Bayesian Inference for Gaussians, Working With Gaussians

Piyush Rai

Topics in Probabilistic Modeling and Inference (CS698X)

Jan 14, 2019

Recap: Bayesian Inference for Mean of a Gaussian

- Consider N i.i.d. observations $\mathbf{X} = \{x_1, \ldots, x_N\}$ drawn from a one-dim Gaussian $\mathcal{N}(x|\mu, \sigma^2)$

$$\begin{aligned} p(x_n|\mu,\sigma^2) &= \mathcal{N}(x|\mu,\sigma^2) \propto \exp\left[-\frac{(x_n-\mu)^2}{2\sigma^2}\right] \\ p(\mu) &= \mathcal{N}(\mu|\mu_0,\sigma_0^2) \end{aligned}$$

- Due to conjugacy, posterior is also Gaussian: $p(\mu|\mathbf{X}) \propto \exp\left[-\frac{(\mu-\mu_N)^2}{2\sigma_N^2}\right]$ with

$$\begin{aligned} \frac{1}{\sigma_N^2} &= \frac{1}{\sigma_0^2} + \frac{N}{\sigma^2} \\ \mu_N &= \frac{\sigma^2}{N\sigma_0^2+\sigma^2}\mu_0 + \frac{N\sigma_0^2}{N\sigma_0^2+\sigma^2}\bar{x} \qquad (\text{where } \bar{x} = \frac{\sum_{n=1}^N x_n}{N}) \end{aligned}$$

- Posterior predictive for a new observation x_* is also Gaussian

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

(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 $p(\mu) = \mathcal{N}(\mu|\mu_0, \frac{\sigma^2}{\kappa_0})$. Also, what does κ_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 σ^2 is inverse-gamma: $p(\sigma^2|\alpha,\beta) = \text{IG}(\alpha,\beta)$

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

- Given N i.i.d. observations $\mathbf{X} = \{x_1,\ldots,x_N\}$, the posterior over σ^2 will also be inverse-gamma

$$p(\sigma^2|\mathbf{X}) = IG\left(\alpha + \frac{N}{2}, \beta + \frac{\sum_{n=1}^N (x_n-\mu)^2}{2}\right)$$

- 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 λ is gamma: $p(\lambda|\alpha,\beta) = \text{Gamma}(\alpha,\beta)$

$$p(\lambda) \propto (\lambda)^{(\alpha-1)} \exp\left[-\beta\lambda\right] \qquad \text{(note: note: } \alpha = \text{shape}, \beta = \text{rate; mean of Gamma}(\alpha,\beta) = \frac{\alpha}{\beta}\text{)}$$

  - The posterior is also gamma: $p(\lambda|\mathbf{X}) = \text{Gamma}(\alpha + \frac{N}{2}, \beta + \frac{\sum_{n=1}^N (x_n-\mu)^2}{2})$
- Exercise: Work out (or look up) the posterior predictive $p(x_*|\mathbf{X})$ 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 $\mathbf{X} = \{x_1, \ldots, x_N\}$, the likelihood will be

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

- Let's choose the following joint distribution as the prior (compare its form with $p(\mathbf{X}|\mu, \lambda)$)

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

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

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

  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, $p(\mu,\lambda|\mathbf{X})$ will also be NG: $p(\mu,\lambda|\mathbf{X}) \propto p(\mathbf{X}|\mu,\lambda)p(\mu,\lambda)$

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

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 μ and λ

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

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

- Posterior predictive distribution of a new observation x_*

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

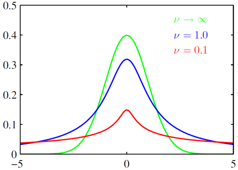
[1] For full derivation, refer to "Conjugate Bayesian analysis of the Gaussian distribution" - Murphy (2007)

An Aside: generalized-t and Student-t distribution

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

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

- $\mu = 0, \sigma^2 = 1$: 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 $\nu \to \infty$, 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 $\boldsymbol{\mu}$ and cov. matrix $\boldsymbol{\Sigma}$

$$\begin{aligned}\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) &= \frac{1}{\sqrt{(2\pi)^D|\boldsymbol{\Sigma}|}}\exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right\}\\ &= \frac{1}{\sqrt{(2\pi)^D|\boldsymbol{\Sigma}|}}\exp\left\{-\frac{1}{2}\text{trace}\left[\boldsymbol{\Sigma}^{-1}\mathbf{S}\right]\right\} \qquad \text{where } \mathbf{S}=(\boldsymbol{x}-\boldsymbol{\mu})(\boldsymbol{x}-\boldsymbol{\mu})^\top\end{aligned}$$

- An alternate representation: The "information form"

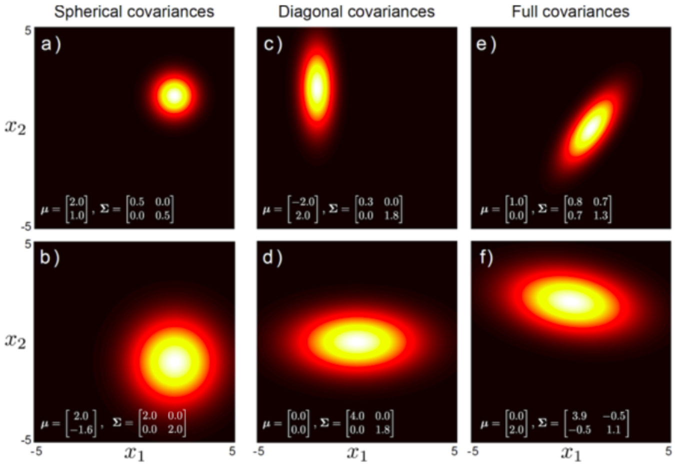
$$\mathcal{N}_c(\boldsymbol{x}|\boldsymbol{\xi},\boldsymbol{\Lambda}) = (2\pi)^{-D/2}|\boldsymbol{\Lambda}|^{1/2}\exp\left\{-\frac{1}{2}\left(\boldsymbol{x}^\top\boldsymbol{\Lambda}\boldsymbol{x}+\boldsymbol{\xi}^\top\boldsymbol{\Lambda}^{-1}\boldsymbol{\xi}-2\boldsymbol{x}^\top\boldsymbol{\xi}\right)\right\}$$

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

- Note that there is a term quadratic in $\boldsymbol{x}$ (involves $\boldsymbol{\Lambda}=\boldsymbol{\Sigma}^{-1}$) and linear in $\boldsymbol{x}$ (involves $\boldsymbol{\xi}=\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}$)
- Information form can help recognize $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ of a Gaussian when doing algebraic manipulations

Multivariate Gaussian: The Covariance Matrix

The covariance matrix can be spherical, diagonal, or full



Picture courtesy: Computer vision: models, learning and inference (Simon Price)

Marginals and Conditionals from Gaussian Joint Distribution

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

$$\boldsymbol{x} = \begin{bmatrix} \boldsymbol{x}_a \\ \boldsymbol{x}_b \end{bmatrix} \qquad \boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_a \\ \boldsymbol{\mu}_b \end{bmatrix}$$

$$\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{aa} & \boldsymbol{\Sigma}_{ab} \\ \boldsymbol{\Sigma}_{ba} & \boldsymbol{\Sigma}_{bb} \end{bmatrix} \qquad \boldsymbol{\Lambda} = \begin{bmatrix} \boldsymbol{\Lambda}_{aa} & \boldsymbol{\Lambda}_{ab} \\ \boldsymbol{\Lambda}_{ba} & \boldsymbol{\Lambda}_{bb} \end{bmatrix}$$

- The marginal distribution of one block, say $\boldsymbol{x}_a$, is a Gaussian

$$p(\boldsymbol{x}_a) = \int p(\boldsymbol{x}_a, \boldsymbol{x}_b) d\boldsymbol{x}_b = \mathcal{N}(\boldsymbol{x}_a | \boldsymbol{\mu}_a, \boldsymbol{\Sigma}_{aa})$$

- The conditional distribution of $\boldsymbol{x}_a$ given $\boldsymbol{x}_b$, is Gaussian, i.e., $p(\boldsymbol{x}_a|\boldsymbol{x}_b) = \mathcal{N}(\boldsymbol{x}_a|\boldsymbol{\mu}_{a|b}, \boldsymbol{\Sigma}_{a|b})$ where

$$\begin{aligned} \boldsymbol{\Sigma}_{a|b} &= \boldsymbol{\Lambda}_{aa}^{-1} = \boldsymbol{\Sigma}_{aa} - \boldsymbol{\Sigma}_{ab}\boldsymbol{\Sigma}_{bb}^{-1}\boldsymbol{\Sigma}_{ba} \\ \boldsymbol{\mu}_{a|b} &= \boldsymbol{\Sigma}_{a|b}\{\boldsymbol{\Lambda}_{aa}\boldsymbol{\mu}_a - \boldsymbol{\Lambda}_{ab}(\boldsymbol{x}_b - \boldsymbol{\mu}_b)\} \\ &= \boldsymbol{\mu}_a - \boldsymbol{\Lambda}_{aa}^{-1}\boldsymbol{\Lambda}_{ab}(\boldsymbol{x}_b - \boldsymbol{\mu}_b) \\ &= \boldsymbol{\mu}_a + \boldsymbol{\Sigma}_{ab}\boldsymbol{\Sigma}_{bb}^{-1}(\boldsymbol{x}_b - \boldsymbol{\mu}_b) \end{aligned}$$

("smaller" than Σ_{aa}; makes sense intuitively)

- Both results are extremely useful when working with Gaussian joint distributions

An Aside: Linear Transformations of Random Variables

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

Linear Gaussian Model

- Consider linear transformation of a Gaussian r.v. $\boldsymbol{z}$ with $p(\boldsymbol{z}) = \mathcal{N}(\boldsymbol{z}|\boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1})$, plus Gaussian noise

$$\boxed{\boldsymbol{x} = \mathbf{A}\boldsymbol{z} + \boldsymbol{b} + \boldsymbol{\epsilon}} \qquad \text{where} \quad p(\boldsymbol{\epsilon}) = \mathcal{N}(\boldsymbol{\epsilon}|\mathbf{0}, \mathbf{L}^{-1})$$

- Easy to see that, conditioned on $\boldsymbol{z}$, $\boldsymbol{x}$ too has a Gaussian distribution

$$p(\boldsymbol{x}|\boldsymbol{z}) = \mathcal{N}(\boldsymbol{x}|\mathbf{A}\boldsymbol{z} + \boldsymbol{b}, \mathbf{L}^{-1})$$

- This is called a Linear Gaussian Model. Very commonly encountered in probabilistic modeling
- The following two distributions are of particular interest. Defining $\boldsymbol{\Sigma} = (\boldsymbol{\Lambda} + \mathbf{A}^\top \mathbf{L}\mathbf{A})^{-1}$, we have

$$p(\boldsymbol{z}|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|\boldsymbol{z})p(\boldsymbol{z})}{p(\boldsymbol{z})} = \mathcal{N}(\boldsymbol{z}|\boldsymbol{\Sigma}\left\{\mathbf{A}^\top \mathbf{L}(\boldsymbol{x} - \boldsymbol{b}) + \boldsymbol{\Lambda}\boldsymbol{\mu}\right\}, \boldsymbol{\Sigma}) \qquad \text{(a Gaussian posterior :-))}$$

$$p(\boldsymbol{x}) = \int p(\boldsymbol{x}|\boldsymbol{z})p(\boldsymbol{z})d\boldsymbol{z} = \mathcal{N}(\boldsymbol{x}|\mathbf{A}\boldsymbol{\mu} + \boldsymbol{b}, \mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{A}^\top + \mathbf{L}^{-1}) \qquad \text{(a Gaussian predictive/marginal :-))}$$

- Exercise: Prove the above two results (MLAPP Chap. 4 and PRML Chap. 2 contain the proof)

Applications?

- Gaussians and Linear Gaussian Models are widely used in probabilistic/Bayesian models
- Some popular applications are
  - Probability density estimation: Given $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$, estimate $p(\boldsymbol{x})$ assuming Gaussian likelihood/noise
  - Given N sensor obs. $\{\boldsymbol{x}_n\}_{n=1}^N$ with $\boldsymbol{x}_n = \boldsymbol{\mu} + \epsilon_n$ (Gaussian noise ϵ_n), estimate the "source" value $\boldsymbol{\mu}$ (possibly along with the variance of the estimate of $\boldsymbol{\mu}$)
  - Estimating missing data: $p(\boldsymbol{x}_n^{miss}|\boldsymbol{x}_n^{obs})$ - can also get other quantities, such as $\mathbb{E}[\boldsymbol{x}_n^{miss}|\boldsymbol{x}_n^{obs}]$
  - Linear Regression with Gaussian Likelihood

$$\boldsymbol{y} = \mathbf{X}\boldsymbol{w} + \boldsymbol{\epsilon} \qquad (\boldsymbol{w} \text{ is Gaussian weight vector, } \boldsymbol{\epsilon} \text{ is } N \times 1 \text{ indep. Gaussian noise})$$

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

$$\boldsymbol{x}_n = \mathbf{W}\boldsymbol{z}_n + \boldsymbol{\epsilon}_n \qquad (\boldsymbol{z}_n \text{ is Gaussian low-dim } K \times 1 \text{ latent var, } \boldsymbol{\epsilon}_n \text{ is } D \times 1 \text{ indep. Gaussian noise})$$

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

$$\boldsymbol{y} = \boldsymbol{f} + \text{noise} \qquad (\text{GP assumes } \boldsymbol{f} = [f(\boldsymbol{x}_1), \ldots, f(\boldsymbol{x}_N)] \text{ is jointly Gaussian})$$

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