Bayesian linear regression with unknown noise precision

Janurary 2, 2023

For Bayesian linear regression with fixed noise precision, we relied on a theorem for linear Gaussian systems to go from the prior and likelihood to the posterior. Here, we first state and prove a new theorem that's very similar to that theorem, which will be crucially for helping us go from the prior and likelihood to the posterior for Bayesian linear regression with unknown noise precision.

Theorem 1. *Given the following prior and likelihood*

$$\begin{aligned} p(\sigma^2) &= \mathrm{IG}(a_\sigma, b_\sigma) \\ p(\boldsymbol{x}) &= \mathcal{N}(\boldsymbol{x}|\, \boldsymbol{\mu}_x, \sigma^2 \boldsymbol{\Sigma}_x), \\ p(\boldsymbol{y}|\, \boldsymbol{x}, \sigma^2) &= \mathcal{N}(\boldsymbol{x}|\, \boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}, \boldsymbol{\Sigma}_y) \end{aligned}$$

the posterior is given by

$$\begin{aligned} p(\boldsymbol{x}, \sigma^2|\, \boldsymbol{y}) &= \mathcal{N}(\boldsymbol{x}|\, \boldsymbol{\mu}_{x|y}, \sigma^2 \boldsymbol{\Sigma}_{x|y})\, \mathrm{IG}(\sigma^2|\, a_{\sigma|y}, b_{\sigma|y}) \\ \boldsymbol{\mu}_{x|y} &= \boldsymbol{\Sigma}_{x|y}(\boldsymbol{\Sigma}_x^{-1}\boldsymbol{\mu}_x + \boldsymbol{A}^T\boldsymbol{y}) \\ \boldsymbol{\Sigma}_{x|y} &= (\boldsymbol{\Sigma}_x^{-1} + \boldsymbol{A}^T\boldsymbol{A})^{-1} \\ a_{\sigma|y} &= a_\sigma + D_y/2 \\ b_{\sigma|y} &= b_\sigma + \frac{1}{2}(\boldsymbol{\mu}_x^T\boldsymbol{\Sigma}_x^{-1}\boldsymbol{\mu}_x + \boldsymbol{y}^T\boldsymbol{y} - \boldsymbol{\mu}_{x|y}^T\, \boldsymbol{\Sigma}_{x|y}^{-1}\, \boldsymbol{\mu}_{x|y}) \end{aligned}$$

Proof. The proof should be analagous to the proof of the Bayes rule for linear Gaussian systems presented in Section 4.4.3 of Murphy. □

Likelihood.

$$p(\boldsymbol{y}|\, \boldsymbol{X}, \boldsymbol{w}, \sigma^2) = \prod_{i=1}^{N} \mathcal{N}(y_i|\, \mu_i, \sigma^2) = \mathcal{N}(\boldsymbol{y}|\, \boldsymbol{X}\boldsymbol{w}, \sigma^2\boldsymbol{I})$$

with $\boldsymbol{\mu} = \boldsymbol{X}\boldsymbol{w}$.

Prior.

$$p(\boldsymbol{w}, \sigma^2) = \mathcal{N}(\boldsymbol{w}|\, \boldsymbol{w}_0, \sigma^2\boldsymbol{V}_0)\, \mathrm{IG}(\sigma^2|\, a_0, b_0)$$

Posterior. Applying Theorem 1, we obtain:

$$\begin{aligned} p(\boldsymbol{w}, \sigma^2|\, \boldsymbol{y}, \boldsymbol{X}) &= \mathcal{N}(\boldsymbol{w}|\, \boldsymbol{w}_N, \sigma^2\boldsymbol{V}_N)\, \mathrm{IG}(\sigma^2|\, a_N, b_N) \\ \boldsymbol{w}_N &= \boldsymbol{V}_N(\boldsymbol{V}_0^{-1}\boldsymbol{w}_0 + \boldsymbol{X}^T\boldsymbol{y}) \\ \boldsymbol{V}_N &= (\boldsymbol{V}_0^{-1} + \boldsymbol{X}^T\boldsymbol{X})^{-1} \\ a_N &= a_0 + N/2 \\ b_N &= b_0 + \frac{1}{2}(\boldsymbol{w}_0^T\boldsymbol{V}_0^{-1}\boldsymbol{w}_0 + \boldsymbol{y}^T\boldsymbol{y} - \boldsymbol{w}_N^T\boldsymbol{V}_N^{-1}\boldsymbol{w}_N), \end{aligned}$$

which is exactly the same as Equation 7.69 to 7.73 from Murphy.

Marginal posterior.

$$\begin{aligned}
p(\boldsymbol{w}|\,\boldsymbol{y},\boldsymbol{X}) &= \int p(\boldsymbol{w},\sigma^2|\,\boldsymbol{y},\boldsymbol{X})\,d\sigma^2 \\
&= \int \mathcal{N}(\boldsymbol{w}|\,\boldsymbol{w}_N,\sigma^2\boldsymbol{V}_N)\,\mathrm{IG}(\sigma^2|\,a_N,b_N)\,d\sigma^2 \\
&= \int \mathcal{N}(\boldsymbol{w}|\,\boldsymbol{w}_N,(1/\sigma^2)\boldsymbol{V}_N)\,\mathrm{Ga}(1/\sigma^2|\,a_N,b_N)\,d\sigma^2 \\
&= \mathcal{T}(\boldsymbol{w}|\,\boldsymbol{w}_N,(b_N/a_N)\boldsymbol{V}_N,2a_N) \quad \text{(by Eq 11.61)}
\end{aligned}$$

$$\begin{aligned}
p(\sigma^2|\,\boldsymbol{y},\boldsymbol{X}) &= \int p(\boldsymbol{w},\sigma^2|\,\boldsymbol{y},\boldsymbol{X})\,d\boldsymbol{w} \\
&= \int \mathcal{N}(\boldsymbol{w}|\,\boldsymbol{w}_N,\sigma^2\boldsymbol{V}_N)\,\mathrm{IG}(\sigma^2|\,a_N,b_N)\,d\boldsymbol{w} \\
&= \int \mathcal{N}(\boldsymbol{w}|\,\boldsymbol{w}_N,\sigma^2\boldsymbol{V}_N)\;d\boldsymbol{w}\,\mathrm{IG}(\sigma^2|\,a_N,b_N) \\
&= \mathrm{IG}(\sigma^2|\,a_N,b_N)
\end{aligned}$$

Posterior predictive.

$$\begin{aligned}
p(\tilde{\boldsymbol{y}}|\,\tilde{\boldsymbol{X}},\boldsymbol{X},\boldsymbol{y}) &= \iint \mathcal{N}(\tilde{\boldsymbol{y}}|\,\tilde{\boldsymbol{X}}^T\boldsymbol{w},\sigma^2\boldsymbol{I}_m)\,\mathrm{NIW}(\boldsymbol{w},\sigma^2|\,\boldsymbol{w}_N,\boldsymbol{V}_N,a_N,b_N)\,d\boldsymbol{w}\,d\sigma^2 \\
&= \int\Big[\int \mathcal{N}(\tilde{\boldsymbol{y}}|\,\tilde{\boldsymbol{X}}^T\boldsymbol{w},\sigma^2\boldsymbol{I}_m)\,\mathcal{N}(\boldsymbol{w}|\,\boldsymbol{w}_N,\sigma^2\boldsymbol{V}_N)\,d\boldsymbol{w}\Big]\,\mathrm{IG}(\sigma^2|\,a_N,b_N)\;d\sigma^2 \\
&= \int \mathcal{N}(\tilde{\boldsymbol{y}}|\,\tilde{\boldsymbol{X}}^T\boldsymbol{w}_N,\sigma^2(\boldsymbol{I}_m+\tilde{\boldsymbol{X}}^T\boldsymbol{V}_N\tilde{\boldsymbol{X}}))\,\mathrm{IG}(\sigma^2|\,a_N,b_N)\;d\sigma^2 \quad \text{(by Eq 4.126)} \\
&= \int \mathcal{N}(\tilde{\boldsymbol{y}}|\,\tilde{\boldsymbol{X}}^T\boldsymbol{w}_N,(1/\sigma^2)(\boldsymbol{I}_m+\tilde{\boldsymbol{X}}^T\boldsymbol{V}_N\tilde{\boldsymbol{X}}))\,\mathrm{Ga}(1/\sigma^2|\,a_N,b_N)\;d\sigma^2 \\
&= \mathcal{T}\left(\tilde{\boldsymbol{y}}\middle|\,\tilde{\boldsymbol{X}}^T\boldsymbol{w}_N,\frac{b_N}{a_N}(\boldsymbol{I}_m+\tilde{\boldsymbol{X}}^T\boldsymbol{V}_N\tilde{\boldsymbol{X}}),2a_N\right) \quad \text{(by Eq 11.61)}
\end{aligned}$$