05 - Conjugate models - 01

Bayesian Statistics

Spring 2022-2023

Josep Fortiana

Matemàtiques - Informàtica UB

Monday, March 20, 2023

05 - Conjugate models - 01

What is a Bayesian conjugate model?

Statistics on the mean of a normal variable

Gamma, chi-squared et cætera

Statistics on the variance of a normal variable

Normal data with both parameters unknown

05 - Conjugate models - 01

What is a Bayesian conjugate model?

Statistics on the mean of a normal variable

Gamma, chi-squared et cætera

Statistics on the variance of a normal variable

Normal data with both parameters unknown

Bayesian information flow with conjugation

Prior		Likelihood		Posterior
Beta(a, b)		Binomial		Beta$(\tilde{a}, \tilde{b})$
Normal(θ, γ^2)	$+$	Normal$(\mu, \,\cdot\,)$	$\rightarrow$	Normal$(\tilde{\theta}, \tilde{\gamma}^2)$
IG(a, b)		Normal$(\,\cdot\,, \sigma^2)$		IG$(\tilde{a}, \tilde{b})$

05 - Conjugate models - 01

What is a Bayesian conjugate model?

Statistics on the mean of a normal variable

Gamma, chi-squared et cætera

Statistics on the variance of a normal variable

Normal data with both parameters unknown

Prior distribution of the mean

$X \sim \text{Normal}(\mu, \sigma^2)$, σ is known (constant),

$\mu \sim \text{Normal}(\theta, \gamma^2)$, prior of μ is also a gaussian.

This prior distribution is tailored to the information we have on μ before the experiment.

Prior variance, γ^2, is set larger (more uncertainty) when initial information is scarce.

μ's posterior properties

$$\mathsf{E}(\mu \mid x) = \qquad \theta_x \overset{\text{def}}{=\!=} \frac{\gamma^2}{\sigma^2+\gamma^2}\,x + \frac{\sigma^2}{\sigma^2+\gamma^2}\,\theta$$

Convex combination of prior θ, and observed x.

$$\operatorname{var}(\mu \mid x) = \qquad \tau^2 \overset{\text{def}}{=\!=} \frac{\sigma^2\,\gamma^2}{\sigma^2+\gamma^2} = \frac{1}{\dfrac{1}{\sigma^2}+\dfrac{1}{\gamma^2}}$$

Precision $=1/(\text{Variance})$

Relative weight of x is: $$\frac{\gamma^2}{\sigma^2+\gamma^2}=\frac{\frac{1}{\sigma^2}}{\frac{1}{\sigma^2}+\frac{1}{\gamma^2}}=\frac{\psi_D}{\psi_D+\psi_I},$$

$$\psi_D=\frac{1}{\sigma^2},$$ data *precision,*

$$\psi_I=\frac{1}{\gamma^2},$$ prior precision.

Posterior precision

Precision is *additive*:

$$\psi_F \equiv \frac{1}{\tau^2} = \frac{1}{\sigma^2} + \frac{1}{\gamma^2} = \psi_I + \psi_D.$$

Posterior precision $=$ sum of prior and data precisions.

Computation

Posterior pdf of μ, for a given x, with Bayes:

1. Joint pdf of (x, μ):

$$h(x, \mu) = f(x \mid \mu) \cdot h(\mu),$$

2. Integrate out μ, to give $f(x)$.
3. Then the posterior:

$$h(\mu \mid x) = \frac{h(x, \mu)}{f(x)}.$$

Computation

Likelihood of x, given μ:

$$f(x \mid \mu) = \frac{1}{\sqrt{2\pi}\,\sigma}\,\exp\left\{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right\}$$

Prior pdf of μ, given θ, γ^2:

$$h(\mu) = \frac{1}{\sqrt{2\pi}\,\gamma}\,\exp\left\{-\frac{1}{2}\frac{(\mu-\theta)^2}{\gamma^2}\right\}$$

Exponent in the product $h(x, \mu)$

$$\left\{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}-\frac{1}{2}\frac{(\mu-\theta)^2}{\gamma^2}\right\}$$

$$=-\frac{1}{2}\left\{\frac{\gamma^2(x^2-2x\mu+\mu^2)+\sigma^2(\mu^2-2\theta\mu+\theta^2)}{\sigma^2\gamma^2}\right\}$$

$$=-\frac{1}{2}\left\{\frac{\mu^2(\sigma^2+\gamma^2)-2\mu(\theta\sigma^2+x\gamma^2)+(x^2\gamma^2+\theta^2\sigma^2)}{\sigma^2\gamma^2}\right\}$$

Obtaining the marginal $f(x)$

Divide both numerator and denominator by $(\sigma^2 + \gamma^2)$,

$$= -\frac{1}{2}\left\{ \frac{\mu^2 - 2\mu\theta_x + \dfrac{(x^2\gamma^2 + \theta^2\sigma^2)}{\sigma^2 + \gamma^2}}{\tau^2} \right\}$$

"Complete the square" $\rightarrow$ A first summand:

$$-\frac{1}{2}\left\{\frac{\mu^2 - 2\mu\theta_x + \theta_x^2}{\tau^2}\right\} = -\frac{1}{2}\left\{\frac{(\mu - \theta_x)^2}{\tau^2}\right\},$$

Obtaining the marginal $f(x)$

And a second summand which, simplifying, gives:

$$-\frac{1}{2}\left\{\frac{(x-\theta)^2}{\sigma^2+\gamma^2}\right\}.$$

The exp of the first part is almost a normal pdf for μ.

Needs multiplying by $1/\sqrt{2\,\pi\,\tau^2}$.

The marginal = Prior predictive pdf

We do this (and compensate, multiplying by $\sqrt{2\,\pi\,\tau^2}$).

Integral of the first part with respect to μ gives 1, thus:

$$f(x) = (2\pi)^{-1/2}\,(\sigma^2+\gamma^2)^{-1/2}\exp\left\{-\frac{1}{2}\left[\frac{(x-\theta)^2}{\sigma^2+\gamma^2}\right]\right\},$$

a Normal$(\theta, (\sigma^2+\gamma^2))$.

Average of $f(x \mid \mu)$ over all possible values of μ,

each with relative weight proportional to the prior $h(\mu)$.

Joint distribution of (x, μ)

$h(x, \mu) = f(x \mid \mu) \cdot h(\mu)$ is a bivariate normal.

Joint distribution of (x, μ)

$h(x, \mu) = f(x \mid \mu) \cdot h(\mu)$ is a bivariate normal.

However, it is <u>NOT</u> the product of its two marginals!!

Joint distribution of (x, μ)

$h(x, \mu) = f(x \mid \mu) \cdot h(\mu)$ is a bivariate normal.

However, it is NOT the product of its two marginals!!

The correlation coefficient is:

$$\rho^2(x, \mu) = \frac{\gamma^2}{\gamma^2 + \sigma^2}.$$

Posterior pdf of μ, given x

Dividing $h(x,\mu)$ by $f(x)$, we obtain the posterior pdf:

$$h(\mu|x) = (2\pi)^{-1/2}\,\tau^{-1}\exp\left\{-\frac{1}{2}\frac{(\mu-\theta_x)^2}{\tau^2}\right\},$$

a normal distribution, with expectation:

$$\theta_x = \frac{\sigma^2}{\sigma^2+\gamma^2}\,\theta + \frac{\gamma^2}{\sigma^2+\gamma^2}\,x,$$

and variance: $\tau^2 = \dfrac{\sigma^2\gamma^2}{\sigma^2+\gamma^2}$.

Posterior predictive pdf

$f(\tilde{x} \mid x)$, pdf of a new observation $\tilde{x}$, given the previously observed value x.

By definition, $f(\tilde{x} \mid x)$ is the average of $f(\tilde{x} \mid \mu)$ over all possible values of μ, each with relative weight, now proportional to $h(\mu \mid x)$, the posterior pdf of μ given x.

Same computation as with the prior predictive pdf.

The posterior predictive pdf

Result: the posterior predictive pdf of a new $\tilde{x}$, given x, is a normal distribution:

$$(\tilde{x} \mid x) \sim \operatorname{Normal}(\theta_x, \sigma^2 + \tau^2), \qquad \text{where, as above,}$$

$$\theta_x = \frac{\sigma^2}{\sigma^2 + \gamma^2}\,\theta + \frac{\gamma^2}{\sigma^2 + \gamma^2}\,x,$$

$$\tau^2 = \frac{\sigma^2\gamma^2}{\sigma^2 + \gamma^2}.$$

Case of an *n*-sample

An n-sample,

$$X_1, \ldots, X_n, \quad \text{i.i.d.} \sim \mathrm{Normal}(\mu, \sigma^2),$$

is equivalent to a single observation of:

$$\bar{X} \sim \mathrm{Normal}(\mu, \sigma^2/n).$$

(according to the *Principle of Sufficiency*).

Case of an n-sample

Thus, the posterior parameters of μ:

$$\begin{aligned} \mathrm{E}(\mu \mid x) &= \theta_x \stackrel{\text{def}}{=\!=} \frac{\gamma^2}{\sigma^2/n + \gamma^2}\, x + \frac{\sigma^2/n}{\sigma^2/n + \gamma^2}\, \theta \\ \operatorname{var}(\mu \mid x) &= \tau^2 \stackrel{\text{def}}{=\!=} \frac{\sigma^2\, \gamma^2}{\sigma^2 + n\, \gamma^2} \end{aligned}$$

05 - Conjugate models - 01

Gamma distribution pdf

The Gamma(α, β) probability distribution with *shape* parameter α and *rate* parameter β has pdf:

$$f(x|\alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} \cdot x^{\alpha-1}\, e^{-\beta\, x}, \quad x \geq 0, \qquad \alpha, \beta > 0.$$

See the Wikipedia article for alternative parameterizations.

Additivity property

The sum of $X_1, \dots, X_n$, independent r.v. :

$$X_i \sim \text{Gamma}(\alpha_i, \beta), \qquad 1 \leq i \leq n,$$

with the same β, is also Gamma-distributed,

$$S = \textstyle\sum_{i=1}^n X_i \sim \text{Gamma}(\textstyle\sum_{i=1}^n \alpha_i, \beta),$$

with shape equal to the sum of the shape parameters.

Expectation, variance, mode of Gamma(α, β)

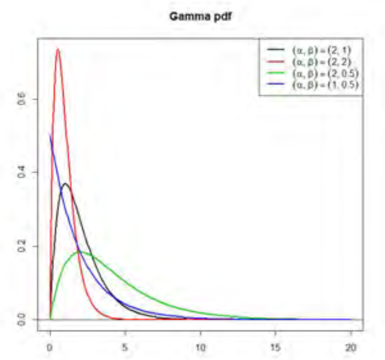
For $X \sim \operatorname{Gamma}(\alpha, \beta)$,

$$\operatorname{E}(X) = \frac{\alpha}{\beta}, \quad \operatorname{var}(X) = \frac{\alpha}{\beta^2}.$$

The mode is:

$$\frac{\alpha - 1}{\beta}, \quad \text{for } \alpha > 1.$$

Example Gamma(α, β) pdf's



Chi squared with *k degrees of freedom:* $\chi^2(k)$

The $\chi^2(k)$, or χ^2_k, distribution is a $\operatorname{Gamma}(\alpha,\beta)$ with $\alpha = \dfrac{k}{2}$ and $\beta = \dfrac{1}{2}$.

Its pdf:

$$f(x \mid k) = \frac{1}{2^{\frac{k}{2}} \cdot \Gamma\left(\frac{k}{2}\right)} \cdot x^{\frac{k}{2}-1} \cdot e^{-\frac{x}{2}}, \qquad x > 0, k > 0.$$

The $\chi^2(k)$ probability distribution

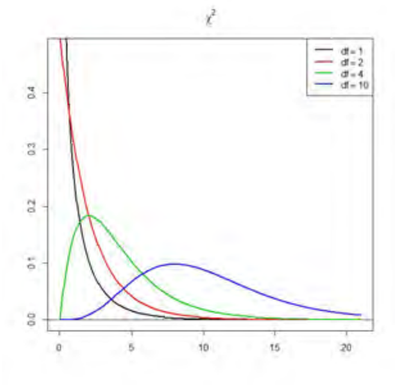
Historical name origin: $\chi =$ a normal variate.

$X \sim \text{Normal}(0, 1), Q = X^2 \sim \chi^2(1).$

$X_1, \ldots, X_n$ i.i.d.$\sim$ Normal$(0, 1)$,

$$Q_n \equiv \sum_{i=1}^{n} X_i^2 \sim \chi^2(n).$$

Example $\chi^2(k)$ pdf's



Gamma distribution is a *scaled* χ^2

$X \sim \text{Gamma}(\alpha, \beta)$. The new r.v. :

$$Z = 2\,\beta\,X, \qquad X = \frac{1}{2\,\beta}\,Z,$$

has pdf:

$$f_Z(z) = \frac{1}{2^{\alpha} \cdot \Gamma(\alpha)} \cdot z^{\alpha-1} \cdot e^{-\frac{z}{2}}, \qquad z > 0,$$

a χ^2, with $k = 2\,\alpha$ degrees of freedom.

The inverse gamma distribution

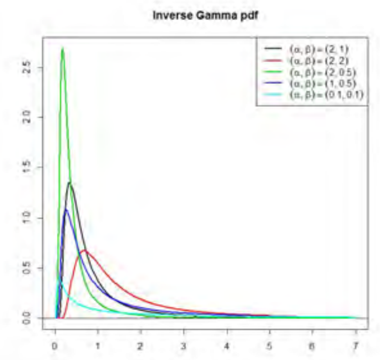
$Y = \dfrac{1}{X}$ is an *inverse gamma* $\mathrm{IG}(\alpha, \beta)$,

when $X \sim \mathrm{Gamma}(\alpha, \beta)$. Its pdf is:

$$f_Y(y|\alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} \cdot \frac{1}{y^{\alpha+1}} \cdot e^{-\frac{\beta}{y}}, \quad y > 0,\ \beta > 0.$$

Warning: parameter β in the $\mathrm{IG}(\alpha, \beta)$ is called the <u>scale</u> parameter, the converse nomenclature of that in the $\mathrm{Gamma}(\alpha, \beta)$ distribution.

Example Inverse Gamma pdf's



Expectation, variance, mode of $\mathrm{IG}(\alpha, \beta)$

For $Y \sim \mathrm{IG}(\alpha, \beta)$,

$$\begin{aligned}
\mathrm{E}(Y) &= \frac{\beta}{\alpha - 1}, \quad \text{for } \alpha > 1, \\
\operatorname{var}(Y) &= \frac{\beta^2}{(\alpha - 1)^2\,(\alpha - 2)}, \quad \text{for } \alpha > 2. \\
\mathrm{Mode}(Y) &= \frac{\beta}{\alpha + 1}.
\end{aligned}$$

Inverse chi squared distribution

The *inverse chi squared distribution* with k degrees of freedom, Inv-$\chi^2(k)$, is an $\mathrm{IG}(\alpha = \frac{k}{2}, \beta = \frac{1}{2})$.

Its pdf is:

$$f(z) = \frac{2^{-k/2}}{\Gamma(k/2)}\, z^{-k/2-1}\, e^{-1/(2z)}, \qquad z > 0.$$

Expectation, variance, mode of Inv-$\chi^2(k)$

For $Z \sim \text{Inv-}\chi^2(k)$,

$$\begin{aligned}
\mathrm{E}(Z) &= \frac{1}{k-2}, \quad \text{for } k > 2,\\
\operatorname{var}(Z) &= \frac{2}{(k-2)^2\,(k-4)}, \quad \text{for } k > 4.\\
\operatorname{Mode}(Z) &= \frac{1}{k+2}.
\end{aligned}$$

Scaled inverse chi squared distribution

Gelman *et al.*, (BDA3) write an $\operatorname{IG}(\alpha, \beta)$ as a Scaled-Inv-$\chi^2(\nu, \tau^2)$ distribution, where:

$$
\begin{aligned}
\nu &= 2\,\alpha, & \tau^2 &= \frac{2\,\beta}{\nu} = \frac{\beta}{\alpha}, \\
\alpha &= \frac{\nu}{2}, & \beta &= \alpha\,\tau^2 = \frac{\nu\,\tau^2}{2}.
\end{aligned}
$$

pdf of the Scaled-Inv-χ^2 distribution

$$f(x \mid \nu, \tau^2) = \frac{(\tau^2\,\nu/2)^{\nu/2}}{\Gamma(\nu/2)} \cdot x^{-\frac{\nu+2}{2}} \cdot \exp\left\{-\frac{\nu\,\tau^2}{2\,x}\right\},$$

Often we just need that:

$$f(x \mid \nu, \tau^2) \;\propto\; x^{-\frac{\nu+2}{2}} \cdot \exp\left\{-\frac{\nu\,\tau^2}{2\,x}\right\},$$

05 - Conjugate models - 01

Likelihood (θ known)

$\mathbf{x} = (x_1, \ldots, x_n)$ i.i.d.$\sim$ Normal(θ, σ^2), with unknown σ^2 but known θ, assumed 0. Likelihood:

$$f(\mathbf{x} \mid \psi) = (2\pi)^{-n/2} \cdot \psi^{n/2} \cdot \exp\left\{-\frac{n\, s^2}{2} \cdot \psi\right\},$$

where $\psi = \dfrac{1}{\sigma^2}$ is the precision parameter, and

$s^2 = \dfrac{1}{n} \displaystyle\sum_{i=1}^{n} x_i^2$ is the empirical variance.

Conjugate prior for precision and variance

The conjugate prior for ψ is $\mathrm{Gamma}(\alpha, \beta)$.

$$h(\psi \mid \alpha, \beta) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} \cdot \psi^{\alpha-1} \cdot \exp\{-\beta\,\psi\}.$$

The conjugate prior for $\sigma^2 = 1/\psi$, is an $\mathrm{IG}(\alpha, \beta)$.

Joint pdf = Likelihood times prior

$h(x\,,\psi) =$

$$(2\pi)^{-n/2}\cdot\frac{\beta^{\alpha}}{\Gamma(\alpha)}\cdot\psi^{\left(\frac{n}{2}+\alpha-1\right)}\cdot\exp\left\{-\left(\frac{n\,s^2}{2}+\beta\right)\cdot\psi\right\}.$$

Define:

$$\left\{\begin{array}{rcl}\widetilde{\alpha} & = & \alpha+\dfrac{n}{2},\\[2ex] \widetilde{\beta} & = & \beta+\dfrac{n\,s^2}{2}.\end{array}\right.$$

Preparing marginalization

Multiply and divide by: $\dfrac{\widetilde{\beta}^{\,\widetilde{\alpha}}}{\Gamma(\widetilde{\alpha})}$.

The second half is a $\operatorname{Gamma}(\widetilde{\alpha},\widetilde{\beta})$ pdf, integrates to 1.

The remaining expression is the marginal of x :

$$f(x\,) = (2\pi)^{-n/2}\cdot\frac{\Gamma\left(\alpha+\frac{n}{2}\right)}{\Gamma(\alpha)}\cdot\frac{\beta^{\alpha}}{\left(\beta+n\,\frac{s^2}{2}\right)^{\left(\alpha+\frac{n}{2}\right)}}.$$

Marginal pdf - Prior predictive pdf

$f(x)$ depends on x through $s^2 = \dfrac{1}{n} \sum_{i=1}^{n} x_i^2$.

(Sufficiency)

Define: $k = 2\,\alpha + n - 1$, $t^2 = k \cdot \dfrac{n\,s^2}{2\,\beta}$.

Marginal pdf - Prior predictive pdf

The marginal pdf, in terms of t^2, is proportional to:

$$f(t) = \frac{\Gamma\left(\frac{k+1}{2}\right)}{\sqrt{k\,\pi}\cdot\Gamma\left(\frac{k}{2}\right)}\cdot\left(1+\frac{t^2}{k}\right)^{-\frac{k+1}{2}}, \quad -\infty < t < \infty,$$

a Student's $t(k)$ pdf.

Posterior pdf of $(\psi|x)$

From Bayes' rule we see that:

$$(\psi|x) \sim \text{Gamma}(\widetilde{\alpha}, \widetilde{\beta}),$$

where:

$$\begin{cases} \widetilde{\alpha} & = & \alpha + \dfrac{n}{2}, \\ \widetilde{\beta} & = & \beta + \dfrac{n\,s^2}{2}. \end{cases}$$

05 - Conjugate models - 01

Likelihood

The likelihood function for n i.i.d.$\sim$ Normal$(\mu, 1/\psi)$ normal observations, $x = (x_1, \ldots, x_n)$, $\psi = 1/\sigma^2$, is:

$$f(x \,|\mu, \psi) \propto \psi^{n/2} \cdot \exp\left\{-\frac{\psi}{2} \sum_{i=1}^{n} (x_i - \mu)^2\right\},$$

We assume now that both parameters (μ, ψ) are unknown, hence we must provide prior pdf's for both of them.

Joint prior pdf

New feature when there is more than one parameter: we need a joint prior pdf for (μ, ψ).

We could try to assume that μ and ψ are independent, by posing a prior pdf:

$$h(\mu, \psi) = h_1(\mu) \cdot h_2(\psi),$$

but then we would not obtain a conjugate prior.

A pair of dependent priors

We propose:

$$\psi \sim \mathrm{Gamma}(\alpha, \beta),$$

$$\mu \mid \psi \sim \mathrm{Normal}(\theta_0, 1/(n_0\,\psi)).$$

n_0 is a scaling factor, thought of as the number of observations in a virtual “prior sample".

The mean of n_0 observations, each with variance σ^2, has variance σ^2/n_0, corresponding to the precision $n_0\,\psi$.

Posterior for θ, given x and ψ

We already did this computation:

$$\mu \mid (x\,,\psi) \sim \mathrm{Normal}(\theta_x, \psi_x),$$

where:

$$\begin{aligned}\theta_x &= \frac{n}{n+n_0}\,\bar{x} + \frac{n_0}{n+n_0}\,\theta_0,\\ \psi_x &= (n+n_0)\cdot\psi.\end{aligned}$$

Posterior for ψ, given x

$$(\psi|x) \sim \operatorname{Gamma}(\widetilde{\alpha}, \widetilde{\beta}), \qquad \text{where:}$$

$$\begin{cases} \widetilde{\alpha} &= \alpha + \dfrac{n}{2}, \\ \widetilde{\beta} &= \beta + \dfrac{n\, s^2}{2} + \dfrac{n \cdot n_0}{2\,(n + n_0)}\,(\bar{x} - \theta_0)^2. \end{cases}$$

Here $s^2 = \sum_{i=1}^{n}(x_i - \bar{x})^2/n$.

In terms of σ^2

If $x = (x_1, \ldots, x_n)$, each $x_i \sim \text{Normal}(\mu, \sigma^2)$, with prior:

$$\begin{cases} \sigma^2 & \sim \quad \text{IG}(\alpha, \beta), \\ \mu \mid \sigma^2 & \sim \quad \text{Normal}(\theta_0, \dfrac{\sigma^2}{n_0}). \end{cases}$$

As above, n_0 is a scaling factor, thought of as the number of observations in a virtual “prior sample".

Posterior for μ, given x and σ^2

$$\mu \mid (x\,, \sigma^2) \sim \operatorname{Normal}(\theta_x, \sigma_x^2),$$

where:

$$\begin{aligned} \theta_x &= \frac{n}{n+n_0}\,\bar{x} + \frac{n_0}{n+n_0}\,\theta_0, \\ \sigma_x^2 &= \frac{\sigma^2}{n+n_0}. \end{aligned}$$

Posterior for σ^2, given $\boldsymbol{x}$

$$(\sigma^2|\boldsymbol{x}) \sim \mathrm{IG}(\widetilde{\alpha}, \widetilde{\beta}), \qquad \text{where:}$$

$$\begin{cases} \widetilde{\alpha} &=& \alpha + \dfrac{n}{2}, \\[2ex] \widetilde{\beta} &=& \beta + \dfrac{n\,s^2}{2} + \dfrac{n \cdot n_0}{2\,(n+n_0)}\,(\bar{x}-\theta_0)^2. \end{cases}$$

Here $s^2 = \displaystyle\sum_{i=1}^{n}(x_i - \bar{x})^2/n.$