Recap: Bayesian Inference for Mean of a Gaussian

- Consider \( N \) i.i.d. observations \( X = \{x_1, \ldots, x_N\} \) drawn from a one-dim Gaussian \( \mathcal{N}(x|\mu, \sigma^2) \)
  
  \[
p(x_n|\mu, \sigma^2) = \mathcal{N}(x|\mu, \sigma^2) \propto \exp\left[-\frac{(x_n - \mu)^2}{2\sigma^2}\right]
  \]
  
  \[
p(\mu) = \mathcal{N}(\mu|\mu_0, \sigma_0^2)
  \]

- Due to conjugacy, posterior is also Gaussian: \( p(\mu|X) \propto \exp\left[-\frac{(\mu-\mu_N)^2}{2\sigma_N^2}\right] \)
  with
  
  \[
  \frac{1}{\sigma_N^2} = \frac{1}{\sigma_0^2} + \frac{N}{\sigma^2} \\
  \mu_N = \frac{\sigma^2}{N\sigma_0^2 + \sigma^2} \mu_0 + \frac{N\sigma_0^2}{N\sigma_0^2 + \sigma^2} \bar{x} \\
  \text{(where } \bar{x} = \frac{\sum_{n=1}^{N} x_n}{N} \text{)}
  \]

- Posterior predictive for a new observation \( x_* \) is also Gaussian
  
  \[
p(x_*|X) = \int p(x_*|\mu, \sigma^2)p(\mu|X)d\mu = \int \mathcal{N}(x_*|\mu, \sigma^2)\mathcal{N}(\mu|\mu_N, \sigma_N^2)d\mu = \mathcal{N}(x_*|\mu_N, \sigma^2 + \sigma_N^2)
  \]
  
  (can also obtain the above by noting that \( x_* = \mu + \epsilon_* \) where \( \mu \sim \mathcal{N}(\mu_N, \sigma_N^2) \) and \( \epsilon_* \sim \mathcal{N}(0, \sigma^2) \))

- Exercise: Compute the posterior if \( p(\mu) = \mathcal{N}(\mu|\mu_0, \sigma_0^2) \). Also, what does \( \kappa_0 \) mean intuitively?
Recap: Bayesian Inference for Variance/Precision of a Gaussian

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

- Conjugate prior for variance \( \sigma^2 \) is inverse-gamma: \( p(\sigma^2|\alpha, \beta) = IG(\alpha, \beta) \)
  
  \[ p(\sigma^2) \propto (\sigma^2)^{-(\alpha+1)} \exp\left[-\frac{\beta}{\sigma^2}\right] \quad \text{(note: } \alpha = \text{shape}, \beta = \text{scale}; \text{mean of } IG(\alpha, \beta) = \frac{\beta}{\alpha - 1}) \]

- Given \( N \) i.i.d. observations \( X = \{x_1, \ldots, x_N\} \), the posterior over \( \sigma^2 \) will also be inverse-gamma
  
  \[ p(\sigma^2|X) = IG\left(\alpha + \frac{N}{2}, \beta + \frac{1}{2} \sum_{n=1}^{N}(x_n - \mu)^2\right) \]

- Likewise, we can infer the posterior over the precision parameter (say \( \lambda = 1/\sigma^2 \))
  
  - The Gaussian likelihood in precision notation: \( p(x_n|\mu, \lambda) = \mathcal{N}(x|\mu, \lambda) = \sqrt{\frac{\lambda}{2\pi}} \exp\left[-\frac{\lambda}{2}(x_n - \mu)^2\right] \)
  
  - Conjugate prior for precision \( \lambda \) is gamma: \( p(\lambda|\alpha, \beta) = \Gamma(\alpha, \beta) \)
    
    \[ p(\lambda) \propto (\lambda)^{\alpha-1} \exp[-\beta\lambda] \quad \text{(note: note: } \alpha = \text{shape}, \beta = \text{rate}; \text{mean of } \Gamma(\alpha, \beta) = \frac{\alpha}{\beta}) \]

  - The posterior is also gamma: \( p(\lambda|X) = \Gamma(\alpha + \frac{N}{2}, \beta + \frac{1}{2} \sum_{n=1}^{N}(x_n - \mu)^2) \)

- Exercise: Work out (or look up) the posterior predictive \( p(x_*|X) \) in these cases (isn’t Gaussian)
Bayesian Inference for Both Parameters of a Gaussian

Goal: Infer the mean and precision of a univariate Gaussian \( \mathcal{N}(x|\mu, \lambda^{-1}) \)

Given \( N \) i.i.d. observations \( X = \{x_1, \ldots, x_N\} \), the likelihood will be

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

Let’s choose the following joint distribution as the prior (compare its form with \( p(X|\mu, \lambda) \))

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

The above is known as the Normal-gamma (NG) distribution (product of a Normal and a gamma)

\[
p(\mu, \lambda) = \mathcal{N}(\mu|\mu_0, (\kappa_0 \lambda)^{-1})\text{Gamma}(\lambda|\alpha_0, \beta_0) = \text{NG}(\mu, \lambda|\mu_0, \kappa_0, \alpha_0, \beta_0) \quad \text{(note: } \mu \text{ and } \lambda \text{ are coupled in the Gaussian part)}
\]

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

NG is conjugate to Gaussian when both mean & precision are unknown
Bayesian Inference for Both Parameters of a Gaussian

Due to conjugacy, \( p(\mu, \lambda | X) \) will also be NG: 
\[
p(\mu, \lambda | X) \propto p(X|\mu, \lambda)p(\mu, \lambda)
\]

\[
p(\mu, \lambda | X) = \text{NG}(\mu_N, \kappa_N, \alpha_N, \beta_N) = \mathcal{N}(\mu|\mu_N, (\kappa_N \lambda)^{-1})\text{Gamma}(\lambda|\alpha_N, \beta_N)
\]

where the updated posterior hyperparameters are given by\(^1\)

\[
\mu_N = \frac{\kappa_0 \mu_0 + N\bar{x}}{\kappa_0 + N}, \quad \kappa_N = \kappa_0 + N
\]

\[
\alpha_N = \alpha_0 + N/2, \quad \beta_N = \beta_0 + \frac{1}{2} \sum_{n=1}^{N} (x_n - \bar{x})^2 + \frac{\kappa_0 N(\bar{x} - \mu_0)^2}{2(\kappa_0 + N)}
\]

Note: The above is the joint posterior. We can also get marginal posteriors for \( \mu \) and \( \lambda \)

\[
p(\lambda | X) = \int p(\mu, \lambda | X) d\mu = \text{Gamma}(\lambda|\alpha_N, \beta_N)
\]

\[
p(\mu | X) = \int p(\mu, \lambda | X) d\lambda = \int p(\mu | \lambda, X)p(\lambda | X) d\lambda = t_{2\alpha_N}(\mu|\mu_N, \beta_N/(\alpha_N \kappa_N))
\]

Posterior predictive distribution of a new observation \( x_* \)

\[
p(x_* | X) = \int \frac{p(x_* | \mu, \lambda) p(\mu, \lambda | X)}{p(\mu | X) p(\lambda | X)} d\mu d\lambda = t_{2\alpha_N} \left( x_* | \mu_N, \frac{\beta_N(\kappa_N + 1)}{\alpha_N \kappa_N} \right)
\]

\(^1\) For full derivation, refer to “Conjugate Bayesian analysis of the Gaussian distribution” - Murphy (2007)
An Aside: generalized-t and Student-t distribution

- Obtained if we integrate out the precision of a Gaussian using a conjugate gamma prior

\[
p(x|\mu, a, b) = \int \mathcal{N}(x|\mu, \lambda^{-1})\text{Gamma}(\lambda|a, b) d\lambda
\]

\[
= t_{2a}(x|\mu, b/a) = t_\nu(x|\mu, \sigma^2) \quad \text{(generalized-t distribution)}
\]

- \(\mu = 0, \sigma^2 = 1\): Student-t distribution \((t_\nu(0, 1))\). Note: If \(x \sim t_\nu(\mu, \sigma^2)\) then \(\frac{x-\mu}{\sigma} \sim t_\nu(0, 1)\)

- The t-distribution has a “fatter” tail than a Gaussian and also sharper around the mean
  - Also a useful prior for sparsity prior (e.g., for weights in regression/classification)
  - For \(\nu \rightarrow \infty\), it is equivalent to a Gaussian
Bayesian Inference for Multivariate Gaussian?

- The parameters are now the mean vector and the covariance/precision matrix.
- Posterior updates for these have forms similar to that in the univariate case.
- For the mean, commonly a multivariate Gaussian prior is used.
  - Posterior is also Gaussian due to conjugacy.
- For the covariance matrix (with mean fixed), commonly an inverse-Wishart prior is used.
  - Posterior is also inverse-Wishart due to conjugacy.
- For the precision matrix (with mean fixed), commonly a Wishart prior is used.
  - Posterior is also Wishart due to conjugacy.
- When both parameters are unknown, there still exist conjugate joint priors.
  - Normal-Inverse Wishart for mean + cov matrix, Normal-Wishart for mean + precision matrix.
- For further details (e.g., full equations, posterior predictive, etc), refer to “Conjugate Bayesian analysis of the Gaussian distribution” by Murphy (2007), or MLAPP Chapter 4.
Some Useful Properties of Gaussians
Multivariate Gaussian: Some Alternative Representations

- The (multivariate) Gaussian with mean $\mu$ and cov. matrix $\Sigma$

\[
\mathcal{N}(x|\mu, \Sigma) = \frac{1}{\sqrt{(2\pi)^D |\Sigma|}} \exp \left\{ -\frac{1}{2} (x - \mu)^\top \Sigma^{-1} (x - \mu) \right\}
\]

\[
= \frac{1}{\sqrt{(2\pi)^D |\Sigma|}} \exp \left\{ -\frac{1}{2} \text{trace} \left[ \Sigma^{-1} S \right] \right\}
\]

where $S = (x - \mu)(x - \mu)^\top$.

- An alternate representation: The "information form"

\[
\mathcal{N}_c(x|\xi, \Lambda) = (2\pi)^{-D/2} |\Lambda|^{1/2} \exp \left\{ -\frac{1}{2} \left( x^\top \Lambda x + \xi^\top \Lambda^{-1} \xi - 2x^\top \xi \right) \right\}
\]

where $\Lambda = \Sigma^{-1}$ and $\xi = \Sigma^{-1} \mu$ are the "natural parameters" (more when we discuss exp. family).

- Note that there is a term quadratic in $x$ (involves $\Lambda = \Sigma^{-1}$) and linear in $x$ (involves $\xi = \Sigma^{-1} \mu$)

- Information form can help recognize $\mu$ and $\Sigma$ of a Gaussian when doing algebraic manipulations
Multivariate Gaussian: The Covariance Matrix

The covariance matrix can be spherical, diagonal, or full
Marginals and Conditionals from Gaussian Joint Distribution

Given \( x \) having multivariate Gaussian distribution \( N(x|\mu, \Sigma) \) with \( \Lambda = \Sigma^{-1} \). Suppose

\[
\begin{align*}
x &= \begin{bmatrix} x_a \\ x_b \end{bmatrix} \\
\mu &= \begin{bmatrix} \mu_a \\ \mu_b \end{bmatrix} \\
\Sigma &= \begin{bmatrix} \Sigma_{aa} & \Sigma_{ab} \\ \Sigma_{ba} & \Sigma_{bb} \end{bmatrix} \\
\Lambda &= \begin{bmatrix} \Lambda_{aa} & \Lambda_{ab} \\ \Lambda_{ba} & \Lambda_{bb} \end{bmatrix}
\end{align*}
\]

The marginal distribution of one block, say \( x_a \), is a Gaussian

\[
p(x_a) = \int p(x_a, x_b) dx_b = N(x_a|\mu_a, \Sigma_{aa})
\]

The conditional distribution of \( x_a \) given \( x_b \), is Gaussian, i.e., \( p(x_a|x_b) = N(x_a|\mu_a|x_b, \Sigma_a|x_b) \) where

\[
\begin{align*}
\Sigma_{a|x_b} &= \Lambda_{aa}^{-1} = \Sigma_{aa} - \Sigma_{ab} \Sigma^{-1} \Sigma_{ab} \\
\mu_{a|x_b} &= \Sigma_{a|x_b} \{ \Lambda_{aa} \mu_a - \Lambda_{ab} (x_b - \mu_b) \} \\
&= \mu_a - \Lambda_{aa}^{-1} \Lambda_{ab} (x_b - \mu_b) \\
&= \mu_a + \Sigma_{ab} \Sigma_{bb}^{-1} (x_b - \mu_b)
\end{align*}
\]

Both results are extremely useful when working with Gaussian joint distributions.
An Aside: Linear Transformations of Random Variables

- Suppose \( x = f(z) = Az + b \) be a linear function of an r.v. \( z \) (not necessarily Gaussian)
- Suppose \( \mathbb{E}[z] = \mu \) and \( \text{cov}[z] = \Sigma \)
  - Expectation of \( x \)
    \[
    \mathbb{E}[x] = \mathbb{E}[Az + b] = A\mu + b
    \]
  - Covariance of \( x \)
    \[
    \text{cov}[x] = \text{cov}[Az + b] = A\Sigma A^T
    \]
- Likewise if \( x = f(z) = a^T z + b \) is a scalar-valued linear function of an r.v. \( z \):
  - \( \mathbb{E}[x] = \mathbb{E}[a^T z + b] = a^T \mu + b \)
  - \( \text{var}[x] = \text{var}[a^T z + b] = a^T \Sigma a \)
- These properties are often helpful in obtaining the marginal distribution \( p(x) \) from \( p(z) \)
Linear Gaussian Model

- Consider linear transformation of a Gaussian r.v. $z$ with $p(z) = \mathcal{N}(z|\mu, \Lambda^{-1})$, plus Gaussian noise
  \[
  x = Az + b + \epsilon
  \]
  where $p(\epsilon) = \mathcal{N}(\epsilon|0, L^{-1})$

- Easy to see that, conditioned on $z$, $x$ too has a Gaussian distribution
  \[
  p(x|z) = \mathcal{N}(x|Az + b, L^{-1})
  \]

- This is called a Linear Gaussian Model. Very commonly encountered in probabilistic modeling

- The following two distributions are of particular interest. Defining $\Sigma = (\Lambda + A^T L A)^{-1}$, we have
  \[
  p(z|x) = \frac{p(x|z)p(z)}{p(z)} = \mathcal{N}(z|\Sigma \left\{ A^T L (x - b) + \Lambda \mu \right\}, \Sigma)
  \]
  (a Gaussian posterior :-))

  \[
  p(x) = \int p(x|z)p(z)dz = \mathcal{N}(x|A\mu + b, A\Lambda^{-1}A^T + L^{-1})
  \]
  (a Gaussian predictive/marginal :-))

- **Exercise:** Prove the above two results (MLAPP Chap. 4 and PRML Chap. 2 contain the proof)
Gaussians and Linear Gaussian Models are widely used in probabilistic/Bayesian models

Some popular applications are

- Probability density estimation: Given \( x_1, \ldots, x_N \), estimate \( p(x) \) assuming Gaussian likelihood/noise
- Given \( N \) sensor obs. \( \{ x_n \}_{n=1}^N \) with \( x_n = \mu + \epsilon_n \) (Gaussian noise \( \epsilon_n \)), estimate the “source” value \( \mu \) (possibly along with the variance of the estimate of \( \mu \))
- Estimating missing data: \( p(x_{n}^{\text{miss}}|x_{n}^{\text{obs}}) \) - can also get other quantities, such as \( \mathbb{E}[x_{n}^{\text{miss}}|x_{n}^{\text{obs}}] \)
- Linear Regression with Gaussian Likelihood

\[
y = Xw + \epsilon \quad (w \text{ is Gaussian weight vector, } \epsilon \text{ is } N \times 1 \text{ indep. Gaussian noise})
\]

- Linear latent variable models (probabilistic PCA, factor analysis, Kalman filters) and their mixtures

\[
x_n = Wz_n + \epsilon_n \quad (z_n \text{ is Gaussian low-dim } K \times 1 \text{ latent var, } \epsilon_n \text{ is } D \times 1 \text{ indep. Gaussian noise})
\]

- Gaussian Processes (GP) extensively use Gaussian conditioning and marginalization rules

\[
y = f + \text{noise} \quad (\text{GP assumes } f = [f(x_1), \ldots, f(x_N)] \text{ is jointly Gaussian})
\]

- More complex models where parts of the model use Gaussian likelihoods/priors