Bayesian Inference for Gaussian Linear Model

Aleksandar Dogandžić

2017-05-13

Contents

Reading: §10 in the textbook.

✻ Theorem 10.3 in the textbook. Consider the linear model:

$$\boldsymbol{X} = H\boldsymbol{\Theta} + \boldsymbol{W}$$

where

- $\boldsymbol{X}$ is an $N \times 1$ measurement vector,
- H is a known $N \times p$ regression matrix,
- $\boldsymbol{\Theta}$ is an unknown $p \times 1$ random vector that we wish to estimate, with Gaussian prior probability density function (pdf)

$$\boldsymbol{\Theta} \sim \mathcal{N}(\boldsymbol{\mu}_{\boldsymbol{\Theta}}, C_{\boldsymbol{\Theta}}),$$

- $\boldsymbol{W}$ is an $N \times 1$ noise vector with Gaussian prior pdf

$$\boldsymbol{W} \sim \mathcal{N}(\mathbf{0}, C_{\boldsymbol{W}}),$$

- $\boldsymbol{W}$ and $\boldsymbol{\Theta}$ are independent and $C_{\boldsymbol{W}}$, $\boldsymbol{\mu}_{\boldsymbol{\Theta}}$, and $C_{\boldsymbol{\Theta}}$ are known hyperparameters.

Then, the posterior pdf $f_{\boldsymbol{\Theta}|\boldsymbol{X}}(\boldsymbol{\theta} \mid \boldsymbol{x})$ is Gaussian with mean vector $\mathrm{E}_{\boldsymbol{\Theta}|\boldsymbol{X}}(\boldsymbol{\theta} \mid \boldsymbol{x})$ and covariance matrix $\mathrm{cov}_{\boldsymbol{\Theta}|\boldsymbol{X}}(\boldsymbol{\theta} \mid \boldsymbol{x})$:

$$f_{\boldsymbol{\Theta}|\boldsymbol{X}}(\boldsymbol{\theta} \mid \boldsymbol{x}) = \mathcal{N}\big(\boldsymbol{\theta} \mid \mathrm{E}_{\boldsymbol{\Theta}|\boldsymbol{X}}(\boldsymbol{\theta} \mid \boldsymbol{x}), \mathrm{cov}_{\boldsymbol{\Theta}|\boldsymbol{X}}(\boldsymbol{\theta} \mid \boldsymbol{x})\big) \tag{1a}$$

where

$$\begin{aligned}
\mathrm{E}_{\boldsymbol{\Theta}|\boldsymbol{X}}(\boldsymbol{\Theta} \mid \boldsymbol{x}) &= (H^T C_{\boldsymbol{W}}^{-1} H + C_{\boldsymbol{\Theta}}^{-1})^{-1}(H^T C_{\boldsymbol{W}}^{-1} \boldsymbol{x} + C_{\boldsymbol{\Theta}}^{-1}\boldsymbol{\mu}_{\boldsymbol{\Theta}}) && \text{(1b)}\\
&= \boldsymbol{\mu}_{\boldsymbol{\Theta}} + C_{\boldsymbol{\Theta}} H^T (H C_{\boldsymbol{\Theta}} H^T + C_{\boldsymbol{W}})^{-1}(\boldsymbol{x} - H\boldsymbol{\mu}_{\boldsymbol{\Theta}}) && \text{(1c)}\\
\mathrm{cov}_{\boldsymbol{\Theta}|\boldsymbol{X}}(\boldsymbol{\Theta} \mid \boldsymbol{x}) &= (H^T C_{\boldsymbol{W}}^{-1} H + C_{\boldsymbol{\Theta}}^{-1})^{-1} && \text{(1d)}\\
&= C_{\boldsymbol{\Theta}} - C_{\boldsymbol{\Theta}} H^T \big(H C_{\boldsymbol{\Theta}} H^T + C_{\boldsymbol{W}}\big)^{-1} H C_{\boldsymbol{\Theta}}. && \text{(1e)}
\end{aligned}$$

Proof: Before we proceed, recall that $\boldsymbol{Z} \sim \mathcal{N}(\boldsymbol{\mu}, C)$ implies

$$f_{\boldsymbol{Z}}(\boldsymbol{z}) \propto \exp\big(-0.5\boldsymbol{z}^T C^{-1} \boldsymbol{z} + \boldsymbol{z}^T C^{-1} \boldsymbol{\mu}\big).$$

Now, compute the posterior pdf of $\boldsymbol{\theta}$:

$$
\begin{aligned}
f_{\boldsymbol{\Theta}|\boldsymbol{X}}(\boldsymbol{\theta} \mid \boldsymbol{x}) &\propto f_{\boldsymbol{X}|\boldsymbol{\Theta}}(\boldsymbol{x} \mid \boldsymbol{\theta}) f(\boldsymbol{\theta}) \\
&\propto \exp\big[-0.5(\boldsymbol{x} - H\boldsymbol{\theta})^T C_{\boldsymbol{W}}^{-1} (\boldsymbol{x} - H\boldsymbol{\theta})\big] \exp\big[-0.5(\boldsymbol{\theta} - \boldsymbol{\mu}_{\boldsymbol{\Theta}})^T C_{\boldsymbol{\Theta}}^{-1} (\boldsymbol{\theta} - \boldsymbol{\mu}_{\boldsymbol{\Theta}})\big] \\
&\propto \exp\big(-0.5\boldsymbol{\theta}^T H^T C_{\boldsymbol{W}}^{-1} H \boldsymbol{\theta} + \boldsymbol{\theta}^T H^T C_{\boldsymbol{W}}^{-1} \boldsymbol{x}\big) \exp\big(-0.5\boldsymbol{\theta}^T C_{\boldsymbol{\Theta}}^{-1} \boldsymbol{\theta} + \boldsymbol{\theta}^T C_{\boldsymbol{\Theta}}^{-1} \boldsymbol{\mu}_{\boldsymbol{\Theta}}\big) \\
&\propto \exp\big[-0.5\boldsymbol{\theta}^T (H^T C_{\boldsymbol{W}}^{-1} H + C_{\boldsymbol{\Theta}}^{-1})\boldsymbol{\theta} + \boldsymbol{\theta}^T (H^T C_{\boldsymbol{W}}^{-1} \boldsymbol{x} + C_{\boldsymbol{\Theta}}^{-1} \boldsymbol{\mu}_{\boldsymbol{\Theta}})\big]
\end{aligned}
$$

which yields

$$
f_{\boldsymbol{\Theta}|\boldsymbol{X}}(\boldsymbol{\theta} \mid \boldsymbol{x}) = \mathcal{N}\big(\boldsymbol{\theta} \mid (H^T C_{\boldsymbol{W}}^{-1} H + C_{\boldsymbol{\Theta}}^{-1})^{-1} (H^T C_{\boldsymbol{W}}^{-1} \boldsymbol{x} + C_{\boldsymbol{\Theta}}^{-1} \boldsymbol{\mu}_{\boldsymbol{\Theta}}), (H^T C_{\boldsymbol{W}}^{-1} H + C_{\boldsymbol{\Theta}}^{-1})^{-1}\big).
$$

Therefore,

$$
\begin{aligned}
\mathrm{E}_{\boldsymbol{\Theta}|\boldsymbol{X}}(\boldsymbol{\Theta} \mid \boldsymbol{x}) &= (H^T C_{\boldsymbol{W}}^{-1} H + C_{\boldsymbol{\Theta}}^{-1})^{-1} (H^T C_{\boldsymbol{W}}^{-1} \boldsymbol{x} + C_{\boldsymbol{\Theta}}^{-1} \boldsymbol{\mu}_{\boldsymbol{\Theta}}) \\
\mathrm{cov}_{\boldsymbol{\Theta}|\boldsymbol{X}}(\boldsymbol{\Theta} \mid \boldsymbol{x}) &= (H^T C_{\boldsymbol{W}}^{-1} H + C_{\boldsymbol{\Theta}}^{-1})^{-1}.
\end{aligned}
$$

Recall the matrix inversion lemma:

$$
(R + STU)^{-1} = R^{-1} - R^{-1} S (T^{-1} + U R^{-1} S)^{-1} U R^{-1}
$$

and apply it to (1d) as follows:

$$
\mathrm{cov}_{\boldsymbol{\Theta}|\boldsymbol{X}}\big(\boldsymbol{\Theta} \mid \boldsymbol{x}\big) = C_{\boldsymbol{\Theta}} - C_{\boldsymbol{\Theta}} H^T (H C_{\boldsymbol{\Theta}} H^T + C_{\boldsymbol{W}})^{-1} H C_{\boldsymbol{\Theta}}.
$$

Recall the identity:

$$
(R + STU)^{-1} ST = R^{-1} S (T^{-1} + U R^{-1} S)^{-1}
$$

and apply it and the matrix inversion lemma to (1b) as follows:

$$
\begin{aligned}
\mathrm{E}_{\boldsymbol{\Theta}|\boldsymbol{X}}\big(\boldsymbol{\Theta} \mid \boldsymbol{x}\big) &= C_{\boldsymbol{\Theta}} H^T (H C_{\boldsymbol{\Theta}} H^T + C_{\boldsymbol{W}})^{-1} \boldsymbol{x} \\
&\quad + \big[I_p - C_{\boldsymbol{\Theta}} H^T (H C_{\boldsymbol{\Theta}} H^T + C_{\boldsymbol{W}})^{-1} H\big] \boldsymbol{\mu}_{\boldsymbol{\Theta}} \\
&= \boldsymbol{\mu}_{\boldsymbol{\Theta}} + C_{\boldsymbol{\Theta}} H^T (H C_{\boldsymbol{\Theta}} H^T + C_{\boldsymbol{W}})^{-1} (\boldsymbol{x} - H\boldsymbol{\mu}_{\boldsymbol{\Theta}}). \qquad \square
\end{aligned}
$$

✻ COMMENTS:

- DC level estimation in additive white Gaussian noise (AWGN) with known variance is a special case;
- Posterior mean:

$$
\mathrm{E}_{\boldsymbol{\Theta}|\boldsymbol{X}}(\boldsymbol{\Theta} \mid \boldsymbol{x}) = \Big(\underbrace{H^T C_{\boldsymbol{W}}^{-1} H}_{\text{likelihood precision}} + \underbrace{C_{\boldsymbol{\Theta}}^{-1}}_{\text{prior precision}} \Big)^{-1} \Big(\underbrace{H^T C_{\boldsymbol{W}}^{-1} \boldsymbol{x}}_{\text{data-dependent term}} + \underbrace{C_{\boldsymbol{\Theta}}^{-1} \boldsymbol{\mu}_{\boldsymbol{\theta}}}_{\text{prior-dependent term}} \Big);
$$

White noise and noninformative (flat) prior on $\boldsymbol{\theta}$

CONSIDER the Jeffreys' noninformative (flat) prior pdf for $\boldsymbol{\theta}$:

$$f(\boldsymbol{\theta}) \propto 1 \quad (C_{\boldsymbol{\theta}}^{-1} = \mathbf{0})$$

and white noise:

$$C_{\boldsymbol{W}} = \sigma^2 I;$$

Then, $f_{\boldsymbol{\Theta}|\boldsymbol{X}}(\boldsymbol{\theta} \,|\, \boldsymbol{x})$ in (1a) simplifies to

$$f_{\boldsymbol{\Theta}|\boldsymbol{X}}(\boldsymbol{\theta} \,|\, \boldsymbol{x}) = \mathcal{N}\big(\boldsymbol{\theta} \,\big|\, \overbrace{(H^T H)^{-1} H^T \boldsymbol{x}}^{\widehat{\boldsymbol{\theta}}_{\mathrm{LS}}(\boldsymbol{x})}, \sigma^2 (H^T H)^{-1}\big)$$

provided that $H^T H$ is invertible, for which we need H to have full rank equal to p, implying $N \geq p$;

Prediction

CONSIDER predicting a $X_\star$ coming from the following model:

$$X_\star = \boldsymbol{h}_\star^T \boldsymbol{\theta} + W_\star$$

where $W_\star \sim \mathcal{N}(0, \sigma^2)$ is independent from $\boldsymbol{W}$, implying that $X_\star$ and $\boldsymbol{x}$ are conditionally independent given $\boldsymbol{\theta} = \boldsymbol{\theta}$ and, therefore,

$$f_{X_\star|\boldsymbol{\Theta},\boldsymbol{X}}(x_\star \,|\, \boldsymbol{\theta}, \boldsymbol{x}) = f_{X_\star|\boldsymbol{\Theta}}(x_\star \,|\, \boldsymbol{\theta}) = \mathcal{N}\big(x_\star \,|\, \boldsymbol{h}_\star^T \boldsymbol{\theta}, \sigma^2\big).$$

Then, our posterior predictive pdf is[1]

[1] along the lines of (7) in handout `bpred`

$$f_{X_\star|\boldsymbol{X}}(x_\star \,|\, \boldsymbol{x}) = \int \underbrace{f_{X_\star|\boldsymbol{\Theta}}(x_\star \,|\, \boldsymbol{\theta})}_{\mathcal{N}(x_\star \,|\, \boldsymbol{h}_\star^T \boldsymbol{\theta}, \sigma^2)} \; \underbrace{f_{\boldsymbol{\Theta}|\boldsymbol{X}}(\boldsymbol{\theta} \,|\, \boldsymbol{x})}_{\mathcal{N}(\boldsymbol{\theta} \,|\, \widehat{\boldsymbol{\theta}}(\boldsymbol{x}), C_{\mathrm{post}})} \; \mathrm{d}\boldsymbol{\theta}$$

where

$$\widehat{\boldsymbol{\theta}}(\boldsymbol{x}) = (H^T C_{\boldsymbol{W}}^{-1} H + C_{\boldsymbol{\Theta}}^{-1})^{-1} (H^T C_{\boldsymbol{W}}^{-1} \boldsymbol{x} + C_{\boldsymbol{\Theta}}^{-1} \boldsymbol{\mu}_{\boldsymbol{\Theta}})$$
$$C_{\mathrm{post}} = (H^T C_{\boldsymbol{W}}^{-1} H + C_{\boldsymbol{\Theta}}^{-1})^{-1}$$

which implies

$$f_{X_\star|\boldsymbol{X}}(x_\star \,|\, \boldsymbol{x}) = \mathcal{N}\big(x_\star \,|\, \boldsymbol{h}_\star^T \widehat{\boldsymbol{\theta}}(\boldsymbol{x}), \boldsymbol{h}_\star^T C_{\mathrm{post}} \boldsymbol{h}_\star + \sigma^2\big).$$

Acronyms

AWGN additive white Gaussian noise. 2

pdf probability density function. 1–3