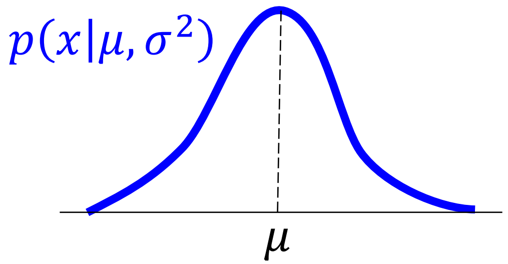
Bayesian Inference for Mean of a Univariate Gaussian

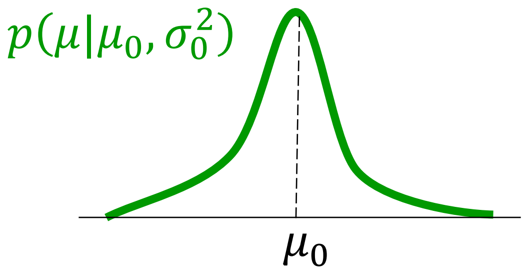
Assumed fixed

- Consider N i.i.d. scalar obs $\mathbf{X} = \{x_1, x_2, \dots, x_N\}$ drawn from $p(x|\mu,\sigma^2) = \mathcal{N}(x|\mu,\sigma^2)$



$$p(x_n|\mu,\sigma^2) = \mathcal{N}(x|\mu,\sigma^2) \propto \exp\left[-\frac{(x_n-\mu)^2}{2\sigma^2}\right]$$

$$p(\mathbf{X}|\mu,\sigma^2) = \prod_{n=1}^{N} p(x_n|\mu,\sigma^2)$$



- Each x_n is a noisy measurement of $\mu \in \mathbb{R}$, i.e., $x_n = \mu + \epsilon_n$ where $\epsilon_n \sim \mathcal{N}(0,\sigma^2)$
- Would like to estimate μ given $\mathbf{X}$ using fully Bayesian inference (not point estimation)
- Need a prior over μ. Let's choose a Gaussian $p(\mu|\mu_0,\sigma_0^2) = \mathcal{N}(\mu|\mu_0,\sigma_0^2)$

  Assume μ_0 and σ_0^2 to be fixed/known

  - The prior basically says that *a priori* μ is close to μ_0
  - The prior's variance σ_0^2 tells us how certain we are about the above assumption
  - Since σ^2 in the likelihood model $\mathcal{N}(x|\mu,\sigma^2)$ is known, the Gaussian prior $\mathcal{N}(x|\mu_0,\sigma_0^2)$ on μ is also conjugate to the likelihood (thus posterior of μ will also be Gaussian)

Bayesian Inference for Mean of a Univariate Gaussian

- The posterior distribution for the unknown mean parameter μ

On conditioning side, skipping all fixed params and hyperparams from the notation

$$p(\mu|\mathbf{X}) = \frac{p(\mathbf{X}|\mu)p(\mu)}{p(\mathbf{X})} \propto \prod_{n=1}^{N} \exp\left[-\frac{(x_n-\mu)^2}{2\sigma^2}\right] \times \exp\left[-\frac{(\mu-\mu_0)^2}{2\sigma_0^2}\right]$$

- Simplifying the above (using completing the squares trick – see note) gives

$$p(\mu|\mathbf{X}) \propto \exp\left[-\frac{(\mu-\mu_N)^2}{2\sigma_N^2}\right]$$

Gaussian posterior (not a surprise since the chosen prior was conjugate to the likelihood)

Gaussian posterior's precision is the sum of the prior's precision and sum of the noise precisions of all the observations

$$\frac{1}{\sigma_N^2} = \frac{1}{\sigma_0^2} + \frac{N}{\sigma^2}$$

Contribution from the prior

Contribution from the data

Also the MLE solution for μ

Gaussian posterior's mean is a convex combination of prior's mean and the MLE solution

$$\mu_N = \frac{\sigma^2}{N\sigma_0^2+\sigma^2}\mu_0 + \frac{N\sigma_0^2}{N\sigma_0^2+\sigma^2}\bar{x} \quad \left(\text{where } \bar{x} = \frac{\sum_{n=1}^N x_n}{N}\right)$$

- What happens to the posterior as N (number of observations) grows very large?
  - Data (likelihood part) overwhelms the prior
  - Posterior's variance σ_N^2 will approximately be σ^2/N (and goes to 0 as $N \to \infty$)
  - The posterior's mean μ_N approaches $\bar{x}$ (which is also the MLE solution)

Meaning, we become very-very certain about the estimate of μ

Bayesian Inference for Mean of a Univariate Gaussian

- Using the inferred posterior $p(\mu|\mathbf{X})$, we can find the posterior predictive distribution

Assumed fixed, only μ is the unknown here

$$p(x_*|\mathbf{X}) = \int p(x_*|\mu,\sigma^2)p(\mu|\mathbf{X})d\mu$$

$$= \int \mathcal{N}(x_*|\mu,\sigma^2)\mathcal{N}(\mu|\mu_N,\sigma_N^2)d\mu$$

$$= \mathcal{N}(x_*|\mu_N,\sigma^2+\sigma_N^2)$$

On conditioning side, skipping all fixed params and hyperparams from the notation

Conditional of x_* given μ is Gaussian, and μ has a Gaussian posterior, so marginal of x_* (after we marginalize μ) will also be a Gaussian

PRML [Bis 06], 2.115, and also mentioned in prob-stats refresher slides

Result follows from properties of Gaussian and noting that a PPD is also a marginal distribution

This "extra" variance is due to the averaging over the posterior's uncertainty

A useful fact: When we have conjugacy, the posterior predictive distribution also has a closed form (will see this result more formally when talking about exponential family distributions)

- For an alternative way to get the above result, note that

$$x_* = \mu + \epsilon \qquad \mu \sim \mathcal{N}(\mu_N,\sigma_N^2) \qquad \epsilon \sim \mathcal{N}(0,\sigma^2)$$

$$\Rightarrow \quad p(x_*|\mathbf{X}) = \mathcal{N}(x_*|\mu_N,\sigma^2+\sigma_N^2)$$

Since both μ and ϵ are Gaussian r.v., and are independence, x_* is also has a Gaussian predictive, and the respective means and variances of μ and ϵ get added up

- In contrast, the plug-in predictive given a point estimate $\hat{\mu}$ will be

$$p(x_*|\mathbf{X}) = \int p(x_*|\mu,\sigma^2)p(\mu|\mathbf{X})d\mu \approx p(x_*|\hat{\mu},\sigma^2) = \mathcal{N}(x_*|\hat{\mu},\sigma^2)$$

Note that PPD had a larger variance $(\sigma^2+\sigma_N^2)$

Bayesian Inference for Variance of a Univariate Gaussian

- Consider N i.i.d. scalar obs $\mathbf{X} = \{x_1, x_2, \dots, x_N\}$ drawn from $\mathcal{N}(x|\mu, \sigma^2)$

$$p(x_n|\mu, \sigma^2) = \mathcal{N}(x|\mu, \sigma^2) \quad \text{and} \quad p(\mathbf{X}|\mu, \sigma^2) = \prod_{n=1}^{N} p(x_n|\mu, \sigma^2)$$

- Assume the variance $\sigma^2 \in \mathbb{R}_+$ to be unknown and mean μ to be fixed/known
- Would like to estimate σ^2 given $\mathbf{X}$ using fully Bayesian inference (not point estimation)
- Need a prior over σ^2. What prior $p(\sigma^2)$ to choose in this case?
- If we want a conjugate prior, it should have the same form as the likelihood

$$p(x_n|\mu, \sigma^2) \propto (\sigma^2)^{-1/2} \exp\left[-\frac{(x_n - \mu)^2}{2\sigma^2}\right]$$

- An inverse-gamma dist $IG(\alpha, \beta)$ has this form (α, β are shape and scale hyperparams)

$$p(\sigma^2) \propto (\sigma^2)^{-(\alpha+1)} \exp\left[-\frac{\beta}{\sigma^2}\right] \qquad \left(\text{note: mean of } IG(\alpha, \beta) = \frac{\beta}{\alpha - 1}\right)$$

- Due to conjugacy, posterior will also be IG: $p(\sigma^2|\mathbf{X}) = IG\left(\alpha + \frac{N}{2}, \beta + \frac{\sum_{n=1}^N (x_n - \mu)^2}{2}\right)$

Working with Gaussians: Variance vs Precision

- Often, it is easier to work with the precision (=1/variance) rather than variance

$$p(x_n|\mu,\lambda^{-1}) = \mathcal{N}(x|\mu,\lambda^{-1}) = \sqrt{\frac{\lambda}{2\pi}}\exp\left[-\frac{\lambda}{2}(x_n-\mu)^2\right]$$

- If mean is known, for precision, $\mathbf{Gamma}(\alpha,\beta)$ is a conjugate prior to Gaussian lik.

PDF of gamma distribution

$$p(\lambda) \propto (\lambda)^{(\alpha-1)}\exp[-\beta\lambda]$$

(Note: mean of $\mathbf{Gamma}(\alpha,\beta) = \frac{\alpha}{\beta}$)

α and β are the shape and rate params, resp., of the Gamma distribution

- (Verify) The posterior $p(\lambda\,|X)$ will be $\text{Gamma}\left(\alpha+\frac{N}{2},\beta+\frac{\sum_{n=1}^N(x_n-\mu)^2}{2}\right)$

- Note: Unlike the case of unknown mean and fixed variance, the PPD for this case (and also the unknown variance case) will not be a Gaussian

- Note: Gamma distribution can be defined in terms of shape and scale or shape and rate parametrization (scale = 1/rate). Likewise, inverse Gamma can also be defined both shape and scale (which we saw) as well as shape and rate parametrizations.

Bayesian Inference for Both Parameters of a Gaussian

- Gaussian with unknown scalar mean and unknown scalar precision (two parameters)
- Consider N i.i.d. scalar obs $\mathbf{X} = \{x_1, x_2, \ldots, x_N\}$ drawn from $\mathcal{N}(x|\mu, \lambda^{-1})$
- Assume both mean μ and precision λ to be unknown. The likelihood can be written as

$$
\begin{aligned}
p(\mathbf{X}|\mu,\lambda) &= \prod_{n=1}^{N} \sqrt{\frac{\lambda}{2\pi}} \exp\left[-\frac{\lambda}{2}(x_n-\mu)^2\right] \\
&\propto \left[\lambda^{1/2}\exp\left(-\frac{\lambda\mu^2}{2}\right)\right]^N \exp\left[\lambda\mu\sum_{n=1}^{N}x_n - \frac{\lambda}{2}\sum_{n=1}^{N}x_n^2\right]
\end{aligned}
$$

- Would like a joint conjugate prior distribution $p(\mu, \lambda)$
  - It must have the same form as the likelihood as written above. Basically, something that looks like

$$
p(\mu,\lambda) \propto \left[\lambda^{1/2}\exp\left(-\frac{\lambda\mu^2}{2}\right)\right]^{\kappa_0} \exp\left[\lambda\mu c - \lambda d\right]
$$

Thankfully, this is a known distribution: normal-gamma (NG) distribution ☺

Called so since it can be written as a product of a normal and a gamma (next slide)

The NG also has a multivariate version called normal-Wishart distribution to jointly model a real-valued vector and a PSD matrix

Detour: Normal-gamma (Gaussian-gamma) Distribution

- We saw that the conjugate prior needed to have the form

$$
\begin{aligned}
p(\mu,\lambda) &\propto \left[\lambda^{1/2}\exp\left(-\frac{\lambda\mu^2}{2}\right)\right]^{\kappa_0}\exp\left[\lambda\mu c-\lambda d\right] \\
&= \underbrace{\exp\left[-\frac{\kappa_0\lambda}{2}(\mu-c/\kappa_0)^2\right]}_{\text{prop. to a Gaussian}}\underbrace{\lambda^{\kappa_0/2}\exp\left[-\left(d-\frac{c^2}{2\kappa_0}\right)\lambda\right]}_{\text{prop. to a gamma}} \qquad \text{(re-arranging terms)}
\end{aligned}
$$

- The above is product of a normal and a gamma distribution

Assuming shape-rate parametrization of the gamma

$$p(\mu,\lambda)=\mathcal{N}(\mu|\mu_0,(\kappa_0\lambda)^{-1})\text{Gamma}(\lambda|\alpha_0,\beta_0)=\text{NG}(\mu_0,\kappa_0,\alpha_0,\beta_0)$$

where $\mu_0=c/\kappa_0$, $\alpha_0=1+\kappa_0/2$, $\beta_0=d-c^2/2\kappa_0$ are prior's hyperparameters

- The NG $p(\mu,\lambda)=\text{NG}(\mu_0,\kappa_0,\alpha_0,\beta_0)$ is conjugate to a Gaussian distribution if both its mean and precision parameters are unknown and are to be estimated
  - Thus a useful prior in many problems involving Gaussians with unknown mean and precision

Bayesian Inference for Both Parameters of a Gaussian

- Due to conjugacy, the joint posterior $p(\mu, \lambda|\mathbf{X})$ will also be normal-gamma

Skipping all hyperparameters on the conditioning side

$$p(\mu, \lambda|\mathbf{X}) \propto p(\mathbf{X}|\mu, \lambda)p(\mu, \lambda)$$

- Plugging in the expressions for $p(\mathbf{X}|\mu, \lambda)$ and $p(\mu, \lambda)$, we get

$$p(\mu, \lambda|\mathbf{X}) = \mathrm{NG}(\mu_N, \kappa_N, \alpha_N, \beta_N) = \mathcal{N}(\mu|\mu_N, (\kappa_N\lambda)^{-1})\mathrm{Gamma}(\lambda|\alpha_N, \beta_N)$$

- The above's posterior's parameters will be

$$\begin{aligned} \mu_N &= \frac{\kappa_0\mu_0 + N\bar{x}}{\kappa_0 + N} \\ \kappa_N &= \kappa_0 + N \\ \alpha_N &= \alpha_0 + N/2 \\ \beta_N &= \beta_0 + \frac{1}{2}\sum_{n=1}^{N}(x_n - \bar{x})^2 + \frac{\kappa_0 N(\bar{x} - \mu_0)^2}{2(\kappa_0 + N)} \end{aligned}$$

(For full derivation of posterior, refer to "Conjugate Bayesian analysis of the Gaussian distribution" - Murphy (2007))

Other Quantities of Interest

- We saw that the joint posterior for mean and precision is NG

$$p(\mu, \lambda|\mathbf{X}) = \mathrm{NG}(\mu_N, \kappa_N, \alpha_N, \beta_N) = \mathcal{N}(\mu|\mu_N, (\kappa_N\lambda)^{-1})\mathrm{Gamma}(\lambda|\alpha_N, \beta_N)$$

- From the above, we can also obtain the marginal posteriors for $\boldsymbol{\mu}$ and $\boldsymbol{\lambda}$

$$p(\lambda|\mathbf{X}) = \int p(\mu, \lambda|\mathbf{X})d\mu = \mathrm{Gamma}(\lambda|\alpha_N, \beta_N)$$

$$p(\mu|\mathbf{X}) = \int p(\mu, \lambda|\mathbf{X})d\lambda = \int p(\mu|\lambda, \mathbf{X})p(\lambda|\mathbf{X})d\lambda = \underbrace{t_{2\alpha_N}(\mu|\mu_N, \beta_N/(\alpha_N\kappa_N))}_{\text{t distribution}}$$

- Marginal likelihood of the model

$$p(\mathbf{X}) = \frac{\Gamma(\alpha_N)}{\Gamma(\alpha_0)}\frac{\beta_0^{\alpha_0}}{\beta_N^{\alpha_N}}\left(\frac{\kappa_0}{\kappa_N}\right)^{\frac{1}{2}}(2\pi)^{-N/2}$$

Marginal lik has closed form expression (for conjugate lik and prior, the marginal lik has closed form – more when we see exp-family distributions)

- PPD of a new observation x_*

$$p(x_*|\mathbf{X}) = \int \underbrace{p(x_*|\mu, \lambda)}_{\text{Gaussian}} \underbrace{p(\mu, \lambda|\mathbf{X})}_{\text{Normal-Gamma}} d\mu d\lambda = t_{2\alpha_N}\left(x_*|\mu_N, \frac{\beta_N(\kappa_N+1)}{\alpha_N\kappa_N}\right)$$

(For full derivation of posterior, refer to "Conjugate Bayesian analysis of the Gaussian distribution" - Murphy (2007))

An Aside: Student-t distribution

- An infinite sum of Gaussian distributions, with same means but different precisions

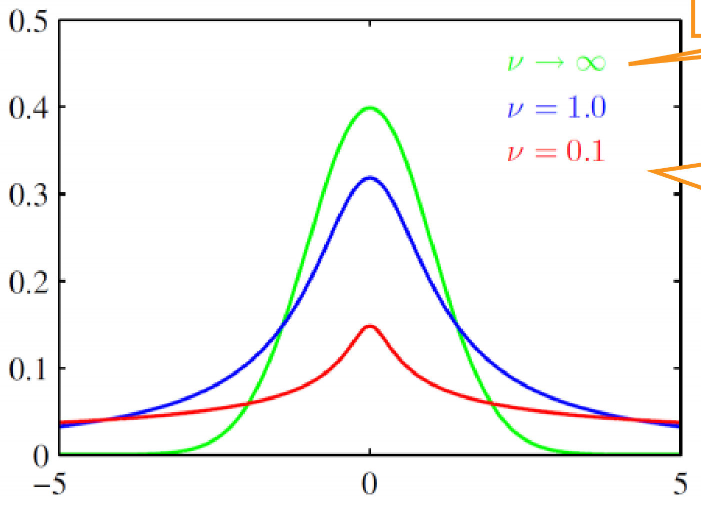
$$
\begin{aligned}
p(x|\mu, a, b) &= \int \mathcal{N}(x|\mu, \lambda^{-1})\text{Gamma}(\lambda|a, b)d\lambda \\
&= t_{2a}(x|\mu, b/a) = t_{\nu}(x|\mu, \sigma^2)
\end{aligned}
$$

Same as saying that we are integrating out the precision parameter of a Gaussian with the mean held as fixed

- $\nu > 0$ is called the degree of freedom, μ is the mean, and σ^2 is the scale

$$
\begin{aligned}
t_{\nu}(x|\mu, \sigma^2) &= c\left[1 + \frac{1}{\nu}\left(\frac{x-\mu}{\sigma}\right)^2\right]^{-\left(\frac{\nu+1}{2}\right)} \\
c &= \frac{\Gamma(\nu/2 + 1/2)}{\Gamma(\nu/2)}\frac{1}{\sqrt{\nu\pi}\sigma}
\end{aligned}
$$

$$
\begin{aligned}
\text{mean} &= \mu, \ \nu > 1 \\
\text{mode} &= \mu \\
\text{var} &= \frac{\nu\sigma^2}{(\nu-2)}, \ \nu > 2
\end{aligned}
$$



As ν tends to infinity, student-t becomes a Gaussian

Has fatter tail than Gaussian and is sharper around the mean

Zero-mean Student-t (and other such "infinite sum of Gaussians" are useful priors for modeling sparse weights

Inferring Params of Gaussian: Some Other Cases

- We only considered parameter estimation for univariate Gaussian distribution
  - The approach also extends to inferring parameters of a multivariate Gaussian
  - For the unknown mean and precision matrix, normal-Wishart can be used as prior
- Posterior updates have forms similar to that in the univariate case
- When working with mean-variance, can use normal-inverse gamma as conjugate prior
  - For multivariate Gaussian, can use normal-inverse Wishart for mean-covariance pair
- Other priors can also be used as well when inferring parameters of Gaussians, e.g.,
  - normal-Inverse χ^2 commonly used in Statistics community for scalar mean-variance estimation
- May also refer to "Conjugate Bayesian analysis of the Gaussian distribution" - Murphy (2007) for various examples and more detailed derivations

Coming Up Next

- Working with multivariate Gaussians
- Exponential Family distributions
- Conditional Models (e.g., supervised learning – regression and classification)