STAT 220 NOTES - BAYESIAN DATA ANALYSIS

GREG TAM

Contents

1. September 4th, 2014

2. September 8th, 2014

Using Bayes' Theorem does not make you a Bayesian. Anybody can use Bayes' Theorem in the proper setting. Using Bayes' Theorem when you shouldn't be using it makes you a Bayesian.

In probability, we have $(\Omega, \mathcal{F}, \mathbb{P})$ where

(1) The outcome is an element $\omega \in \Omega$
(2) An event is a set of ω in Ω
(3) Probability is defined on $\mathcal{F}$
(4) $\forall A \in \mathcal{F}$, $\mathbb{P}(A)$ is defined with

$$\begin{cases} \mathbb{P}(\emptyset) = 0 \\ \mathbb{P}(\Omega) = 1 \\ \mathbb{P}(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} \mathbb{P}(A_i) \qquad \text{for } A_i \cap A_j = \emptyset, i \neq j \end{cases}$$

Reasons for needing $\mathcal{F}$:

(1) Technical
(2) Practical: "Information Structure" and "Resolution"

Suppose we have a table we want to measure and have a ruler that only measures by inches (by integer value). If the table is 22.5 inches, then we cannot measure that, so it is not a measurable event.

Theorem 1 (Bayes' Theorem)**.**

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(B|A)\,\mathbb{P}(A)}{\mathbb{P}(B|A)\,\mathbb{P}(A) + \mathbb{P}(B|A^c)\,\mathbb{P}(A^c)} = \frac{\mathbb{P}(B \cap A)}{\mathbb{P}(B)}$$

2.0.1. *Random Variables.* Suppose we have random variables X, Y, Z, $(X : \Omega \to \mathbb{R})$. For discrete random variables, we have

$$\mathbb{P}(X = x|Y = y) = \frac{\mathbb{P}(Y = y|X = x)\,\mathbb{P}(X = x)}{\sum_{x'} \mathbb{P}(Y = y|X = x')\,\mathbb{P}(X = x')}$$

In the continuous case, we have

$$f_{X|Y}(x|y) = \frac{f_{XY}(x,y)}{f_Y(y)} = \frac{f_{Y|X}(y|x) f_X(x)}{\int f_{Y|X}(y|x') f_X(x')\,\mathrm{d}x'}$$

As a Bayesian, we should do something "inappropriate". Assume a statistical model

$$X|\theta \sim f_\theta(X)$$

where X is the observation, θ is the parameter, and f is the parametric model. We want to infer θ.

Example 1. *If $X \sim \mathrm{Bin}(20, \theta)$, the "estimate" of θ is*

$$\hat{\theta} = \frac{X}{20}$$

How do we quantify "uncertainty"? We want to infer θ by giving a prior $\mathbb{P}_0(\theta)$ on θ. Then

$$\mathbb{P}(\theta|X) = \frac{f(x|\theta)\mathbb{P}_0(\theta)}{\int f(x|\theta')\mathbb{P}_0(\theta')\,\mathrm{d}\theta'}$$

Example 2.

$$
\begin{aligned}
X &\sim \mathrm{Bin}(20, \theta) \\
f(x|\theta) &= \binom{20}{x} \theta^x (1-\theta)^{20-x} \\
\mathbb{P}_0\left(\theta\right) &= \mathrm{Beta}(\theta; a, b) \equiv \frac{\Gamma(a,b)}{\Gamma(a)\Gamma(b)} \theta^{a-1}(1-\theta)^{b-1}
\end{aligned}
$$

Then we have (for $a > 0, b > 0$)

$$
\begin{aligned}
\mathbb{P}(\theta|x) &\propto f(x|\theta)\mathbb{P}_0\left(\theta\right) \\
&\propto \theta^{x+a-1}(1-\theta)^{20-x+b-1} \\
&= \mathrm{Beta}(\theta; x+a, 20-x+b)
\end{aligned}
$$

- Easy use of prior information if you have some
- Quantifying uncertainty using probability theory is the only known coherent framework
- Conditional analysis is automatically done
- Optimality (Decision-theoretical support)
- Coherence (Axiom-based arguments)
- Operational advantage

Example 3. *Suppose we toss a coin toss C and a result X. If $C = 1$, $X \sim \mathcal{N}(\theta, 1)$. If $C = 2$, $X \sim \mathcal{N}(\theta, 100)$. Suppose we have an observation $(C = 1, X = 21.5)$. What is the confidence interval for θ?*

3. September 11th, 2014

Bayesian Inference:

- Probabilistic modeling of data and knowledge
- Probabilistic statement (description) about uncertainties
- Using probability as a direct measure of uncertainty
  - Probability Interval: $\mathbb{P}(\theta \in I_X | X)$
  - Confidence Interval: $\mathbb{P}(I(X) \ni \theta | \theta)$

Frequentists give a method of checking a confidence interval, but no constructive method of creating one.

3.1. **Confidence Intervals.**

3.1.1. *Binomial Distribution.*

$$X \sim \mathrm{Bin}(n, \theta)$$

In the classic case, we have $\hat{\theta} = \frac{X}{n}$. To get the confidence interval, we use the Normal approximation to get

$$\hat{\theta} \pm 1.96\sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{n}}$$

In the Bayesian case, we give a prior distribution

$$\mathbb{P}_0\left(\theta\right) \propto \theta^{a-1}(1-\theta)^{b-1}$$

which is the density for a Beta(a, b) variable. This is the conjugate prior. The posterior distribution is

$$
\begin{aligned}
\mathbb{P}(\theta|X) &\propto \mathbb{P}_0\left(\theta\right)\mathbb{P}(X|\theta) \\
&\propto \theta^{x+a-1}(1-\theta)^{n-x+b-1} \\
&= \mathrm{Beta}(x+a, n-x+b)
\end{aligned}
$$

which has density

$$f(\theta|x) = \frac{\Gamma(n+a+b)}{\Gamma(x+a)\Gamma(n-x+b)} \theta^{n+a-1}(1-\theta)^{n-x+b-1}$$

To get the 95% probability interval, we simply take off the 2.5% tails off each side of the distribution. In `R`, we can do this simply with `qbeta(0.025,x+a,n-x+b)`

3.1.2. *Normal Distribution.*

(1) Known Variance (σ_0^2 is known). If $X_1, \ldots, X_n \overset{iid}{\sim} \mathcal{N}(\mu, \sigma_0^2)$, then

$$\begin{aligned} f(x|\mu) &= \prod_{i=1}^{n} \left(\frac{1}{\sqrt{2\pi}\sigma_0} e^{-\frac{(x_i-\mu)^2}{2\sigma_0^2}} \right) \\ &\propto \exp\left(-\frac{\sum_{i=1}^n x_i^2 + 2\mu \sum_{i=1}^n x_i + n\mu^2}{2\sigma_0^2} \right) \\ &\propto \exp\left(-\frac{n\mu^2 - 2n\mu\bar{x} + \bar{x}^2}{2\sigma_0^2} \right) \\ &= \exp\left(-\frac{n(\mu - \bar{x})^2}{2\sigma_0^2} \right) \end{aligned}$$

So $\bar{X}$ is the sufficient statistic. This shows that the probability is equivalent to

$$\bar{X} \sim \mathcal{N}\left(\mu, \frac{\sigma_0^2}{n}\right)$$

In the classic case, $\hat{\mu} = \bar{X}$, so

$$\bar{X} \pm 1.96\frac{\sigma_0}{\sqrt{n}}$$

is the "perfect confidence interval" i.e.

$$\mathbb{P}\bigg(\bar{X} - 1.96\frac{\sigma_0}{\sqrt{n}} \leq \mu \leq \bar{X} + 1.96\frac{\sigma_0}{\sqrt{n}} \bigg) = 0.95$$

which is equivalent to

$$\mathbb{P}\bigg(\bar{X} - \mu \in \left(-1.96\frac{\sigma_0}{\sqrt{n}}, 1.96\frac{\sigma_0}{\sqrt{n}}\right) \bigg| \mu \bigg) = 0.95$$

In the Bayesian case,

$$\mathbb{P}_0(\mu) \sim \mathcal{N}(\mu_0, \delta_0^2)$$

which is the conjugate prior. Then

$$\mathbb{P}(\mu|X) = \mathcal{N}(\mu^*, {\sigma^*}^2)$$

where the parameters are defined by

$$\begin{aligned} \mu^* &= \frac{\frac{\mu_0}{\delta_0^2} + \frac{n\bar{x}}{\sigma_0^2}}{\frac{1}{\delta_0^2} + \frac{n}{\sigma_0^2}} \\ {\sigma^*}^2 &= \frac{1}{\frac{1}{\delta_0^2} + \frac{n}{\sigma_0^2}} \end{aligned}$$

1) If $\mu_0 = 0$, then $\delta_0^2 = \sigma_0^2$ which implies

$$\begin{aligned} \mu^* &= \frac{n}{n+1}\bar{x} \\ {\sigma^*}^2 &= \frac{1}{n+1}\sigma_0^2 \end{aligned}$$

2) As $\delta_0^2 \longleftarrow \infty$, we have

$$\begin{aligned} \mu^* &\longrightarrow \bar{x} \\ {\sigma^*}^2 &\longrightarrow \frac{\sigma_0^2}{n} \end{aligned}$$

The probability interval is

$$\mu^* \pm 1.96\sigma^*$$

As $\delta_0 \longrightarrow \infty$, we get

$$\bar{x} \pm 1.96\frac{\sigma_0}{\sqrt{n}}$$

which is the same as before. What is also interesting is that we have

$$\begin{aligned} \bar{X} - \mu|\mu &\sim \mathcal{N}\left(0, \frac{\sigma_0^2}{n}\right) \\ \bar{X} - \mu|X &\sim \mathcal{N}\left(0, \frac{\sigma_0^2}{n}\right) \end{aligned}$$

when $\delta_0^2 \longrightarrow \infty$. This is in fact a pivotal quantity.

The prior $\mathbb{P}_0\left(\mu\right) \propto C$, which is the "flat prior" is also called the noninformative prior ("Objective Bayesian").

4. September 15th, 2014

The density function for a Beta(a, b) distribution is

$$\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1}$$

where $a > 0, b > 0$.

$$\mathbb{E}[\theta] = \frac{a}{a+b}$$

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

The form of the density implies that

$$\int_0^1 \theta^{a-1}(1-\theta)^{b-1}\,\mathrm{d}\theta = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$$

because we know densities integrate to 1. Similarly, we know that

$$\int_{-\infty}^{\infty} e^{-\frac{x^2}{2\sigma^2}}\,\mathrm{d}x = \sqrt{2\pi\sigma^2}$$

Example 4. *If we have*

$$Y_1, \ldots, Y_n \overset{iid}{\sim} \mathcal{N}(\mu, \sigma_0^2)$$

$$\mu \overset{prior}{\sim} \mathcal{N}(\mu_0, \tau_0^2)$$

Then we have

$$\begin{aligned} f_Y(y) &\propto \left(\frac{1}{\sigma_0}\right)^n \exp\left(-\sum_{i=1}^n \frac{(y_i-\mu)^2}{2\sigma_0^2}\right) \times \exp\left(-\frac{(\mu-\mu_0)^2}{2\tau_0^2}\right) \\ &\propto \exp\left\{-\frac{n\mu^2 - 2n\bar{y}\mu}{2\sigma_0^2} - \frac{\mu^2 - 2\mu\mu_0}{2\tau_0^2}\right\} \exp\left[\frac{1}{2}\left\{\left(\frac{n}{\sigma_0^2} + \frac{1}{\tau_0^2}\right)\mu^2 - 2\left(\frac{n\bar{y}}{\sigma_0^2} + \frac{\mu_0}{\tau_0^2}\right)\mu\right\}\right] \end{aligned}$$

Example 5. $Y_1, \ldots, Y_n \overset{iid}{\sim} \mathcal{N}(\mu_0, \sigma^2)$ *where μ_0 is known. We can rewrite this as*

$$Y_1 - \mu_0, \ldots, Y_n - \mu_0 \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$$

WLOG assume $\mu_0 = 0$. Then

$$\mathbb{P}\big(Y_1, \ldots, Y_n | \sigma^2\big) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n e^{-\sum_{i=1}^n \frac{y_i^2}{2\sigma^2}}$$

We put a prior on σ^2 which has a distribution Inverse χ^2, which has the form

$$\pi_0(\sigma^2) \propto \left(\frac{1}{\sigma^2}\right)^x e^{-\frac{x}{2\sigma^2}}$$

The density of $\theta \sim \text{Inv-}\chi^2(\nu, s^2)$ is

$$\frac{(\nu/2)^{-\nu/2}}{\Gamma(\nu/2)} s^\nu \theta^{-\nu/2+1} e^{-\frac{\nu s^2}{2\theta}}$$

So we have $\sigma^2 \sim \text{Inv-}\chi^2(\nu, s^2)$ and

$$\pi_0(\sigma^2) \propto (\sigma^2)^{-\nu/2+1} e^{-\frac{\nu s^2}{2\sigma^2}}$$

which gives

$$\mathbb{P}\big(\sigma^2 | \boldsymbol{Y}\big) \propto (\sigma^2)^{-\frac{n+\nu}{2}+1} \exp\left(-\frac{\sum_{i=1}^n y_i^2 + \nu s^2}{2\sigma^2}\right)$$

Using the fact that

$$\sum_{i=1}^n y_i^2 = nS_y^2$$

we have

$$\sigma^2 \sim \text{Inv-}\chi^2\left(n+\nu, \frac{\sum_{i=1}^n y_i^2 + \nu s^2}{n+\nu}\right)$$

that is

$$\frac{\sum_{i=1}^n Y_i^2 + \nu s^2}{\sigma^2} \sim \chi^2_{n+\nu}$$

Note 1: How do we simulate from $\mathbb{P}\big(\sigma^2 | \boldsymbol{Y}\big)$?
Note 2: If $\nu = 0$, do we have

$$\frac{\sum_{i=1}^n Y_i^2}{\sigma^2}\bigg| \boldsymbol{Y} \sim \chi^2_n$$

This is similar to

$$\frac{\sum_{i=1}^n Y_i^2}{\sigma^2}\bigg| \sigma^2 \sim \chi^2_n$$

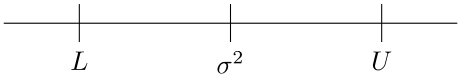
So if $\nu = 0$,

$$\pi_0(\sigma^2) \propto \left(\frac{1}{\sigma^2}\right) \mathcal{N}\left(n, \frac{\sum_{i=1}^n y_i^2}{n}\right)???$$

4.1. **Confidence Intervals (Frequentist Case).**

$$\frac{\sum_{i=1}^n y_i^2}{\sigma^2}\bigg| \sigma^2 \sim \chi^2_n$$

Then we wish to find points U and L



that is finding a U such that

$$\mathbb{P}\bigg(\frac{1}{U} \le \frac{\sum_{i=1}^n Y_i^2}{\sigma^2}\bigg|\sigma^2\bigg) = 0.95$$

and so

$$0.05 = 1 - F(1/U)$$
$$U = \frac{1}{1 - F^{-1}(0.05)}$$

so that

$$\sigma^2 \le \frac{U}{\sum_{i=1}^n y_i^2} = \frac{1}{(1 - F^{-1}(0.05) \sum_{i=1}^n y_i^2}$$

To find the other side, we find L such that

$$\mathbb{P}\bigg(\frac{\sum_{i=1}^n Y_i^2}{\sigma^2} \le \frac{1}{L}\bigg|\sigma^2\bigg) = 0.95$$

In general, to find A such that

$$\mathbb{P}\big(\sigma^2 \le A | \boldsymbol{Y}\big) = 0.95$$

we can use the fact that

$$\begin{aligned}\mathbb{P}\big(\sigma^2 \le A|\boldsymbol{Y}\big) &= \mathbb{P}\bigg(\frac{1}{A} \le \frac{1}{\sigma^2}\bigg|\boldsymbol{Y}\bigg) \\ &= \mathbb{P}\bigg(\frac{\sum_{i=1}^n Y_i^2}{A} \le \frac{\sum_{i=1}^n Y_i^2}{\sigma^2}\bigg|\boldsymbol{Y}\bigg) \\ &= 1 - F\left(\frac{\sum_{i=1}^n Y_i^2}{A}\right)\end{aligned}$$

- Confidence Interval:

$$\mathbb{P}\bigg(a \le \frac{\sum_{i=1}^n Y_i^2}{\sigma^2} \le A\bigg|\sigma^2\bigg) = 0.95$$

- Probability Interval:

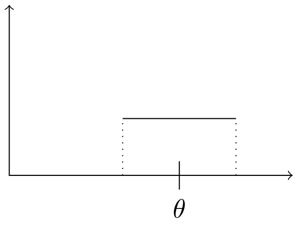
$$\mathbb{P}\bigg(a' \le \frac{\sum_{i=1}^n Y_i^2}{\sigma^2} \le A'\bigg|\boldsymbol{Y}\bigg) = 0.95$$

These in fact end up being the same because $\frac{\sum_{i=1}^n Y_i^2}{\sigma^2}$ is a pivotal quantity.

4.2. **Location and Scale Family.** The location family is a distribution of the form

$$Y_1, \ldots, Y_n \overset{iid}{\sim} f(y-\theta)$$

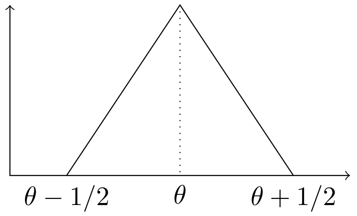
Example 6. *If* $Y \sim \text{Uniform}\left(\theta - \frac{1}{2}, \theta + \frac{1}{2}\right)$



The ad hoc way to estimate this is done by using

$$\hat{\theta} = \frac{Y_1 + Y_2}{2}$$

but then the sampling distribution of $\hat{\theta}$ *is "wrong"*



If instead we use $d_2 = Y_2 - Y_1, \ldots, d_n = Y_n - Y_{n-1}$, *we have that* $(d_2, \ldots, d_n)$ *is something called an ancillary statistic. We want to consider*

$$Y_1 | d_2, \ldots, d_n$$

and want to make inference based on this conditional distribution.

$$\mathbb{P}(Y_1, d_2, \ldots, d_n|\theta) = f(y_1 - \theta) f(y_1 + d_2 - \theta) \cdots f(y_1 + d_n - \theta)$$

so

$$\begin{aligned}
\mathbb{P}(y_1 | d_2, \ldots, d_n, \theta) &= \frac{f(y_1 - \theta) f(y_1 + d_2 - \theta) \cdots f(y_1 + d_n - \theta)}{\int f(z-\theta) f(z + d_2 - \theta) \cdots f(z + d_n - \theta) \, \mathrm{d}z} \\
&= \frac{f(y_1 - \theta) f(y_1 + d_2 - \theta) \cdots f(y_1 + d_n - \theta)}{\int f(u) f(u + d_2) \cdots f(u + d_n) \, \mathrm{d}u} \\
&\overset{u=\underline{y_1}-\tau}{=} \frac{f(y_1 - \theta) f(y_1 + d_2 - \theta) \cdots f(y_1 + d_n - \theta)}{\int f(y_1 - \tau) f(y_2 - \tau) \cdots f(y_n - \tau) \, \mathrm{d}\tau}
\end{aligned}$$

and the numerator here can be expressed as a $g(y_1 - \theta)$ *for some function* g. *If we integrate out the* θ, *we get something called the* ***Pitman estimator***.

5. September 18th, 2014

5.1. **Location Family.** $Y_1, \ldots, Y_n \sim f(y - \theta)$ where $\theta \in (-\infty, \infty)$.

Example 7. $Y_i \overset{iid}{\sim} \text{Cauchy}(\theta)$

$$f(y|\theta) = \frac{1}{\pi} \frac{1}{1 + (y-\theta)^2}$$

(1) If we only have $y_1 \sim f(y - \theta)$, *then how do we give a* 90% *confidence interval for* θ*?*

$$I(y_1) = (y_1 - a, y_1 + A)$$

where $-a = F^{-1}(0.05)$ *and* $A = F^{-1}(0.95)$. *This interval is constructed in a way such that*

$$\begin{aligned}
\mathbb{P}(I(y_1) \ni \theta | \theta) &= \mathbb{P}(y_1 - a \leq \theta \leq y + 1A|\theta) \\
&= \mathbb{P}(-a \leq \theta + y_1 \leq A|\theta) \\
&= 90\%
\end{aligned}$$

(2) If $y_1, \ldots, y_n \overset{iid}{\sim} f(y-\theta)$, what is

$$\mathbb{P}(y_1 | \underbrace{y_2 - y_1}_{d_2}, \ldots, \underbrace{y_n - y_1}_{d_n}, \theta)?$$

We have

$$\mathbb{P}(y_1 | d_2, \ldots, d_n, \theta) = \frac{f(y_1-\theta)f(d_2+y_1-\theta)\cdots f(d_n+y_1-\theta)}{\int f(y_1-\varphi)\cdots f(y_n-\varphi)\,\mathrm{d}\varphi}$$

What is a confidence interval for this? By Bayes' Theorem, we have

$$\mathbb{P}(\theta | y_1, \ldots, y_n) = \frac{f(y_1-\theta)\cdots f(y_n-\theta)f_0(\theta)}{\int f(y_1-\tau)\cdots f(y_n-\tau)f_0(\tau)\,\mathrm{d}\tau}$$

$f_0(\tau) \propto C$, which is called a "flat prior". Here, $y-\theta$ is a pivot since it does not rely on any other parameters. The prior distribution does not change the distribution of $y-\theta$.

$$\begin{aligned}\mathbb{P}(\theta | y_1, \ldots, y_n) &= \frac{f(y_1-\theta)f(d_2+y_1-\theta)\cdots f(d_n+y_1-\theta)}{\int f(y_1-\tau)\cdots f(y_n-\tau)\,\mathrm{d}\tau} \\ &= g(y_1-\theta)\end{aligned}$$

The difference here is it treats θ as a random variable instead of the y's. The Bayes estimator is

$$\begin{aligned}\hat{\theta} &= \mathbb{E}[\theta | y_1, \ldots y_n] \\ &= \frac{\int \theta f(y_1-\theta)\cdots f(y_n-\theta)\,\mathrm{d}\theta}{\int f(y_1-\tau)\cdots f(y_n-\tau)\,\mathrm{d}\tau}\end{aligned}$$

5.2. **Noninformative Prior (1 dimension).**

(a) For the location family, the noninformative prior is uniform. This keeps the pivotal quantity's distribution $y-\theta$ unchanged.
(b) General Case:

$$y_1, \ldots, y_n \sim \mathbb{P}(y|\theta)$$

Likelihood Function:

$$\mathcal{L}(\theta | y_1, \ldots, y_n) \propto \prod_{i=1}^{n} \mathbb{P}(y_i|\theta)$$

Log-Likelihood Function:

$$\ell(\theta | y_1, \ldots, y_n) = \sum_{i=1}^{n} \log \mathbb{P}(y_1|\theta)$$

We have $\hat{\theta} = \arg\max_\theta \ell(\theta | y_1, \ldots, y_n)$. Then

$$\begin{aligned}\ell(\theta | y_1, \ldots, y_n) &= \ell(\hat{\theta} | y_1, \ldots, y_n) + \ell'(\hat{\theta} | y_1, \ldots, y_n)(\theta-\hat{\theta}) + \frac{1}{2}\ell''(\hat{\theta} | y_1, \ldots, y_n)(\theta-\hat{\theta})^2 + o(\bullet)(\theta-\hat{\theta})^3 \\ &= \ell(\hat{\theta} | y_1, \ldots, y_n) - n\hat{\mathcal{I}}_n(\hat{\theta})(\theta-\hat{\theta})^2 + o(\bullet)\end{aligned}$$

We know

$$\begin{aligned}\ell''(\theta | y_1, \ldots, y_n) &= \sum_{i=1}^{n} \frac{\partial^2}{\partial\theta^2} \log \mathbb{P}(y_1|\theta)) \\ &= -n\hat{\mathcal{I}}_n(\theta) \qquad \text{(observed Fisher information)}\end{aligned}$$

The Fisher Information is

$$\begin{aligned}\mathcal{I}(\theta) &= -\mathbb{E}\left[\frac{\partial^2}{\partial\theta^2} \log \mathbb{P}(y|\theta) \middle| \theta\right] \\ &= \operatorname{Var}\left(\frac{\partial}{\partial\theta} \log \mathbb{P}(y|\theta) \middle| \theta\right)\end{aligned}$$

and the observed version is

$$\hat{\mathcal{I}}_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} \frac{\partial^2}{\partial\theta^2} \log \mathbb{P}(y_i|\theta) \overset{\text{as n} \to \infty}{\longrightarrow} I(\theta)$$

and so

$$\begin{aligned}\ell(\theta|y_1,\dots,y_n) &\approx C + \frac{n\hat{\mathcal{I}}(\hat\theta_n)}{2}(\theta-\hat\theta)^2\\ \mathcal{L}(\theta|y_1,\dots,y_n) &\approx C\mathcal{N}\left(\hat\theta, \frac{1}{n\mathcal{I}(\hat\theta)}\right)\end{aligned}$$

Question: Find a parameterization $\xi = \varphi(\theta)$ such that

$$\mathcal{I}(\xi) = c$$

Then we can put the flat prior on ξ

$$\mathcal{I}(\xi) = \mathcal{I}(\theta)\left(\frac{\mathrm{d}\theta}{\mathrm{d}\xi}\right)^2$$

so

$$\left(\frac{\mathrm{d}\xi}{\mathrm{d}\theta}\right)^2 \propto \mathcal{I}(\theta)$$

and

$$\left|\frac{\mathrm{d}\xi}{\mathrm{d}\theta}\right| \propto \sqrt{\mathcal{I}(\theta)}$$

If we give flat prior on ξ,

$$\begin{aligned}C\mathrm{d}\xi &= C\frac{\mathrm{d}\xi}{\mathrm{d}\theta}\,\mathrm{d}\theta\\ &\propto \sqrt{\mathcal{I}(\theta)}\,\mathrm{d}\theta\end{aligned}$$

So the prior on θ should be $\propto \sqrt{\mathcal{I}(\theta)}$

Example 8 (Example of noninformative priors)**.**
(1) Gaussian with unknown mean: $\pi(\theta) \propto C$
(2) Gaussian with unknown variance:

$$\begin{aligned}\pi(\sigma) &\propto \frac{1}{\sigma}\\ \pi(\sigma^2) &\propto \frac{1}{\sigma^2}\end{aligned}$$

(3) Binomial: $Y \sim \mathrm{Bin}(n,\theta)$*, so the pmf is*

$$\theta^y(1-\theta)^{n-y}$$

and so

$$\begin{aligned}\mathcal{I}(\theta) &= -\mathbb{E}\left[-\frac{y}{\theta^2} - \frac{n-y}{(1-\theta)^2}\right]\\ &= \frac{n}{\theta} + \frac{n}{1-\theta}\\ &= \frac{n}{\theta(1-\theta)}\end{aligned}$$

This means

$$\pi_0(\theta) \propto \theta^{-1/2}(1-\theta)^{-1/2}$$

and the noninformative prior is $\mathrm{Beta}\left(\frac{1}{2},\frac{1}{2}\right)$

6. September 22nd, 2014

6.1. Likelihood Function. The log-likelihood is defined as

$$\ell(\theta) = \sum_{i=1}^n \log f(y_i|\theta)$$

In the Bayesian sense, the prior is $\pi_0(\theta)$ and the posterior distribution is proportional to

$$\pi_0(\theta)\prod_{i=1}^n f(y_i|\theta)$$

Then

$$\log \mathbb{P}(\theta|y_1,\ldots,y_n) = C + \log \pi_0(0) + \sum_{i=1}^n \log f(y_i|\theta)$$

We can approximate this by

$$\begin{aligned}\ell(\theta) &\approx \ell(\hat{\theta}) + \frac{1}{2}\sum_{i=1}^n \frac{\partial^2}{\partial\theta^2}\log f(y_i|\hat{\theta})(\theta-\hat{\theta})^2 \\ &= \ell(\hat{\theta}) + \frac{1}{2}n\hat{\mathcal{I}}(\hat{\theta})(\theta-\hat{\theta})^2\end{aligned}$$

and so

$$\log \mathbb{P}(\theta|y_1,\ldots,y_n) \approx C + \log\pi_0(\theta) - \frac{1}{2}n\hat{\mathcal{I}}(\hat{\theta})(\theta-\hat{\theta})^2 + o((\theta-\hat{\theta})^3)$$

This implies this is approximately Normal $\mathcal{N}\left(\hat{\theta}, \sqrt{n\hat{\mathcal{I}}(\hat{\theta})^{-1}}\right)$ and so

$$\mathbb{P}(\theta|y) \approx \mathcal{N}\left(\hat{\theta}, \sqrt{n\hat{\mathcal{I}}(\hat{\theta})}\right)$$

which gives

$$\sqrt{n\hat{\mathcal{I}}(\hat{\theta})} \approx \mathcal{N}(0,1)$$

If $\hat{\mathcal{I}}(\hat{\theta}) \approx \mathcal{I}(\theta_0)$ (smoothly varying), The probability interval is

$$\hat{\theta} \pm z_{\alpha/2}\frac{1}{\sqrt{n\mathcal{I}(\theta_0)}}$$

where θ_0 is the “true value” and the asymptotic confidence interval is

$$\hat{\theta} \pm z_{\alpha/2}\frac{1}{\sqrt{n\mathcal{I}(\theta_0)}}$$

MLE: $\hat{\theta} = \arg\max \ell(\theta)$.

$$\ell'(\theta) = \ell'(\theta_0) + \ell''(\theta_0)(\theta-\theta_0) + o((\theta-\theta_0)^2)$$

Plugging in $\hat{\theta}$, we get

$$\ell'(\hat{\theta}) = \ell'(\theta_0) + \ell'(\theta_0)(\theta-\theta_0) + o((\hat{\theta}-\theta_0)^2)$$

$\ell'(\hat{\theta}) = 0$ since it is the MLE. Rearranging, we get

$$\begin{aligned}\hat{\theta}-\theta_0 &= -\frac{\ell'(\theta_0)}{\ell''(\theta_0)} + o((\hat{\theta}-\theta_0)^2) \\ &= \frac{\sum_{i=1}^n \frac{\partial}{\partial\theta}\log f(y_i|\theta_0)}{\sum_{i=1}^n \frac{\partial^2}{\partial\theta^2}\log f(y_i|\theta_0)} + o((\hat{\theta}-\theta_0)^2)\end{aligned}$$

Dividing both sides by $\sqrt{n}$, we get

$$\sqrt{n}(\hat{\theta}-\theta_0) = \frac{\frac{1}{\sqrt{n}}\sum_{i=1}^n \frac{\partial}{\partial\theta}\log f(y_i|\theta_0)}{\frac{1}{n}\sum_{i=1}^n \frac{\partial^2}{\partial\theta^2}\log f(y_i|\theta_0)} + o((\hat{\theta}-\theta_0)^2)$$

Recall that

$$\mathbb{E}\left[\frac{\partial}{\partial\theta}f(y|\theta)\right] = 0$$

since

$$\begin{aligned}\int \frac{f'(y|\theta)}{f(y|\theta)}f(y|\theta)\,\mathrm{d}y &= \int \frac{\partial}{\partial\theta}f(y|\theta)\,\mathrm{d}y \\ &= \frac{\partial}{\partial\theta}\int f(y|\theta)\,\mathrm{d}y \\ &= 0\end{aligned}$$

This shows that

$$\frac{1}{\sqrt{n}}\sum_{i=1}^n \frac{\partial}{\partial\theta}\log f(y_i|\theta_0) \xRightarrow{\text{CLT}} \mathcal{N}(0,\mathcal{I}(\theta_0))$$

and

$$-\frac{1}{n}\sum_{i=1}^n \frac{\partial^2}{\partial\theta}\log f(y_i|\theta_0) \xrightarrow{\text{SLLN}} \mathcal{I}(\theta_0)$$

6.2. **Poisson Distribution.** If $Y_1, \dots, Y_n \overset{iid}{\sim} \text{Pois}(\lambda)$, then

$$\mathbb{P}(y|\lambda)\prod_{i=1}^{n} \frac{\lambda^y}{y_i} e^{y_i}$$

So the posterior distribution is

$$\mathbb{P}(\lambda|y_1, \dots, y_n) \propto \pi_0(\lambda)\lambda^{\sum_{i=1}^n y_i} e^{-n\lambda}$$

If we take the log of a single y_i, we get

$$\log f(y_i|\lambda) = y_i \log \lambda - \lambda - \log y_i!$$

Differentiating this gives us

$$\frac{\partial}{\partial \lambda} \log f(y|\lambda) = \frac{y_i}{\lambda} - 1$$

and so

$$V \sim (\qquad) = \frac{1}{\lambda}$$

The noninformative prior is

$$\pi_0(\lambda) \propto \left(\frac{1}{\lambda}\right)^{1/2} = \lambda^{-1/2}$$

The conjugate prior is

$$\pi_0(\lambda) = \frac{\beta^\alpha}{\Gamma(\alpha)} \lambda^{\alpha-1} e^{-\beta\lambda}$$

which is the Gamma(α, β) distribution. The posterior distribution is Gamma$(\alpha + \sum_{i=1}^n y_i, n + \beta)$.

It is also possible to integrate out the λ. Suppose that $Y \sim \text{Pois}(\lambda)$ and $\lambda \sim \pi_0(\lambda) = \text{Gamma}(\alpha, \beta)$.

$$\begin{aligned}
\mathbb{P}(Y = y|\alpha, \beta) &= \int \frac{\lambda^y}{y!} e^{-\lambda} \lambda^{\alpha-1} e^{-\beta\lambda} \, \mathrm{d}\lambda \\
&\propto \frac{1}{y!} \frac{\Gamma(\alpha + y)}{(\beta+1)^{\alpha+y}} \frac{\beta^\alpha}{\Gamma(\alpha)} \\
&= \frac{\Gamma(\alpha + y)}{\Gamma(\alpha)\Gamma(y+1)} \left(\frac{\beta}{\beta+1}\right)^\alpha \left(\frac{1}{\beta+1}\right)^y
\end{aligned}$$

which is the pdf of the Negative Binomial distribution.

Example 9. *Suppose* $Y_1 \sim \text{Pois}(y_1|\lambda_1), \dots, Y_k \sim \text{Pois}(y_k|\lambda_k)$ *where* $\lambda_i \overset{iid}{\sim} \text{Gamma}(\alpha, \beta)$. *An alternative approach is to have* $Y_1, \dots, Y_n \overset{iid}{\sim} \text{Pois}(\lambda)$ *where* $\lambda \sim \text{Gamma}(\alpha, \beta)$.

6.3. **Poisson Regression.** $Y_i \sim \text{Pois}(X_i\lambda)$ for $i = 1, \dots, n$, which has likelihood function

$$\mathbb{P}(Y_i|X_i\lambda) = \frac{(x_i\lambda)^{y_i}}{y_i!} e^{-x_i\lambda}$$

6.4. **Exponential Family.** The general form of an exponential family is

$$f(y)g(\theta)e^{\phi(\theta)u(y)}$$

Example 10 (Normal distribution)**.**

$$\begin{aligned}
\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y-\mu)^2}{2\sigma^2}} &= \left(\frac{1}{\sigma})\right) e^{-\frac{y^2-2y\mu+\mu^2}{2\sigma^2}} \\
&= \left(\frac{1}{\sigma} e^{-\frac{\mu^2}{2\sigma^2}}\right) \exp\left(-\frac{1}{2}\begin{pmatrix} y^2 \\ -2y \end{pmatrix}\left(\frac{1}{\sigma^2}, \frac{\mu}{\sigma^2}\right)\right)
\end{aligned}$$

7. September 25th, 2014

7.1. **The notion of i.i.d.** Suppose you have a coin and toss it 10 times and you get 9 heads and 1 tail. What is the probability of heads on the next toss? In a typical frequentist model with a fair coin, we would expect this to be $\frac{1}{2}$. However, how do we know that this is a fair coin? We do not know the true probability of heads.

Theorem 2 (De Finiti Theorem)**.** *$\{X_1, X_2, \ldots\}$ is an infinite sequence of exchangeable binary random variables, that is for $i_1 \neq i_2 \neq \cdots \neq i_k$*

$$\mathcal{L}(X_{i_1}, X_{i_2}, \ldots, X_{ik}) = \mathcal{L}(X_1, \ldots, X_k)$$

Then $\exists$ a probability measure $F(\bullet)$ on $[0,1]$ such that $\forall n$

$$\mathbb{P}(X_1, \ldots, X_n) = \int_0^1 \prod_{i=1}^n \left\{\theta^{X_i}(1-\theta)^{1-X_i}\right\} F(\mathrm{d}\theta)$$

Proof. Let $X_1 + \cdot + X_n = y_n$. Then

$$\begin{aligned}\mathbb{P}(X_1 = x_1, \ldots, X_n = x_n | X_1 + \cdots + X_N = y_N) &= \frac{\mathbb{P}(X_1 = x_1, \ldots, X_n = x_n, X_{n+1} + \cdots + X_N = y_N - y_n)}{\mathbb{P}(X_1 + \cdots + X_N = y_N)} \\ &= \frac{\binom{N-n}{y_N - y_n}}{\binom{N}{y_N}} \qquad \text{because this is hypergeometric} \\ &= \frac{(Y_N)_{y_n}(N - y_N)_{N-y_n}}{(N)_n}\end{aligned}$$

where $(N)_n = N(N-1)\cdots(N-n+1)$. Then

$$\begin{aligned}\mathbb{P}(X_1 = x_1, \ldots, X_n = x_n) &= \sum_{y_N} \frac{(Y_N)_{y_n}(N - Y_N)_{n-y_n}}{(N)_n} \mathbb{P}\big(\tfrac{X_1 + \cdots + X_N}{N} = \tfrac{Y_N}{N}\big) \\ &\xrightarrow{N \to \infty} \int \theta^{y_n}(1-\theta)^{n-y_n} F(\mathrm{d}\theta)\end{aligned}$$

where the probability on the right-hand side converges to a probability measure as $N \to \infty$. □

7.2. **Simple Multiparameter Models.**

(1) $Y_1, \ldots, Y_n \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$ with μ, σ^2 unknown. For the MLE, $\hat{\mu} = \bar{y}$ and

$$\hat{\sigma}^2 | \mu = \frac{\sum_{i=1}^n (y_i - \mu)^2}{n}$$

How do we get rid of σ^2 if we are only interested in μ?

1. "Plug-in":

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^n (y_i - \bar{y})^2}{n} \longrightarrow \hat{\mu} = \bar{y} \Rightarrow \bar{y} \pm Z_{\alpha/2}\frac{\hat{\sigma}}{\sqrt{n}}$$

2. Profile-likelihood: How do we account for uncertainty?
3. Conditional Distribution: $\sum_{i=1}^n (y_i - \bar{y})^2$ is a sufficient statistic for σ^2.

$$\sqrt{n}\frac{\bar{y} - \mu}{\sqrt{\frac{1}{n-1}\sum_{i=1}^n (y_i - \bar{y})^2}} \sim T$$

and so

$$\begin{aligned}\mathcal{L}\left(\bar{y}\middle|\sum_{i=1}^n (y_i - \bar{y})^2\right) &= \mathcal{L}\left(T\middle|\sum_{i=1}^n (y_i - \bar{y})^2\right) \\ &= \mathcal{L}(T)\end{aligned}$$

4. Bayes: if

$$\sigma^2 \sim \text{Inv-}\chi^2(\nu_0, \sigma_0^2)$$

$$\mu | \sigma^2 \sim \mathcal{N}\left(\mu_0, \frac{\sigma^2}{\kappa_0}\right)$$

which implies

$$\frac{\nu_0 \sigma_0^2}{\sigma^2} \sim \chi^2(\nu_0)$$

Letting $\kappa_0 \longrightarrow 0$, we get

$$\mathbb{P}(\sigma) \propto \frac{1}{\sigma^2}$$
$$\mathbb{P}\big(\mu, \sigma^2\big) \propto (\sigma^2)^{-1}$$

and

$$\begin{aligned}\mathbb{P}\big(\mu, \sigma^2 | y_1, \dots, y_n\big) &= (\sigma^{-1})(\sigma^2)^{-\frac{n+\nu_0+1}{2}} \exp\left(-\frac{1}{2\sigma^2}\left(\nu_0\sigma_0^2 + (n-1)s^2 + n(\bar{y}-\mu)^2 + \kappa_0(\mu-\mu_0)^2\right)\right) \\ &\sim N\text{-Inv-}\chi^2\left(\mu_n, \frac{\sigma_n^2}{\kappa_n}, \nu_n, \sigma_n\right) \\ \mathbb{P}\big(\sigma^2 | y_1, \dots, y_n\big) &= \text{Inv-}\chi^2(\nu_n, \sigma_n^2)\end{aligned}$$

Jun will likely post notes on this later...this lecture was a bit of a mess.

8. September 29th, 2014

Say we have parameters μ and σ^2 and we are interested in μ only. The Bayesian approach to getting rid of the nuisance parameter (σ^2) is to integrate.

8.1. **Two-Sample Problem.**

$$\begin{array}{cc} y_{11} & y_{12} \\ \vdots & \vdots \\ y_{n_1 1} & y_{n_2 2} \\ \hline \bar{y}_{\bullet 1} & \bar{y}_{\bullet 2} \end{array}$$

We have

$$y_{i1} \sim \mathcal{N}(\mu_1, \sigma_1^2)$$
$$y_{i2} \sim \mathcal{N} * \mu_2, \sigma_2^2)$$

Of interest is $\delta = \mu_2 - \mu_!$. Take $\hat{\delta} = y_{\bullet 2} - \bar{y}_{\bullet 1}$

(1) General approximation (not ideal)

$$\mathbb{E}\Big[\hat{\delta} | \mu_1, \mu_2\Big] = \mu_2 - \mu_1$$

which is unbiased, and

$$\text{Var}(\hat{\delta} | \mu_1, \mu_2) = \frac{\sigma_2^2}{n_2} + \frac{\sigma_1^2}{n_1}$$

so

$$\hat{\delta} \sim \mathcal{N}\left(\mu_2 - \mu_1, \frac{\sigma_2^2}{n_2} + \frac{\sigma_1^2}{n_1}\right)$$

Our confidence interval would then be

$$\delta \pm 1.96\sqrt{\frac{\hat{\sigma}_2^2}{n_2} + \frac{\hat{\sigma}_1^2}{n_1}}$$

where

$$\hat{\sigma}_2^2 = \frac{1}{n_2 - 1}\sum_{i=1}^{n_2}(y_{i2} - \bar{y}_{\bullet 2})^2$$
$$\hat{\sigma}_1^2 = \frac{1}{n_1 - 1}\sum_{i=1}^{n_1}(y_{i1} - \bar{y}_{\bullet 1})^2$$

(2) Well known: If $\sigma_1 = \sigma_2$.

$$\hat{\sigma}_p^2 = \frac{1}{n_1 + n_2 - 2}\left\{\sum_{i=1}^{n_2}(y_{i2} - \bar{y}_{\bullet 2})^2 + \sum_{i=1}^{n_1}(y_{i1} - \bar{y}_{\bullet 1})^2\right\}$$

and

$$\frac{\hat{\delta}}{\sqrt{\frac{1}{n_1} + \frac{1}{n_2}\hat{\sigma}_p}} \sim t_{n_1 + n_2 - 2}$$

where this is exact and not an approximation. If $\sigma_1 \neq \sigma_2$, then the answer above can be wrong (too aggressive).

(3) Welsh-t distribution

If we have

$$s_1^2 = \sum_{i=1}^{n_1}(y_{i1} - \bar{y}_{\bullet 1})^2$$
$$s_2^2 = \sum_{i=1}^{n_2}(y_{i2} - \bar{y}_{\bullet 2})^2$$

and look at

$$\hat{\delta}|s_1^2, s_2^2, \mu_1, \mu_2$$

there is no pivotal quantity with $\sigma_1^2 \neq \sigma_2^2$.

8.1.1. *Bayesian Case.*

$$y_{11}, \dots, y_{n_1 1} \overset{iid}{\sim} \mathcal{N}(\mu_1, \sigma_1^2)$$

implies

$$\mathbb{P}\big(\mu_1, \sigma_1^2 | y_{11}, \dots, y_{n_1 1}\big) \sim \text{N-Inv-}\chi^2()$$

Similarly, if

$$y_{12}, \dots, y_{n_2 2} \overset{iid}{\sim} \mathcal{N}(\mu_2, \sigma_2^2)$$

then

$$\mathbb{P}\big(\mu_2, \sigma_2^2 | y_{12}, \dots, y_{n_2 2}\big) \sim \text{N-Inv-}\chi^2()$$

and so

$$\mathbb{P}\big(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2 | \boldsymbol{y}\big) \propto \mathbb{P}\big(\mu_1, \sigma_1^2 | \boldsymbol{y}\big) \times \mathbb{P}\big(\mu_2, \sigma_2^2 | \boldsymbol{y}\big)$$

8.1.2. *Strategies.* Simulate from the above posterior distribution.

$$\begin{pmatrix} \mu_1 \\ \mu_2 \\ \sigma_1 \\ \sigma_2 \end{pmatrix}^{(1)} \dots \begin{pmatrix} \mu_1 \\ \mu_2 \\ \sigma_1 \\ \sigma_2 \end{pmatrix}^{(M)}$$

where $M = 10000$ say. Then we have

$$\delta^{(1)} = \mu_2^{(1)} - \mu_1^{(1)}, \dots, \delta^{(M)} = \mu_2^{(M)} - \mu_1^{(M)}$$

and then we can look at the histogram of $\delta^{(s)}$.

8.1.3. $\sigma_1^2 = \sigma_2^2$.

$$\mathbb{P}\big(\mu_1, \mu_2, \sigma^2 | \boldsymbol{y}\big) \propto \left(\frac{1}{\sigma}\right)^{n_1+n_2} \exp\left[-\frac{1}{2\sigma^2}\left\{\sum_{i=1}^{n_1}(y_{i1} - \mu_1)^2 + \sum_{i=1}^{n_2}(y_{i2} - \mu_2)^2\right\}\right] \times \mathbb{P}\big(\mu_1, \mu_2, \sigma^2\big)$$

9. October 2nd, 2014

Example 11 (Sampling Example)**.** *Suppose*

$$X_1, \dots, X_n \overset{iid}{\sim} \text{Beta}(\alpha, \beta)$$

which has density

$$\prod_{i=1}^{n}\left\{x_i^{\alpha-1}(1-x_i)^{\beta-1}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\right\} = \frac{\{\Gamma(\alpha+\beta)\}^n}{\Gamma(\alpha)^n\Gamma(\beta)^n}\left(\prod_{i=1}^{n}x_i\right)^{\alpha-1}\left\{\prod_{i=1}^{n}(1-x_i)\right\}^{\beta-1}$$

Taking logs, we get

$$\log \ell(\alpha, \beta) = n\log\Gamma(\alpha+\beta) - n\log\Gamma(\alpha) - n\log\Gamma(\beta) + (\alpha-1)\log\left(\prod_{i=1}^{n}x_i\right) + (\beta-1)\log\left(\prod_{i=1}^{n}(1-x_i)\right)$$

How do we choose initial values for α and β if we are only given our data? We can use the method of moments to do this

$$\mathbb{E}[X] = \frac{\alpha}{\alpha+\beta}$$
$$\text{Var}(X) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$$
$$= (\mathbb{E}[X])(1-\mathbb{E}[X])\frac{1}{\alpha+\beta+1}$$

9.1. **Metropolis-Hastings Algorithm.** One major advantage of the Metropolis-Hastings algorithm is that it allows sampling from a distribution when you don't know the normalizing constant. In real life applications with unusual distributions, the normalizing constant is very hard to find analytically.

Example 12 (Poisson)**.** *How do we sample from a Poisson distribution?*
(1) Use Exponential random variables (Poisson Process)
(2) Use the inverse CDF. This is fairly easy because it is a discrete random variable

How do we determine where to cut off the initial tail of the Metropolis-Hastings?

- We can use an R statistic to determine the point
- We can look at multiple chains and see where they group togehter
- We can do it by eye
- We can use an ACF. Say the autocorrelation disappears at lag 25. Then we can sample for every 25 values.

We have that $\bar{X} \to \hat{\mu}$. What is $\operatorname{Var}(\bar{X})$? In the iid case $\operatorname{Var}(\bar{X}) = \frac{\sigma^2}{n}$. In the MCMC case,

$$\begin{aligned}\operatorname{Var}\left(\frac{X_1 + \cdots + X_n}{m}\right) &= \frac{1}{m^2}\left(m \operatorname{Var}(X) + 2(m-1)\operatorname{Cov}(X_1, X_2) + 2(m-2)(X_1, X_3) + \cdots\right) \\ &= \frac{1}{m}\left\{\sigma^2 + 2\rho_1\sigma^2\left(1 - \frac{1}{m}\right) + 2\rho_2\sigma^2\left(1 - \frac{2}{m}\right) + \cdots\right\} \\ &\approx \frac{\sigma^2}{m}(1 + 2\rho_1 + 2\rho_2 + 2\rho_3 + \cdots) \qquad \text{for large } m\end{aligned}$$

In an iid case, $\rho_i = 0$ for all i and the variance is simply $\frac{\sigma^2}{m}$.

10. October 6th, 2014

10.1. **Basic Decision-Theoretic Concepts and Arguments.**

- $\Theta =$ parameter space or "states of nature"
- $A =$ action space
- $\mathcal{X} =$ the sample space of a random variable X
- $\mathbb{P}(X = x|\theta)$ is the parameter model
- $L : \Theta \times A \to \mathbb{R}^+$. loss function $L(\theta, a)$
- $d : A \times \mathcal{X} \longrightarrow [0, 1]$. $d(a, x) = d(a|x)$
- D :set of all decision functions.
- Risk of $d \in D$ is

$$\begin{aligned}R(\theta, d) &= \int L(\theta, d(\bullet|X))\mathbb{P}(X|\theta)\ \mathrm{d}x \\ &= \mathbb{E}[L(\theta, d)|\theta]\end{aligned}$$

It is important to note that risk is a funtion,

$$R : \Theta \times D \longrightarrow \mathbb{R}^+$$

The loss function is non-negative and unbounded.

Definition 1 (Admissibility)**.** *We say $d \in D$ is admissible if $\nexists d'$ such that*

$$R(\theta, d') \leq R(\theta, d)$$

for all θ and $\exists\theta_0$ such that

$$R(\theta_0, d') < R(\theta, d)$$

Definition 2 (Minimax)**.** *d_M is "minimax" if*

$$\inf_{d \in D} \sup_{\theta \in \Theta} R(\theta, d) = \sup_{\theta \in \Theta} R(\theta, d_M)$$

Definition 3 (Bayes Risk)**.** *For any prior Λ and $d \in D$ the Bayes risk of d with respect to Λ is*

$$R(\Lambda, d) = \int R(\theta, d)\, \Lambda(d\theta)$$

Definition 4 (Bayes Estimator/Bayes Decision Rule)**.** *A Bayes decision rule with respect to* Λ*,* $\mathrm{d}\Lambda$ *is*

$$R(\Lambda, d_\Lambda) = \inf_{d \in D} R(\Lambda, d)$$

We have

$$\begin{aligned} R(\Lambda, d) &= \iint L(\theta, d(X))\mathbb{P}(X|\theta) \ \mathrm{d}X\Lambda(\mathrm{d}\theta) \\ &= \iint L(\theta, d(X))\mathbb{P}(X|\theta) \ \Lambda(\mathrm{d}\theta)\mathrm{d}X \\ &= \int \left\{ \int L(\theta, d(X))\mathbb{P}(\theta|X) \ \mathrm{d}\theta \right\} \mathbb{P}(X) \ \mathrm{d}X \end{aligned}$$

since

$$\mathbb{P}(x) = \int \mathbb{P}(X|\theta) \ \Lambda(\mathrm{d}\theta)$$

and where

$$L(\theta, d) = \int L(\theta, a)\mathbb{P}(a|X) \ \mathrm{d}a$$

Since $d_\Lambda(x) = \arg\inf_d \mathbb{E}[L(\theta, d)|X]$, we have
(1) $L(\theta, d) = (\theta - d)^2$ we have

$$\mathbb{E}\big[(\theta - d(x))^2 \big| X\big] \Rightarrow d(X) = \mathbb{E}[\theta|X]$$

(2) $L(\theta, d) = |\theta - d|$

$$\int |\theta - d(x)|\mathbb{P}(\theta|x) \ \mathrm{d}\theta \Rightarrow d(x) = \text{post median}(\theta|X)$$

(3)

$$L(\theta, d) = \begin{cases} 0 & \theta = d \\ 1 & \theta \neq d \end{cases} \Rightarrow d_\Lambda(x) = \text{post mode}(\theta|X)$$

Consider the case

$$\begin{aligned} \Theta &= \{\theta_1, \dots, \theta_l\{ \\ X &= \{X_1, \dots, X_m\} \\ A &= \{a_1, \dots, a_k\} \end{aligned}$$

Definition 5 (Risk Body)**.** *The risk body is*

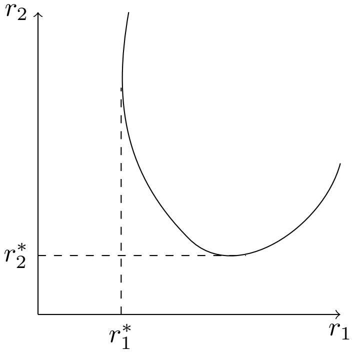
$$R = \{(R(\theta_1, d), \dots, R_l(\theta, d)), d \in D\}$$

Theorem 3. *The risk body is convex.*

Proof. For any $\alpha \in [0, 1]$,

$$\begin{aligned} &\alpha\big(R(\theta_1, d_1), \dots, R(\theta_l, d_1)\big) + (1 - \alpha)\big(R(\theta_1, d_2, \dots, R(\theta_l, d_2)\big) \\ &= \big(\alpha R(\theta_1, d_1) + (1 - \alpha)R(\theta_1, d_2), \dots, \alpha R(\theta_l, d_1) + (1 - \alpha)R(\theta_l, d_2)\big) \end{aligned}$$

□



Theorem 4. *If the prior is*

$$\Lambda(\mathrm{d}\theta) = (\lambda_1, \lambda_2, \dots, \lambda_l)$$

where $\sum_j \lambda_j = 1$ *and* $\lambda_j \geq 0$*, and* d_0 *is the Bayes rule with respect to* Λ*, then* d_0 *has to be on the boundary of* R*.*

As a motivation, imagine this in 2 dimensions.

$$R(\Lambda, d) = \lambda_1 R(\theta_1, d) + \lambda_2 R(\theta_2, d)$$

Definition 6 (Generalized Bayes Rule)**.** *If d_i is the Bayes rule with respect to Λ_i and $\Lambda_i \longrightarrow \Lambda$ where Λ might not necessarily be proper, then $d_i \longrightarrow d^*$ is called the generalized Bayes rule.*

10.1.1. *Counter-example.* The generalized Bayes rule is not neessarily admissible! Suppose we have

$$(X_1, Y_1, Z_1) \overset{iid}{\sim} \mathcal{N}((\mu_1, \mu_2, \mu_3), I)$$
$$\vdots$$
$$(X_n, Y_n, Z_n) \overset{iid}{\sim} \mathcal{N}((\mu_1, \mu_2, \mu_3), I)$$

then $(\bar{X}, \bar{Y}, \bar{Z})$ is not admissible under L^2 loss

In the mulivariate Normal density, we have a term $v'\Sigma^{-1}v$. Through properties of the trace, we have

$$\begin{aligned} v'\Sigma^{-1}v &= \operatorname{tr}(v'\Sigma^{-1}v) \\ &= \operatorname{tr}(\Sigma^{-1}vv') \end{aligned}$$

And so

$$\begin{aligned} \mathbb{P}(Y|\mu, \Sigma) &\propto |\Sigma|^{-1/2} \exp\left\{-\frac{1}{2}(y-\mu)'\Sigma^{-1}(y-\mu)\right\} \\ &\propto |\Sigma|^{-1/2} \exp\left\{-\frac{1}{2} \operatorname{tr}\left(\Sigma^{-1}(y-\mu)(y-\mu)'\right)\right\} \end{aligned}$$

and so

$$\mathbb{P}(Y_1, \dots, Y_n|\mu, \Sigma) \propto |\Sigma|^{-n/2} \exp\left\{-\frac{1}{2} \operatorname{tr}\left(\Sigma^{-1}\sum_{i=1}^n (y_i-\mu)(y_i-\mu)'\right)\right\}$$

Note that

$$\sum_{i=1}^n (y_i-\mu)(y_i-\mu)' = \sum_{i=1}^n (y_i-\bar{y})(y_i-\bar{y})'$$

and so

$$\mathbb{P}(Y_1, \dots, Y_n|\mu, \Sigma) \propto |\Sigma|^{-n/2} \exp\left\{-\frac{1}{2}\left[\operatorname{tr}(\Sigma^{-1}S) + n(\bar{y}-\mu)'\Sigma^{-1}(\bar{y}-\mu)\right]\right\}$$

where $S = \sum_{i=1}^n (y_i-\bar{y})(y_i-\bar{y})'$

11. October 9th, 2014

$$\boldsymbol{X} \sim \operatorname{MVN}(\boldsymbol{\mu}, \Sigma)$$

If we decompose this into

$$\begin{pmatrix} \boldsymbol{X}_1 \\ \boldsymbol{X}_2 \end{pmatrix} \sim \operatorname{MVN}\left(\begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix}, \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}\right)$$

Then

$$\boldsymbol{X}_1|\boldsymbol{X}_2 = \boldsymbol{x}_2 \sim \operatorname{MVN}(\boldsymbol{\mu}_1^*, \Sigma_{11}^*)$$

where

$$\begin{aligned} \boldsymbol{\mu}_1^* &= \boldsymbol{\mu}_1 + \Sigma_{12}\Sigma_{22}^{-1}(\boldsymbol{X}_2 - \boldsymbol{\mu}_2) \\ \Sigma_{11}^* * &= \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} \end{aligned}$$

If $\boldsymbol{X}_1, \dots, \boldsymbol{X}_n \overset{iid}{\sim} \operatorname{MVN}(\boldsymbol{\mu}_p, \Sigma_{p\times p})$, we can get the MLEs $\hat{\boldsymbol{\mu}}_p, \hat{\Sigma}_{p\times p}$.

11.1. **Multinomial and Dirichlet.**

$$y_i, \dots, y_n \overset{iid}{\sim} \operatorname{Mult}(1, \theta_1, \dots, \theta_k)$$

where $\theta_1 + \cdots + \theta_k = 1$ and $\theta_j > 0$. We can think of this as a bunch of bins side by side, with width θ_j and we randomly drop a ball into a bin randomly.

Write $y_i = (y_{i1}, \dots, y_{ik})$. Then the likelihood is

$$\begin{aligned} \mathcal{L}(\theta|y) &= \prod_{i=1}^n \{\theta_1^{y_{i1}} \cdots \theta_k^{y_{ik}}\} \\ &= \theta_1^{\sum_{i=1}^n y_{i1}} \cdots \theta_k^{\sum_{i=1}^n y_{ik}} \\ &= \theta_1^{n_1} \cdots \theta_k^{n_k} \end{aligned}$$

where $n_j = \#\{y_i = j\}$ and $\sum_{i=1}^n n_j = n$.

Suppose $(s_1, \ldots, s_p)$ is a partition of $\{1, 2, \ldots, k\}$. Then let

$$X_i = \begin{cases} 1 & y_i \in s_1 \\ \vdots & \vdots \\ p & y_i \in s_p \end{cases}$$

Then $X_i \sim \text{Mult}(1; \theta_{s_1}, \ldots, \theta_{s_p})$ where $\theta_{s_j} = \sum_{i \in s_j} \theta_i$.

We also have by "**merging**"

$$\begin{pmatrix} N_1 \\ \vdots \\ N_k \end{pmatrix} \sim \text{Mult}\left(n, (\theta_1, \ldots, \theta_k)\right)$$

where $(N_1, \ldots, N_k)$ is a count vector pf $(y_1 \ldots, y_n)$. Then we have

$$\mathbb{P}\left(\begin{pmatrix} N_1 \\ \vdots \\ N_k \end{pmatrix} = \begin{pmatrix} n_1 \\ \vdots \\ n_k \end{pmatrix}\right) = \frac{n!}{n_1! \cdots n_k!} \theta_1^{n_1}, \cdots \theta_k^{n_k}$$

For the first case, the MLE is $\hat{\theta}_j = \frac{n_j}{n}$. If we set the MLE as

$$\hat{\theta}_j = \frac{n_j + \delta}{n + k\delta}$$

then we have a "pseudo count".

In Bayesian statistics, we have a prior on $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$.

$$\mathbb{P}_0\left(\boldsymbol{\theta}\right) \propto \theta_1^{\alpha_1 - 1} \cdots \theta_k^{\alpha_k - 1}$$

which is the Dirichlet distribution. The Dirichlet distribution is a probability distribution defined on a probability distribution, that is it is the distribution for a probability parameter.

$$\mathbb{P}_0\left(\boldsymbol{\theta}\right) = \frac{\Gamma(\alpha_1 + \cdots + \alpha_k)}{\Gamma(\alpha_1) \cdots \Gamma(\alpha_k)} \theta_1^{\alpha_1 - 1} \cdots \theta_k^{\alpha_k - 1}$$

where $0 \leq \theta_j \leq 1$ and $\sum_j \theta_j = 1$. This also requires $\alpha_j > 0$.

11.1.1. *Posterior Distribution.*

$$\begin{aligned} \mathbb{P}(\boldsymbol{\theta} | \boldsymbol{Y}) &\propto \theta_1^{n_1 + \alpha_1 - 1} \cdots \theta_k^{n_k + \alpha_k - 1} \\ &= \text{Dirichlet}(n_1 + \alpha_1, \ldots, n_k + \alpha_k) \end{aligned}$$

where we have

$$\begin{aligned} \mathbb{E}[\boldsymbol{\theta} | \boldsymbol{y}] &= \left(\mathbb{E}[\theta_1 | \boldsymbol{y}], \ldots, \mathbb{E}[\theta_k | \boldsymbol{y}]\right) \\ &= \left(\frac{n_1 + \alpha_1}{n + \|\alpha\|}, \ldots, \frac{n_j + \alpha_j}{n + \|\alpha\|}\right) \end{aligned}$$

where $\|\alpha\| = \alpha_1 + \cdots + \alpha_k$.

11.1.2. *Merging Property.* If $\theta_1' = \theta_1 + \cdots + \theta_\ell$, then

- $\boldsymbol{\theta}^* = (\theta_1', \theta_{\ell+1}, \ldots, \theta_k) \sim \text{Dirichlet}(\alpha_1 + \cdots + \alpha_\ell, \alpha_{\ell+1}, \ldots, \alpha_k)$
- $\theta_j \sim \beta(\alpha_j, \|\alpha\| - \alpha_j)$

Sketch of Proof. Suppose we have $\theta_1 + \theta_2 = \gamma$. Then

$$\begin{aligned} \theta_1 &= \gamma\delta \\ \theta_2 &= \gamma(1 - \delta) \end{aligned}$$

where $\delta = \frac{\theta_1}{\theta_1 + \theta_2}$. Then our density is

$$(\gamma\delta)^{\alpha_1 - 1}(\gamma(1 - \delta))^{\alpha_2 - 1} = \delta^{\alpha_1 - 1}(1 - \delta)^{\alpha_2 - 1}\gamma^{\alpha_1 + \alpha_2 - 2}$$

We must do a change of variables,

$$\begin{aligned}\frac{\mathrm{d}\theta_1\mathrm{d}\theta_2}{\mathrm{d}\gamma\mathrm{d}\delta} &= \begin{vmatrix}\delta & \gamma \\ 1-\delta & -\gamma\end{vmatrix} \\ &= |-\delta\gamma - \gamma + \delta\gamma| \\ &= \gamma\end{aligned}$$

and so our density is simply

$$\delta^{\alpha_1-1}(1-\delta)^{\alpha_2-1}\gamma^{\alpha_1+\alpha_2-2}\gamma$$

□

If $(s_1, \dots, s_p)$ is a partition of $\{1, \dots, k\}$, that is

$$\begin{gathered} s_i \cap s_j = \emptyset \\ s_1 \cup \cdots \cup s_p = \{1, \dots, k\}\end{gathered}$$

then

$$(\theta_{s_1}, \dots, \theta_{s_p}) \sim \text{Dirichlet}(\alpha_{s_1}, \dots, \alpha_{s_p})$$

where

$$\begin{aligned}\theta_{s_j} &= \sum_{\ell \in s_j} \theta_\ell \\ \alpha_{s_j} &= \sum_{\ell \in s_j} \alpha_\ell\end{aligned}$$

If $(\boldsymbol{\theta} = (\theta_1, \dots, \theta_k) \sim \text{Dirichlet}(\alpha_1, \dots, \alpha_k)$, then

$$\begin{aligned}\mathbb{E}[\theta_j] &= \frac{\alpha_j}{\|\alpha\|} \\ \text{Var}(\theta_j) &= \frac{\alpha_j(\|\alpha\| - \alpha_j)}{\|\alpha\|^2(\|\alpha\|+1)}\end{aligned}$$

To get $\text{Cov}(\theta_j, \theta_k)$, we can use the fact that

$$\text{Var}(\theta_j + \theta_k) = \frac{(\alpha_j + \alpha_k)(\|\alpha\| - \alpha_j - \alpha_k)}{\|\alpha\|^2(\|\alpha\|+1)}$$

which gives

$$\text{Cov}(\theta_j, \theta_k) = -\frac{\alpha_j\alpha_k}{\|\alpha\|(\|\alpha\|+1)}$$

11.1.3. *Dirichlet Process.* Suppose D is the domain of P, an unknown probability distribution. We say $P \sim \mathcal{D}(\alpha)$, where α is a nonnegative finite measure on D, then this is called a Dirichlet process with base measure α if $\forall$ measurable finite partition of D, $\{A_1, \dots, A_k\}$, i.e. $A_i \cap A_j = \emptyset$, $\bigcup_j A_j = D$

$$(P(A_1), \dots, P(A_k)) \sim \text{Dirichlet}(\alpha(A_1), \dots, \alpha(A_k))$$

12. October 16th, 2014

12.1. **Dirichlet Distribution.** $\theta = (\theta_1, \dots, \theta_k)$, with $\sum_j \theta_j = 1$ with $\theta_j \geq 0$. We can think of this by imagining an interval on $[0,1]$ and dividing it into slots, where n_j is the number of points in slot j. Then we have

$$\hat{\theta}_j = \frac{n_j}{n}$$

and the prior on $\boldsymbol{\theta}$ is

$$\boldsymbol{\theta} \sim \text{Dirichlet}(\boldsymbol{\alpha})$$

And so

$$\boldsymbol{\theta}|\text{Data} \sim \text{Dirichlet}(\boldsymbol{\alpha} + \boldsymbol{n})$$

where $\boldsymbol{n} = (n_1, \dots, n_k)$.

12.2. **Dirichlet Process.**

Definition 7 (Dirichlet Process)**.** *Suppose P is a probability distribution on $\mathcal{R}$. We say $P \in \mathcal{D}(\alpha)$ if for any measurable finite parttion of $\mathcal{R}$, $(A_1, \ldots, A_k)$, we have*

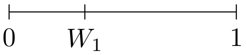
$$(P(A_1), \ldots, P(A_k)) \sim \text{Dirichlet}(\alpha(A_1), \ldots, \alpha(A_k))$$

Example 13. *if $A \subset \mathcal{R}$, then*

$$P(A) \sim \text{Beta}(\alpha(A), \|\alpha\| - \alpha(A))$$

12.2.1. *How to Simulate $P \sim \mathcal{D}(\alpha)$.*

1. Draw $X_1, X_2, \ldots, X_n, \ldots \overset{iid}{\sim} \frac{\alpha}{\|\alpha\|}$.
2. This is our so-called "stick breaking" process. Draw $W_1 \sim \text{Beta}(1, \|\alpha\|)$.



3. Draw $W_2 = (1 - W_1)Z_2$ where $Z_2 \sim \text{Beta}(1, \|\alpha\|)$.
4. Draw $W_3 = (1 - W_2 - W_2)Z_3$ where $Z_3 \sim \text{Beta}(1, \|\alpha\|)$
5. $P = \sum_{j=1}^{\infty} W_j \delta_{X_j}$

Example 14. *If we draw $\theta \sim \text{Dirichlet}(\boldsymbol{\alpha})$ and then draw $X \sim \text{Mult}(1, \theta)$, then*

$$\mathbb{E}[\mathbb{P}(X = j|\theta)] = \mathbb{E}[\theta_j] = \frac{\alpha_j}{\|\alpha\|}$$

The reasoning for this is as follows:

$$\begin{aligned} \mathbb{P}(X \in A) &= \mathbb{E}[\mathbb{1}_{X \in A}] \\ &= \mathbb{E}[\mathbb{E}[\mathbb{1}_{X \in A}|P]] \\ &= \mathbb{E}[P(A)] \\ &= \frac{\alpha(A)}{\|\alpha\|} \end{aligned}$$

Now, if we have $X_1, \ldots, X_n \overset{iid}{\sim} P$ with $P \sim \mathcal{D}(\alpha)$ then

$$\mathbb{P}(P|X_1, \ldots, X_N) = \mathcal{D}(\alpha + \delta_{X_1} + \delta_{X_2} + \cdots + \delta_{X_n})$$

What is $[X_2|X_1]$? We know that

$$P|X_1 \sim \mathcal{D}(\alpha + \delta_{X_1})$$

So this gives us

$$X_2 \sim \frac{\alpha + \delta_{X_1}}{\|\alpha\| + 1}$$

and so we know

$$\mathbb{P}(X_1 = X_2) \geq \frac{1}{\|\alpha\| + 1}$$

Now we know that

$$X_{n+1}|X_1, \ldots, X_n \sim \frac{\alpha + \delta_{X_1} + \cdots + \delta_{X_n}}{\|\alpha\| + n}$$

Definition 8 (Chinese Restaurant Process)**.** *Each new person has a probability $\frac{\|\alpha\|}{\|\alpha\|+n}$ to take a new table and a probability proportional to size to going to an existing table.*

$$\begin{aligned} \hat{p} &= \mathbb{E}[P|X_1, \ldots, X_n] \\ &= \frac{\alpha + \sum_{j=1}^{n} \delta_{X_j}}{\|\alpha\| + n} \end{aligned}$$

Note that if $\alpha \to 0$, we get the empirical distribution.

12.3. **Bayesian Variable Selection & Hypothesis Testing.** Suppose we have data D and we are choosing between candidate models, $M_1, \ldots, M_k$. Each model may have its own set of parameters, θ_j for M_j. The different θ_j's may have different numbers of dimensions. Under the Bayesian framework, we have proper priors $p_j(\theta_j)$ for each model and also prior probability π_j for M_j, then

$$\begin{aligned}\mathbb{P}(M_j|D) &= \frac{\mathbb{P}(D|M_j)\,\mathbb{P}(M_j)}{\sum_{h=1}^{k}\mathbb{P}(D|M_h)\,\mathbb{P}(M_h)} \\ &\propto \pi_j \int \mathbb{P}_{M_j}\left(D|\theta_j\right)\mathbb{P}_j\left(\theta_j\right)\,\mathrm{d}\theta_j\end{aligned}$$

where $\pi_j = \mathbb{P}(M_j)$.

Sometimes we wish to consider

$$\frac{\int \mathbb{P}_{M_j}\left(D|\theta_j\right)\mathbb{P}_j\left(\theta_j\right)\,\mathrm{d}\theta_j}{\int \mathbb{P}_{M_\ell}\left(D|\theta_\ell\right)\mathbb{P}_\ell\left(\theta_\ell\right)\,\mathrm{d}\theta_\ell} \equiv \frac{\mathbb{P}(D|M_j)}{\mathbb{P}(D|M_\ell)}$$

which is called the "Bayes Factor".

Example 15. *Suppose we have $Y \sim \mathcal{N}\left(\theta, \frac{1}{n}\right)$ and we are testing for $H_0 : \theta = 0$, $H_1 : \theta \neq 0$. In the frequentist case, we reject if Y is outside of $[-1.96/\sqrt{n}, 1.96/\sqrt{n}]$.*

For the Bayes factor case, we have

$$\begin{aligned}\mathbb{P}(Y|H_0) &= \frac{\sqrt{n}}{\sqrt{2\pi}}e^{-\frac{ny^2}{2}} \\ \mathbb{P}(Y|H_A) &= \int_{-\infty}^{\infty}\frac{\sqrt{n}}{\sqrt{2\pi}}e^{-\frac{n(y-\theta)^2}{2}}\frac{1}{\sqrt{2\pi\sigma_0^2}}e^{-\frac{\theta-\mu_0)^2}{2\sigma_0^2}}\,\mathrm{d}\theta \\ &= \frac{1}{\sqrt{2\pi}}\frac{\sqrt{n}}{\sqrt{n\sigma_0^2+1}}e^{-\frac{1}{2}\frac{(y-\mu_0)^2}{\sigma_0^2+\frac{1}{n}}} \\ \frac{\mathbb{P}(Y|H_A)}{\mathbb{P}(Y|H_0)} &= \frac{1}{\sqrt{1+n\sigma_0^2}}e^{\frac{1}{2}\frac{n\sigma_0^2y^2}{1+n\sigma_0^2}} \qquad (\mu_0 = 0)\end{aligned}$$

(These calculations were messy... Jun wasn't sure if we was right.)

Suppose we observe $\sqrt{n}y = 2$, that is we will reject H_0 in the frequentist case. Note that $\frac{\mathbb{P}(Y|H_A)}{\mathbb{P}(Y|H_0)}$ depends on n.

13. October 20th, 2014

13.1. **Model Selection.** In the notes, we should have

$$\tilde{B} = \frac{\mathbb{P}(X|H_1)}{\mathbb{P}(X|H_0)} = O\left(\frac{1}{\sqrt{n}}\right)$$

n_{11}	n_{12}	$n_{1\cdot}$
n_{21}	n_{22}	$n_{2\cdot}$
$n_{\cdot 1}$	$n_{\cdot 2}$	

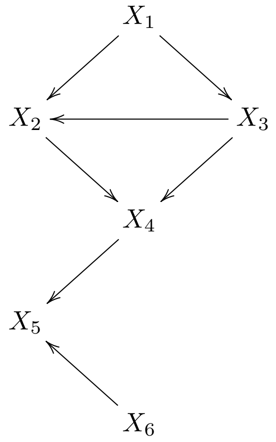
Then we have a $\mathrm{Bin}(n_1, p_1)$ and $\mathrm{Bin}(n_2, p_2)$ and testing of independence is equivalent to testing $p_1 = p_2$.

13.2. **Brief Introduction of condition Independent Graph (Graphical Model, BN).** A set of random variables $X_1, X_2, \ldots, X_p$. An undirected graph with p nodes satisfies the property that

$$X_k|X_{-k}X_k|X_{\partial k}$$

where ∂k is the neighbours of X_k.

13.3. **Directed Acyclic Graph (DAG).**



Then the joint distribution can be written as

$$\mathbb{P}(X_1, \dots, X_6) = \mathbb{P}(X_1)\,\mathbb{P}(X_2|X_1, X_3)\,\mathbb{P}(X_3|X_1)\,\mathbb{P}(X_4|X_2, X_3)\,\mathbb{P}(X_5|X_4, X_6)\,\mathbb{P}(X_6)$$

Definition 9 (Moralization)**.** *In the above example X_6 is not conditionally independent on the rest. A simple example of this is in families. Parents are not independent once they have children.*

If we have the undirected graph

$$X_1 \longrightarrow\!\!\!\!\!\!- X_2 - X_3$$

Then this is equivalent to all of the following

$$X_1 \longrightarrow X_2 \longrightarrow X_3$$

$$X_1 \longleftarrow X_2 \longrightarrow X_3$$

$$X_1 \longleftarrow X_2 \longleftarrow X_3$$

13.4. **Regression Model Selection/Variable Selection.** Suppose we have Y, $X = (X_1, \dots, X_p)$. Let $S \subset \{1, \dots, p\}$ and $X_S = \{X_j, j \in S\}$. Our model, M_S is

$$Y = \begin{pmatrix} \mathbf{1} & X_S \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_S \end{pmatrix} + \varepsilon$$

Then, given a prior $\pi(\beta_0, \beta_S, \sigma^2)$, we can calculate $\mathbb{P}(Y|X, M_S)$.

Practically some methods for this are

(1) Stepwise regression
(2) Lasso, "penalized regression"
(3) MCMC: Introduce $I(S) = (0, 0, 1, 1, 0, \dots)$ where

$$I(S)_j = \begin{cases} 0 & j \in S \\ 1 & \text{otherwise} \end{cases}$$

$Y|X, I_S, \beta)$

14. October 23rd, 2014

Suppose we have θ, $\boldsymbol{Y}$, and $\mathbb{P}(\theta|\boldsymbol{Y})$. We want to mimic the classic hypothesis testing procedure is to form a test statistic $T(\boldsymbol{Y})$.

Definition 10 (Posterior Predictive Testing)**.**

$$\boldsymbol{y}^{(1)}, \dots, \boldsymbol{y}^{(n)} \sim \int f(\boldsymbol{y}|\boldsymbol{\theta})\mathbb{P}(\boldsymbol{\theta}|\boldsymbol{y}_{obs})\,\mathrm{d}\boldsymbol{\theta}$$

In real life, this is a difficult thing to find, so computationally, we find

$$\boldsymbol{\theta}^{(i)} \sim \mathbb{P}(\boldsymbol{\theta}|\boldsymbol{y})$$

$$\boldsymbol{y}^{(i)} \sim f(\boldsymbol{y}|\boldsymbol{\theta}^{(i)})$$

Then compute

$$T(\boldsymbol{y}^{(1)}), \dots, T(\boldsymbol{y}^{(n)})$$

with $T(\boldsymbol{y}_{obs})$. Then the posterior predictive p-value is

$$\mathbb{P}(T(\boldsymbol{y}) \geq T(\boldsymbol{y}_{obs})|\boldsymbol{y}_{obs})$$

Definition 11 (Deviance)**.** *The deviance is defined as*

$$D_{obs}(\theta) = -2\log \mathbb{P}(y_{obs}|\theta)$$

It measures the prediction accuracy of the model.

The goal is to estimate the prediction accuracy for new y. But, $D_{\hat{\theta}}(\boldsymbol{y}) = -2\log \mathbb{P}\Big(\boldsymbol{y}|\hat{\theta}\Big)$ is too optimistic. We may instead use:

- AIC: $-2\log \mathbb{P}\Big(y_{obs}|\hat{\theta}\Big) + 2p$
- BIC: $-2\log \mathbb{P}\Big(y_{obs}|\hat{\theta}\Big) + p\log n$

Suppose we have $\|\boldsymbol{\theta}\|_{L_0}$. If $X \sim \mathcal{N}(\theta, \frac{1}{n})$ with $\theta > 0$, then how many parameters do we have? It depends. The effective number of parameters is

$$p_D^{(1)} = \hat{D}_{ave}(y_{obs}) - D_{\hat{\theta}}(y_{obs})$$

where

$$\hat{D}_{ave}(y_{obs}) = \mathbb{E}[D_\theta(y_{obs})|y_{obs}]$$
$$D_{\hat{\theta}}(y_{obs}) = \min_\theta D_\theta(y_{obs})$$

Then, for somethign we cal DIC, we have the post. is

$$D_{\hat{\theta}}(y_{obs}) + 2\left\{\hat{D}_{ave}(y) - D_{\hat{\theta}}(y_{obs})\right\} = 2\hat{D}_{ave}(y) - D_{\hat{\theta}}(y_{obs})$$

Suppose we have

$$X_1, \ldots, X_n \overset{iid}{\sim} p\mathcal{N}(\mu_1, \sigma_1^2) + (1-p)\mathcal{N}(\mu_2, \sigma_2^2)$$

We can estimate these parameters using the EM algorithm or even by doing an MLE.

15. October 27th, 2014

If we have a mixture

$$X_1, \ldots, X_n \sim \rho\mathcal{N}(\mu_1, \sigma_1^2) + (1-\rho)\mathcal{N}(\mu_2, \sigma_2^2)$$

how would we compare this to a $\mathcal{N}(\mu, \sigma^2)$? One particular way to determine which distribution the data come from, we would perform a likelihood-ratio test. Note that the mixture model is equivalent to a single Normal random variable if we have any of

- $\mu_1 = \mu_2$, $\sigma_1^2 = \sigma_2^2$
- $\rho = 0$
- $\rho = 1$

We have some Bayesian approaches:

(1) Bayes factor approach: we would need to calculate

$$\mathbb{P}(X_1, \ldots, X_n|\text{mixture}) = \int \mathbb{P}\big(X|\rho, \mu, \sigma^2\big)\, \pi_0(\rho, \mu\sigma^2)\, \mathrm{d}\rho \mathrm{d}\mu \mathrm{d}\sigma^2$$

, which is difficult to do, vs

$$\mathbb{P}(X_1, \ldots, X_n|\text{null}) = \int \mathbb{P}\big(X|\mu, \sigma^2\big)\, \pi_0(\mu, \sigma^2)\, \mathrm{d}\mu \mathrm{d}\sigma^2$$

(2) Posterior predictive checking: we consider $\mathbb{P}(\boldsymbol{X}|X_{obs})$ and look at some test statistic $T(\boldsymbol{X})$ and compare it with $T(X_{obs})$.

15.0.1. *Estimation of the Mixture Distribution.* To do estimate of the mixture distribution, it can be helpful to introduce an indicator random variable I_i, since we are sampling from exactly one Normal distribution for each data point. Let

$$X_i|I_i = k \sim \mathcal{N}(\mu_k, \sigma_k^2)$$
$$\mathbb{P}(I_i = 1) = \rho$$

Suppose we have

$\boldsymbol{Y}_{obs}$	"observed data"
$\boldsymbol{Y}_{mis}$	"missing data"
$\boldsymbol{Y}_{comp}$	$(\boldsymbol{Y}_{obs}, \boldsymbol{Y}_{mis})$

Calculating $\mathbb{P}(\boldsymbol{Y}_{comp}|\Theta)$ is "easy". Of interest is to find

(1) Observed data likelihood and MLE, $\mathbb{P}(\boldsymbol{Y}_{obs}|\theta)$
(2) Posterior distribtuion $\mathbb{P}(\theta|Y_{obs})$.

Example 16. *Suppose $Y_1, \ldots, Y_n \overset{iid}{\sim} \mathcal{N}(\theta, \sigma^2)$ and $(Y^*_{n+1}, \ldots, Y^*_{n+m})$ are missing. If you believe that the estimate for each of the missing data are $\bar{y}$, then the overall variance is reduced.*

Another way around this is to compute $\mathbb{P}\Big(Y^{(t)}_{mis} | Y_{obs}, \hat{\theta}_t\Big)$ and then find $\hat{\theta}_{t+1} | Y_{obs}, Y^{(t)}_{mis}$. However, this was found to be wrong in many cases. In the EM paper, it was found that to solve this, you use

$$Q = \mathbb{E}\Big[\log \mathbb{P}(Y_{obs}, Y_{mis} | \theta) \, | \theta^{(t)}\Big]$$

Then the next step is to maximize $\theta^{(t+1)}$.

Back to the above mixture model, we have

$$\mathbb{P}\Big(I_i = 1 | X_i, \theta^{(t)}\Big) = \frac{\rho \times \varphi(x_i; \hat{\mu}^{(t)}_1, \{\hat{\sigma}^2_1\}^{(t)})}{\rho \times \varphi(x_i; \hat{\mu}^{(t)}_1, \{\hat{\sigma}^2_1\}^{(t)}) + (1 - \rho) \times \varphi(x_i; \hat{\mu}^{(t)}_2, \{\hat{\sigma}^2_2\}^{(t)})}$$

We start with $\rho^{(0)}, \mu^{(0)}_1, \sigma^{(0)}_1, \mu^{(0)}_2, \sigma^{(0)}$. Then we iterate

(1) $\mathbb{P}\big(I_i | \theta^{(t)}\big) = \omega^{(t)}_i$

(2) $\mu^{(t+1)}_1 = \frac{\sum \omega^{(t)}_i x_i}{\sum \omega^{(t)}_i}$, $\sigma_1^{(t+1^2} = \ldots$

In the Bayesian case, we have priors onto

$$(\rho, \mu^2_1, \sigma^2_1, \mu^2_2, \sigma^2_2) \propto \underbrace{\pi_{01}(\rho)}_{\text{Beta}} \underbrace{\pi_{02}(\mu_1, \sigma^2_1)}_{\text{N-Inv-}\chi^2} \underbrace{\pi_{03}(\mu_2, \sigma^2_2)}_{\text{N-Inv-}\chi^2}$$

(1) Try to use proper priors

(2) Posterior sampling? Brute force

$$\mathbb{P}(\boldsymbol{\theta} | X_1, \ldots, X_n) \propto \prod_{i=1}^{n} \{\rho \mathcal{N}(\ldots) + (1 - \rho) \mathcal{N}(\ldots)\} \, \pi_0(\ldots)$$

15.1. **Data Augmentation.** This makes use of the missing data structure.

$$\mathbb{P}(\theta | \boldsymbol{Y}_{obs}) = \int \mathbb{P}(\theta | \boldsymbol{Y}_{obs}, \boldsymbol{Y}_{mis}) \, \mathbb{P}(\boldsymbol{Y}_{mis} | \boldsymbol{Y}_{obs}) \, \mathrm{d}\boldsymbol{Y}_{mis}$$

i.e., if we can draw $Y^{(1)}_{mis}, \ldots, Y^{(m)}_{mis} \sim \mathbb{P}(\boldsymbol{Y}_{mis}, \boldsymbol{Y}_{obs})$, then

$$\hat{\mathbb{P}}(\theta | \boldsymbol{Y}_{obs}) = \frac{1}{m} \sum_{j=1}^{m} \mathbb{P}\Big(\theta | \boldsymbol{Y}_{obs}, \boldsymbol{Y}^{(j)}_{mis}\Big)$$

Once we have this, we get

$$\mathbb{P}(\boldsymbol{Y}_{mis} | \boldsymbol{Y}_{obs}) = \int \mathbb{P}(\boldsymbol{Y}_{mis} | \boldsymbol{Y}_{obs}, \theta) \, \mathbb{P}(\theta | \boldsymbol{Y}_{obs}) \, \mathrm{d}\theta$$

Similarly, if we can draw $\theta^{(1)}, \ldots, \theta^{(m)} \sim \mathbb{P}(\theta | \boldsymbol{Y}_{obs})$, then

$$\hat{\mathbb{P}}(\boldsymbol{Y}_{mis} | \boldsymbol{Y}_{obs}) = \frac{1}{m} \sum_{j=1}^{m} \mathbb{P}\Big(\boldsymbol{Y}_{mis} \Big| \boldsymbol{Y}_{obs}, \theta^{(j)}\Big)$$

When $m = 1$, the data augmentation algorithm is a two-component Gibbs sampler.

$$\begin{tikzcd}
\theta^{(0)} \arrow[dr] & & \theta^{(1)} \arrow[dr] & & {} \\
& \boldsymbol{Y}^{(1)}_{mis} \arrow[ur] & & \boldsymbol{Y}^{(2)}_{mis} \arrow[ur] &
\end{tikzcd}$$

Each update follows the corresponding conditional distribution. More generally, we want to draw $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_d) \sim \pi(\boldsymbol{\theta})$. We can iterate the following conditional updates

$$\theta^{(t)}_1 \sim \pi(\theta_1 | \theta^{(t)}_{-1})$$
$$\theta^{(t+1)}_2 \sim \pi(\theta_2 | \theta^{(t+1)}_1, \theta^{(t)}_{-1-2})$$
$$\vdots$$

that is, we draw $i \in \{1, \ldots, d\}$ at random and update

$$\theta^{(t+1)}_i \in \pi(\theta_i, \theta^{(t)}_{-i})$$

15.1.1. *Mixture Case.* We can draw $I_i^{(t)} \sim \mathbb{P}\Big(I_i | X_i, \boldsymbol{\theta}^{(t)}\Big)$, then draw $\boldsymbol{\theta}^{(t+1)} \sim \mathbb{P}\Big(\theta | x_i, I_i^{(t)}\Big)$.

16. October 20th, 2014

Try to use EM or Gibbs sampler to fit a mixture model. It is important to know how to do this.

16.1. **Model-based Clustering.** Another name for model-based clustering is "unsupervised learning". Examples of this are:

- Hierarchical Clustering
- k-means

Some criteria are

(a) From Calinski and Hartigen ('74), we wish to maximize

$$\max \mathrm{CH}(k) = \frac{B(k)/(k-1)}{W(k)/(n-k)}$$

(b) In Hartigen ('75),

$$H(k) = \left(\frac{W(k)}{w(k+1)} - 1\right)(n-k-1)$$

Stop when $H(k) < 0$.

(c) Average silhouette criterion

$$s(\cdot) = \frac{b(i) - a(i)}{\max(a(i), b(i))}$$

(d) AIC, BIC

(e) Posterior Predictive Checking

(f) Full Bayesian

where (d), (e), and (f) are for parametric mixture models.

16.2. **Full Bayes Model.** Assume M is the number of components, which follows some prior distribution $\pi_0(M)$. Then

$$\begin{aligned} X_1, \dots, X_n &\overset{iid}{\sim} \lambda_1 f(X|\theta_1) + \cdots + \lambda_M f(X|\theta_M) \\ (\lambda_1, \dots, \lambda_M) &\overset{iid}{\sim} \mathrm{Dirichlet}(\alpha_1, \dots, \alpha_M) \\ \theta_m &\sim \mathbb{P}_m(\theta_m) \end{aligned}$$

So we have

$$\mathbb{P}(\boldsymbol{X}, M, \Lambda_M, \Theta_M) = \mathbb{P}(\boldsymbol{X}|M, \Lambda_M, \Theta_M)\,\pi_0(M)\pi(\Lambda_M|M)\mathbb{P}(\Theta_M|M)$$

16.3. **Dirichlet Process Mixture.**

$$\begin{aligned} y &\sim f(y|\theta) \\ \theta &\sim G(\cdot) \\ G &\sim \mathrm{Dirichlet}(\boldsymbol{\alpha}) \end{aligned}$$

Then the mixture can be written as

$$\int f(y|\theta) G(\mathrm{d}\theta)$$

Fun fact: the t distribution is a continuous mixture of Normal distributions.
Proof: `http://www.johndcook.com/t_normal_mixture.pdf`

16.4. **Empirical Bayes.** Suppose we have

$$\begin{aligned} \mathbb{P}(X|\lambda) &= \mathbb{P}(X = x|\Lambda = \lambda) \\ \Lambda &\sim G(\cdot) \end{aligned}$$

By Bayes theorem, we have that our "prior" (which is where the empirical part comes from) is

$$\hat{\Lambda}_G = \frac{\int \lambda \mathbb{P}(x|\lambda)\,G(\mathrm{d}\lambda)}{\int \mathbb{P}(x|\lambda)\,G(\mathrm{d}\lambda)}$$

Notation:

$$\mathbb{P}_G(x) = \int \mathbb{P}(X = x|\Lambda = \lambda)\,G(\mathrm{d}\lambda)$$

which is the marginal distribution. In practice, we don't have any observations Λ. However, we have

$$\begin{aligned}
X_1 &\sim \mathbb{P}(x_1|\lambda_1)\\
X_2 &\sim \mathbb{P}(x_2|\lambda_2)\\
&\vdots\\
X_k &\sim \mathbb{P}(x_k|\lambda_k)
\end{aligned}$$

where $\lambda_j \overset{iid}{\sim} G$ and are unobserved.

Question: How do we estimate λ_{k+1} given X_{k+1}?

Example 17. *Suppose $X_i \sim \mathrm{Pois}(\lambda_i)$ and $\lambda_i \sim G$. Then we have that*

$$\begin{align}
\hat{\Lambda}_G(x) &= \frac{\int \lambda \frac{\lambda^x}{x!} e^{-\lambda} G(\mathrm{d}\lambda)}{\int \frac{\lambda^x}{x!} e^{-\lambda} G(\mathrm{d}\lambda)} \tag{1}\\
&= \frac{(x+1)\int \frac{\lambda^{x+1}}{(x+1)!} e^{-\lambda} G(\mathrm{d}\lambda)}{\mathbb{P}_G\,(x)} \tag{2}\\
&= \frac{(x+1)\mathbb{P}_G\,(x+1)}{\mathbb{P}_G\,(x)} \tag{3}\\
&\approx \frac{(x+1)\hat{\mathbb{P}}_G(x+1)}{\hat{\mathbb{P}}_G(x)} \tag{4}
\end{align}$$

where in the last line we estimate (3) by the data using (4), that is

$$\hat{\mathbb{P}}_G(x) = \frac{\{\$X_i = x, i = 1, \dots, k+1\}}{k+1}$$

We can set n_x as the number of observations equal to x, so we can rewrite this as

$$\hat{\Lambda}_G(x) = \frac{(x+1)n_{x+1}}{n_x}$$

17. November 3rd, 2014

17.1. **Shakespeare Unseen Species Problem.** Let n_x be the number of distinct words that appeared exactly x times.

$$\begin{aligned}
n_1 &= 14375\\
n_2 &= 4343\\
&\vdots
\end{aligned}$$

The number of total words is

$$\sum_{x\geq 1} x n_x = 884647$$

and the number of distinct words is

$$\sum_{x\geq 1} n_x = 31534$$

Suppose in total there are S words, each with a parameter λ_s, $s = 1, 2, \dots, S$.

$$\lambda_s \sim G(\lambda)$$

We let X_s be the number of times words are being used, $X_s \sim \mathrm{Pois}(\lambda_s)$. Then we have

$$\begin{aligned}
\mathbb{P}(X_s = x) &= \int \frac{\lambda^x}{x!} e^{-\lambda}\, \mathrm{d}G(\lambda)\\
\mathbb{E}[\lambda_s | X_s = x] &= \int \frac{\lambda \frac{\lambda^x}{x!} e^{-\lambda}\, G(\mathrm{d}\lambda)}{\int \frac{\lambda^x}{x!} e^{-\lambda}\, G(\mathrm{d}\lambda)}\\
&= \frac{\int \frac{\lambda^{x+1}}{x!} e^{-\lambda}\, G(\mathrm{d}\lambda)}{p_x}\\
&= (x+1)\frac{p_{x+1}}{p_x}
\end{aligned}$$

Then we can estimate

$$\hat{\lambda}_s = (x+1)\frac{n_{x+1}}{n_x}$$

If Shakespeare were to write another "t" volumes, how many new words would he have used? If $X_s \sim \text{Pois}(\lambda_s)$, then we can think of a variable Y_s, which is independent of X_s such that $Y_s \sim \text{Pois}(\lambda_s t)$, which is the number of times word "s" would have been used.

$$\mathbb{E}[n_x] = s \cdot p_x = s \int \frac{\lambda^x}{x!} e^{-\lambda} \, G(\mathrm{d}\lambda)$$

The key is to find

$$\mathbb{P}(X_s = 0, Y_s > 0) = \int e^{-\lambda}(1 - e^{-\lambda t}) \, G(\mathrm{d}\lambda)$$

Recall that

$$e^{-\lambda t} = 1 - \lambda t + \frac{\lambda^2 t^2}{2!} - \frac{\lambda^3 t^3}{3!} + \cdots$$

And so

$$\begin{aligned}
\mathbb{P}(X_s = 0, Y_s > 0) &= \int e^{-\lambda} \left(\lambda t - \frac{\lambda^2 t^2}{2!} + \frac{\lambda^3 t^3}{3!} - \cdots \right) G(\mathrm{d}\lambda) \\
&= t \int \lambda e^{-\lambda} \, G(\mathrm{d}\lambda) - t^2 \int \frac{\lambda^2}{2!} e^{-\lambda} \, G(\mathrm{d}\lambda) + t^3 \int \frac{\lambda^3}{3!} e^{-\lambda} \, G(\mathrm{d}\lambda) - \cdots \\
&= tp_1 - t^2 p_2 + t^3 p_3 - t^4 p_4 + \cdots
\end{aligned}$$

If we let $m_0 = \sum_s I_{X_s = 0, Y_s > 0}$, then

$$\mathbb{E}[m_0] = s(tp_1 - t^2 p_2 + t^3 p_3 - \cdots)$$

so an estimate for m_0 is

$$\hat{m}_0 = tn_1 - t^2 n_2 + t^3 n_3 - \cdots$$

In the case where $t = 1$, then

$$\begin{aligned}
\hat{\Delta}(1) &= n_1 - n_2 + n_3 - \cdots \\
&= 11430 \\
\text{Var}(\hat{\Delta}(1)) &\approx n_1 + n_2 + n_3 + \cdots
\end{aligned}$$

and so we have

$$n_x = \sum_s I_{X_0 = x_0} \mathrel{\dot{\sim}} \text{Pois}(sp_x)$$

17.2. **Parametric Approach.** Assume $G \sim \Gamma(\alpha, \beta)$, that is

$$g(\lambda) = \frac{\lambda^{\alpha-1}}{\beta^\alpha \Gamma(\alpha)} e^{-\lambda/\beta}$$

Then we have

$$\begin{aligned}
p_x &= \mathbb{P}(X_s = x) \\
&= \int \frac{\lambda^x e^{-\lambda}}{x!} \frac{\lambda^{\alpha-1}}{\beta^\alpha \Gamma(\alpha)} e^{-\lambda/\beta} \, \mathrm{d}\lambda \\
&= \int \frac{\lambda^{x+\alpha-1}}{x! \beta^\alpha \Gamma(\alpha)} e^{-\left(\frac{1+\beta}{\beta}\right)\lambda} \, \mathrm{d}\lambda
\end{aligned}$$

which is the Beta-Binomial distribution. So we have

$$\begin{aligned}
\Gamma &= \frac{\beta}{1+\beta} \\
\Rightarrow \beta &= \frac{\gamma}{1-\gamma} \\
\Rightarrow \int \lambda^{x+\alpha-1} e^{-\lambda/\gamma} \, \mathrm{d}\gamma &= \gamma^{x+\alpha} \Gamma(x+\alpha)
\end{aligned}$$

and so

$$\begin{aligned}
p_x &= \frac{\gamma^{x+\alpha}\Gamma(x+\alpha)}{x!\Gamma(\alpha)} \frac{(1-\gamma)^\alpha}{\gamma^\alpha} \\
&= \frac{\Gamma(x+\alpha)}{x!\Gamma(\alpha)} \gamma^x (1-\gamma)^\alpha
\end{aligned}$$

which is where we get the name Negative Binomial. Then, we have

$$(n_1, \dots, n_{x_0}) \Bigg| \sum_{x=1}^{x_0} n_x = N_{x_0} \sim \text{Mult}(N_{x_0}, (p_1, \dots, p_{x_0}))$$

and we can estimate (α, γ) using the MLE. Then we get an estimate

$$\begin{aligned} \hat{\Delta}(1) &= S(p_1 - p_2 + \cdots) \\ &= 11483 \end{aligned}$$

which is a very accurate estimate.

17.3. **Gaussian Hierarchical Bayes Model.**

(1) Empirical Bayes: started by Stein's estimator. Suppose we have

$$X_i | \theta_i \sim \mathcal{N}(\theta_I, 1)$$

for $i = 1, \dots, k$. We wish to estimate $\Theta = (\theta_1, \dots, \theta_k)$. The MLE is simply $\hat{\Theta}_{MLE} = (x_1, \dots, x_k)$. Under the loss function $L(\Theta, \hat{\Theta}) = \sum_{i=1}^k (\theta_i - \hat{\theta}_i)^2$, we have

$$R(\Theta, \hat{\Theta}) = k$$

However, Stein showed for $k \geq 3$, then $\hat{\Theta}$ is not admissible.

$$\delta_i(x) = \mu_i + \left(1 - \frac{k-2}{\sum_j (x_j - \mu_j)^2}\right)(x_i - \mu_i)$$

Suppose $\mu = (\mu_1, \dots, \mu_k)$ is a fixed vector given in advance. Then

$$R(\Theta, \delta_i) \leq k - \frac{(k-2)^2}{(k-2) + \sum_{i=1}^k (\theta_i - \mu_i)^2}$$

18. November 6th, 2014

Art Dempster's lecture

19. November 10th, 2014

Art Dempster's lecture

20. November 13th, 2014

Suppose we have

$$\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\theta}_k, I_k)$$

The MLE $\hat{\boldsymbol{\theta}} = \boldsymbol{X}$. This is inadmissible under L^2 loss:

$$\begin{aligned} L(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}) &= \|\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}\|^2 \\ &= \sum_{i=1}^k (\theta_i - \hat{\theta}_i)^2 \end{aligned}$$

The expectation is

$$\mathbb{E}\Big[L(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}_{mle} | \boldsymbol{\theta}\Big] = k$$

20.1. **James-Stein Estimator.**

$$\delta_i(X) = \mu_i + \left(1 - \frac{k-2}{\sum_{j=1}^k (X_j - \mu_j)^2}\right)(X_i - \mu_i)$$

where $(\mu_1, \dots, \mu_k)$ is any vector. They showed that

$$R(\theta, \delta) \leq k - \frac{(k-2)^2}{k - 2 + \sum_{j=1}^k (\theta_j - \mu_j)^2}$$

and this value is less than k for $k > 2$. The improved version of this is to take

$$\delta_i(X) = \mu_i + \left(1 - \frac{k-2}{\sum_{j=1}^k (X_j - \mu_j)^2}\right)^+ (X_i - \mu_i)$$

that is, we take the positive part of $\left(1 + \frac{k-2}{\sum_{j=1}^k (X_j - \mu_j)^2}\right)$.

A more Bayesian view:

$$\begin{aligned}X_i &\sim \mathcal{N}(\theta_i, 1)\\ \theta_i &\sim \mathcal{N}(\mu_i, V)\end{aligned}$$

The posterior distribution is given by

$$\theta_i | X_i \sim \mathcal{N}(\mu_i^*, \sigma^{*2})$$

and

$$\begin{aligned}\hat{\theta}_i &= \mu_i^*\\ &= \frac{\frac{\mu_i}{V} + X_i}{1 + \frac{1}{V}}\\ &= \mu_i + \frac{V}{1+V}(X_i - \mu_i)\\ &= \mu_i + \left(1 - \frac{1}{1+V}\right)(X_i - \mu_i)\end{aligned}$$

20.2. **Empirical Bayes.** $X_i \sim \mathcal{N}(\theta_i, 1)$, where $\theta_i \sim \mathcal{N}(\mu, V)$. We have

$$\mathbb{E}[\theta_i | X_i, \mu, V] = \mu + \frac{V}{1+V}(X_i - \mu)$$

If we take the marginal of X_i, we have

$$X_i \sim \mathcal{N}(\mu, 1 + V)$$

and so we have

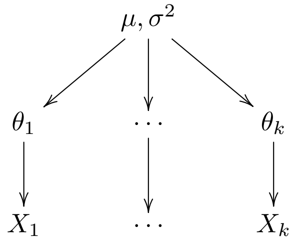
$$\begin{aligned}\hat{\mu} &= \bar{X}\\ \widehat{1+V} &= \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2\\ &= S_x^2\end{aligned}$$

so we can set $\hat{V} = (S_x^2 - 1)^+$.

20.2.1. *Empirical Bayes Estimator.* So now we have

$$\begin{aligned}\hat{\theta}_i &= \bar{X} + \frac{\hat{V}}{1+\hat{V}}(X_i - \bar{X})\\ &= \frac{\bar{X} + (S_x^2 - 1)^+ X_i}{1 + (S_x^2 - 1)^+}\end{aligned}$$

20.2.2. *Full Bayesian Approach.* In the full Bayesian approach, we assume a prior on (μ, V), namely $\pi(\mu, V)$. In the previous example, if we were to form a confidence interval for $\theta_i | X_i, \mu, V$, it would be too narrow because it does not account for the variability in μ and V. How do we pick a prior? In this case, we cannot use the Jeffrey's prior since it will lead to an improper posterior distribution. However, if we have $\pi(\mu, V) \propto c$ for some constant c, then the posterior distribution will be proper.



$$\pi(\theta_1, \ldots, \theta_j, \mu, \sigma^2) \propto \prod_{j=1}^{k}\left\{\varphi(x_j - \theta_j)\phi(\theta_j, \mu, \sigma^2)\right\} \times \pi(\mu, \sigma^2)$$

Then this gives

$$\boldsymbol{\theta}|\mu,\sigma^2,\boldsymbol{X} \sim \mathcal{N}\left(\frac{\frac{\mu}{\sigma^2}+\boldsymbol{X}}{\frac{1}{\sigma^2}+1}, \frac{1}{1+\frac{1}{\sigma^2}}\right)$$
$$\mu,\sigma^2|\boldsymbol{\theta},\boldsymbol{X} \sim \mathbb{P}\big(\mu|\sigma^2,\boldsymbol{\theta}\big)\,\mathbb{P}\big(\sigma^2|\boldsymbol{\theta}\big)$$

So this is a simple example which we can try Gibbs sampler on. Since

$$X_i \overset{iid}{\sim} \mathcal{N}(\mu, V+1)$$

we get that

$$\mathbb{P}(\mu,V) \propto \left(\frac{1}{\sqrt{2\pi(V+1)}}\right)^K \exp\left\{\frac{\sum_{i=1}^k (X_i-\mu)^2}{2(V+1)}\right\} \times c$$

21. November 17th, 2014

If $Y \sim \operatorname{Bin}(n,p)$, then we have

$$\mathbb{E}\left[\frac{Y}{n}\right] = p$$
$$\operatorname{Var}\left(\frac{Y}{n}\right) = \frac{p(1-p)}{n}$$

so if we have a transformation $\varphi\left(\frac{Y}{n}\right)$ so that it has a stabilized variance, that is, we want $\operatorname{Var}\left(\phi\left(\frac{Y}{n}\right)\right) \approx c$, for some constant c. So, we have

$$\operatorname{Var}\left(\varphi\left(\frac{Y}{n}\right)\right) \approx \{\varphi'(p)\}^2 \operatorname{Var}\left(\frac{Y}{n}\right)$$
$$\approx c$$

so we have

$$\{\varphi'(p)\}^2 \frac{p(1-p)}{n} \approx c$$
$$\varphi'(p) \propto \frac{1}{\sqrt{p(1-p)}}$$
$$\varphi(p) \propto \int \frac{1}{\sqrt{p(1-p)}}\,\mathrm{d}p$$
$$\propto \arcsin(p)$$

which is why in a Binomial hierarchical model, we have the transformation

$$X_i = \sqrt{n_i}\arcsin\left(\frac{2Y_i}{n_i}-1\right)$$

Now, suppose for $i=1,\dots,k$,

$$Y_i \sim \operatorname{Bin}(n_i,\theta_i)$$
$$\theta_i \sim \operatorname{Beta}(\alpha,\beta)$$

We could marginalize Y_i, to get

$$\mathbb{P}(Y_i=y_i|\alpha,\beta) = \int \frac{n_i!}{y_i!(n_i-y_i)!}\theta_i^{y_i+\alpha-1}(1-\theta_i)^{n_i-y_i+\beta-1}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\,\mathrm{d}\theta_i$$
$$\propto \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\frac{\Gamma(y_i+\alpha)\Gamma(n_i-y_i+\beta)}{\Gamma(n_i+\alpha+\beta)}$$

and so

$$\prod_{i=1}^k \mathbb{P}(Y_i=y_i|\alpha,\beta) = \prod_{i=1}^k \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\frac{\Gamma(y_i+\alpha)\Gamma(n_i-y_i+\beta)}{\Gamma(n_i+\alpha+\beta)}$$

We can sample from this using Metropolis-Hastings. The biggest challenge of this is picking the proposal. How far should each step go? Clearly it has to depend on the data. We can possibly determine this by looking at the moments.

Another strategy is to keep the θ's and we can do full Gibbs.

$$\theta_i|\alpha,\beta,\boldsymbol{Y} \sim \text{Beta}$$
$$\alpha,\beta|\theta_i,\boldsymbol{Y} \sim \left(\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\right)^k \prod_{i=1}^k \theta_i^{\alpha+Y_i-1}(1-\theta_i)^{\beta+n_i+Y_i-1}$$

In general, it is better to marginalize over θ to have fewer parameters to deal with.

Note that once we draw $(\alpha^{(m)}, \beta^{(m)})$, we can draw θ_j's from the Beta distribution. To get $\mathbb{E}[\theta_i|\boldsymbol{Y}]$, we can either look at $\theta_i^{(1)}, \ldots, \theta_i^{(m)}$ and take an empircal mean, or we can look at $(\alpha^{(m)}, \beta^{(m)})$. The latter case is better.

21.1. **Nonparametric Hierarchical Bayes.** In this, we have for $i = 1, \ldots, k$.

$$\begin{aligned} Y_i &\sim \mathrm{Bin}(n_i, \theta_i) \\ \theta_i &\sim F \\ F &\sim \mathcal{D}(m(\cdot)) \end{aligned}$$

where m is some finite measure on $[0,1]$, that is $m = c\pi_0$, where π_0 is a probability measure on $[0,1]$ and c are the "pseudo-counts". Now, it is much harder here to marginalize out the parameters.

What if we tried to do Gibbs?

$$\begin{aligned} \theta_i|F, Y_i &\propto \theta_i^{Y_i}(1-\theta_i)^{n_i - Y_i} F(\mathrm{d}\theta_i) \\ F|\theta_1, \ldots, \theta_k &\sim \mathcal{D}\left(m + \sum_{i=1}^k \delta_{\theta_i}\right) \\ &\sim \mathcal{D}\left(c\pi_0 + \sum_{i=1}^k \delta_{\theta_i}\right) \end{aligned}$$

From here, we have an infinite stick-breaking process. We need to find a spot where we cut off. However, we wouldn't know if that would affect the GIbbs. It turns out this actually works okay according to Jun.

An alternate strategy is to look at

$$\begin{aligned} \theta_1|\theta_2, \ldots, \theta_k &\propto m + \sum_{i=2}^k \delta_{\theta_i} \\ &= c\pi_0 + \sum_{i=2}^k \delta_{\theta_i} \end{aligned}$$

which is the Chinese restaurant process. So the posterior distribution is

$$\theta_1|\boldsymbol{Y}, \theta_2, \ldots, \theta_k \propto \theta_1^{y_i}(1-\theta_1)^{n_1 - y_1}\left(c\pi_0 + \sum_{i=2}^k \delta_{\theta_i}\right)$$

Now, take $\pi_0(\cdot) = \mathrm{Beta}(\alpha, \beta)$. Then

$$\begin{aligned} \theta_1|\boldsymbol{Y}, \theta_2, \ldots, \theta_k &= c\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta_1^{\alpha+y_i-1}(1-\theta_1)^{n_1-y_1+\beta-1} + \sum_{i=2}^k \theta_1^{y_1}(1-\theta_1)^{n_1-y_1}\delta_{\theta_i}(\theta_1) \\ &= c\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\frac{\Gamma(y_1+\alpha)\Gamma(n_1-y_1+\beta)}{\Gamma(n_1+\alpha+\beta)}\frac{\Gamma(n_1+\alpha+\beta)}{\Gamma(y_1+\alpha)\Gamma(n_1-y_1+\beta)}\theta_1^{y_1+\alpha-1}(1-\theta_1)^{n_1-y_1+\beta-1} \\ &\quad + \sum_{i=2}^k \theta_i^{y_1}(1-\theta_i)^{n_1-y_1}\delta_{\theta_i}(\theta_1) \end{aligned}$$

The coefficients $\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\frac{\Gamma(y_1+\alpha)\Gamma(n_1-y_1+\beta)}{\Gamma(n_1+\alpha+\beta)}$ and $\theta_i^{y_1}(1-\theta_i)^{n_1-y_1}$ are weights in this case, and we simulate a Uniform random variable to tell you which component to take.

When we land on the same θ, we can't change the location, so instead of recording the location of θ_i, we can instead record when it stays.

22. November 20th, 2014

Recall that the mixture model is

$$\lambda_1 F_1(x) + \lambda_2 F_2(x) + \cdots + \lambda_k F_k(x)$$

where $\lambda_1 + \cdots + \lambda_k = 1$ and $\lambda_k > 0$.

Recall that a t distribution can be formulated with

$$\begin{aligned} V &\sim \chi_n^2 \\ Z &\sim \mathcal{N}(0,1) \end{aligned}$$

with V and Z independent. Then

$$X = \frac{Z}{\sqrt{V}/n} \sim t_n$$

We can take $(Z, V) \longrightarrow X$. Then

$$\int F_V(x)\,\mathrm{d}\mu(v) = \int \mathcal{N}\left(0, \frac{V}{n}\right) g(v)\,\mathrm{d}v$$

where $g(v)$ is the density for a χ^2_n random variable.

22.1. **Potential Data.**

$$\begin{aligned}\boldsymbol{Y} &= (Y_1, \ldots, Y_N)\\ \boldsymbol{I} &= (I_1, \ldots, I_N)\end{aligned}$$

where

$$I_{ij} = \begin{cases} 1 & \text{if } y_{ij} \text{ is observed} \\ 0 & \text{if } y_{ij} \text{ is not observed} \end{cases}$$

We require a "stability assumption": y_{ij} is not influenced by $\boldsymbol{I}$. Our full model is then

$$\mathbb{P}(\boldsymbol{Y}, \boldsymbol{I} | \boldsymbol{X}, \theta, \varphi) = \mathbb{P}(\boldsymbol{Y} | \boldsymbol{X}, \theta)\, \mathbb{P}(\boldsymbol{I} | \boldsymbol{X}, \boldsymbol{Y}, \varphi)$$

We then have $\boldsymbol{Y} = (\boldsymbol{Y}_{obs}, \boldsymbol{Y}_{mis})$.

$$\begin{aligned}\mathbb{P}(\theta, \varphi | \boldsymbol{X}, \boldsymbol{Y}_{obs}, \boldsymbol{I}) &\propto \mathbb{P}(\theta, \varphi | \boldsymbol{X})\, \mathbb{P}(\boldsymbol{Y}_{obs}, I | \boldsymbol{X}, \boldsymbol{\theta}, \boldsymbol{\varphi})\\ &= \mathbb{P}(\theta, \varphi | \boldsymbol{X}) \int \mathbb{P}(\boldsymbol{Y}_{obs}, \boldsymbol{Y}_{mis} | \boldsymbol{X}, \theta)\, \mathbb{P}(\boldsymbol{I} | \boldsymbol{X}, \boldsymbol{Y}_{obs}, \boldsymbol{Y}_{mis}, \varphi)\ \mathrm{d}\boldsymbol{Y}_{mis}\end{aligned}$$

- Assumption 1: We have distinct parameters,

$$\mathbb{P}(\varphi | \boldsymbol{X}, \theta) = \mathbb{P}(\varphi | \boldsymbol{X})$$

- Assumption 2: We have missing at random data,

$$\mathbb{P}(\boldsymbol{I} | \boldsymbol{X}, \boldsymbol{Y}_{obs}, \boldsymbol{Y}_{mis}, \varphi) = \mathbb{P}(\boldsymbol{I} | \boldsymbol{X}, \boldsymbol{Y} obs, \boldsymbol{\varphi})$$

and so,

$$\mathbb{P}(\theta | \boldsymbol{X}, \boldsymbol{Y}_{obs}, I) \propto \mathbb{P}(\theta | \boldsymbol{X}) \int \mathbb{P}(\boldsymbol{Y}_{obs}, \boldsymbol{Y}_{mis} | \boldsymbol{X}, \theta)\ \mathrm{d}\boldsymbol{Y}_{mis}$$

We call this "ignorable".

Definition 12 (Strongly Ignorable)**.**

$$\mathbb{P}(\boldsymbol{I} | \boldsymbol{X}, \boldsymbol{Y}, \varphi) = \mathbb{P}(\boldsymbol{I} | \boldsymbol{X})$$

22.2. **Causal Inference.**

treatment	control
Y_{11}	Y_{10}
Y_{21}	Y_{20}
⋮	⋮
Y_{n1}	Y_{n0}

we have $\mathbb{E}[Y_{11} - Y_{10}] = \mathbb{E}[Y_1] - \mathbb{E}[Y_0]$. We can never have people in treatment AND control.

23. November 24th, 2014

We have our data $\boldsymbol{Y}$ which can be partitioned into $\boldsymbol{y}_{obs}$ and $\boldsymbol{Y}_{mis}$ as well as $\boldsymbol{I}$, which is our design. Parameters for these are θ, φ.

23.1. **"Ignorable Design".**

$$\mathbb{P}(\boldsymbol{Y}, \boldsymbol{I} | \theta, \varphi, \boldsymbol{X}) = \mathbb{P}(\boldsymbol{Y} | \boldsymbol{X}, \theta)\, \mathbb{P}(\boldsymbol{I} | \boldsymbol{X}, \boldsymbol{Y}, \varphi)$$

The observed data likelihood is

$$\mathbb{P}(\boldsymbol{Y}_{obs}, \boldsymbol{I} | \boldsymbol{X}, \theta, \varphi) = \int \mathbb{P}(\boldsymbol{Y}_{obs}, \boldsymbol{Y}_{mis} | \boldsymbol{X}, \theta)\, \mathbb{P}(\boldsymbol{I} | \boldsymbol{X}, \boldsymbol{Y}_{obs}, \boldsymbol{Y}_{mis}, \varphi)\ \mathrm{d}y_{mis}$$

We have

(i) Distinct parameters θ, φ
(ii) $\mathbb{P}(\boldsymbol{I} | \boldsymbol{X}, \boldsymbol{Y}_{obs}, \boldsymbol{Y}_{mis}, \varphi) = \mathbb{P}(\boldsymbol{I} | \boldsymbol{X}, \boldsymbol{Y}_{obs}, \varphi)$

23.2. **Causality.** For unit i, we have treated (Y_{1i}) vs control (Y_{0i}), where we can only observe one of them, but we are interested in $Y_{1i} - Y_{0i}$, which is unobservable. Instead, we can look at $\mathbb{E}[Y_{1i} - Y_{0i}] = \mathbb{E}[Y_{1i}] - \mathbb{E}[Y_{0i}]$.

$c_i(1) = 1$	$c_i(0) = 0$	Compliers
$c_i(1) = 0$	$c_i(0) = 0$	Never-taker
$c_i(1) = 1$	$c_i(0) = 1$	Always taker
$c_i(1) = 0$	$c_i(0) = 1$	Defiers

24. December 1st, 2014