{-1}\\ \nu_0=0\end{array}}$$

this requires X to be of full rank

this makes the prior improper

The effect of c can be seen in the posterior:

$$E[\beta|\sigma^2,\underline{y},X]=\frac{1}{c+1}\underline{b}_0+\frac{c}{c+1}\hat{\beta}_{MLE} \Rightarrow$$ c weights the contribution of MLE vs B_0, b_0, the magic number is eg $\ln(n)$

(4) Reference prior: the prior is improper, but the posterior comes back to the usage of the previous ones

$$\pi(\beta,\sigma^2)\propto\frac{1}{\sigma^2}\mathbb{1}_{(0,+\infty)}(\sigma^2) \Rightarrow \begin{array}{l}\beta|\sigma^2,\underline{y},X\sim N_p(\dots)\\ \sigma^2|\underline{y},X\sim\text{inv-Gamma}(\dots)\end{array}$$

(5) The most frequent/common choice is however a semi-conjugate priors, we we choose a closed form for the full conditionals, and then we can use $\boxed{\text{GS}}$.

$$\begin{array}{l}\underline{\beta}\sim N_p(b_0,B_0)\\ \tau\sim\text{Gamma}(a,b)\\ (\text{with } \beta\perp\tau)\end{array} \Rightarrow$$ the full conditionals are
$$\underline{\beta}|\tau,\dots\sim N_p(\dots)$$
$$\tau|\underline{\beta},\dots\sim\text{Gamma}(\dots)$$

GENERALIZED LINEAR MODELS

Here we assume that the y's are extracted from an Exp-family distr, and the mean μ of y depends on the covariates $\underline{x}$ not necessarily as $\mu=\underline{x}^T\underline{\beta}$ as we did before, but we now are more flexible, and we allow $\mu=h(\underline{x}^T\underline{\beta})=h(\eta)$.

So in this context we have:

- the random component, i.e. the distribution of $Y_i | \underline{x}_i$, which must be from the Exp family, and we call $\mu_i = E(Y_i | \underline{x}_i)$
- the linear (classical) predictor $\eta_i = \underline{x}_i^T \underline{\beta}$ (i.e. the covariates choice)
- the link function $g(\mu_i) = \eta_i$ (the generalization of $g(x)=x$) or the inverse $h(\eta_i) = \mu_i$, called response function.

The Exp-family distr have the form

$$f(y_i | \theta_i, \varphi) = e^{c(y_i, \varphi)} \cdot e^{\left(\frac{y_i \theta_i - b(\theta_i)}{\varphi} \right)}$$ θ_i: natural param, φ: scale param

So the params we have are β and φ (as θ_i can enter in $\underline{x}_i^T \beta$). The presence of φ is according to if we model it as a random effect or not (mixed). We usually have

cond priors: $\pi(\underline{\beta} | \varphi) \cdot \pi(\varphi)$, and usually we set
$\perp$ priors: $\pi(\beta) \cdot \pi(\varphi)$ $\quad \underline{\beta} \sim N_p(\underline{b}_0, B_0)$

Focusing on the binary regression, i.e. $Y_i | \underline{x}_i \overset{\perp}{\sim} Ber(\mu_i = g(\eta_i = \underline{x}_i^T \underline{\beta}))$. We could have different choices of the link function:

$\mu_i = \Phi(\eta_i)$ [probit model]

$\mu_i = \frac{e^{\eta_i}}{1 + e^{\eta_i}}$ [logit model]

$\mu_i = 1 - e^{-e^{\eta_i}}$ [complementary log-log model]

GIBBS SAMPLER FOR THE PROBIT MODEL

See the notes made [illegible], [illegible] [illegible] [illegible] [illegible] [illegible]. Anyway, the idea is that we have the probit model

$$Y_i | \underline{x}_i, \beta \overset{\perp}{\sim} Ber(\mu_i), \quad \mu_i = \Phi(\eta_i) = \Phi(\underline{x}_i^T \underline{\beta})$$
prior: $\beta \sim N_p(\underline{b}_0, B_0)$

The idea is to introduce some latent variables as

$$z_i = \underline{x}_i^T \beta + \varepsilon_i \quad (\varepsilon_i \overset{iid}{\sim} N(0,1)) \quad \rightarrow \quad Y_i = \begin{cases} 1 & \text{if } z_i > 0 \\ 0 & \text{if } z_i \leq 0 \end{cases} \quad \forall i = 1, \dots, n$$

which is an equivalent formulation since

$$P(Y_i = 1) = P(z_i > 0) = P\left(\frac{z_i - \underline{x}_i^T \underline{\beta}}{1} > \frac{0 - \underline{x}_i^T \underline{\beta}}{1} \right) = P(Z \geq -\underline{x}_i^T \beta) =$$
$$= 1 - \Phi(-\dots) = \Phi(\dots) = \Phi(\underline{x}_i^T \beta) = \mu_i$$

Then for the GS we start finding the joint law and from that the two full conditionals for β and $\underline{z}$.

$$f(\underline{y}, \underline{z}, \beta) = f(\underline{y} | \underline{z} \cancel{\beta}) \cdot f(\underline{z}) \cdot f(\beta) = \quad \text{as } \underline{\beta} \perp \underline{z} \text{ a priori}$$
$$= \left[\prod_{i=1}^n \underbrace{f(y_i | z_i)}_{\mathbb{1}_{\{y_i = 1\}} \mathbb{1}_{\{z_i > 0\}} + \mathbb{1}_{\{y_i = 0\}} \mathbb{1}_{\{z_i \leq 0\}}} \right] \cdot \left[\prod_{i=1}^n \underbrace{f(z_i)}_{\sim N(\underline{x}_i^T \underline{\beta}, 1)} \right] \cdot \underbrace{f(\underline{\beta})}_{\sim N(\underline{b}_0, B_0)}$$

and then we get the full conditionals $f(\underline{\beta} | \underline{z}, \underline{y})$ and $f(\underline{z} | \underline{\beta}, \underline{y})$. Check the notes for this part.

HIERARCHICAL MODELS

We now move to models which can account for latent variables, i.e. variables complex or even impossible to measure or that were not taken into account: i.e. variables somehow out of the model but that should be in.

So we will have data y that depends on latent variables $\underline{z}$ and params $\underline{\theta}$, and latent will depend on params. So we have a multilevel modelling

$$f(y | \underline{z}, \underline{\theta}) \quad , \quad f(\underline{z} | \underline{\theta}) \quad , \quad f(\underline{\theta})$$

An example of this approach (towards Bayesian mixed effect models) is when we have grouped data, since for the levels we have the groups and the units inside the groups.

Suppose we have J groups, and $\underline{y}_1, \dots, \underline{y}_J$ with $\underline{y}_j = (y_{1j}, \dots, y_{n_j j})$ representing the units inside each group. The groups are neither ⊥ nor equal: it is an half-way is letting them be exchangeable.

Similarly, the groups-characteristic params are not ⊥ nor equal, so also for them we suppose exchangeability. This leads to the multilevel model:

$$y_{1j}, \dots, y_{n_j j} \mid \theta_j \overset{iid}{\sim} f(y \mid \theta_j) \qquad \text{within-group model}$$

$$\theta_1, \dots, \theta_J \mid \phi \overset{iid}{\sim} f(\theta_j \mid \phi) \qquad \text{between-group model}$$

$$\phi \sim \pi(\phi) \qquad \text{prior distr}$$

Actually we call them hierarchical or multilevel or random effects models, are the same. And the θ_j are not ⊥ as we wish to exchange information among the groups, for example if we need to do prediction on a new group.

Taking that above hierarchical structure, we can resort on covariates and build a

LINEAR MIXED EFFECT MODEL

Let y_{ij} the r.v. corresponding to an observation, where it could be
- unit i from group j, or
- unit i measurement taken at time j

Then for each unit we have its vector of covariates, $\underline{x}_{ij}$ for $i = 1, \dots, n_j$ (# units in group j) and $j = 1, \dots, J$ (# of groups). We want to model the covariates through vectors β_j as in a classical linear model, so we get

$$y_{ij} = \underline{x}_{ij}^T \beta_j + \underbrace{\varepsilon_{ij}}_{\overset{iid}{\sim} N(0, \sigma^2)} \quad \text{or} \quad \boxed{\underline{y}_j \mid X_j, \beta_j, \sigma^2 \overset{\perp}{\sim} N_{n_j}(X_j \beta_j, \sigma^2 I_{n_j})}$$

(1) WITHIN

This for the within-group model. For the between-group we need to specify the distr of β_js. Again, they are supposed exchangeable, and we get a multilevel

BETWEEN (2) $\beta_j \mid \underline{\theta}, \Sigma \overset{iid}{\sim} N_p(\underline{\theta}, \Sigma)$

PRIOR (3) $\underline{\theta}, \Sigma \sim \pi(\theta) \cdot \pi(\Sigma)$

(and $\sigma^2 \sim \pi(\sigma^2)$)

This is a linear mixed effect model. The ME can be better seen through a re-parametrization.

$$\beta_j = \underline{\theta} + \underbrace{\underline{\gamma}_j}_{\overset{iid}{\sim} N(0, \Sigma)} \implies y_{ij} = \underline{x}_{ij}^T \beta_j + \varepsilon_{ij} = \boxed{\underline{x}_{ij}^T \underline{\theta}} + \boxed{\underline{x}_{ij}^T \underline{\gamma}_j} + \varepsilon_{ij}$$

$\underline{\theta}$: fixed effect part, as it is constant across groups

$\underline{\gamma}_j$: random effect part, as it is group specific

This can also bring the extension of different covariates choice for the two parts, as:

$$y_{ij} = \underline{x}_{ij}^T \underline{\theta} + \underline{z}_{ij}^T \underline{\gamma}_j + \varepsilon_{ij}$$

∼ with $\underline{\theta}$ and $\underline{\gamma}_j$ also of different dim maybe (like p and r)

$$\underline{\gamma}_j \mid \Sigma \overset{iid}{\sim} N_r(\underline{0}, \Sigma)$$

$$\underline{\theta}, \Sigma, \sigma^2 \sim \pi(\underline{\theta}) \cdot \pi(\Sigma) \, \pi(\sigma^2)$$

typically we set

$$\underline{\theta} \sim N_p(\mu_0, \Lambda_0)$$

$$\Sigma \sim \text{inv Wishart}(\dots)$$

$$\sigma^2 \sim \text{inv Gamma}(\dots)$$

MODEL ASSESSMENT

We now see the chapters
- model selection: which model among these is the best?
- model checking: does our model fit well enough the data?

MODEL SELECTION

(1) Compute the posterior probabilities that each model is correct and select the model(s) with the largest probability.

- case of two models comparison: special case of the next one
- case of $K = J+1$ models comparison. We have

$$M_j: \underline{y} | \theta_j, M_j \sim f(\underline{y} | \theta_j, M_j), \quad \pi(\theta_j | M_j) = \pi(\theta_j | m=j) \qquad j = 0 \to J$$

for j in 0:J
- set $P(m=j)$ the prior prob of choosing model j (typically a uniform, so $1/K$)
- compute the posterior of θ in model M_j

$$\pi(\theta_j | \underline{y}, M_j) = \frac{f(\underline{y} | \theta_j, M_j) \cdot \pi(\theta_j | M_j)}{\underbrace{\int_{\Theta_j} f(\underline{y} | \theta_j, M_j)\, \pi(\theta_j | M_j)\, d\theta_j}_{f(\underline{y} | M_j)}}$$

- compute the posterior prob mass of model M_j

$$\underbrace{P(m=j | \underline{y})}_{\text{"we choose model } M_j\text{"}} = \frac{f(\underline{y} | M_j) \cdot P(m=j)}{\underbrace{\sum_{j=0}^{J} f(\underline{y} | M_j) \cdot P(m=j)}_{f(\underline{y})}}$$

← like a step further above

(2) Compute for each model a score about how good the model is at predicting future observations, and choose the best-scored one.

- KLD_j (Kullback-Leibler divergence): if we knew the true law $p(\tilde{y})$ that generates the data, we could compare it with the $f(\tilde{y} | \underline{y}, M_j)$ through the KLD, for each model M_j. But not knowing $p(\tilde{y})$ we have to use a proxy.

$$LPPD_j = \sum_{i=1}^{n} \ln(f(y_i | \underline{y}, M_j))$$

- $LPPD_j$ (log pointwise predictive density): select the largest one. But here we are using data twice, as we predict by using all the vector $\underline{y}$. So the ideas are

removing data to avoid conditioning ↙

$$LPML_j = \sum_{i=1}^{n} \ln(f(y_i | \underline{y}_{-i}, M_j)) = \sum_{i=1}^{n} \ln(CPO_i | M_j)$$

(log pseudo marginal likelihood)

CPO_i: conditional predictive ordinate (of unit i)

↘ add a penalization:

$$WAIC_j = -2(LPPD_j) + 2 p_{W_j}$$

$$p_{W_j} = \sum_{i=1}^{n} \mathrm{var}_{\theta_j | \underline{y}} \left[\ln(f(y_i | \theta_j, M_j)) \right]$$

p_{W_j}: (model's effective number of parameters)

$\mathrm{var}_{\theta_j | \underline{y}}[\ldots]$: computable from the post values of the MCMC

$$CPO_i = f(y_i | \underline{y}_{(-i)}) = \binom{\text{posterior}}{\text{pred. dens.}} = \int_{\Theta} f(y_i | \theta) \cdot \pi(\theta | \underline{y}_{(-i)})\, d\theta =$$

$$= \int_{\Theta} f(y_i | \theta) \cdot \left[\frac{\prod_{l=1, l \neq i}^{n} f(y_l | \theta) \cdot \pi(\theta)}{\int_{\Theta} \prod_{l=1, l \neq i}^{n} f(y_l | \theta)\, \pi(\theta)\, d\theta} \right] d\theta$$

$$\Rightarrow \frac{1}{CPO_i} = \int_{\Theta} \frac{\boxed{\prod_{l=1, l \neq i}^{n} f(y_l | \theta) \cdot \pi(\theta) \cdot [f(y_i | \theta)}\, \frac{1}{f(y_i | \theta)}]}{\underbrace{\boxed{\int_{\Theta} \prod_{l=1}^{n} f(y_l | \theta)\, \pi(\theta)\, d\theta}}_{= \pi(\theta | \underline{y})}}\, d\theta =$$

$$= \int_{\Theta} \frac{1}{f(y_i | \theta)}\, \pi(\theta | \underline{y})\, d\theta$$

→ actually we don't need n different MCMC samples (from each posterior $\pi(\theta_j | \underline{y}, M_j)$)

MODEL CHECKING

Once we have selected some models we want to be sure that they are able to correctly fit the data. We can do it in:

(1) we could use the bayesian model as a data generation mechanism. So from the MCMC we take eg the "fitted values" and compare their distr/quantiles to the ones of the real data.

(2) we can study the outlier detection, to see if the real data are probable for the model we chose. We call them posterior predictive tail probabilities:

$$p_i = \min\{ \mathbb{P}(\tilde{y} > y_i \mid \underline{y}, M) , \mathbb{P}(\tilde{y} < y_i \mid \underline{y}, M) \} \quad i = 1, \dots, n$$

if this is too low then it is a bad news

COVARIATE SELECTION

(1) a first idea for covariate selection is to convert this task to a model selection, where we evaluate all possible models. But this scan of the entire model space becomes unfeasible when we have lots of covariates (for k regressors we get 2^k models).

Anyway the goal is to select the best set of βs to build a model in the form

$$g(\mu_i) = g(E[Y \mid \underline{x}, \underline{\beta}]) = \eta_i = \underline{x}^T \underline{\beta} = \beta_1 x_1 + \dots + \beta_k x_k$$

Idea: we start with the full covariate models, but we assume a sparse prior on the β_i to allow a regularization effect. In this way the useless covariate will get zeroed out.

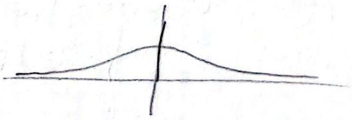
(2) Spike and slab we define $\underline{\gamma} = (\gamma_1, \dots, \gamma_k)$ the vector to describe a certain model choice, so $\gamma_i \in \{0,1\}$ and $\gamma_i = 1 \iff \beta_i \neq 0$ (we include/use the covariate i).

Then we can define a hierarchical model treating the β as a function of γ, so $\pi(\beta, \gamma) = \pi(\beta \mid \gamma) \cdot \pi(\gamma)$. We can have

$$\beta_j \mid \gamma_j \overset{\perp}{\sim} (1-\gamma_j)\delta_0 + \gamma_j \mathcal{N}(0, \sigma_j^2)$$
$$\gamma_j \mid \theta_j \overset{\perp}{\sim} \mathrm{Ber}(\theta_j)$$
$$\theta_j \overset{\perp}{\sim} \pi(\theta_j)$$

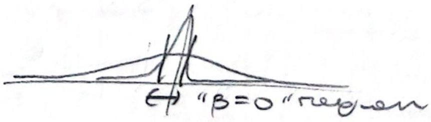
eg $\mathcal{U}([0,1])$, or $\theta_j = 1/2 \ \forall j$ to set no prior preferences

[spike & slab]



$$\beta_j \mid \sigma_j^2 \overset{\perp}{\sim} \mathcal{N}(0, \sigma_j^2)$$
$$\sigma_j^2 \mid \dots, \gamma_j \overset{\perp}{\sim} (1-\gamma_j)\delta_{\tau_j^2} + \gamma_j \delta_{c_j^2 \tau_j^2}$$
$$\gamma_j \mid \theta_j \overset{\perp}{\sim} \mathrm{Ber}(\theta_j)$$

[SSVS: stochastic search variable selection]



And we can select the best choice by looking at the posteriors:

$$\pi(\underline{\gamma}_0 \mid \underline{y}) = \frac{f(\underline{y} \mid \underline{\gamma}_0) \cdot \pi(\underline{\gamma}_0)}{f(\underline{y})} = \frac{f(\underline{y} \mid \underline{\gamma}_0) \cdot \pi(\underline{\gamma}_0)}{\sum_{\tilde{\underline{\gamma}}_0} f(\underline{y} \mid \tilde{\underline{\gamma}}_0) \cdot \pi(\tilde{\underline{\gamma}}_0)}$$

Or if we have samples (like from a MCMC chain) we can select:

- HPD (highest posterior probability): choose the set of γ which occurred the most in the simulation

$$\arg\max_{\underline{\gamma}_0} \frac{1}{m} \sum_{t=1}^{m} \mathbb{1}_{\{\underline{\gamma}^{(t)} = \underline{\gamma}_0\}}$$

- MPM (median probability model): pick all the covariates for which the posterior inclusion probability is high

$$\text{all } j : \pi(\gamma_j = 1 \mid \underline{y}) \approx$$
$$\approx \frac{1}{m} \sum_t \mathbb{1}_{(\gamma_j^{(t)} = 1)} > \frac{1}{2}$$

- HS (hard shrinkage): pick all the covariates for which $0 \notin$ to the 95% posterior CI of β_j

$$\text{all } j : 0 \notin CI_{\beta_j}^{0.95}$$

SURVIVAL ANALYSIS

Here we have data arising from studying the time until the occurrence of a certain event (an item breaks, a patient dies). So the target is studying a random time span T. We assume

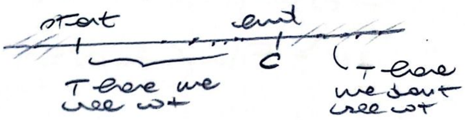
- : $T \geq 0$ a.s.
- : T (r.v.) is absolutely continuous
- · data are censored

item starts working
item stops working
T

About censoring we have

- right censoring: the occurring event may happen after a certain time limit, c, so we don't observe it but we just know that $T > c$. We get

  $y_i = T_i \wedge c_i \qquad \delta_i = \begin{cases} 1 & T_i \leq c_i \\ 0 & T_i > c_i \end{cases} \quad \forall i$



- left censoring: when we may have events which occurred before the start of the analysis. So we just know "(the) happened before" (as the opposite of right censoring) so we may get $T < t^*$.

- interval censoring: here we can only observe the data at intervals, so we may get $T \in (t^*, t^{**})$

Anyway, we focus on right censored data, and we assume to have
(1) independent censoring, i.e. $T \perp C$ (both are r.v.)
(2) non-informative censoring, i.e. the censoring distribution, so $g_C(\cdot)$, does not depend on params of the law of T, $f_T(\cdot)$

To build models we need to define two functions:

- survival function: $S(t) = 1 - F(t) = \mathbb{P}(T > t)$
- hazard function (or failure rate):

$$h(t) = \lim_{\Delta t \to 0} \mathbb{P}(t \leq T \leq t + \Delta t \mid T > t) = \frac{f(t)}{S(t)}$$

which characterizes a distribution, since $F(t) = 1 - e^{-\int_0^t h(u)\,du}$

Now we suppose to have just data (y_i, δ_i) for $i = 1, \dots, n$ (no covariates for now), and so we can compute the likelihood:

$\underbrace{(T_i, C_i)}\mid f(\cdot), g(\cdot) \overset{iid}{\sim} f(\cdot)g(\cdot)$

$\delta_i = \mathbb{1}\{T_i \leq C_i\} \qquad \Rightarrow \qquad p((y,\delta) \mid f, g) = \prod_{i=1}^n p((y_i, \delta_i) \mid \underset{\theta}{\underline{f, g}})$

$y_i = T_i \wedge C_i =$
$= T_i^{\delta_i} C_i^{1-\delta_i}$

- if $\delta_i = 1$ (so $y_i = T_i$) $\Rightarrow p(y_i, \delta_i = 1 \mid \theta) = p(T_i, T_i \leq C_i \mid \theta) =$
  $= p(y_i, C_i \geq y_i \mid \theta) = f(y_i)\cancel{(1 - G(y_i))}$
- if $\delta_i = 0$ ($y_i = C_i$) $\Rightarrow p(y_i, \delta_i = 0 \mid \theta) = p(C_i, T_i > C_i \mid \theta) =$
  $= p(y_i, T_i > y_i \mid \theta) = \cancel{g(y_i)}(1 - F(y_i))$

by non informative censoring

$$\Rightarrow p(y, \delta \mid \theta) \propto \prod_{i=1}^n (f(y_i))^{\delta_i}(1 - F(y_i))^{1-\delta_i} =$$
$$= \prod_{i=1}^n (f(y_i))^{\delta_i}(S(y_i))^{1-\delta_i} =$$
$$= \prod (h(y_i))^{\delta_i} S(y_i)$$

PARAMETRIC MODELS

We want to make inference on the law of T, so $f(\cdot)$. A way is to let it be a parametric distribution, like $f_{T|\theta}(\cdot \mid \theta)$, and make inference on θ (like $E(T \mid \theta)$, $med(T \mid \theta)$, ecc).

(1) Exponential model: $T_1, \dots, T_n \mid \theta \overset{iid}{\sim} Exp(\theta)$.

$f(t) = \theta e^{-\theta t}\, \mathbb{1}\{t \geq 0\} \quad \rightarrow \quad h(t) = \theta$
$S(t) = 1 - F(t) = e^{-\theta t}$

$n_u = \#\{\delta_i = 1\}$

$$\Rightarrow p(y, \delta \mid \theta) \propto \prod_{i=1}^n (h(y_i))^{\delta_i} S(y_i) = (h(y_i))^{n_u} \prod S(y_i)$$
$$= \theta^{n_u} \prod (e^{-\theta y_i}) = \theta^{n_u} e^{-\theta \sum y_i}$$

REGRESSION MODELS (IE WITH COVARIATES)
Now as data we get (y_i, δ_i, x_i) $\forall i$, still under right censoring and the two assumptions. Being the $T_i \geq 0$ we can model more easily this:

$$\boxed{\ln(T_i) = x_i^T\beta + \sigma\varepsilon_i \quad , \quad \varepsilon_i \overset{iid}{\sim} F_\varepsilon \text{ (a known distribution)}}$$

and we write $\boxed{T_i \overset{ind}{\sim} AFT(F_\varepsilon, \beta, \sigma \mid x_i)}$, ie accelerated failure time. The name is because

$$F_T(t) = \mathbb{P}(T \leq t) = \mathbb{P}(e^{x_i^T\beta} \underbrace{e^{\sigma\varepsilon_i}}_{W_\sigma} \leq t) = \mathbb{P}(W_\sigma \leq te^{-x_i^T\beta}) =$$

$$= F_{W_\sigma}(t \cdot e^{-x_i^T\beta}) \sim \text{a cdf evaluated in a scaled time}$$

Common choices for F_ε
(1) $\boxed{F_\varepsilon = N(0,1)}$, in this way we get $\ln(T_i) \overset{ind}{\sim} N(x_i^T\beta, \sigma^2)$ which is called log-normal AFT. For example this leads to:

$t^*: F_T(t^*) = 1/2?$ (the median survival time)

$$F_T(t^*) = \mathbb{P}(T \leq t^*) = \mathbb{P}(\ln T \leq \ln T^*) =$$
$$= \mathbb{P}\left(Z \leq \frac{\ln t^* - x^T\beta}{\sigma}\right) = \Phi\left(\frac{\ln t^* - x^T\beta}{\sigma}\right) \overset{!}{=} \frac{1}{2}$$
$$\Rightarrow \frac{\ln t^* - x^T\beta}{\sigma} = 0 \Rightarrow t^* = e^{x^T\beta}$$

which allows to study the relative median: the ratio of t^*_1 and t^*_2 for two patients of equal covariates but one

$$RM(1,2) = \frac{t^*_1}{t^*_2} = \frac{\cancel{e^{x_{11}\beta_1}} \cdots e^{x_{1k}\beta_k}}{\cancel{e^{x_{21}\beta_1}} \cdots e^{x_{2k}\beta_k}} = e^{(x_{1k} - x_{2k})\beta_k}$$

the only different covariate, suppose

to study the impact/effect of a certain covariate.

SPATIAL MODELS

In lots of contexts (environment, ecology, climate, etc) we have to work on data which are multivariate (response and covariates), temporal and spatial.

For spatial data we divide into:
- point-referenced data (geostatistical data): where $Y(s)$ the response is at location $s \in D \subset \mathbb{R}^r$ and s varies continuously over D
- areal data: now $s \in D$ with D partitioned into a finite collection of areal units, with well-defined boundaries
- point-pattern data: when D is random

General rule: there should be in data a spatial pattern, ie units closer in space will tend to be similar.

POINT-REFERENCED DATA
We have an underlying stoch process $\{Y(s) : s \in D \subset \mathbb{R}^r = \mathbb{R}^2\}$ at n locations $s_1, \dots, s_n$. We define
- a weak stationary process: $E[Y(s)] = \mu \ \forall s$; $Cov[Y(s), Y(s+h)] = C(h) \ \forall s \ \forall h$
- a strictly stationary process: $\mathcal{L}(Y(s_1), \dots, Y(s_n)) = \mathcal{L}(Y(s_1+h), \dots, Y(s_n+h))$
- semivariogram and covariance ($C = cov$): $2\gamma(h) = var[Y(s+h) - Y(s)]$; $\gamma(h) = C(0) - C(h)$
- isotropic process: $\gamma(h) = \gamma(\|h\|)$

GAUSSIAN REGRESSION MODEL

Regression as we have covariates at each location. The model works as the following

$$\boxed{Y(s) = x^T_{(s)} \beta + \underbrace{w(s)}_{} + \underbrace{\varepsilon(s)}_{}}$$

a spatial residual: a Gaussian process (like a random effect)

pure (non spatial) residual: like the usual err) of $Y(\cdot)$

$\varepsilon(s_i) \overset{iid}{\sim} N(0, \tau^2)$ (nugget)

$\{W(s)\} \sim GP(0, C(h) = \sigma^2 \rho(h; \phi))$

possible covariance models:

$C(h) = \begin{cases} \sigma^2 e^{-\phi h} & h > 0 \\ \sigma^2 + \tau^2 & h = 0 \end{cases}$ [exponential]

$C(h) = \begin{cases} \sigma^2 e^{-|\phi h|^p} \\ \sigma^2 + \tau^2 \end{cases}$ [powered exp]

$C(h) = \{ \dots$ [Matern]

So for a set of locations we get $Y(s_i)$, the $x(s_i)$ which will form a matrix X of $n \times p$ size, and the full model is:

$y | w, \beta, \tau^2 \sim N_n(\boxed{X\beta + w}, \tau^2 I_n)$

$w | \theta \sim N_n(0, \Sigma(\theta))$, where

$w = \begin{pmatrix} w(s_1) \\ \vdots \\ w(s_n) \end{pmatrix}$ $\Sigma(\theta)$ has entries $\sigma^2 \cdot \rho(\underbrace{\|s_i - s_j\|}_{h}, \theta)$

$\beta \sim N_p(\mu_\beta, \Sigma_\beta)$

$\theta = (\sigma^2, \phi, \tau^2) \sim$ uninformative!

of w) (of ρ \ of y

For predictions (we bayesian kriging) we can exploit the fact that $Y(s)$ is a stoch (Gaussian) process, so a new location so will have $Y(s_0)$ Gaussian.

$f(y_0 | y, X, x_0) = \int f(y_0 | y, \beta, \theta, x_0) \cdot f(\beta, \theta | y, X) \, d\beta \, d\theta$

$\binom{y_0}{y} \in \mathbb{R}^{n+1} | \beta, \theta \sim N_{n+1}\left(\tilde{X}\beta = \begin{pmatrix} x_0 \\ \boxed{X} \end{pmatrix} \beta, \tilde{\Sigma}_{augmented}\right)$

AREAL DATA

Here we start defining a proximity matrix $W = [w_{ij}]$, which tells us if locations i and j are close, and so $w_{ij} = 1$ (eg these areas share a boundary) or not, and so $w_{ij} = 0$.

This will allow us to treat/describe more easily the full conditional laws, as we will then get $f(y_i | y_j : j \in \partial i)$, with ∂i the set of neighbours of i.

From specifying these full conditionals actually we are able to recover the joint law of $(Y_1, \dots, Y_n)$. This procedure is called MRF (Markov Random Field).

The CAR model (conditional auto-regressive) is an example of MRF as the corresponding joint is a Gibbs distr (it exists but can be improper). We model

(the full conditionals) $y_i | y_{(-i)} \sim N\left(\frac{\sum_j w_{ij} y_j}{\sum_j w_{ij}}, \frac{\tau^2_i}{\sum_j w_{ij}} \right)$

$\Rightarrow$ (the joint distr) $p(y_1, \dots, y_n) \propto e^{[-\frac{1}{2\tau^2} y^T (\underbrace{D_w - W}_{}) y]}$

$D = \begin{pmatrix} \tau_1^2 & & 0 \\ & \ddots & \\ 0 & & \tau_n^2 \end{pmatrix}$ $W = \begin{pmatrix} w_{1\cdot} & & 0 \\ & \ddots & \\ 0 & & w_{n\cdot} \end{pmatrix}$

(this matrix actually is singular, so the joint is improper (as we can't invert it))

Solution: use $\Sigma_y^{-1} = D_w - \rho W$

- choose a ρ to make it non singular
- put a prior on ρ in $(0,1)$