Stat260: Bayesian Modeling and Inference **Lecture Date: February 8^{th}, 2010**

The Conjugate Prior for the Normal Distribution

Lecturer: Michael I. Jordan *Scribe: Teodor Mihai Moldovan*

We will look at the Gaussian distribution from a Bayesian point of view. In the standard form, the likelihood has two parameters, the mean μ and the variance σ^2:

$$P(x_1, x_2, \cdots, x_n \mid \mu, \sigma^2) \propto \frac{1}{\sigma^n} \exp\left(-\frac{1}{2\sigma^2} \sum (x_i - \mu)^2\right) \tag{1}$$

Our aim is to find conjugate prior distributions for these parameters. We will investigate the hyper-parameter (prior parameter) update relations and the problem of predicting new data from old data: $P(x_{\text{new}} \mid x_{\text{old}})$.

1 Fixed variance (σ^2), random mean (μ)

Keeping σ^2 fixed, the conjugate prior for μ is a Gaussian.

$$P(\mu \mid \underset{\substack{\swarrow \\ \text{typically } 0}}{\mu_0}\ ,\ \underset{\substack{\searrow \\ \text{typically large}}}{\sigma_0^2}\ ) \propto \frac{1}{\sigma_0} \exp\left(-\frac{1}{2\sigma_0^2}(\mu - \mu_0)^2\right) \tag{2}$$

Remark 1. In practice, when little is known about μ, it is common to set the location hyper-parameter to zero and the scale to some large value.

1.1 Posterior for single measurement ($n = 1$)

We want to put together the prior (2) and the likelihood (1) to get the posterior $(\mu \mid x)$. For now, assume we have only one measurement ($n = 1$);

There are several ways to do this:

- We could multiply the two distributions directly and complete the square in the exponent.
- Note that μ and x have a joint Gaussian distribution. Then the conditional $\mu \mid x$ is also a Gaussian for whose parameters we know formulas:

Lemma 2. *Assume (z_1, z_2) is distributed according to a bivariate Gaussian. Then $z_1 \mid z_2$ is Gaussian distributed with parameters:*

$$E(z_1 \mid z_2) = E(z_1) + \frac{Cov(z_1, z_2)}{Var(z_2)}(z_2 - E(z_2)) \tag{3}$$

$$Var(z_1 \mid z_2) = Var(z_1) - \frac{Cov^2(z_1, z_2)}{Var(z_2)} \tag{4}$$

Remark 3. These formulas are extremely useful so you should memorize them. They are easily derived based on the notion of a Schur complement of a matrix.

We apply this lemma with the correspondence: $x \to z_2$, $\mu \to z_1$

$$\begin{aligned} x &= \mu + \sigma\varepsilon & \varepsilon &\sim \mathcal{N}(0,1) \\ \mu &= \mu_0 + \sigma_0\delta & \delta &\sim \mathcal{N}(0,1) \end{aligned}$$

$$E(x) = \mu_0 \tag{5}$$

$$\operatorname{Var}(x) = E(\operatorname{Var}(x \,|\, \mu)) + \operatorname{Var}(E(x \,|\, \mu)) = \sigma^2 + \sigma_0^2 \tag{6}$$

$$\operatorname{Cov}(x, \mu) = E(x - \mu_0)(\mu - \mu_0) = \sigma_0^2 \tag{7}$$

Using equations 3 and 4:

$$E(\mu \,|\, x) = \mu_0 + \frac{\sigma_0^2}{\sigma^2 + \sigma_0^2}(x - \mu_0) = \frac{\sigma_0^2}{\sigma^2 + \sigma_0^2}\ \underset{\substack{\searrow \\ \text{MLE}}}{x}\ + \frac{\sigma^2}{\sigma^2 + \sigma_0^2}\ \underset{\substack{\searrow \\ \text{prior mean}}}{\mu_0} \tag{8}$$

$$\operatorname{Var}(\mu \,|\, x) = \frac{\sigma^2\sigma_0^2}{\sigma^2 + \sigma_0^2} = \frac{1}{\frac{1}{\sigma_0^2} + \frac{1}{\sigma^2}} = (\tau_{\text{prior}} + \tau_{\text{data}})^{-1} \tag{9}$$

Definition 4. $1 \,/\, \sigma^2$ is usually called the *precision* and is denoted by τ

The posterior mean is usually a convex combination of the prior mean and the MLE.

The posterior precision is, in this case, the sum of the prior precision and the data precision

$$\tau_{\text{post}} = \tau_{\text{prior}} + \tau_{\text{data}}$$

We summarize our results so far:

Lemma 5. *Assume* $x \,|\, \mu \sim \mathcal{N}(\mu, \sigma^2)$ *and* $\mu \sim \mathcal{N}(\mu_0, \sigma_0^2)$. *Then:*

$$\mu \,|\, x \sim \mathcal{N}\left(\frac{\sigma_0^2}{\sigma^2 + \sigma_0^2}\ x\ + \frac{\sigma^2}{\sigma^2 + \sigma_0^2}\ \mu_0\ , \left(\frac{1}{\sigma_0^2} + \frac{1}{\sigma^2}\right)^{-1}\right)$$

1.2 Posterior for multiple measurements $(n \geq 1)$

Now look at the posterior update for multiple measurements. We could adapt our previous derivation, but that would be tedious since we would have to use the multivariate version of Lemma 2. Instead we will reduce the problem to the univariate case, with the sample mean $\bar{x} = (\sum x_i)/n$ as the new variable.

$$x_i \,|\, \mu \sim \mathcal{N}(\mu, \sigma^2) \text{ i.i.d.} \quad \Rightarrow \quad \bar{x} \,|\, \mu \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right) \tag{10}$$

$$\begin{aligned} P(x_1, x_2, \cdots, x_n \,|\, \mu) &\propto_\mu \frac{1}{\sigma}\exp\left(-\frac{1}{2\sigma^2}\sum (x_i - \mu)^2\right) \\ &\propto_\mu \exp\left(-\frac{1}{2\sigma^2}\left(\sum x_i^2 - 2\mu\sum x_i + n\mu^2\right)\right) \\ &\propto_\mu \exp\left(-\frac{n}{2\sigma^2}\left(-2\mu\bar{x} + \mu^2\right)\right) \propto_\mu \exp\left(-\frac{n}{2\sigma^2}(\bar{x} - \mu)^2\right) \\ &\propto_\mu P(\bar{x} \,|\, \mu) \end{aligned} \tag{11}$$

Then for the posterior probability, we get

$$\begin{aligned} P(\mu \,|\, x_1, x_2, \cdots, x_n) &\propto P(x_1, x_2, \cdots, x_n \,|\, \mu) P(\mu) \propto P(\bar{x} \,|\, \mu) P(\mu) \\ &\propto P(\mu \,|\, \bar{x}) \end{aligned} \tag{12}$$

We can now plug $\bar{x}$ into our previous result and we get:

Lemma 6. *Assume $x_i \,|\, \mu \sim \mathcal{N}(\mu, \sigma^2)$ i.i.d.and $\mu \sim \mathcal{N}(\mu_0, \sigma_0^2)$. Then:*

$$\mu \,|\, x_1, x_2, \cdots, x_n \sim \mathcal{N}\left(\frac{\sigma_0^2}{\frac{\sigma^2}{n} + \sigma_0^2} x + \frac{\sigma^2}{\frac{\sigma^2}{n} + \sigma_0^2} \mu_0, \left(\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2} \right)^{-1} \right)$$

2 Random variance (σ^2), fixed mean (μ)

2.1 Posterior

Assuming μ is fixed, then the conjugate prior for σ^2 is an inverse Gamma distribution:

$$z \,|\, \alpha, \beta \sim \mathrm{IG}(\alpha, \beta) \qquad P(z \,|\, \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} z^{-\alpha-1} \exp\left(-\frac{\beta}{z} \right) \tag{13}$$

For the posterior we get another inverse Gamma:

$$\begin{aligned} P(\sigma^2 \,|\, \alpha, \beta) &\propto (\sigma^2)^{-(\alpha + \frac{n}{2}) - 1} \exp\left(-\frac{\beta + \frac{1}{2} \sum (x_i - \mu)}{\sigma^2} \right) \\ &\propto (\sigma^2)^{-\alpha_{\text{post}} - 1} \exp\left(-\frac{\beta_{\text{post}}}{\sigma^2} \right) \end{aligned} \tag{14}$$

Lemma 7. *If $x_i \,|\, \mu, \sigma^2 \sim \mathcal{N}(\mu, \sigma^2)$ i.i.d.and $\sigma^2 \sim \mathrm{IG}(\alpha, \beta)$. Then:*

$$\sigma^2 \,|\, x_1, x_2, \cdots, x_n \sim \mathrm{IG}\left(\alpha + \frac{n}{2}, \beta + \frac{1}{2} \sum (x_i - \mu) \right)$$

If we re-parametrize in terms of precisions, the conjugate prior is a Gamma distribution.

$$\tau \,|\, \alpha, \beta \sim \mathrm{Ga}(\alpha, \beta) \qquad P(\tau \,|\, \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} \tau^{\alpha-1} \exp(-\tau\beta) \tag{15}$$

And the posterior is:

$$P(\tau \,|\, \alpha, \beta) \propto \tau^{(\alpha + \frac{n}{2}) - 1} \exp\left(-\tau \left(\beta + \frac{1}{2} \sum (x_i - \mu) \right) \right) \tag{16}$$

Lemma 8. *If $x_i \,|\, \mu, \tau \sim \mathcal{N}(\mu, \tau)$ i.i.d.and $\tau \sim \mathrm{Ga}(\alpha, \beta)$. Then:*

$$\tau \,|\, x_1, x_2, \cdots, x_n \sim \mathrm{Ga}\left(\alpha + \frac{n}{2}, \beta + \frac{1}{2} \sum (x_i - \mu) \right)$$

Remark 9. Should we prefer working with variances or precisions? We should prefer both:

- Variances add when we marginalize
- Precisions add when we condition

2.2 Prediction

We might want to compute the probability of getting some new data given old data. This can be done by marginalizing out parameters:

$$\begin{aligned} P(x_{\text{new}} \,|\, x, \mu, \alpha, \beta) &= \int P(x_{\text{new}} \,|\, x, \mu, \tau, \alpha, \beta) P(\tau \,|\, x, \alpha, \beta) d\tau \\ &= \int P(x_{\text{new}} \,|\, \mu, \tau) P(\tau \,|\, x, \alpha, \beta) d\tau \\ &= \int P(x_{\text{new}} \,|\, \mu, \tau) P(\tau \,|\, \alpha_{\text{post}}, \beta_{\text{post}}) d\tau \qquad (17) \end{aligned}$$

This integral "smears" the Gaussian into a heavier tailed distribution, which will turn out to be a student's t-distribution:

$$\begin{aligned} \tau \,|\, \alpha, \beta &\sim \text{Ga}(\alpha, \beta) \\ x \,|\, \tau, \mu &\sim \mathcal{N}(\mu, \tau) \end{aligned}$$

$$\begin{aligned} P(x \,|\, \mu, \alpha, \beta) &= \int \frac{\beta^\alpha}{\Gamma(\alpha)} \tau^{\alpha-1} e^{-\tau\beta} \left(\frac{\tau}{2\pi}\right)^{\frac{1}{2}} \exp\left(-\frac{\tau}{2}(x-\mu)^2\right) d\tau \\ &= \frac{\beta^\alpha}{\Gamma(\alpha)} \frac{1}{\sqrt{2\pi}} \int \tau^{\left(\alpha+\frac{1}{2}\right)-1} e^{-\tau\left(\beta+(x-\mu)^2\right)/2}\, d\tau \\ & \qquad\qquad\qquad\qquad\qquad\qquad\qquad \swarrow \\ & \qquad\qquad \text{Gamma integral; use memorized normalizing constant} \\ &= \frac{\beta^\alpha}{\Gamma(\alpha)} \frac{1}{\sqrt{2\pi}} \frac{\Gamma\left(\alpha+\frac{1}{2}\right)}{\left(\beta+\frac{1}{2}(x-\mu)^2\right)^{\alpha+\frac{1}{2}}} \\ &= \frac{\Gamma\left(\alpha+\frac{1}{2}\right)}{\Gamma(\alpha)} \frac{1}{(2\pi\beta)^{\frac{1}{2}}} \frac{1}{\left(1+\frac{1}{2\beta}(x-\mu)^2\right)^{\alpha+\frac{1}{2}}} \qquad (18) \end{aligned}$$

Remark 10. The student-t density has three parameters: μ, α, β and is symmetric around μ. When α is an integer or a half-integer we get simplifications using the formulas $\Gamma(k+1) = k\Gamma(k)$ and $\Gamma(1/2) = \sqrt{\pi}$

The following is another useful parametrization for the student's t-distribution:

$$p = 2\alpha \qquad\qquad \lambda = \frac{\alpha}{\beta}$$

$$P(x \,|\, \mu, p, \lambda) = \frac{\Gamma\left(\frac{p+1}{2}\right)}{\Gamma\left(\frac{p}{2}\right)} \left(\frac{\lambda}{\pi p}\right)^{\frac{1}{2}} \frac{1}{\left(1+\frac{\lambda}{p}(x-\mu)^2\right)^{\frac{p+1}{2}}} \qquad (19)$$

with two interesting special cases:

- If $p = 1$ we get a Cauchy distribution
- If $p \to \infty$ we get a Gaussian distribution

Remark 11. We might want to sample from a student's t-distribution. We would sample $\tau \sim \text{Ga}(\alpha, \beta)$, then sample $x_i \sim \mathcal{N}(\mu, \tau)$, collect x_i and repeat.

3 Both variance (σ^2) and mean (μ) are random

Now, we want to put a prior on μ and σ^2 together. We could simply multiply the prior densities we obtained in the previous two sections, implicitly assuming μ and σ^2 are independent. Unfortunately, if we did that, we would not get a conjugate prior. One way to see this is that if we believe that our data is generated according to the graphical model in Figure 1, we find that, conditioned on x, the two parameters μ and σ^2 are, in fact, dependent and this should be expressed by a conjugate prior.

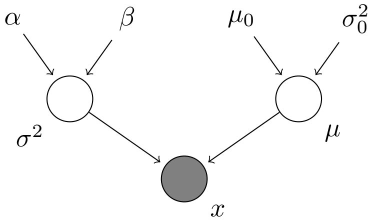

Figure 1: μ and σ^2 are dependent conditioned on x

We will use the following prior distribution which, as we will show, is conjugate to the Gaussian likelihood:

$$\begin{gathered} x_i \,|\, \mu, \tau \sim \mathcal{N}(\mu, \tau) \quad \text{i.i.d.} \\ \mu \,|\, \tau \sim \mathcal{N}(\mu_0, n_0 \tau) \\ \tau \sim \text{Ga}(\alpha, \beta) \end{gathered}$$

3.1 Posterior

First look at $\mu \,|\, x, \tau$. This is the simpler part, as we can use Lemma 8:

$$\mu \,|\, x, \tau \sim \mathcal{N}\left(\frac{n\tau}{n\tau + n_0\tau}\bar{x} + \frac{n_0\tau}{n\tau + n_0\tau}\mu_0 \quad , \quad n\tau + n_0\tau \right) \tag{20}$$

Next, look at $\tau \,|\, x$. We get this by expressing the joint density $P(\tau, \mu \,|\, x)$ and marginalizing out μ:

$$\begin{aligned} P(\tau, \mu \,|\, x) &\propto P(\tau) \cdot P(\mu \,|\, \tau) \cdot P(x \,|\, \tau, \mu) && (21) \\ &\propto \tau^{\alpha-1} e^{-\beta\tau} \tau^{1/2} \exp\left(-\frac{n_0\tau}{2}(\mu - \mu_0)^2 \right) \tau^{n/2} \exp\left(-\frac{\tau}{2}\sum (x_i - \mu)^2 \right) \\ &\qquad \text{trick: } x_i - \bar{x} + \bar{x} - \mu \\ &\propto \tau^{\alpha + \frac{n}{2} - 1} \exp\left(-\tau\left(\beta + \frac{1}{2}\sum (x_i - \bar{x})^2 \right) \right) \tau^{1/2} \exp\left(-\frac{\tau}{2}(n_0(\mu - \mu_0)^2 + n(\bar{x} - \mu)^2) \right) && (22) \end{aligned}$$

As we integrate out μ we get the normalization constant:

$$\tau^{-\frac{1}{2}} \exp\left(\frac{n n_0 \tau}{2(n + n_0)}(\bar{x} - \mu_0)^2 \right)$$

Which leads to a Gamma posterior for τ:

$$P(\tau \,|\, x) \propto \tau^{\alpha + \frac{n}{2} - 1} \exp\left(-\tau\left(\beta + \frac{1}{2}\sum (x_i - \bar{x})^2 + \frac{n n_0}{2(n + n_0)}(\bar{x} - \mu_0)^2 \right) \right) \tag{23}$$

To summarize:

Lemma 12. *If we assume:*

$$\begin{gathered} x_i \mid \mu, \tau \sim \mathcal{N}(\mu, \tau) \quad \textit{i.i.d.} \\ \mu \mid \tau \sim \mathcal{N}(\mu_0, n_0\tau) \\ \tau \sim \mathrm{Ga}(\alpha, \beta) \end{gathered}$$

Then the posterior is:

$$\begin{aligned} \mu \mid \tau, x &\sim \mathcal{N}\left(\frac{n\tau}{n\tau + n_0\tau}\bar{x} + \frac{n_0\tau}{n\tau + n_0\tau}\mu_0 \quad , \quad n\tau + n_0\tau\right) \\ \tau \mid x &\sim \mathrm{Ga}\left(\alpha + \frac{n}{2} \quad , \quad \beta + \frac{1}{2}\sum(x_i - \bar{x})^2 + \frac{nn_0}{2(n+n_0)}(\bar{x} - \mu_0)^2\right) \end{aligned}$$

3.2 Prediction

$$\begin{aligned} P(x_{\mathrm{new}} \mid x) &= \int\int \underset{\tau \mid x}{\mathrm{Gamma}} \cdot \underset{\mu \mid \tau, x}{\mathrm{Gaussian}} \cdot \underset{x_{\mathrm{new}} \mid \tau, \mu}{\mathrm{Gaussian}} \quad d\tau d\mu \\ P(x_{\mathrm{new}} \mid x) &= \int \underset{\tau \mid x}{\mathrm{Gamma}} \int \underset{\mu \mid \tau, x}{\mathrm{Gaussian}} \cdot \underset{x_{\mathrm{new}} \mid \tau, \mu}{\mathrm{Gaussian}} \quad d\tau d\mu \\ P(x_{\mathrm{new}} \mid x) &= \int \underset{\tau \mid x}{\mathrm{Gamma}} \cdot \underset{x_{\mathrm{new}} \mid \tau, x}{\mathrm{Gaussian}} \quad d\tau \\ P(x_{\mathrm{new}} \mid x) &= \underset{x_{\mathrm{new}} \mid x}{\text{student-t}} \end{aligned}$$