Gaussian models

Bayesian Methodology in Biostatistics (BST 249)

Jeffrey W. Miller

Department of Biostatistics
Harvard T.H. Chan School of Public Health

Outline

Univariate normal model

Conjugate prior for the mean

Example: Is human height bimodal?

Conjugate prior for the mean and precision

Example: The Pygmalion effect

Other common priors for normal parameters

Multivariate normal

Outline

Univariate normal model

Conjugate prior for the mean
- Example: Is human height bimodal?

Conjugate prior for the mean and precision
- Example: The Pygmalion effect

Other common priors for normal parameters

Multivariate normal

Univariate normal distribution

- The normal (a.k.a. Gaussian) distribution $\mathcal{N}(\mu, \sigma^2)$ with mean $\mu \in \mathbb{R}$ and variance $\sigma^2 > 0$ has p.d.f.

$$\mathcal{N}(x \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\Big(-\frac{1}{2\sigma^2}(x-\mu)^2\Big)$$

  for $x \in \mathbb{R}$.

- It is often convenient to work with the precision $\lambda = 1/\sigma^2$ rather than the variance. In this parametrization, the p.d.f. is

$$\mathcal{N}(x \mid \mu, \lambda^{-1}) = \sqrt{\frac{\lambda}{2\pi}} \exp\big(-\tfrac{1}{2}\lambda(x-\mu)^2\big).$$

Univariate normal distribution

- The normal distribution has special properties that give it a unique position in probability and statistics.

- Central limit theorem (CLT)
  - CLT: The sum of a large number of independent random variables is approximately normal.
  - Consequently, many real-world quantities tend to be normally distributed.
  - When designing models, the CLT helps us understand when a normal model would be appropriate.

- Analytic tractability
  - Posterior computations can often be done in closed form, making normal models computationally convenient.
  - Normal distributions can be combined to build complex models that are still tractable.

Outline

Univariate normal model

Conjugate prior for the mean

- Example: Is human height bimodal?

Conjugate prior for the mean and precision

- Example: The Pygmalion effect

Other common priors for normal parameters

Multivariate normal

Conjugate prior for the mean

- Consider an i.i.d. normal model:

$$X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathcal{N}(\theta, \lambda^{-1}).$$

- Assume the precision $\lambda = 1/\sigma^2$ is known and fixed.
- Assume the prior on the mean is $p(\theta) = \mathcal{N}(\theta \mid \mu_0, \lambda_0^{-1})$, i.e.,

$$\boldsymbol{\theta} \sim \mathcal{N}(\mu_0, \lambda_0^{-1}).$$

- This is sometimes referred to as a *Normal–Normal model*.

- The posterior is $p(\theta | x_{1:n}) = \mathcal{N}(\theta \mid M, L^{-1})$, i.e.,

$$\boldsymbol{\theta} | x_{1:n} \sim \mathcal{N}(M, L^{-1}) \tag{1}$$

  where $L = \lambda_0 + n\lambda$ and

$$M = \frac{\lambda_0 \mu_0 + \lambda \sum_{i=1}^{n} x_i}{\lambda_0 + n\lambda}.$$

- Thus, the normal distribution is a conjugate prior for the mean of a normal distribution with known precision.

Derivation of the Normal-Normal posterior

(Whiteboard activity)

Derivation of the Normal-Normal posterior (1/2)

- For any $x \in \mathbb{R}$, $\ell > 0$,

$$\begin{aligned} \mathcal{N}(x \mid \theta, \ell^{-1}) &= \sqrt{\frac{\ell}{2\pi}} \exp\left(-\tfrac{1}{2}\ell(x-\theta)^2\right) \\ &\underset{\theta}{\propto} \exp\left(-\tfrac{1}{2}\ell(x^2 - 2x\theta + \theta^2)\right) \\ &\underset{\theta}{\propto} \exp\left(\ell x\theta - \tfrac{1}{2}\ell\theta^2\right). \end{aligned} \tag{2}$$

- Due to the symmetry of the normal p.d.f.,

$$\begin{aligned} \mathcal{N}(\theta \mid \mu_0, \lambda_0^{-1}) &= \mathcal{N}(\mu_0 \mid \theta, \lambda_0^{-1}) \\ &\underset{\theta}{\propto} \exp\left(\lambda_0\mu_0\theta - \tfrac{1}{2}\lambda_0\theta^2\right) \end{aligned} \tag{3}$$

by Equation 2 with $x = \mu_0$ and $\ell = \lambda_0$.

Derivation of the Normal-Normal posterior (2/2)

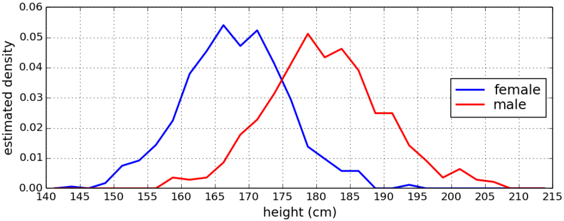
- Therefore,

$$
\begin{aligned}
p(\theta|x_{1:n}) &\propto p(\theta)p(x_{1:n}|\theta) \\
&= \mathcal{N}(\theta \mid \mu_0, \lambda_0^{-1}) \prod_{i=1}^{n} \mathcal{N}(x_i \mid \theta, \lambda^{-1}) \\
&\overset{\text{(a)}}{\propto} \exp\left(\lambda_0 \mu_0 \theta - \tfrac{1}{2}\lambda_0 \theta^2\right) \exp\left(\lambda (\textstyle\sum x_i)\theta - \tfrac{1}{2} n \lambda \theta^2\right) \\
&= \exp\Big((\lambda_0 \mu_0 + \lambda \textstyle\sum x_i)\theta - \tfrac{1}{2}(\lambda_0 + n\lambda)\theta^2\Big) \\
&= \exp(LM\theta - \tfrac{1}{2}L\theta^2) \\
&\overset{\text{(b)}}{\propto} \mathcal{N}(M \mid \theta, L^{-1}) = \mathcal{N}(\theta \mid M, L^{-1})
\end{aligned}
$$

  where $L = \lambda_0 + n\lambda$ and $M = (\lambda_0 \mu_0 + \lambda \sum x_i)/L$.
- Step (a) uses Equations 2 and 3, and step (b) uses Equation 2 with $x = M$ and $\ell = L$. This proves Equation 1.

Example: Is human height bimodal?

Heights of Dutch women ($n = 695$) and men ($n = 562$)



- Human height is a classic example of a normal distributed quantity, when separated by sex. (And it is actually remarkable close to normal.)
- This is probably due to the CLT, since it seems that many independent genetic factors contribute to height.
- Meanwhile, when pooling women and men, height is often said to be bimodal (i.e., has two modes). But is it really?

Example: Is human height bimodal?

- This example illustrates:
  - Bayesian analysis with a normal model
  - computing a posterior quantity of interest
  - prior selection
  - a simple but interesting application

Example: Is human height bimodal?

"Living histogram" of 143 UConn students

x-axis = height, color = sex (female/male)

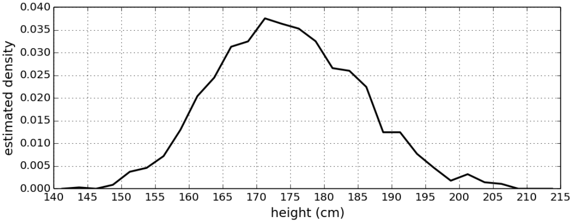
The Hartford Courant (1996)

- Crow (1997) writes, "Since both sexes are included, the distribution is bimodal."

Example: Is human height bimodal?

Heights of Dutch women and men, combined

(assuming equal proportions of women and men in the population)



- Visually, the combined distribution does not look bimodal, but maybe we don't have enough data yet.
- How could we test whether the population distribution is actually bimodal, accounting for uncertainty?

Height bimodality example: Likelihood/model

- Assume the female heights are

$$X_1, \ldots, X_k \overset{\text{iid}}{\sim} \mathcal{N}(\theta_f, \sigma^2),$$

where $k = 695$, and the male heights are

$$Y_1, \ldots, Y_\ell \overset{\text{iid}}{\sim} \mathcal{N}(\theta_m, \sigma^2),$$

where $\ell = 562$.

- Assume the p.d.f. of the combined distribution of heights is

$$\tfrac{1}{2}\mathcal{N}(x \mid \theta_f, \sigma^2) + \tfrac{1}{2}\mathcal{N}(x \mid \theta_m, \sigma^2).$$

- This a two-component *mixture* distribution with equal weights.

Height bimodality example: Target of inference

- By Helguerro (1904), the combined distribution is bimodal if and only if

  $$|\theta_f - \theta_m| > 2\sigma,$$

  i.e., if the difference in means is greater than twice the standard deviation.

- So, to address our question of interest ("Is human height bimodal?"), we would like to compute the posterior probability of this event:

  $$\mathbb{P}(\text{bimodal} \mid \text{data}) = \mathbb{P}\big(|\boldsymbol{\theta}_f - \boldsymbol{\theta}_m| > 2\sigma \;\big|\; x_{1:k}, y_{1:\ell}\big).$$

- To make this probability well-defined, we need to put priors on the parameters.

Group activity: Height bimodality example

Go to breakout rooms and work together to answer these questions:
https://forms.gle/WYkeHyRZFAxmYyEHA

(Two people per room, randomly assigned. 15 minutes.)

Height bimodality example: Prior

- Let's put independent normal priors on θ_f and θ_m:

$$\boldsymbol{\theta}_f \sim \mathcal{N}(\mu_{0,f}, \sigma_0^2) \qquad \boldsymbol{\theta}_m \sim \mathcal{N}(\mu_{0,m}, \sigma_0^2).$$

- Soon we will consider priors on σ^2, but for now let's assume σ^2 is fixed and known:

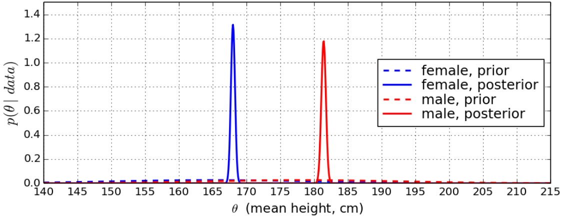
$$\sigma = 8 \text{ cm (about 3 inches).}$$

- Based on common knowledge of typical human heights, let's set the hyperparameters as follows:

$$\mu_{0,f} = 165 \text{ cm } (\approx 5 \text{ feet, } 5 \text{ inches})$$
$$\mu_{0,m} = 178 \text{ cm } (\approx 5 \text{ feet, } 10 \text{ inches})$$
$$\sigma_0 = 15 \text{ cm } (\approx 6 \text{ inches})$$

- Note that σ_0 represents our prior uncertainty about the mean heights, not about the heights of individuals.

Height bimodality example: Posterior



- The details here are not important – just the general idea.
- Since women and men are modeled independently,

$$p(\theta_f, \theta_m \mid x_{1:k}, y_{1:\ell}) = p(\theta_f | x_{1:k}) p(\theta_m | y_{1:\ell}).$$

- Equation 1 gives us $p(\theta_f | x_{1:k})$ and $p(\theta_m | y_{1:\ell})$:

$$\boldsymbol{\theta}_f | x_{1:k} \sim \mathcal{N}(M_f, L_f^{-1}) \qquad \boldsymbol{\theta}_m | y_{1:\ell} \sim \mathcal{N}(M_m, L_m^{-1})$$

where

$$M_f = 168.0 \text{ cm (5' 6.1'')} \qquad 1/\sqrt{L_f} = 0.30 \text{ cm}$$

$$M_m = 181.4 \text{ cm (5' 11.4'')} \qquad 1/\sqrt{L_m} = 0.34 \text{ cm}$$

Height bimodality example: Result

- Since a linear combination of independent normals is normal,

$$\begin{aligned}\boldsymbol{\theta}_m - \boldsymbol{\theta}_f \mid x_{1:k}, y_{1:\ell} &\sim \mathcal{N}(M_m - M_f,\, L_m^{-1} + L_f^{-1}) \\ &= \mathcal{N}(13.4, 0.45^2).\end{aligned}$$

- So we can compute $\mathbb{P}(\text{bimodal} \mid \text{data})$ using $\Phi(x \mid \mu, \sigma^2)$, the c.d.f. of $\mathcal{N}(\mu, \sigma^2)$:

$$\begin{aligned}&\mathbb{P}(\text{bimodal} \mid \text{data}) = \mathbb{P}\big(|\boldsymbol{\theta}_m - \boldsymbol{\theta}_f| > 2\sigma \bigm| x_{1:k}, y_{1:\ell}\big) \\ &= \Phi(-2\sigma \mid 13.4, 0.45^2) + \big(1 - \Phi(2\sigma \mid 13.4, 0.45^2)\big) \\ &= 6.1 \times 10^{-9}.\end{aligned}$$

- The posterior probability of bimodality is close to zero since the posteriors are about 13 or 14 cm apart, which is under the $2\sigma = 16$ cm threshold for bimodality, and they are sufficiently concentrated.

- **Critical thinking: How sensitive is this result to our assumptions?**

Outline

Univariate normal model

Conjugate prior for the mean

- Example: Is human height bimodal?

Conjugate prior for the mean and precision

- Example: The Pygmalion effect

Other common priors for normal parameters

Multivariate normal

Conjugate prior for the mean and precision

- Model: $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \lambda^{-1})$.
- Suppose that both μ and λ are unknown.

- The $\text{NormalGamma}(m, c, a, b)$ distribution, with $m \in \mathbb{R}$ and $c, a, b > 0$, is a joint distribution on (μ, λ) obtained by letting

$$\begin{aligned} \boldsymbol{\lambda} &\sim \text{Gamma}(a, b) \\ \boldsymbol{\mu}|\lambda &\sim \mathcal{N}(m, (c\lambda)^{-1}). \end{aligned}$$

- In other words, the joint p.d.f. is

$$p(\mu, \lambda) = p(\mu|\lambda)p(\lambda) = \mathcal{N}(\mu \mid m, (c\lambda)^{-1})\,\text{Gamma}(\lambda \mid a, b)$$

  which we will denote by $\text{NormalGamma}(\mu, \lambda \mid m, c, a, b)$.

- It turns out that this is a conjugate prior on (μ, λ).

Conjugate prior for the mean and precision

- Indeed, the posterior is

$$\boldsymbol{\mu}, \boldsymbol{\lambda} | x_{1:n} \sim \mathrm{NormalGamma}(M, C, A, B) \tag{4}$$

where

$$\begin{aligned} M &= \frac{cm + \sum_{i=1}^n x_i}{c+n} \\ C &= c + n \\ A &= a + n/2 \\ B &= b + \tfrac{1}{2}\big(cm^2 - CM^2 + \textstyle\sum_{i=1}^n x_i^2\big). \end{aligned}$$

- For interpretation, B can also be written (by rearranging terms) as

$$B = b + \frac{1}{2}\sum_{i=1}^n (x_i - \bar{x})^2 + \frac{1}{2}\frac{cn}{c+n}(\bar{x} - m)^2. \tag{5}$$

Interpretation of posterior parameters

- $M =$ posterior mean of μ.
  - Convex combination of the prior mean and the sample mean:

$$M = \frac{c}{c+n}m + \frac{n}{c+n}\bar{x}.$$

- $C =$ "sample size" for estimating μ.
  - The standard deviation of $\mu|\lambda$ is $\lambda^{-1/2}/\sqrt{C}$.
- $A =$ shape parameter of the posterior on λ.
  - Grows linearly with sample size.
- $B =$ rate parameter (inverse scale) of the posterior on λ.
  - See Equation 5 for decomposition of B.

Derivation of the posterior (1/2)

(Whiteboard activity)

Derivation of the posterior (1/2)

- Multiplying out $(\mu - m)^2 = \mu^2 - 2\mu m + m^2$ and collecting terms, we have

$$\begin{aligned} &\mathrm{NormalGamma}(\mu, \lambda \mid m, c, a, b) \\ &= \sqrt{\frac{c\lambda}{2\pi}} \exp\Big(-\tfrac{1}{2} c\lambda(\mu - m)^2\Big) \frac{b^a}{\Gamma(a)} \lambda^{a-1} \exp(-b\lambda) \\ &\underset{\mu,\lambda}{\propto} \lambda^{a-1/2} \exp\Big(-\tfrac{1}{2}\lambda(c\mu^2 - 2cm\mu + cm^2 + 2b)\Big). \end{aligned} \tag{6}$$

- Similarly, for any x,

$$\mathcal{N}(x \mid \mu, \lambda^{-1}) \underset{\mu,\lambda}{\propto} \lambda^{1/2} \exp\Big(-\tfrac{1}{2}\lambda(\mu^2 - 2x\mu + x^2)\Big). \tag{7}$$

Derivation of the posterior (2/2)

- Using Equations 6 and 7, we get

$$
\begin{aligned}
p(\mu, \lambda | x_{1:n}) &\underset{\mu,\lambda}{\propto} \mathrm{NormalGamma}(\mu, \lambda \mid m, c, a, b) \prod_{i=1}^{n} \mathcal{N}(x_i \mid \mu, \lambda) \\
&\underset{\mu,\lambda}{\propto} \lambda^{a+n/2-1/2} \exp\Big(-\tfrac{1}{2}\lambda\big((c+n)\mu^2 - 2(cm + \textstyle\sum x_i)\mu \\
&\qquad\qquad + cm^2 + 2b + \textstyle\sum x_i^2\big)\Big) \\
&\overset{\text{(a)}}{=} \lambda^{A-1/2} \exp\Big(-\tfrac{1}{2}\lambda\big(C\mu^2 - 2CM\mu + CM^2 + 2B\big)\Big) \\
&\overset{\text{(b)}}{\propto} \mathrm{NormalGamma}(\mu, \lambda \mid M, C, A, B).
\end{aligned}
$$

- Step (b) is by Equation 6, and step (a) holds if

$$
\begin{aligned}
A &= a + n/2 & CM &= (cm + \textstyle\sum x_i) \\
C &= c + n & CM^2 + 2B &= cm^2 + 2b + \textstyle\sum x_i^2.
\end{aligned}
$$

- Solving for M and B, we get the result in Equation 4.

Example: The Pygmalion effect

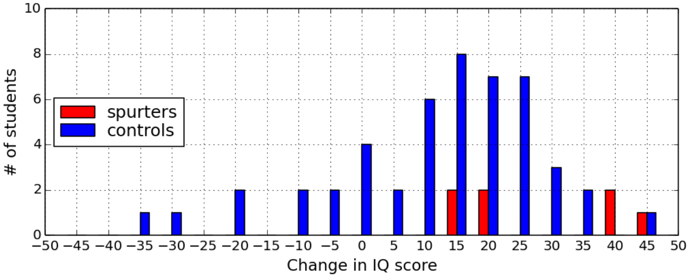
- Do a teacher's expectations influence student achievement?

- Rosenthal and Jacobson (1968) performed a famous experiment to address this question.
  - At the beginning of the year, all students were given an IQ test.
  - The researchers randomly selected around 20% of the students in each class.
  - They told teacher these students were "spurters" (outstanding students).
  - At the end of the year, all students were given another IQ test.

- The changes in IQ score for the first-grade students were:*
  - spurters (S): $x =$ (18, 40, 15, 17, 20, 44, 38)
  - controls (C): $y =$ (–4, 0, –19, 24, 19, 10, 5, 10, 29, 13, –9, –8, 20, –1, 12, 21, –7, 14, 13, 20, 11, 16, 15, 27, 23, 36, –33, 34, 13, 11, –19, 21, 6, 25, 30, 22, –28, 15, 26, –1, –2, 43, 23, 22, 25, 16, 10, 29)

*Note: The original data are not available. These data are from the `ex1321` dataset of the `R` package `Sleuth3`, which was constructed to match the summary statistics and conclusions of the original study.

Example: The Pygmalion effect



- Summary statistics:

spurters:	$n_S = 7$	$\bar{x} = 27.4$	$\hat{\sigma}_x = 11.7$
controls:	$n_C = 48$	$\bar{y} = 12.0$	$\hat{\sigma}_y = 16.1$

- The average increase in IQ score is larger for the spurters.

- How strongly do these data support the hypothesis that the teachers' expectations caused the spurters to perform better than their classmates?

Pygmalion example: Model & Target of inference

- IQ tests are calibrated to make the scores normally distributed, so it makes sense to use a normal model.

$$\text{spurters: } X_1, \ldots, X_{n_S} \overset{\text{iid}}{\sim} \mathcal{N}(\mu_S, \lambda_S^{-1})$$
$$\text{controls: } Y_1, \ldots, Y_{n_C} \overset{\text{iid}}{\sim} \mathcal{N}(\mu_C, \lambda_C^{-1}).$$

- We are interested in the difference between the means—in particular, is $\mu_S > \mu_C$?

- The Bayesian approach is simply to compute the posterior probability that $\mu_S > \mu_C$:

$$\mathbb{P}(\boldsymbol{\mu}_S > \boldsymbol{\mu}_C \mid x_{1:n_S}, y_{1:n_C}).$$

- As before, to make this well-defined we need to assume priors on the parameters.

Pygmalion example: Prior

- We don't know the precisions λ_S and λ_C, and the sample seems too small to estimate λ_S very well.
- Thus, it is important to account for uncertainty in λ_S.

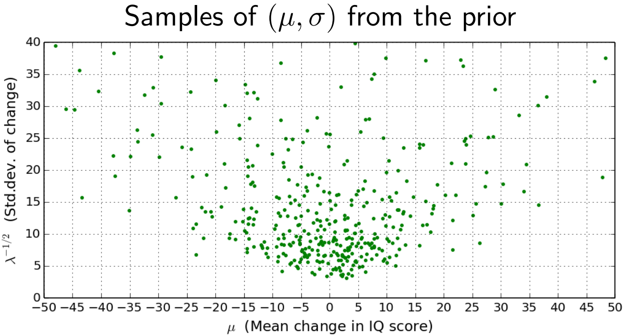
- Let's use independent NormalGamma priors:

$$\text{spurters: } (\boldsymbol{\mu}_S, \boldsymbol{\lambda}_S) \sim \mathrm{NormalGamma}(m, c, a, b)$$
$$\text{controls: } (\boldsymbol{\mu}_C, \boldsymbol{\lambda}_C) \sim \mathrm{NormalGamma}(m, c, a, b).$$

- Choose hyperparameters based on subjective prior knowledge:

$m = 0$	Don't know if students will improve or not, on average.
$c = 1$	Unsure how big the mean change will be.
	(Prior certainty is equivalent to info in c datapoints.)
$a = 1/2$	Unsure how big the stddev of the changes will be.
	(Prior certainty is equivalent to info in $2a$ datapoints.)
$b = 10^2 a$	Expect stddev of changes to be $\approx 10 = \sqrt{b/a} = \mathrm{E}(\lambda)^{-1/2}$.

Pygmalion example: Prior

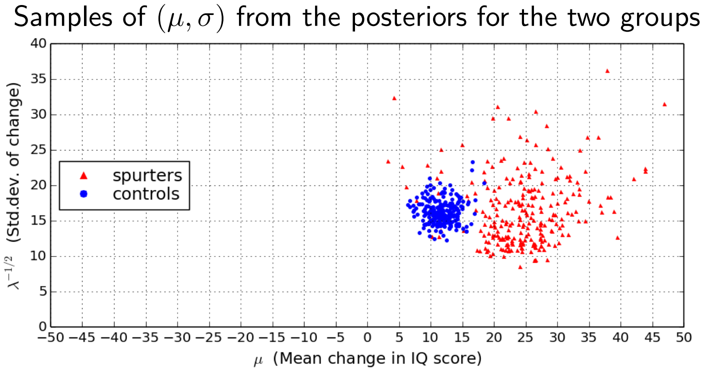


- Does the prior conform to our beliefs? Some ways to check:
  1. Look at samples drawn from the prior. (RECOMMENDED)
  2. Check prior moments, but beware—they can be misleading.
  3. Look at hypothetical datasets $X_{1:n}$ from the prior+model.
  4. Plot the prior c.d.f. and check various quantiles.
  5. Plot the prior p.d.f., but beware—it can be misleading.

Pygmalion example: Posterior

Answer these questions individually (5 minutes):
https://forms.gle/9tVWd73Pp7gvRdtz8

Pygmalion example: Posterior



- By Equation 4, the posteriors are

$$\boldsymbol{\mu}_S, \boldsymbol{\lambda}_S \mid x_{1:n_S} \sim \text{NormalGamma}(24.0,\ 8,\ 4,\ 855)$$
$$\boldsymbol{\mu}_C, \boldsymbol{\lambda}_C \mid y_{1:n_C} \sim \text{NormalGamma}(11.8,\ 49,\ 24.5,\ 6344).$$

Pygmalion example: Results

- Now, what is the posterior probability that $\mu_S > \mu_C$?

- Easiest way: Generate samples from the posterior and calculate how frequently $\mu_S > \mu_C$.
  - This is a Monte Carlo approximation ... more to come on this!

- To do this, we draw $N = 10^6$ i.i.d. samples from the posterior:

  $$(\mu_S^{(i)}, \lambda_S^{(i)}) \overset{\text{iid}}{\sim} \text{NormalGamma}(24.0, 8, 4, 855)$$
  $$(\mu_C^{(i)}, \lambda_C^{(i)}) \overset{\text{iid}}{\sim} \text{NormalGamma}(11.8, 49, 24.5, 6344)$$

  for $i = 1, \ldots, N$, and calculate the approximation

  $$\mathbb{P}(\boldsymbol{\mu}_S > \boldsymbol{\mu}_C \mid x_{1:n_S}, y_{1:n_C}) \approx \frac{1}{N} \sum_{i=1}^{N} \mathrm{I}\big(\mu_S^{(i)} > \mu_C^{(i)}\big) = 0.97.$$

- Interpretation: These data seem to support the hypothesis that the teachers' expectations did in fact play a role.

Outline

Conditionally conjugate prior for the mean and precision

- The NormalGamma prior induces a strong dependency between μ and λ, which can be undesirable.
  - See plot of samples from the prior in Pygmalion example.

- It is often more natural to make them independent *a priori*:

$$\boldsymbol{\lambda} \sim \mathrm{Gamma}(a, b) \text{ and } \boldsymbol{\mu} \sim \mathcal{N}(m, s^2), \text{ independently.}$$

- This is not a conjugate prior, but it is *conditionally conjugate* in the sense that:
  - for any fixed λ, it is conjugate for μ, and
  - for any fixed μ, it is conjugate for λ.

- Conditionally conjugate priors are easy to work with in MCMC and variational inference algorithms.

Conjugate prior for the variance

- So far, we've used a Gamma prior on the precision $\lambda = 1/\sigma^2$.
- What if we wanted to work directly with the variance σ^2?

- If $X \sim \text{Gamma}(a, b)$ then $1/X \sim \text{InvGamma}(a, b)$.
- So, putting a $\text{Gamma}(a, b)$ prior on λ is equivalent to putting an $\text{InvGamma}(a, b)$ prior on σ^2.
- The p.d.f. of the Inverse Gamma distribution is

$$\text{InvGamma}(y|a, b) = \frac{b^a}{\Gamma(a)} y^{-a-1} \exp(-b/y).$$

- Similarly to the NormalGamma prior on (μ, λ), the Normal-InvGamma is a conjugate prior on (μ, σ^2):

$$\boldsymbol{\sigma}^2 \sim \text{InvGamma}(a, b) \qquad \boldsymbol{\mu}|\sigma^2 \sim \mathcal{N}(m, \sigma^2/c).$$

Outline

Multivariate normal distribution

- Let $\mu \in \mathbb{R}^d$, let $C \in \mathbb{R}^{d \times d}$ symmetric positive definite (SPD).

- The multivariate normal distribution $\mathcal{N}(\mu, C)$ has p.d.f.

$$\mathcal{N}(x \mid \mu, C) = \frac{1}{(2\pi)^{d/2}|C|^{1/2}} \exp\big(-\tfrac{1}{2}(x-\mu)^{\mathtt{T}} C^{-1}(x-\mu)\big)$$

  for $x \in \mathbb{R}^d$. Here, $|C| = |\det C|$.

- In terms of the precision matrix $\Lambda = C^{-1}$, the p.d.f. is

$$\mathcal{N}(x \mid \mu, \Lambda^{-1}) = \frac{|\Lambda|^{1/2}}{(2\pi)^{d/2}} \exp\big(-\tfrac{1}{2}(x-\mu)^{\mathtt{T}} \Lambda (x-\mu)\big).$$

- Note that C is SPD if and only if Λ is SPD.

Conjugate prior for the mean

- Consider the model $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \Lambda^{-1})$ given μ, Λ.

- Similar to the univariate case, a conjugate prior on μ is

$$\boldsymbol{\mu} \sim \mathcal{N}(m, L^{-1})$$

  for $m \in \mathbb{R}^d$ and $L \in \mathbb{R}^{d \times d}$ SPD.

- If Λ is fixed, then the resulting posterior is

$$\boldsymbol{\mu} | x_{1:n} \sim \mathcal{N}(m_n, L_n^{-1})$$

  where

$$\begin{aligned} L_n &= L + n\Lambda \\ m_n &= L_n^{-1}(Lm + \Lambda \textstyle\sum_{i=1}^n x_i). \end{aligned}$$

Conjugate prior for the precision matrix

- Univariate: $\mathrm{Gamma}(a, b)$ is conjugate prior on λ.
- Multivariate: $\mathrm{Wishart}(S^{-1}, \nu)$ is conjugate prior on Λ.

- Given $S \in \mathbb{R}^{d\times d}$ SPD and $\nu > d - 1$, the Wishart distribution with inverse scale S and ν degrees of freedom has density

$$W_d(X \mid S^{-1}, \nu) = \frac{|S|^{\nu/2}|X|^{(\nu-d-1)/2}\exp(-\frac{1}{2}\mathrm{tr}(SX))}{2^{\nu d/2}\Gamma_d(\nu/2)}$$

for $X \in \mathbb{R}^{d\times d}$ SPD.

- Here, $\Gamma_d(\nu/2)$ is the multivariate gamma function, and tr is the trace, i.e., $\mathrm{tr}(A) = \sum_{i=1}^{d} A_{ii}$.

Conjugate prior for the precision matrix

- Consider the model $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \Lambda^{-1})$ given μ, Λ.

- If μ is fixed and $\Lambda \sim \text{Wishart}(S^{-1}, \nu)$, then the posterior is

  $$\Lambda | x_{1:n} \sim \text{Wishart}(S_n^{-1}, \nu_n)$$

  where $\nu_n = \nu + n$ and

  $$S_n = S + \sum_{i=1}^{n} (x_i - \mu)(x_i - \mu)^{\mathtt{T}}.$$

- Equivalently, one can put an Inverse Wishart prior on the covariance matrix C.

Joint priors on the mean and precision matrix

- Generalizing from the univariate case, the NormalWishart distribution is a conjugate prior on (μ, Λ).
- Likewise, the Normal-InvWishart is conjugate for (μ, C).

- However, as in the univariate case, we often prefer to place independent priors on μ and Λ (or μ and C).

- Thus, we often prefer the conditionally conjugate prior:

$$\mu \sim \mathcal{N}(m, L^{-1}) \qquad \Lambda \sim \mathrm{Wishart}(S^{-1}, \nu)$$

independently.

History

Gauss

Maxwell

Quetelet

- In 1809, C.F. Gauss introduced the normal distribution as a model for the errors made in astronomical measurements, to justify the method of least squares in linear regression.
- Laplace proved the central limit theorem in 1810 and calculated the normalization constant of the normal.
- James Clerk Maxwell (1831–1879) showed that the normal distribution arose naturally in physics, particularly in thermodynamics.
- Adolphe Quetelet (1796–1874) pioneered the use of the normal distribution in the social sciences.

References and supplements

- Schilling, M. F., Watkins, A. E., & Watkins, W. (2002). Is human height bimodal? The American Statistician, 56(3), 223-229.
- Krul, A. J., Daanen, H. A., & Choi, H. (2011). Self-reported and measured weight, height and body mass index (BMI) in Italy, the Netherlands and North America. The European Journal of Public Health, 21(4), 414-419.
- Crow, J. F. (1997), "Birth Defects, Jimson Weeds and Bell Curves," Genetics, 147, 1-6.
- The Hartford Courant (1996), "Reaching New Heights," November 23, 1996; photo by K. Hanley.
- Helguerro, F. (1904), Sui Massimi Delle Curve Dimorfiche. Biometrika, 3, 85-98.
- Rosenthal, R., & Jacobson, L. (1968). Pygmalion in the classroom. The Urban Review, 3(1), 16-20.

Individual activity: Exit ticket

Answer these questions individually:
https://forms.gle/SAm9W25RzDNP9W1GA