3. Topics in Parametric Bayesian Statistics

Basics, Bayesian Framework, Why Bayesian, Decision Analysis, Conjugacy, Normal Model, Bayesian Network Model, Hierarchical Model, Bayesian Linear Regression, Model Selection

Sun Woo Lim

Mar 6, 2022

Basics: Conditional density, Sum rule, Product rule and Bayes rule

Conditional density $p(x|y) = \frac{p(x,y)}{p(y)}$

$e.g, p(x|y,z) = \frac{p(x,y|z)}{p(y|z)}$: conditioning on z throughout ($\because$ compound fraction with common denominator $p(z)$)

Sum rule $p(x) = \int_y p(x,y)dy$

$e.g, p(x|w) = \int_y \int_z p(x,y,z|w)dzdy$: conditioning on w throughout.

Product rule $p(x,y) = p(x)p(y|x)$ e.g, $p(x,y|z) = p(x|z)p(y|x,z)$: conditioning on z throughout.

✓ **Chain rule (general product rule)** $p(A,B,\dots,Z) = p(A|B,\dots,Z)p(B|C,\dots,Z)\dots p(Z)$: helpful in Bayesian network.

Bayes rule $p(x|y) = \frac{p(x)p(y|x)}{\int_x p(x)p(y|x)dx}$

$e.g, p(x|y,z) = \frac{p(y|x,z)p(x|z)}{p(y|z)}$: conditioning on z throughout

Some examples of practices of conditional density

1) $p(\theta,\sigma^2|y) = \frac{p(\theta,\sigma^2,y)}{\int_{\theta,\sigma^2} p(\theta,\sigma^2,y)d\theta d\sigma^2}$

2) $p(\sigma^2|\theta,y) \propto p(y,\theta,\sigma^2) = p(y|\theta,\sigma^2)p(\theta|\sigma^2)p(\sigma^2)$

Using "proportional to" $\propto$, consider the conditional density is function of which quantity.
e.g, $p(x|y,z)$ is a function of x, proportional to $g(y,z)$, an arbitrary function of y and z.

Basics: Univariate Normal Distribution

$X \in \mathbb{R} \sim N(\mu, \sigma^2)$ if $p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}$.

Useful Facts of Normal Distribution

Suppose $X \sim N(\mu, \sigma^2)$, $X_1 \sim N(\mu_1, \sigma_1^2), X_2 \sim N(\mu_2, \sigma_2^2)$. Then,

1. $\psi_X(t) = exp(\mu t + \sigma^2 t^2/2)$: MGF

2. $Y := aX + b, a \in \mathbb{R} - \{0\}, b \in \mathbb{R} \rightarrow Y \sim N(a\mu + b, a^2\sigma^2)$: Affine transformation of normal is normal

3. Let $X_1, ..., X_k \sim \ indep\ N(\mu_i, \sigma_i^2), i \in \{1, ..., k\}$. Then, $Y := \sum_{i=1}^k a_i X_i \sim N(\sum_{i=1}^k a_i \mu_i, \sum_{i=1}^k a_i^2 \sigma^2)$
: Linear combination of independent normal is normal.

4. $E(X) = median(X) = mode(X) = \mu$ and $f_X(x)$ is symmetric w.r.t. $x = \mu$.

5. Let $f_Y(y) \propto exp\{-\frac{1}{2}(ay^2 - 2by + c)\} \propto exp\{-\frac{1}{2}(ay^2 - 2by)\} \propto exp\{-\frac{1}{2}a(y - \frac{b}{a})^2\} = exp\{-\frac{1}{2}(\frac{y-b/a}{1/\sqrt{a}})^2\}$.
This means that $Y \sim N(\mu = \frac{b}{a}, \sigma^2 = \frac{1}{a})$ obtained by completing the squares.
: to specify normal parameters, only need exponential term.

Applying Fact 5 & technique of combining quadratic forms helps in getting posterior for normal mean with known variance.

6. $X_1 \perp\!\!\!\perp X_2 \leftrightarrow Cov(X_1, X_2) = 0$: In normal distribution, uncorrelated same as independence.

Basics: Multivariate Normal Distribution

$X = (X_1, ..., X_n)^T \sim MVN_n(\mu, \Sigma) \leftrightarrow \exists A \in \mathbb{R}^{n \times m}, \mu \in \mathbb{R}^n$ s.t. $X = AZ + \mu$ where $Z = (Z_1, ..., Z_m)$ with $Z_1, ..., Z_m \sim \ iid\ N(0, 1)$ and $\Sigma = AA^T$.

Density function of MVN

Suppose $X = (X_1, ..., X_n)^T \sim MVN(\mu, \Sigma)$ for $\Sigma \in \mathbb{S}^n_{++}$. Then, $f(x) = (2\pi)^{-\frac{n}{2}}(det(\Sigma))^{-\frac{1}{2}} exp[-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)]$

pf) Multivariate change of variables using 1) $X = AZ + \mu \leftrightarrow Z = A^{-1}(X - \mu)$ and 2) $AA^T = \Sigma$

$\rightarrow f_X(x) = f_Z(A^{-1}(x-\mu))|J|$ where $|J| = |det(A^{-1})| = |\frac{1}{det(A)}|$

$= (2\pi)^{\frac{-n}{2}}|\frac{1}{det(A)}|exp[-\frac{1}{2}((A^{-1}(x-\mu))^T(A^{-1}(x-\mu))]$ using $f_Z(z) = \prod_{i=1}^n\{f_{Z_i}(z_i)\} = (2\pi)^{\frac{-n}{2}} exp(-\frac{1}{2}z^Tz)$.

$= (2\pi)^{\frac{-n}{2}}|\frac{1}{det(A)}|exp[-\frac{1}{2}(x-\mu)^T(A^T)^{-1}A^{-1}(x-\mu)]$.

$= (2\pi)^{-\frac{n}{2}}(det(\Sigma))^{-\frac{1}{2}} exp[-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)]$ using $det(A) = \sqrt{det(\Sigma)}$ and $(A^{-1})^TA^{-1} = (AA^T)^{-1}$

Useful Facts of MVN

Suppose $X \sim MVN_n(\mu, \Sigma)$. Partition X, μ, Σ into $X = (X_1, X_2)^T$ $\mu = (\mu_1, \mu_2)^T$ and $\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}$. Then,

1. $\psi_X(t) = e^{\mu^T t + \frac{1}{2}t^T\Sigma t}$: MGF always exists (not requiring $\Sigma \in \mathbb{S}^n_{++}$)
2. $Y := CX + d, C \in \mathbb{R}^{r \times n} \rightarrow Y \sim MVN_r(C\mu + d, C\Sigma C^T)$: Affine transformation of MVN is MVN

3. Let $N_1, ..., N_k \sim \ indep\ MVN_n(\mu_i, \Sigma_i), i \in \{1, ..., k\}$. Then, $Y := \sum_{i=1}^k a_i N_i \sim MVN_n[\sum_{i=1}^k a_i \mu_i, \sum_{i=1}^k a_i^2 \Sigma_i]$: Linear combination of **independent** MVN is MVN.

(Used different notation $N_1, ..., N_k$ because X_1, X_2 is already defined above.)

4. $E(X) = median(X) = mode(X) = \mu$

5. Let $Y = (Y_1, ..., Y_n)^T$ has pdf $f_Y(y) \propto exp[-\frac{1}{2}(y-\mu)^T \Sigma^{-1}(y-\mu)]$. Then, $Y \sim MVN_n(\mu, \Sigma)$ by completing squares. : to specify MVN parameters, only need exponential term.

Applying Fact 5 & technique of **combining quadratic forms** helps in obtaining conditional posterior for θ in MVN model.

Technique) Combining quadratic forms
Let $x, \mu_1, \mu_2 \in \mathbb{R}^n$ and $A_1, A_2 \in \mathbb{R}^{n \times n}$ where $A_1 + A_2$ is invertible.

Then, the sum of two quadratic forms
$(x-\mu_1)^T A_1 (x-\mu_1) + (x-\mu_2)^T A_2 (x-\mu_2) = [(x-\mu_*)^T (A_1 + A_2)(x-\mu_*)] + [\mu_1^T A_1 \mu_1 + \mu_2^T A_2 \mu_2] - [\mu_*^T (A_1 + A_2)\mu_*]$
where $\mu_* := (A_1 + A_2)^{-1}(A_1 \mu_1 + A_2 \mu_2)$.

6. $X_1 \perp\!\!\!\perp X_2 \leftrightarrow \Sigma_{12} = 0$: in MVN, uncorrelated same as independence

7. Partition $X = (X_1, X_2)^T$ where $X_1 \in \mathbb{R}^r$ and $X_2 \in \mathbb{R}^{n-r}$. Then, $X_1 \sim MVN_r(\mu_1, \Sigma_{11}), X_2 \sim MVN_r(\mu_2, \Sigma_{22})$: Marginal of MVN is MVN

8. $X_1 | X_2 = x_2 \sim MVN_r(\mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2 - \mu_2), \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{22})$: conditional of MVN is MVN

Framework of Bayesian Statistics

Steps of Bayesian Data Analysis

1. **Setup full probability model**: $p(\theta, y) = p(\theta)p(y|\theta)$: joint distribution of observable and latent variables.
2. **Obtain posterior**: $p(\theta|y) = \frac{p(\theta,y)}{\int_\theta p(\theta,y)d\theta}$, or $p(\theta|y) \propto p(\theta)p(y|\theta)$: computation. "posterior" $\propto$ "prior" $\times$ "likelihood"
3. **Evaluate model fit, sensitivity analysis, Decision Analysis**: judgment.

Ways to obtain Posterior

1. Analytic calculation due to conjugacy
2. Monte Carlo (Independent MC)
3. Markov Chain Monte Carlo (MCMC) approximation (e.g, Metropolis-Hastings)
4. Deterministic approximation (e.g, Laplace method, Variational Bayes, Expectation propagation)

Prediction: "predictive distribution": for a quantity that is **observable**!

Prior predictive $p(y) = \int_\theta p(y, \theta)d\theta = \int_\theta p(\theta)p(y|\theta)d\theta$.

Posterior predictive $p(\tilde{y}|y)$
Let $\tilde{y}$: "unknown observable" that is conditionally independent given θ.
$p(\tilde{y}|y) = \int_\theta p(\tilde{y}, \theta|y)d\theta = \int_\theta p(\tilde{y}|\theta, y)p(\theta|y)d\theta = p(\tilde{y}|\theta)p(\theta|y)d\theta$.

Example of steps of Bayesian Data Analysis (BDA3 1.4)

Situation: Somebody spelled "radom":seems awkward. Let the intended word is one of "random", "radon", "radom".

1st step: Full probability Model

1. **prior**

   θ:= discrete R.V taking value 1("random"), 2("radon"), or 3("radom") representing **intended word**.

   Obtain prior by frequency of those words in Google database.

   Relative frequency of "random": $7.6e-5$, "radon" $6.1e-6$, "radom":$3.1e-7$. Normalize them to make them $\sum = 1$!

2. **Likelihood**

   Use a Google (contextual) model that infers $p('radom'|\Theta)$.
   $\rightarrow p('radom'|\theta = 1) = 0.00193$, $p('radom'|\theta = 2) = 0.000143$, $p('radom'|\theta = 3) = 0.975$.

2nd step: Posterior
$p(\theta = 1|'radom') = 0.325, p(\theta = 2|'radom') = 0.002, p(\theta = 3|'radom') = 0.673$. What is your decision?

3rd step: Model Judgment depending on domain (context)
1) Your decision? Any reason not to choose the intended word as "radom"?

2) Model fit? Additional information to incorporate in prior?

Why Bayesian statistics? 1. Justification for using prior

1. Bernstein-Von Mises Theorem: Justification in Frequentist point of view

   $||\pi(\theta|y_1, ..., y_n) - N(\theta_{MLE}, \frac{I(\hat{\theta}_{MLE})^{-1}}{n})||_{TV} \to 0$

   where total variation distance between two probability measures on σ-algebra F is $sup_{A \in F}|P(A) - Q(A)|$

   Meaning) Asymptotically, the posterior distribution of θ behaves the same as dist'n of MLE.

   Justification) prior usage (even highly subjective) justified b/c how informative prior you use, asymptotically (data size $n \to \infty$), the Bayesian inference and the frequentist inference is equivalent.

2. De Finetti's Theorem: Justification in Bayesian point of view

   Def) $Y_1, ..., Y_n$ are **exchangeable** if $p(y_{\pi_1}, y_{\pi_2}, ..., y_{\pi_n}) = p(y_1, .., y_n)$ for all permutations π of $\{1, .., n\}$.

   Def) $Y_1, Y_2, ...$ are **infinitely exchangable** if $Y_1, ..., Y_n$ are exchangable $\forall n \in \mathbb{N}$

   Meaning) $Y_1, ..., Y_n$ exchangable if the label (in the subscript) has no information on the outcome.

   De Finetti) $Y_1, Y_2, ...$ is ∞-exchangable $\leftrightarrow \forall n \in \mathbb{N}$, $p(y_1, .., y_n) = \int_\theta [\prod_{i=1}^n p(y_i|\theta)] p(\theta) d\theta$ for some latent RV θ.

   Meaning) (Probably) correlated exchangable observations $Y_1, .., Y_n$ are conditionally indep. given latent variable θ.

   Just assuming exchangability (weaker assumption than $Y_1 \perp\!\!\!\perp Y_2 \perp\!\!\!\perp ... \perp\!\!\!\perp Y_n$), there must exist

   1) Parameter θ with distribution p, 2) Likelihood $p(y|\theta)$

   Justification) prior usage justified simply because it makes sense!

Why Bayesian statistics? 2. Advantages of Bayesian Statistics

1. Simple computational framework for estimation, prediction, and model selection using sum rule, product rule.
2. Wider spectrum of model (then frequentist) by choosing the amount of information in prior.
   highly informative prior, weakly informative prior and nonformative prior.
   Next page: example of advantage for using highly informative prior in case of small sample size.
3. Bayesian network model (e.g, hierarchical model) that contains parameter (R.V) dependence structures
4. Good at predictions for future data and missing data
5. Boosts common sense understanding of parameters
   (especially in interval estimation, Bayesian interval is easier to understand than confidence interval)

Advantage of using Informative Prior : Adapted from section 2.7 of BDA3

Setting) A US county i of population n_i. X_i: number of kidney cancer deaths in county i in 1980's.
Θ_i : underlying death rate of county i I want to estimate.

Frequentist Estimate (MLE) $\hat{\Theta} = \frac{X}{n}$: reliable for large county, unreliable for small county!

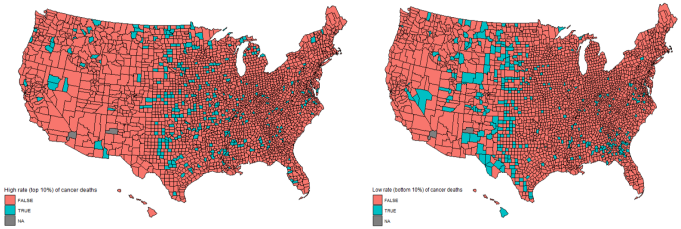


Figure: Frequentist method, left: top 10% counties, right: bottom 10% counties in blue

Sorted all counties by $\hat{\theta}$ and obtained top 10% counties and bottom 10% counties.
Q) Why counties in the middle part of US: both high and low in cancer rate?
A) Central US: low population. e.g, $n_i = 1000$ & $x_i = 0 \rightarrow \hat{\theta} = 0$: unrealistically $\downarrow$, $x_i = 1 : \hat{\theta} = 1e-3$: unrealistically $\uparrow$.

Bayesian Approach :

1. "Likelihood" : $X|\Theta = \theta \sim Bin(n, \theta)$
2. "Prior" : $\Theta \sim Beta(\alpha, \beta)$: Use method of moments estimate for α and β.
   This is empirical Bayesian method using data to choose hyperparameter.
3. "Posterior" : $\Theta|X = x \sim Beta(x + \alpha, n - x + \beta)$.

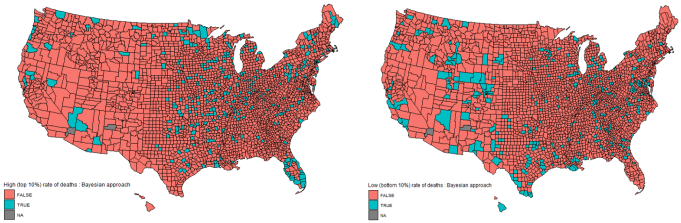


Figure: Bayesian method, left: top 10% counties, right: bottom 10% counties in blue

Sorted all counties by **posterior mean** and obtained top 10% counties and bottom 10% counties : more realistic!
Code: https://github.com/YonseiESC/ESC-21SPRING/blob/main/Week1/Rcode_week1.R
I slightly modified the code provided in Fall2020 UC Berkeley STAT157 class.

Bayesian Decision Analysis using the inference results in decision making

Steps of Bayesian Decision Analysis

1. Consider the space of "Decision (=actions)" d, "space of outcomes (=consequences)" x.
   Note, x is R.V (discrete/continuous).
2. Distribution of x for(=given) each d: $p(x|d)$. Be careful, decision is not random so, no such thing as $p(d)$!
3. "Utility function" $U(x)$: mapping of **outcome** to $\mathbb{R}$
4. Compute $E[U(x)|d] = \int_x U(x)p(x|d)dx$ (or, $\sum_x U(x)p(x|d)$) and obtain $argmax_d E[U(x)|d]$ as the decision!

Bayes Estimate for Bayesian point estimation

Point estimate θ by statistic $\delta(y)$.

Let $L(\theta, \delta(y))$ represent the **loss (function)** when we estimate θ by $\delta(y)$.

Bayes estimate is a decision function δ s.t. $\delta(y)^* = argmin_\delta E[L(\theta, \delta(y))|Y = y] = argmin_\delta \int_\theta L(\theta, \delta(y))p(\theta|y)d\theta$.

Bayes estimate differs by the choice of loss function.

1. For $L(\theta, \delta(y)) = (\theta - \delta(y))^2$: MSE, Bayes estimate of θ $\delta(y) = E[\theta|y] = \int_\theta \theta p(\theta|y)d\theta$: posterior mean
2. For $L(\theta, \delta(y)) = |\theta - \delta(y)|$, $\delta(y) = median(\theta|y)$: posterior median.
3. For $L(\theta, \delta(y)) = I(\delta(y) \neq \theta) = 1 - I(\delta(y) = \theta)$, $\delta(y)^* = argmin_\delta \int_\theta (1 - I(\delta(y) = \theta))p(\theta|y)d\theta = argmin_\delta[1 - \int_\theta I(\delta(y) = \theta) \cdot p(\theta|y)d\theta] = argmax_\delta p(\delta(y)|y) = argmax_\theta p(\theta|y)$: posterior mode

Example of Bayesian Decision Analysis:
https://github.com/avehtari/BDA_course_Aalto/blob/master/slides/slides_ch9.pdf

Setting: Helen in forest. Finds a 14cm footprint which she thinks is either by a big dog or wolf (no other consideration). Thinks she would be definitely attacked confronting a wolf, while no problem with a dog.

Bayesian Decision Analysis

1. $d = \{$Go pick mushroom, Stay home$\}$. $x = \{$0: Nothing happens, 1: half dead, 2: get mushroom$\}$
2. Google search about dist'n of paws of dog/wolf says $\frac{Likelihood\ of\ wolf}{Likelihood\ of\ dog} = \frac{92}{8}$.
   Prior: Helen thinks in her living area, wild dogs are 99 times more than wolf. $p(wolf) = 0.01, p(dog) = 0.99$.
   Posterior: $p(wolf|data) = 0.1, p(dog|data) = 0.9$
   In conclusion, $p(x = 0|d = "home") = 1$, $p(x = 1|d = "go") = 0.1$, $p(x = 2|d = "go") = 0.9$.
   Derive other three probabilities $p(x = 1|"home") = 0$, $p(x = 2|"home") = 0$, $p(x = 0|d = "go") = 0$ easily.
3. $U(0) = 0$, $U(1) = -1000$, $U(2) = 1$.
4. $E[U(x)|d = "home"] = 0, E[U(x)|d = "go"] = -1000 \cdot 0.1 + 1 \cdot 0.9 = -99.1$. $d^* = argmax_d E[U(x)|d] = "home"$

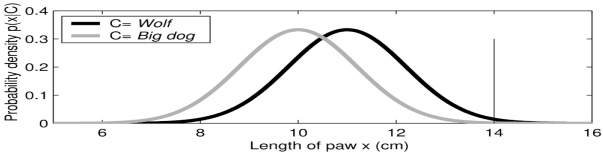


Figure: The distributions about length of paw prints of wolf or dog

Conjugacy and exponential family

Def) When the posterior follows the same parametric form as the prior, the prior has conjugacy with the likelihood.

Thm) Distribution belonging to exponential family has natural conjugate prior.

Note) This case, the prior is **'often'** in exponential family (or a form that the pdf/pmf can be analytically calculated).
Note) Term 'conjugacy' often (mis)used as only the case that conjugate prior dist'n is analytically calculated (posterior too).

DGP in exponential family pdf in canonical form $p(y_i|\eta) = h(y_i)exp(\eta^T T(y_i) - A(\eta))$. One or more parameter(s).

"log normalizer" $A(\eta) = log \int h(y_i)exp(\eta^T T(y_i))dy_i \because \int p(y_i|\eta)dy_i = 1$. Watch out) w.r.t η, $A(\eta)$ is not a normalizer!

Likelihood $p(y_1, ..., y_n|\theta) = \{\prod_{i=1}^n h(y_i)\} \cdot exp[\eta^T(\sum_{i=1}^n T(y_i)) - n \cdot A(\eta)]$, where $(y_1, ..., y_n)$: iid data.

$T(y) := \sum_{i=1}^n T(y_i)$ is sufficient statistic for η.

Conjugate prior $p(\eta|\tau, n_0) = H(\tau, n_0)exp[\eta^T\tau - n_0 \cdot A(\eta)] \propto exp[\eta^T\tau - n_0 \cdot A(\eta)]$: mimics the likelihood form.

τ and n_0: hyperparameters. τ called prior guess and n_0 called prior sample size.

Can write the prior as $p(\eta)$ too unless it is a hierarchical model. $H(\tau, n_0)$: normalizing constant for prior.

posterior
$p(\eta|y_1, ..., y_n) \propto p(\eta) \cdot p(y_1, ..., y_n|\eta) = H(\tau, n_0)exp[\eta^T\tau - n_0 \cdot A(\eta)] \times \{\prod_{i=1}^n h(y_i)\} \cdot exp[\eta^T(\sum_{i=1}^n T(y_i)) - n \cdot A(\eta)]$
$\propto exp[\eta^T\{\tau + \sum_{i=1}^n T(y_i)\} - (n_0 + n)\eta]$: retains the form of the prior = conjugacy!

One Parameter Conjugacy model with examples

Poisson Model

Let $y_1, ..., y_n|\lambda \sim \ iid\ Pois(\lambda)$. Let $\mathcal{D} := (y_1, ..., y_n)$.

Then, **likelihood** $p(\mathcal{D}|\lambda) = \prod_{i=1}^{n}[\frac{\lambda^{y_i}e^{-\lambda}}{y!}] = \prod_{i=1}^{n}\frac{1}{y!} \cdot e^{log\lambda \cdot \sum y_i - n\lambda}$. $T(y) := \sum y_i$ is sufficient statistic for λ.

conjugate prior $p(\lambda) \propto e^{log\lambda \cdot a - n_0\lambda}$ that mimics the likelihood. This is $\Gamma(a, n_0)$.

posterior $p(\lambda|\mathcal{D}) \propto e^{log\lambda \cdot (a + \sum y_i) - (n_0 + n)\lambda} \rightarrow \lambda|y \sim \Gamma(a + \sum y_i, n_0 + n)$

Binomial Model

Likelihood $Y|\theta \sim B(n, \theta)$: $p(y|\theta) = \binom{n}{y}\ \theta^y(1-\theta)^{n-y} = \binom{n}{y}\ exp\{log(\frac{\theta}{1-\theta}) \cdot y + nlog(1-\theta)\}$

conjugate prior $p(\theta) \propto exp\{log(\frac{\theta}{1-\theta}) \cdot a + b \cdot log(1-\theta)\} = (\frac{\theta}{1-\theta})^a \cdot (1-\theta)^b = \theta^a(1-\theta)^{b-a} = \theta^{\alpha-1}(1-\theta)^{\beta-1}$
for $\alpha := a + 1$ and $\beta := b - a + 1$ that mimics the likelihood. This is Beta(α, β).

posterior $p(\theta|y) \propto \theta^{y+\alpha-1}(1-\theta)^{n-y+\beta-1}$. This is Beta($\alpha + y, \beta + n - y$).

$E[\theta|y] = \frac{a+b}{a+b+n}\frac{a}{a+b} + \frac{n}{a+b+n}\frac{y}{n} = \frac{a+b}{a+b+n}\times$ prior mean $+\frac{n}{a+b+n}\times$ MLE : weighted avg between prior avg and data mean!

Univariate Normal Model with one parameter known

Setting Let $y_1, ..., y_n | \theta, \sigma^2 \sim \ iid \ N(\theta, \sigma^2)$. Let $\mathcal{D} := (y_1, ..., y_n)$.

Likelihood $p(\mathcal{D}|\theta, \sigma^2) = \prod_{i=1}^n p(y_i|\theta, \sigma^2) = \prod_{i=1}^n \{\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}(\frac{y_i - \mu}{\sigma})^2}\} \propto (\sigma^2)^{-\frac{n}{2}} exp[-\frac{1}{2}\{\frac{\sum y_i^2}{\sigma^2} - 2\frac{\theta}{\sigma^2}\sum y_i + n\frac{\theta^2}{\sigma^2}\}]$.
The part "proportional to" can change w.r.t. what parameter is of interest.

Situation 1) Known variance, inference for mean (ch 5.2 in Hoff (2009))

$p(\theta|\mathcal{D}, \sigma^2) \propto p(\theta|\sigma^2) \times p(\mathcal{D}|\theta, \sigma^2)$: conditioning on σ^2 (constant) throughout!

Assuming a conjugate prior $\theta \sim N(\mu_0, \tau_0^2)$ (here, $p(\theta|\sigma^2) = p(\theta)$),

$p(\theta|\mathcal{D}, \sigma^2) \propto exp\{-\frac{1}{2\tau_0^2}(\theta - \mu_0)^2\} \times exp\{-\frac{1}{2\sigma^2}\sum(y_i - \theta)^2\} \propto exp(a\theta^2 - 2b\theta)$ where $a = \frac{1}{\tau_0^2} + \frac{n}{\sigma^2}$, $b = \frac{\mu_0}{\tau_0^2} + \frac{\sum y_i}{\sigma^2}$.

Using normal distribution fact, $\theta|\mathcal{D}, \sigma^2 \sim N(\mu_n = \frac{b}{a} = \frac{\frac{\mu_0}{\tau_0^2} + \frac{\sum y_i}{\sigma^2}}{\frac{1}{\tau_0^2} + \frac{n}{\sigma^2}}, \tau_n^2 = \frac{1}{a} = \frac{1}{\frac{1}{\tau_0^2} + \frac{n}{\sigma^2}})$.

Intuitional Understanding of posterior parameters

Let $\tilde{\sigma}^2 := \frac{1}{\sigma^2}, \tilde{\tau}_0^2 := \frac{1}{\tau_0^2}, \tilde{\tau}_n^2 := \frac{1}{\tau_n^2}$ meaning the precision (= inverse variance). Then,

1. $\mu_n = \frac{\tilde{\tau}_0^2}{\tilde{\tau}_0^2 + n\tilde{\sigma}^2} \cdot \mu_0 + \frac{n\tilde{\sigma}^2}{\tilde{\tau}_0^2 + n\tilde{\sigma}^2} \cdot \bar{y}$: weighted average of prior mean and data average.
2. $\frac{1}{\tau_n^2} = \frac{1}{\tau_0^2} + \frac{n}{\sigma^2} \leftrightarrow \tilde{\tau}_n^2 = \tilde{\sigma}^2 + n \cdot \tilde{\tau}_0^2$: posterior precision = prior prec + n $\cdot$ sampling prec = prior prec + data prec.

Situation 2) Known mean, inference for variance (ch 2.6 in BDA3)

$p(\mathcal{D}|\sigma^2, \beta) \propto (\sigma^2)^{-\frac{n}{2}} exp[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\theta)^2] = (\sigma^2)^{-\frac{n}{2}} exp(-\frac{n}{2\sigma^2}v)$ where $v := \frac{1}{n}\sum(y_i-\theta)^2$ is sufficient stat for σ^2.

Assuming a conjugate prior $p(\sigma^2) \propto (\sigma^2)^{-(\frac{\nu_0}{2}+1)} e^{-\frac{\nu_0\sigma_0^2}{2\sigma^2}}$: scaled inverse-$\chi^2(\nu_0, \sigma_0^2)$
(here, $p(\sigma^2|\theta) = p(\sigma^2)$ too!)

$p(\sigma^2|\mathcal{D}, \theta) \propto p(\sigma^2)p(\mathcal{D}|\sigma^2, \theta) \propto (\sigma^2)^{-(\frac{\nu_0}{2}+1)} e^{-\frac{\nu_0\sigma_0^2}{2\sigma^2}} \times (\sigma^2)^{-n/2} e^{-\frac{nv}{2\sigma^2}} \propto (\sigma^2)^{-(\frac{\nu_0+n}{2}+1)} exp\{-\frac{\nu_0\sigma_0^2+nv}{2\sigma^2}\}$
which is scaled inverse-$\chi^2(\nu_0+n, \frac{\nu_0\sigma_0^2+nv}{\nu_0+n})$.

Intuitional Understanding of posterior parameters

1. Degree of freedom $= \nu_0 + n =$ prior information + sample size
2. scale $= \frac{\nu_0\sigma_0^2+nv}{\nu_0+n}$ is degrees-of-freedom-weighted avg of prior and data scales

Multiparameter Full Conjugacy and Semiconjugacy model with examples

In multiparameter setting, the conjugacy concept is more complicated.

Classified into 1) full-conjugacy model and 2) semi-conjugacy model.

1. Full conjugacy is when the joint prior has same parametric form with the posterior.

   ✓ Requirement: Dependency of parameters $e.g, p(\theta_1, \theta_2) = p(\theta_1) \cdot p(\theta_2|\theta_1) \neq p(\theta_1)p(\theta_2)$ and 1) $p(\theta_1)$ and $p(\theta_1|\mathcal{D})$, 2) $p(\theta_2|\theta_1)$ and $p(\theta_2|\theta_1, \mathcal{D})$ needs to have parametric form.

   ✓ Independent Monte Carlo method to sample from first, $p(\theta_1|\mathcal{D})$, then $(\theta_2|\theta_1, \mathcal{D})$ using $p(\theta_1, \theta_2|y) = p(\theta_1|\mathcal{D}) \cdot p(\theta_2|\theta_1, \mathcal{D})$.

   ✓ Independent because I have independent sample $\Theta_{i=1}^{n} := ((\theta_1(1), \theta_2(1)), \ldots, (\theta_1(n), \theta_2(n)))$ although each θ_1, θ_2 is dependent.

2. Semi-conjugacy (= conditional conjugacy) is when prior of each parameter and "conditional" posterior has same parametric form distribution.

   ✓ This does not fit the original definition of conjugacy, while full conjugacy does.

   ✓ Requires independence of parameters $e.g, p(\theta, \sigma^2) = p(\sigma^2) \cdot p(\theta)$

   ✓ Gibbs sampling method to sequentially sample from full conditional posterior.

Situation 3) Univariate Normal Model with unknown mean and variance (excellent explantion in ch 5.3 in Hoff)

Likelihood $p(\mathcal{D}|\theta,\sigma^2) = \prod_{i=1}^n p(y_i|\theta,\sigma^2) = \prod_{i=1}^n \{\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}(\frac{y_i-\mu}{\sigma})^2}\}$

Choice 1. Full conjugate Prior

$p(\theta,\sigma^2) = p(\sigma^2)p(\theta|\sigma^2)$

1. Conditioning on σ^2, normal prior on θ is conjugate to the likelihood (Easy. Invested earlier as situation1)
   In situation 1) assumed $\theta \sim N(\mu_0, \tau_0^2)$. Here, consider a particular case of
   $\tau_0^2 = \sigma^2/\kappa_0 \rightarrow p(\theta|\sigma^2) = pnorm(\theta, mean = \mu_0, var = \frac{\sigma^2}{\kappa_0})$. κ_0 interpreted as prior sample size.
   $p(\theta|y,\sigma^2) = dnorm(\mu_n, \frac{\sigma^2}{\kappa_n})$ where $\kappa_n := \kappa_0 + n$ and $\mu_n = \frac{\kappa_0\mu_0 + n\bar{y}}{\kappa_n}$ just by plugging $\frac{\sigma^2}{\kappa_0}$ in τ_0^2.

2. Need a selection of $p(\sigma^2)$ which has same parametric form as
   $p(\sigma^2|\mathcal{D}) \propto p(\sigma^2)p(\mathcal{D}|\sigma^2) = p(\sigma^2) \times \int_\theta p(\mathcal{D}|\theta,\sigma^2)p(\theta|\sigma^2)d\theta$ (way harder...)
   $p(\sigma^2) = dinvgamma(\frac{\nu_0}{2}, \frac{\nu_0}{2}\sigma_0^2)$ satisfies the condition with
   $p(\sigma^2|\mathcal{D}) = dinvgamma(\frac{\nu_n}{2}, \frac{\nu_n}{2}\sigma_n^2)$ where $\nu_n := \nu_0 + n$ and $\sigma_n^2 := \frac{1}{\nu_n}[\nu_0\sigma_0^2 + (n-1)s^2 + \frac{\kappa_0 n}{\kappa_n}(\bar{y} - \mu_0)^2]$

Independent MC using $p(\theta,\sigma^2|\mathcal{D}) = p(\sigma^2|\mathcal{D}) \times p(\theta|\sigma^2,\mathcal{D})$ just as $p(\theta,\sigma^2) = p(\sigma^2) \times p(\theta|\sigma^2)$.

Note) In each sample, have to sample σ^2 first and then sample θ using sampled σ^2

Choice 2. Semi conjugate prior

$p(\theta, \sigma^2) = p(\theta)p(\sigma^2)$.

$p(\theta)$ has to have same distributional form as $p(\theta|y, \sigma^2)$ and so does $p(\sigma^2)$ with $p(\sigma^2|y, \theta)$.

The choice is $p(\theta) = dnorm(\theta, \mu_0, \tau_0^2)$ and $p(\sigma^2) = dinvgamma(\frac{\nu_0}{2}, \frac{\nu_0\sigma_0^2}{2})$ (Note the difference btwn previous page !).

1. $p(\theta|\mathcal{D}, \sigma^2) = dnorm(\theta, \mu_n = \frac{\frac{\mu_0}{\tau_0^2} + \frac{\sum y_i}{\sigma^2}}{\frac{1}{\tau_0^2} + \frac{n}{\sigma^2}}, \tau_n^2 = \frac{1}{\frac{1}{\tau_0^2} + \frac{n}{\sigma^2}})$: already dealt in situation 1.
2. Letting $\tilde{\sigma}^2 := \frac{1}{\sigma^2}$, due to independence of θ and $\tilde{\sigma}^2$, $p(\tilde{\sigma}^2|\mathcal{D}, \theta) \propto p(\mathcal{D}|\theta, \tilde{\sigma}^2) \times p(\tilde{\sigma}^2)$.
   $\propto (\tilde{\sigma}^2)^{n/2} exp\{-\tilde{\sigma}^2 \sum (y_i - \theta)^2/2\} \times (\tilde{\sigma}^2)^{\frac{\nu_0}{2}-1} exp\{\frac{-\tilde{\sigma}^2\nu_0\sigma_0^2}{2}\}$
   $= (\tilde{\sigma}^2)^{\frac{\nu_0+n}{2}-1} exp\{-\tilde{\sigma}^2 \cdot [\frac{\nu_0\sigma_0^2 + \sum(y_i-\theta)^2}{2}]\}$

   which means that $\sigma^2|\mathcal{D}, \theta = dinvgamma(\sigma^2, \frac{\nu_n}{2}, \frac{\nu_n\sigma_n^2(\theta)}{2})$ where $\nu_n := \nu + n$ and $\sigma_n^2(\theta) := \frac{1}{\nu_n}[\nu_0\sigma_0^2 + \sum(y_i - \theta)^2]$

Gibbs Sampling

1. I have current sample (θ_t, σ_t^2)
2. Sample $\theta_{t+1} \sim p(\theta|\sigma_t^2, \mathcal{D})$. Current sample $(\theta_{t+1}, \sigma_t^2)$
3. Sample $\sigma_{t+1}^2 \sim p(\sigma^2|\theta_{t+1}, \mathcal{D})$. Current sample $(\theta_{t+1}, \sigma_{t+1}^2)$.

Multiparameter Models: Multivariate Gaussian model : ch 7.2, 7.3 in Hoff (2009)

Let $y = (y_1, ..., y_p)^T \sim MVN_p(\theta, \Sigma)$ for $\theta \in \mathbb{R}^p, \Sigma \in \mathbb{S}^p_{++}$. Then, pdf (not likelihood! y is single vector obs)
$p(y|\theta, \Sigma) = (2\pi)^{-\frac{p}{2}}(det(\Sigma))^{-\frac{1}{2}} exp[-\frac{1}{2}(y-\theta)^T\Sigma^{-1}(y-\theta)]$.
Let $y_1, ..., y_n|\theta, \Sigma \sim\ iid\ MVN_p(\theta, \Sigma)$. Let $\mathcal{D} := (y_1, ..., y_n)$ be the collection of vector observations of size n.

Semiconjugate prior distribution

$\theta \sim MVN(\mu_0, \Lambda_0)$ and $\Sigma \sim$Inverse-Wishart(ν_0, S_0^{-1}) has semi-conjugacy with the likelihood.

1. Semiconjugate prior for the mean θ

$p(\theta) \propto exp\{-\frac{1}{2}(\theta-\mu_0)^T\Lambda_0^{-1}(\theta-\mu_0)\} = exp\{-\frac{1}{2}\theta^T\Lambda^{-1}\theta + \theta^T\Lambda_0^{-1}\mu_0 - \frac{1}{2}\mu_0^T\Lambda_0^{-1}\mu_0\}$
$\propto exp\{-\frac{1}{2}\theta^T\Lambda^{-1}\theta + \theta^T\Lambda_0^{-1}\mu_0\} = exp\{-\frac{1}{2}\theta^T A_0\theta + \theta^T b_0\}$ where $A_0 := \Lambda_0^{-1}$ and $b_0 := \Lambda_0^{-1}\mu_0$.

Likelihood $p(\mathcal{D}|\theta, \Sigma) = \prod_{i=1}^n [(2\pi)^{-p/2}|\Sigma|^{-1/2} exp\{-\frac{1}{2}(y_i-\theta)^T\Sigma^{-1}(y_i-\theta)\}]$
$\propto exp\{-\frac{1}{2}\theta^T A_1\theta + \theta^T b_1\}$ where $A_1 := n\Sigma^{-1}$ and $b_1 := n\Sigma^{-1}\bar{y}$ also for combining the sum of square.

"Conditional posterior of θ" $p(\theta|\mathcal{D}, \Sigma) \propto exp\{-\frac{1}{2}\theta^T A_0\theta + \theta^T b_0\} \times exp\{-\frac{1}{2}\theta^T A_1\theta + \theta^T b_1\}$
$= exp\{-\frac{1}{2}\theta^T A_n\theta + \theta^T b_n$ where $A_n := A_0 + A_1 = \Lambda_0^{-1} + n\Sigma^{-1}$ and $b_n := b_0 + b_1 = \Lambda_0^{-1}\mu_0 + n\Sigma^{-1}\bar{y}$.
Thus, $\theta|\mathcal{D}, \Sigma \sim MVN(\mu_n = A_n^{-1}b_n = (\Lambda_0^{-1} + n\Sigma^{-1})^{-1}(\Lambda_0^{-1}\mu_0 + n\Sigma^{-1}\bar{y}), \Lambda_n = A_n^{-1} = (\Lambda_0^{-1} + n\Sigma^{-1})^{-1})$
Used multivariate version of Fact 5 + combining sum of squares (multivariate version of 'situation 1').

2. Semiconjugate prior for the covariance matrix Σ

"Prior" $\Sigma \sim$ Inv-Wishart(ν_0, S_0^{-1}):

$p(\Sigma) = [2^{\nu_0 p/2}\pi^{p(p-1)/4}|S_0|^{-\frac{nu_0}{2}}\prod_{i=1}^{p}\Gamma(\frac{\nu_0+1-j}{2})]^{-1} \times |\Sigma|^{-\frac{\nu_0+p+1}{2}}exp\{-tr(S_0\Sigma^{-1})/2\}$

$\propto |\Sigma|^{-\frac{\nu_0+p+1}{2}}exp\{-tr(S_0\Sigma^{-1})/2\}.$

"Likelihood" $p(\mathcal{D}|\theta,\Sigma) = (2\pi)^{-\frac{np}{2}}|\Sigma|^{-\frac{n}{2}}exp\{-\frac{1}{2}\sum_{i=1}^{n}(y_i-\theta)^T\Sigma^{-1}(y_i-\theta)\}$

$\propto |\Sigma|^{-\frac{n}{2}}exp\{-\frac{1}{2}\sum_{i=1}^{n}(y_i-\theta)^T\Sigma^{-1}(y_i-\theta)\} = |\Sigma|^{-\frac{n}{2}}exp\{-tr(S_\theta\Sigma^{-1})/2\}$ where $S_\theta := \sum_{i=1}^{n}(y_i-\theta)(y_i-\theta)^T$.

"Conditional posterior of Σ" $p(\Sigma|\mathcal{D},\theta) \propto p(\Sigma) \times p(\mathcal{D}|\theta,\Sigma)$

$\propto (|\Sigma|^{-\frac{\nu_0+p+1}{2}}exp\{-tr(S_0\Sigma^{-1})/2\}) \times (|\Sigma|^{-\frac{n}{2}}exp\{-tr(S_\theta\Sigma^{-1})/2\})$

$= |\Sigma|^{-\frac{\nu_0+n+p+1}{2}}exp\{-tr([S_0+S_\theta]\Sigma^{-1})/2\}.$

Thus, $\Sigma|\mathcal{D},\theta \sim$ Inverse-Wishart$(\nu_n := \nu_0+n, S_n^{-1} := [S_0+S_\theta]^{-1})$.

Gibbs Sampling

1. I have current sample (θ_t, Σ_t)
2. Sample $\theta_{t+1} \sim p(\theta|\Sigma_t,\mathcal{D})$. Current sample (θ_{t+1}, Σ_t)
3. Sample $\Sigma_{t+1} \sim p(\Sigma|\theta_{t+1},\mathcal{D})$. Current sample $(\theta_{t+1}, \Sigma_{t+1})$.

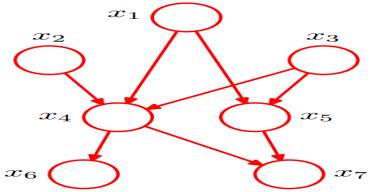
✓ For how to sample from inverse-Wishart distribution, refer to ch7.3 in Hoff(2009).

Bayesian Network Basics

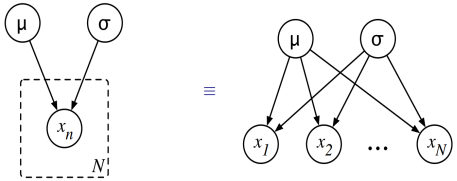
Bayesian network is a method of representing **dependence structures** between **random variables** using **graphical notation**. No directed cycle and not an undirected graph. It is a directed acyclic graph (DAG).

- **Conditional independence** Random variables X is indepenent from Y given Z if $p(x|y,z) = p(x|y)$.
  Bayesian network helps understand **conditional independence / dependence** structure.
- Bayesian network helps factor out the joint pdf using **chain rule**.
- **Node** represents RV's and **arc** represents conditional dependence.
- Write in simple way as $p(x) = \prod_k p(x_k|pa_k)$ where pa_k means parent of R.V x_k.
- **Plate notation**: e.g, write $\{X_i\}_{i=1}^n$ instead of $X_1, ..., X_n$: simple notation.

Exercise) $p(\vec{\beta}, k, \xi_1, ..., \xi_k, \sigma) = p(\vec{\beta}|\xi_1, ..., \xi_k, k, \sigma)p(\xi_1, ..., \xi_k|k)p(k)p(\sigma)$ (DiMatteo et al., 2001). Represent this in graph.



(a) $p(x_1, ..., x_7) = p(x_1)p(x_2)p(x_3)p(x_4|x_1, x_2, x_3)p(x_5|x_1, x_3)p(x_6|x_4)p(x_7|x_4, x_5)$



(b) Plate notation in left makes it more simple.
$p(x_1, ..., x_N, \mu, \sigma = p(\mu)p(\sigma) \prod_{i=1}^n p(x_i|\mu, \sigma)$

Three main Bayesian Network Types: Explain here with 3 RV's for simplicity. Easily generalized with more RV's.

1. **Common parent = common cause** $X \leftarrow Z \rightarrow Y$: when Z is observed, X and Y, which are affected from Z is conditionally independent.
   However, X is dependent on Y because $\exists Z$ that has information about X and Y.
   e.g) Z: temperature, X: # people in the beach, Y: # ice creams sold
   Write: $X \perp Y | Z$.
2. **Cascade (= causal trail or evidential trail)** $X \rightarrow Z \rightarrow Y$: when Z is observed, X and Y are conditionally independent. Z has information that affects Y.
   However, X is dependent on Y because X affects Y through Z.
   Write: $X \perp Y | Z$
   e.g) X: Covid19 confirmed cases, Z: degree of Covid19 regulations, Y: # people protesting against government.
3. **V-structure (=common effect)** $X \rightarrow Z \leftarrow Y$: $X \perp Y$ when Z is unobserved but dependent when Z is observed.
   Write: $\sim (X \perp Y) | Z$
   e.g) Z: minutes late in class, X: traffic jam index, Y: amount of malfunctioning of the alarm clock.

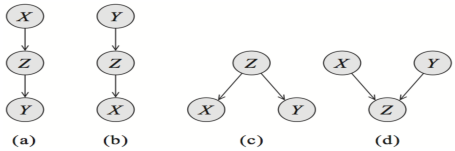


Figure: Cascading (a,b), common parent (c), and V structure (d)

Hierarchical Bayes

Understanding of Hierarchical Bayes by contrast with Empirical Bayes

Setting) **Likelihood** $p(\mathcal{D}|\phi) = \prod_{i=1}^{n} p(y_i|\phi)$, $\phi|\psi \sim p(\phi|\psi)$. Here, $p(\phi|\psi)$ is **prior** of ϕ. ψ is a **hyperparameter**.

- Empirical Bayes: Plug in point estimate of hypermarameters (e.g, Method of Moments, MLE)
- Hierarchical Bayes: Assume **hyperprior** $p(\psi)$

Justification for hierarchical Bayes by De Finetti's theorem

Let $y_1, y_2, ..$ be ∞- exchangeable. Then, for actually observed data $\mathcal{D} = (y_1, y_2, .., y_n)$, by De Finetti's thm, there exist

1) parameter (= latent variable) ϕ with a distribution p.

2) $p(\mathcal{D}|\phi) = \prod_{i=1}^{n} p(y_i|\phi)$ using conditional independence of $Y_1, Y_2, ..., Y_n$ given ϕ

Further, let $\phi = (\phi_1, \phi_2, ...)$ is ∞-exchangeable. Then, for $\phi_1, ..., \phi_n$, by De Finetti's thm, there exist

1) hyperparameter ψ with distribution $p(\psi)$

2) $p(\phi_1, ..., \phi_n|\phi) = \prod_{i=1}^{n} p(\phi_i|\psi)$ using conditional independence of $\phi_1, ..., \phi_n$ given ψ

Can go deeper... So, applying De Finetti's thm several times will justify the hierarchical Bayes!

Complaint of exchangeability assumption) What if $\exists$ info (e.g, explanatory variables) that helps distinguish Y_i's by indices?

Answer) Agglomerate covariate info in exchangeability! i.e, assume $(y_1, y_2, ...)$ and $(\phi_1, \phi_2...)$ ∞-exchangeable given x_j

Advantages of hierarchical Bayes: Appropriate for modeling hierarchical (: complicated) model structure inappropriate with simple nonhierarchical models. Especially helpful in **group comparison** (not the only example though).

Example of Group Comparison using Hierarchical Bayes

Hierarchical Data for group comparison

- $D = \{\vec{Y}_1, \vec{Y}_2, \ldots, \vec{Y}_m\}$: data of m groups
- $Y_j = \{Y_{1,j}, \ldots, Y_{n_j,j}\}, j \in \{1, ..., m\}, n_j \geq 1, \forall j$: and each group has ≥ 1 observation (= member).
- $Y_{i,j}$ means i^{th} observation in group j. $i \in \{1, ..., n_j\}$ and $j \in \{1, ..., m\}$.
- e.g) $Y_{i,j}$: exam grades of i^{th} student (=observation) in school (= group) j.
- Each group j has latent variable ϕ_j that explains $Y_{i,j}$. ϕ_j may be different w.r.t j. e.g, different avg score
- Hyperparameters $\psi \in \mathbb{R}^n$ has **shared** latent information about all of ϕ_j. Normally, $n << m$. e.g, variability of score

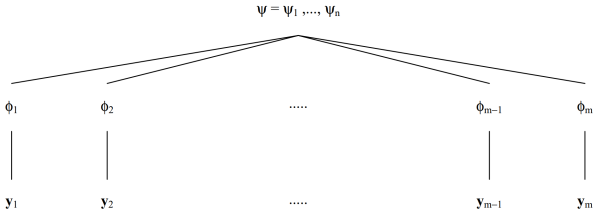


Figure: Hierarchical data for group comparison. hyperprior $\psi \in \mathbb{R}^n$, prior $\phi_j \in \mathbb{R}^q$, data $y_j \in \mathbb{R}^{n_j}$, $j \in \{1, ..., m\}$.

Hierarchical Normal Model for comparing multiple groups : ANOVA in Bayesian perspective (ch 8.3 in Hoff, 2009)

Hierarchical Model

1. Data generating process $y_{1,j}, ..., y_{n_j,j} | \phi_j \sim \ iid \ p(y|\phi_j)$: within-group variability
2. $\phi_1, ..., \phi_m | \psi \sim \ iid \ p(\phi|\psi)$: between group variability
3. Hyperprior $\psi \sim p(\psi)$

Hierarchical Normal Model

1. Data generating process $y_{1,j}, ..., y_{n_j,j} | \phi_j \sim \ iid \ N(\theta_j, \sigma^2)$ where $\phi_j := \{\theta_j, \sigma^2\}$: within-group variability
2. $\phi_1, ..., \phi_m | \psi \sim \ iid \ N(\mu, \tau^2)$ where $\psi := \{\mu, \tau^2\}$: between group variability
3. Hyperpriors $\sigma^2 \sim InvGamma(\frac{\nu_0}{2}, \frac{\nu_0 \sigma_0^2}{2}), \tau^2 \sim InvGamma(\frac{\eta_0}{2}, \frac{\eta_0 \tau_0^2}{2})$, and $\mu \sim N(\mu_0, \gamma_0^2)$.

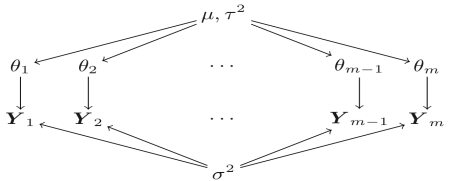


Figure: Network (graph) representation of hierarchical normal model

Inference of hierarchical normal model

For the spatial efficiency of the slide, let $\Theta := (\theta_1, ..., \theta_m)$.

$p(\Theta, \mu, \tau^2, \theta^2|\mathcal{D}) \propto p(\Theta, \mu, \tau^2, \sigma^2, \mathcal{D}) = p(\mu)p(\tau^2)p(\sigma^2)\prod_{j=1}^{m} p(\theta_j|\mu, \tau^2)\prod_{j=1}^{m}\prod_{i=1}^{n_j} p(y_{ij}|\theta_j, \sigma^2)$
: network representation helpful in factorizing joint full (joint) probability into marginals and conditionals (chain rule!)

Full conditional posteriors of each parameter (including hyperparameters) from which we perform Gibbs Sampler

1. $p(\mu|\Theta, \tau^2, \sigma^2, \mathcal{D}) = p(\mu|\Theta, \tau^2) \propto p(\mu)\prod_{j=1}^{m} p(\theta_j|\mu, \tau^2)$ using μ, τ^2, Θ is V structure, $\mu, \Theta, \mathcal{D}$: cascading, $\mu \perp \sigma^2$.
2. $p(\tau^2|\Theta, \mu, \sigma^2, \mathcal{D}) = p(\tau^2|\Theta, \mu) \propto p(\tau^2)\prod_{j=1}^{m} p(\theta_j|\mu, \tau^2)$ using μ, τ^2, Θ is V structure. $\tau^2, \Theta, \mathcal{D}$: cascading, $\tau^2 \perp \sigma^2$.
3. $p(\theta_j|\mu, \tau^2, \sigma^2, \mathcal{D}) \propto p(\theta_j|\mu, \tau^2)\prod_{i=1}^{n_j} p(y_{ij}|\theta_j, \sigma^2)$ using $\theta_j \perp \Theta_{(-j)}|\mu, \tau^2, \sigma^2, \mathcal{D}$
4. $p(\sigma^2|\Theta, \mu, \tau^2, \mathcal{D}) = p(\sigma^2|\Theta, \mathcal{D}) \propto p(\sigma^2)\prod_{j=1}^{m}\prod_{i=1}^{n_j} p(y_{ij}|\theta_j, \sigma^2)$ using $\sigma^2 \perp \{\mu, \tau\}|\{\Theta, \mathcal{D}\}$

1. $\mu|\Theta, \tau^2 \sim N(\frac{m\bar{\theta}/\tau^2 + \mu_0/\gamma_0^2}{m/\tau^2 + 1/\gamma_0^2}, \frac{1}{m/\tau^2 + 1/\gamma_0^2})$
2. $\tau^2|\Theta, \mu \sim inv\Gamma(\frac{\eta_0 + m}{2}, \frac{\eta_0\tau_0^2 + \sum_{j=1}^{m}(\theta_j - \mu)^2}{2})$
3. $\theta_j|\mu, \tau^2, \sigma^2, \mathcal{D} \sim N(\frac{n_j\bar{y}_j/\sigma^2 + 1/\tau^2}{n_j/\sigma^2 + 1/\tau^2}, \frac{1}{n_j/\sigma^2 + 1/\tau^2})$
4. $\sigma^2|\Theta, \mathcal{D} \sim inv\Gamma(\frac{1}{2}[\nu_0 + \sum_j n_j], \frac{1}{2}[\nu_0\sigma_0^2 + \sum_j(\sum_i(y_{ij} - \theta)^2)])$

Bayesian Linear Regression

Data: (y, X) where $y \in \mathbb{R}^n$ and $X \in \mathbb{R}^{n \times p}$. $y|X, \beta, \sigma^2 \sim MVN_n(X\beta, \sigma^2 I)$, where $\beta \in \mathbb{R}^p, \sigma^2 \in \mathbb{R}_{++}$

Then, $p(y|X, \beta, \sigma^2) = |2\pi\sigma^2 I|^{-\frac{1}{2}} exp(-\frac{1}{2}(y - X\beta)^T(\sigma^2 I)^{-1}(y - X\beta)) \propto exp(-\frac{1}{2\sigma^2} SSR(\beta))$

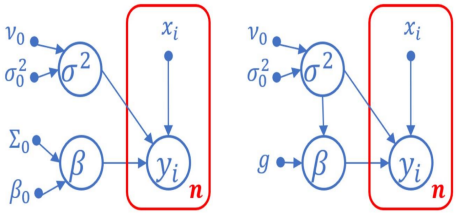
where "Sum Square of Residuals" $SSR(\beta) := ||y - X\beta||_2^2|| = y^T y - 2\beta^T X^T y + \beta^T X^T X \beta$.

Note that in linear model, X is assumed to be given (not considering distribution of X).

In Bayesian full probability model, need joint prior $p(\beta, \sigma^2)$.

Two main choices of prior

1. $p(\beta, \sigma^2) = p(\beta)p(\sigma^2)$: semi-conjugate prior (prior and "conditional" posterior has same parametric form distribution)
2. $p(\beta, \sigma^2) = p(\sigma^2)p(\beta|\sigma^2)$: full-conjugate prior



Bayesian Linear Regression 1) Semi-conjugate prior

1) Obtaining full conditional posterior of β given σ^2, y

Prior: $\beta \sim MVN_p(\beta_0, \Sigma_0)$ is semi-conjugate with the likelihood using the MVN facts mentioned earlier.

Expanding terms with β, $\propto exp[-\frac{1}{2}(\beta^T\Sigma_0^{-1}\beta - 2\beta^T\Sigma_0^{-1}\beta_0 + \beta_0\Sigma_0^{-1}\beta_0)] \propto exp[-\frac{1}{2}(\beta^T\Sigma_0^{-1}\beta - 2\beta^T\Sigma_0^{-1}\beta_0]$.

Conditional posterior of β given σ^2: $p(\beta|y, X, \sigma^2) \propto p(y|X, \beta, \sigma^2) \cdot p(\beta)$.

Expanding terms with β, $\propto exp[-\frac{1}{2}(-2\beta^T X^T y/\sigma^2 + \beta^T X^T X\beta/\sigma^2) - \frac{1}{2}(-2\beta^T\Sigma_0^{-1}\beta_0 + \beta^T\Sigma_0^{-1}\beta)] =$
$exp[-\frac{1}{2}\beta^T(\Sigma_0^{-1} + X^T X/\sigma^2)\beta + \beta^T(\Sigma_0^{-1}\beta_0 + X^T y/\sigma^2)]$.

Solve $Cov(\beta|y, X, \sigma^2)^{-1} = (\Sigma^{-1} + X^T X/\sigma^2)$ & $Cov(\beta|y, X, \sigma^2)^{-1}E(\beta|y, X, \sigma^2) = \Sigma_0^{-1}\beta_0 + X^T y/\sigma^2$

$\rightarrow \beta|y, X, \sigma^2 \sim MVN[(\Sigma^{-1} + X^T X/\sigma^2)^{-1}(\Sigma_0^{-1}\beta_0 + X^T y/\sigma^2), (\Sigma^{-1} + X^T X/\sigma^2)^{-1}]$

Intuition

1. When $|\Sigma_0|^{-1}$: "prior precision magnitude" small $\rightarrow E[\beta|y, X, \sigma^2] \approx (X^T X)^{-1}X^T y$
2. When σ^2 large, $E[\beta|y, X, \sigma^2] \approx \beta_0$

2) Obtaining full conditional posterior of $\gamma := \frac{1}{\sigma^2}$ given β, y

Prior: $\frac{1}{\sigma^2} \sim Gamma(\frac{\nu_0}{2}, \frac{\nu_0\sigma_0^2}{2})$: $\frac{(\frac{\nu_0\sigma_0^2}{2})^{\frac{\nu_0}{2}}}{\Gamma(\frac{\nu_0}{2})}\gamma^{\frac{\nu_0}{2}-1}e^{-\frac{\nu_0\sigma_0^2}{2}\cdot\gamma} \propto \gamma^{\frac{\nu_0}{2}-1}e^{-\frac{\nu_0\sigma_0^2}{2}\cdot\gamma}$ is semi-conjugate with the likelihood.

Conditional posterior of γ given β, y: $p(\gamma|y, X, \beta) \propto p(y|X, \beta, \sigma^2) \cdot p(\gamma)$. By expanding terms with γ,

$\propto \gamma^{\frac{n}{2}} \cdot e^{\gamma\cdot\frac{SSR(\beta)}{2}} \cdot \gamma^{\frac{\nu_0}{2}-1}e^{-\frac{\nu_0\sigma_0^2}{2}\cdot\gamma} \propto \gamma^{\frac{\nu_0+n}{2}-1} \cdot e^{-\gamma\cdot[\frac{\nu_0\sigma_0^2+SSR(\beta)}{2}]} = pgamma(\frac{\nu_0+n}{2}, \frac{\nu_0\sigma_0^2+SSR(\beta)}{2})$

Gibbs Sampler to obtain joint posterior $\beta, \sigma^2|y, X$

Since I have full conditional distributions of β and σ^2, use Gibbs Sampler as follows:

1. I have current sample (β_t, σ_t^2).
2. Update β: Sample β_{t+1} using $\beta|y, X, \sigma^2 \sim MVN[(\Sigma^{-1} + X^TX/\sigma^2)^{-1}(\Sigma_0^{-1}\beta_0 + X^Ty/\sigma^2),\ (\Sigma^{-1} + X^TX/\sigma^2)^{-1}]$
3. Update σ^2: Sample σ_{t+1}^2 using $\sigma^2|y, X, \beta \sim InvGamma(\frac{\nu_0+n}{2}, \frac{\nu_0\sigma_0^2+SSR(\beta)}{2})$

Bayesian Linear Regression 2) Full-conjugate prior

Several choices of full-conjugate prior: $p(\sigma^2, \beta) = p(\sigma^2) \cdot p(\beta|\sigma^2)$

1. Strongly informative prior
2. Unit-information prior (Kass & Wasserman, 1995): $\beta \sim MVN(\beta_0, \Sigma_0)$
   s.t. $\beta_0 = \beta_{MLE}, \Sigma_0^{-1} = \frac{X^T X}{n\sigma^2}$ is a weakly informative prior that has information amount of 1 observation.
3. g-prior (Zellner, 1986) : makes parameter estimation invariant to change in the scaling of predictors (X)
   i.e) $\tilde{X} := XH, \tilde{\beta} := H^{-1}\beta$ for some scaling constant matrix H. Then, the posterior of β and $H\tilde{\beta}$ needs to be equal!

   Thm) This condition is satisfied when $\beta \sim MVN(\beta_0 = 0, \Sigma_0 = c(X^T X)^{-1}), c > 0$.

   pf) $\beta \sim MVN(0, c(X^T X)^{-1}), \tilde{X} := XH, \tilde{\beta} := H^{-1}\beta$.

   Then, $\tilde{\beta} \sim MVN(0, cH^{-1}(X^T X)^{-1}H) = MVN(0, c \cdot (\tilde{X}^T \tilde{X})^{-1})$.

   Also, $y|X, \beta, \sigma^2 \sim MVN(X\beta, \sigma^2 I)$ and $y|X, \tilde{\beta}, \sigma^2 \sim MVN(\tilde{X}\tilde{\beta}, \sigma^2 I)$.

   The prior and likelihood are compatible under scaling $\rightarrow$ posterior is compatible under scaling.

   Amount of information chosen by hyperparameter g: look at the form of $p(\beta|X, y, \sigma^2)$!
   1. $g \rightarrow 0$: $p(\beta|X, y, \sigma^2) \approx p(\beta)$
   2. $g = 1$: prior has equal information as likelihood
   3. $g = n$: unit information prior
   4. $g \rightarrow \infty$: $p(\beta|X, y, \sigma^2) \approx p(\beta_{MLE})$

Zellner's g-prior

$\gamma \sim Gamma(\frac{\nu_0}{2}, \frac{\nu_0 \sigma_0^2}{2})$: same as before.

$\beta|X, \sigma^2 \sim MVN_p(0, g\sigma^2(X^TX)^{-1})$

Marginal posterior of γ given y: $p(\sigma^2|y, X) \propto p(\sigma^2) \cdot p(y|X, \sigma^2)$ where $p(y|X, \sigma^2) = \int_\beta p(y|X, \beta, \sigma^2)p(\beta|X, \sigma^2)d\beta$.

$\propto [\gamma^{\frac{\nu_0}{2}-1} e^{-\gamma \frac{\nu_0 \sigma^2}{2}}] \times [\gamma^{\frac{n}{2}} e^{-\gamma \cdot SSR_g}] = \gamma^{\frac{\nu_0+n}{2}-1} e^{-\gamma \cdot \frac{\nu_0 \sigma_0^2 + SSR_g}{2}}$ where $SSR_g := y^T[I - \frac{g}{g+1}X(X^TX)^{-1}X^T]y$.

Conditional posterior of β given σ^2, y: $\beta|y, X, \sigma^2 \sim MVN(\frac{g}{g+1}(X^TX)^{-1}X^Ty,\ \frac{g}{g+1}\sigma^2(X^TX)^{-1})$

Independent Monte Carlo to obtain joint posterior $p(\beta, \sigma^2|y, X)$

1. Sample σ^2 using $\sigma^2|y, X \sim InvGamma(\frac{\nu_0+n}{2}, \frac{\nu_0 \sigma_0^2 + SSR_g}{2})$
2. Sample β using $\beta|y, X, \sigma^2 \sim MVN(\frac{g}{g+1}(X^TX)^{-1}X^Ty, \frac{g}{g+1}\sigma^2(X^TX)^{-1})$

Be Careful) independent because I have independent sample $\theta_1, \theta_2, ...$ where $\theta := (\beta, \sigma^2)$. But, β and σ^2 are dependent!

So, independent between sample, dependent within sample.

Main principle this slide: $p(\sigma^2, \beta|y, X) = p(\sigma^2|y, X) \times p(\beta|\sigma^2, y, X)$ by conditioning y, X throughout!

Bayesian model selection

Setting

- Data $\mathcal{D}$ and parameter (either continuous or discrete) θ. There are two candidate models M_1, M_2.
- A discrete R.V α: only takes two values 0 (: M_1) or 1 (: M_2). The exact value of α does not matter.

Posterior for the model $p(\alpha|\mathcal{D})$ is the goal of the analysis!

Steps to obtain $p(\alpha|\mathcal{D})$

1. **prior for the model** $p(\alpha)$. Common choice: equal probability.
2. **Prior of** θ **conditional on** α: $\theta|\alpha = 0 \sim p(\theta|\alpha = 0)$, $\theta|\alpha = 1 \sim p(\theta|\alpha = 1)$.
3. **Likelihood of the data**: $p(\mathcal{D}|\theta, \alpha)$. Keep noticing that α can take either 0 or 1.
4. **Calculate Marginal likelihood (= evidence)**: $p(\mathcal{D}|\alpha) = \int_\theta p(\mathcal{D}|\theta, \alpha)p(\theta|\alpha)d\theta$
5. **Posterior of the model**: $p(\alpha|\mathcal{D}) \propto p(\mathcal{D}|\alpha)p(\alpha)$.

The posterior odds ratio tells us relative appropriateness of two models.
$\frac{Pr(\alpha=0|\mathcal{D})}{Pr(\alpha=1|\mathcal{D})} = \frac{Pr(\alpha=0)}{Pr(\alpha=1)} \times \frac{p(\mathcal{D}|\alpha=0)}{p(\mathcal{D}|\alpha=1)}$: **Posterior odds = Prior odds** $\times$ **Bayes factor**

Bayesian Occam's Razor
Complex model has wide support $\rightarrow$ marginal probability of certain event is small.
If $\mathcal{D}$ is in the support of simpler model, marginal likelihood will prefer simpler one!

Example) Chapter 28 in MacKay, Information theory, inference, and learning algorithms
Data: $\mathcal{D} = (-1, 3, 7, 11)$

- Model1 (Linear): $a_n = \beta n + (\alpha - \beta)$ indicating that $\alpha = a_1, n + 1 = a_n + \beta$. Assume $\alpha, \beta \in \mathbb{Z}$
- Model2 (cubic): $a_1 = a, a_{n+1} = ba_n^3 + ca_n^2 + d$. Assume $a \in \mathbb{Z}, b, c, d \in \mathbb{Q}$
- Both models has perfect fit. Which is more plausible?

Prior for the model choice : Assign equal probabilities to two models.

Prior for the parameters for each model

- For M_1, $\alpha, \beta \sim$ $iid\ Unif\{-50, -49, ..., 49, 50\}$
- For M_2, $a \sim Unif\{-50, -49, ..., 49, 50\}$ and b, c, d having form of $\frac{X}{Y}$ where $X \sim Unif\{-50, -49, ..., 49, 50\} \perp Y \sim Unif\{0, 1, ..., 49, 50\}$.

Evidence = marginal likelihood

1. $p(\mathcal{D}|M_1) = \sum_\alpha \sum_\beta [p(\mathcal{D}|\alpha, \beta, M_1) \cdot p(\alpha, \beta|M_1)]$
   Since $p(\mathcal{D}|\alpha, \beta, M_1) = 1$ if $\alpha = -1, \beta = 4$ and $p(\mathcal{D}|\alpha, \beta, M_1) = 0$ for all other (α, β) combinations,
   $p(\mathcal{D}|M_1) = p(\alpha = -1, \beta = 4|M_1) = \frac{1}{101} \frac{1}{101} \approx 1e-4$.
2. $p(\mathcal{D}|M_2) = \sum_a \sum_b \sum_c \sum_d [p(\mathcal{D}|a, b, c, d, M_2) \cdot p(a, b, c, d|M_2)]$
   Since $p(\mathcal{D}|a, b, c, d, M_2) = 1$ if $(a, b, c, d) = (-1, -\frac{1}{11}, \frac{9}{11}, \frac{23}{11})$, and $p(\mathcal{D}|a, b, c, d, M_2) = 0$ for all other (a, b, c, d),
   $p(\mathcal{D}|M_1) = p(a = -1, b = -\frac{1}{11}, c = \frac{9}{11}, d = \frac{23}{11}|M_2) = (\frac{1}{101}) \times (4 \cdot \frac{1}{101} \cdot \frac{1}{50}) \times (4 \cdot \frac{1}{101} \cdot \frac{1}{50}) \times (2 \cdot \frac{1}{101} \cdot \frac{1}{50}) \approx 2.5e-12$.

$\rightarrow$ Overwhelming evidence to choose M_1 over M_2.

Bayesian linear model selection

Question) Ideas to introduce a parameter that represent different linear models?

Answer) Use indicator R.vector $z \in \mathbb{R}^p$ s.t. $y_i = \sum_{j=1}^p z_j b_j x_{ij} + \epsilon_i$. $z_j = I(b_j \neq 0)$: whether j^{th} predictor is selected.

e.g, for $p = 4$, $E(Y|x, b, z = (1, 0, 1, 0)) = b_1 x_1 + b_3 x_3$ and $E(Y|x, b, z = (1, 1, 0, 0)) = b_1 x_1 + b_2 x_2$.

Posterior odds $\frac{p(z_a|y,X)}{p(z_b|y,X)} = \frac{p(z_a)}{p(z_b)} \times \frac{p(y|X,z_a)}{p(y|X,z_b)}$: (y, X) is $\mathcal{D}$ in earlier notation! Common normalizing constant is reduced!

Evidence = Marginal Likelihood $p(y|X, z)$ calculation

$p(y|X, z) = \int_{\sigma^2} \int_\beta p(y, \beta, \sigma^2|X, z) d\beta d\sigma^2 = \int_{\sigma^2} \int_\beta p(y|X, z, \sigma^2, \beta) p(\beta|X, z, \sigma^2) p(\sigma^2) d\beta d\sigma^2$

(X, z) is conditioned throughout. Think of $p(\sigma^2)$ as $p(\sigma^2|X, z)$. The prior for $\gamma := \frac{1}{\sigma^2}$ has no (X, z). Addressed later!

$= \int_{\sigma^2} [\int_\beta p(y|X, z, \sigma^2, \beta) p(\beta|X, z, \sigma^2) d\beta] p(\sigma^2) d\sigma^2 = \int_{\sigma^2} p(y|X, z, \sigma^2) p(\sigma^2) d\sigma^2$ by integrating w.r.t. β first.

With g-prior for β, the inner integral $p(y|X, z, \sigma^2) = \int_\beta p(y|X, z, \sigma^2, \beta) p(\beta|X, z, \sigma^2) d\beta$ is easily calculated.

Let $p_z :=$ # nonzero entries for given z. $X_z \in \mathbb{R}^{n \times p_z}$ for variables for which $z_j = 1$, $\beta_z \in \mathbb{R}^{p_z}$: entries of β with $z_j = 1$.

$p(\beta|X, z, \sigma^2) = p(\beta_z|X_z, \sigma^2) = dMVN_{p_z}(\beta_z, mn = 0, \Sigma = g\sigma^2[X_z^T X_z]^{-1})$: by slight modification of g-prior in p33.

$\therefore$ Inner integral $p(y|X, z, \sigma^2) = (2\pi)^{-\frac{n}{2}} (1 + g)^{-\frac{p_z}{2}} \cdot [\gamma^{n/2} e^{-\gamma SSR_g^z/2}]$ where $SSR_g^z := y^T [I - \frac{g}{g+1} X_z (X_z^T X_z)^{-1} X_z] y$.

Now, time to calculate outer integral $\int_{\sigma^2} p(y|X,z,\sigma^2) \times p(\sigma^2) d\sigma^2$.

Inverse Gamma prior for σ^2: Letting $\gamma := \frac{1}{\sigma^2}$, $\gamma \sim \Gamma(\frac{\nu_0}{2}, \frac{\nu_0 \sigma_0^2}{2}) \leftrightarrow p(\gamma) = \frac{(\nu_0\sigma_0^2/2)^{\nu_0/2}}{\Gamma(\nu_0/2)} \cdot [\gamma^{\nu_0/2-1} e^{-\gamma \nu_0 \sigma_0^2/2}]$.

$p(y|X,z,\sigma^2) \times p(\sigma^2) = p(y|X,z,\gamma) \times p(\gamma) = (2\pi)^{-\frac{n}{2}} (1+g)^{-\frac{p_z}{2}} \cdot [\gamma^{n/2} e^{-\gamma SSR_g^z/2}] \times \frac{(\nu_0\sigma_0^2/2)^{\nu_0/2}}{\Gamma(\nu_0/2)} \cdot [\gamma^{\nu_0/2-1} e^{-\gamma\nu_0\sigma_0^2/2}]$.

$= (2\pi)^{-\frac{n}{2}} (1+g)^{-\frac{p_z}{2}} \cdot \frac{(\nu_0\sigma_0^2/2)^{\nu_0/2}}{\Gamma(\nu_0/2)} \cdot \gamma^{\frac{\nu_0+n}{2}-1} e^{-\gamma \cdot \frac{\nu_0\sigma_0^2 + SSR_g^z}{2}}$.

$\gamma^{\frac{\nu_0+n}{2}-1} e^{-\gamma \cdot \frac{\nu_0\sigma_0^2+SSR_g^z}{2}} = \frac{\Gamma(\frac{\nu_0+n}{2})}{(\frac{\nu_0\sigma_0^2+SSR_g^z}{2})^{\frac{\nu_0+n}{2}-1}} \times dgamma(\gamma, \frac{\nu_0+n}{2}, \frac{\nu_0\sigma_0^2+SSR_g^z}{2}) \because$ gamma density integrates to 1.

$\therefore p(y|X,z) = \int_{\sigma^2} p(y|X,z,\sigma^2) \times p(\sigma^2) d\sigma^2 = \pi^{-n/2} \frac{\Gamma(\frac{\nu_0+n}{2})}{\Gamma(\frac{\nu_0}{2})} (1+g)^{-p_z/2} \frac{(\nu_0\sigma_0^2)^{\nu_0/2}}{(\nu_0\sigma_0^2+SSR_g^z)^{\frac{\nu_0+n}{2}}}$

Intuitive Understanding of Bayes Factor

The **Bayes Factor** is a balance between **Goodness Of Fit (GOF)** and **model variance**.

e.g) Setting $g = n$ and $\nu_0 = 1$ for both models and use s^2_{model}: the estimated residual variance for each model,

$\frac{p(y|X,z_a)}{p(y|X,z_b)} = (1+n)^{\frac{p_{z_b} - p_{z_a}}{2}} (\frac{s^2_{z_a}}{s^2_{z_b}}) \times (\frac{s^2_{z_b} + SSR_g^{z_b}}{s^2_{z_a} + SSR_g^{z_a}})^{\frac{n+1}{2}}$

1. **Model Variance**: $\uparrow p_{z_a} \rightarrow \downarrow \frac{p(y|X,z_a)}{p(y|X,z_b)}$: favors model B.
2. **Goodness of Fit**: $\uparrow SSR_g^{z_b} \rightarrow \uparrow \frac{p(y|X,z_a)}{p(y|X,z_b)}$: favors model A.

References

https://github.com/YonseiESC/ESC-21SPRING/blob/main/Week1/Rcode_week1.R : my example code
https://www.stat.cmu.edu/ larry/=stat705/Lecture24.pdf : Bernstein Von-mises thm
https://www.youtube.com/watch?v=ZF9NxOA3Qeolist=PLFHD4aOUZFp3Fx3rfRkBR0XjP1OCcrYXPindex=12 : De Finetti thm
Gelman, A. (2013). Bayesian Data Analysis, 3rd: Boca Raton. Texts in Statistical Science. : various parts
Hoff, P. D. (2009). A first course in Bayesian statistical methods (Vol. 580). New York: Springer. : various parts
https://www.stat.cmu.edu/ larry/=sml/Bayes.pdf
https://jwmi.github.io/BMS/chapter3-expfams-and-conjugacy.pdf : conjugacy
https://people.eecs.berkeley.edu/ jordan/courses/260-spring10/other-readings/chapter9.pdf : conjugacy
http://www.cs.columbia.edu/ blei/fogm/2016F/doc/exponential_families.pdf : conjugacy
https://github.com/avehtari/BDA_course_Aalto/blob/master/slides/slides_ch9.pdf : Bayesian decision analysis
https://stats.stackexchange.com/questions/90938/what-are-the-definitions-of-semi-conjugate-and-conditional-conjugate-priors/357587 : semi/full conjugacy
https://ermongroup.github.io/cs228-notes/representation/directed/ : network model
http://mlss.tuebingen.mpg.de/2017/speaker_slides/Zoubin3.pdf : network model
https://faculty.cc.gatech.edu/ hic/CS7616/pdf/lecture6.pdf : network model
https://www.cs.cmu.edu/ mgormley/courses/10601-s17/slides/lecture23-bayesnet2.pdf : network model examples
https://www.youtube.com/watch?v=ZF9NxOA3Qeot=337s : Hierarchical Bayes
https://www.youtube.com/watch?v=nNQdvXfW73E : Hierarchical Bayes
MacKay, D. J., Mac Kay, D. J. (2003). Information theory, inference and learning algorithms. Cambridge university press. : Bayesian Model Selection