Bayesian Estimation

Aleksandar Dogandžić

2017-03-21

Contents

READING: §10 in the textbook and [Her15, §4.2].

SUPPOSE that we need to provide a point estimate of the parameter of interest. How to do that in the Bayesian setting? Here, we first consider the most popular *squared-error loss scenario* and then discuss the general scenario with an arbitrary loss.

Construct estimators

$$\widehat{\theta} = \widehat{\theta}(\boldsymbol{x})$$

based on the posterior distribution $f_{\Theta \mid \boldsymbol{X}}(\theta \mid \boldsymbol{x})$. Hence, a *Bayesian approach* to solving the above problem is, say, to obtain $\widehat{\theta}$ by minimizing a *posterior expected (e.g., squared-error) loss*:

$$\begin{aligned}
\rho(\widehat{\theta} \mid \boldsymbol{x}) &= \mathrm{E}_{\Theta \mid \boldsymbol{X}}\{[\widehat{\theta}(\boldsymbol{x}) - \Theta]^2 \mid \boldsymbol{x}\} \\
&= \int \underbrace{\big[\widehat{\theta}(\boldsymbol{x}) - \theta\big]^2}_{\text{squared-error loss}} f_{\Theta \mid \boldsymbol{X}}(\theta \mid \boldsymbol{x})\, \mathrm{d}\theta
\end{aligned}$$

with respect to $\widehat{\theta} = \widehat{\theta}(\boldsymbol{x})$. This is easy to do: Decompose $\rho(\widehat{\theta} \mid \boldsymbol{x})$ as

$$\begin{aligned}
\rho(\widehat{\theta} \mid \boldsymbol{x}) &= \int \big[\widehat{\theta} - \mathrm{E}_{\Theta \mid \boldsymbol{X}}(\Theta \mid \boldsymbol{x}) + \mathrm{E}_{\Theta \mid \boldsymbol{X}}(\Theta \mid \boldsymbol{x}) - \theta\big]^2 f_{\Theta \mid \boldsymbol{X}}(\theta \mid \boldsymbol{x})\, \mathrm{d}\theta \\
&= \int \big[\widehat{\theta} - \mathrm{E}_{\Theta \mid \boldsymbol{X}}(\Theta \mid \boldsymbol{x})\big]^2 f_{\Theta \mid \boldsymbol{X}}(\theta \mid \boldsymbol{x})\, \mathrm{d}\theta \\
&\quad + \underbrace{\int [\theta - \mathrm{E}_{\Theta \mid \boldsymbol{X}}(\Theta \mid \boldsymbol{x})]^2 f_{\Theta \mid \boldsymbol{X}}(\theta \mid \boldsymbol{x})\, \mathrm{d}\theta}_{\text{not a function of } \widehat{\theta}} \\
&\quad + 2[\widehat{\theta} - \mathrm{E}_{\Theta \mid \boldsymbol{X}}(\Theta \mid \boldsymbol{x})] \underbrace{\int \big[\theta - \mathrm{E}_{\Theta \mid \boldsymbol{X}}(\Theta \mid \boldsymbol{x})\big] f_{\Theta \mid \boldsymbol{X}}(\theta \mid \boldsymbol{x})\, \mathrm{d}\theta}_{\mathrm{E}_{\Theta \mid \boldsymbol{X}}(\Theta \mid \boldsymbol{x}) - \mathrm{E}_{\Theta \mid \boldsymbol{X}}(\Theta \mid \boldsymbol{x}) = 0}
\end{aligned}$$

and the optimal $\widehat{\theta}$ follows by minimizing the first term:

Hero refers to this estimator as the conditional-mean estimator [Her15, §4.2.1]

$$\arg\min_{\widehat{\theta}} \rho(\widehat{\theta} \,|\, \boldsymbol{x}) = \mathrm{E}_{\Theta|\boldsymbol{X}}(\Theta \,|\, \boldsymbol{x}).$$

The posterior mean of the parameter θ minimizes its posterior expected squared loss; the minimum posterior expected squared loss is

$$\begin{aligned}\min_{\widehat{\theta}} \rho(\widehat{\theta} \,|\, \boldsymbol{x}) &= \int [\theta - \mathrm{E}_{\Theta|\boldsymbol{X}}(\Theta \,|\, \boldsymbol{x})]^2 f_{\Theta|\boldsymbol{X}}(\theta \,|\, \boldsymbol{x})\, \mathrm{d}\theta \\ &= \mathrm{var}_{\Theta|\boldsymbol{X}}(\Theta \,|\, \boldsymbol{x}).\end{aligned}$$

Bayesian MSE

RECALL mean-square error (MSE) measures:

1. Classical MSE:

$$\begin{aligned}\mathrm{MSE}\{\widehat{\theta}(\boldsymbol{x})\} &= \mathrm{E}_{\boldsymbol{X}|\Theta}([\widehat{\theta}(\boldsymbol{X}) - \theta]^2 \,|\, \theta) \\ &= \int [\widehat{\theta}(\boldsymbol{x}) - \theta]^2 f_{\boldsymbol{X}|\Theta}(\boldsymbol{x} \,|\, \theta)\, \mathrm{d}\boldsymbol{x}\end{aligned} \tag{1}$$

   see also (3b) in handout `est_perf`.

2. Bayesian mean-square error (BMSE) (preposterior MSE):

$$\begin{aligned}\mathrm{BMSE}\{\widehat{\theta}(\boldsymbol{x})\} &= \mathrm{E}_{\boldsymbol{X},\Theta}([\widehat{\theta}(\boldsymbol{X}) - \Theta]^2) \\ &= \int\int [\widehat{\theta}(\boldsymbol{x}) - \theta]^2 f_{\boldsymbol{X}|\Theta}(x \,|\, \theta)\, f(\theta)\, \mathrm{d}x\, \mathrm{d}\theta \\ &\overset{\text{iter. exp.}}{=} \mathrm{E}_\Theta\big[\underbrace{\mathrm{E}_{\boldsymbol{X}|\Theta}\big([\widehat{\theta}(\boldsymbol{X}) - \Theta]^2 \,|\, \Theta\big)}_{\mathrm{MSE}\{\widehat{\theta}(\boldsymbol{X})\},\ \text{see (1)}}\big]\end{aligned} \tag{2}$$

   see also (6) in handout `intro`. The BMSE is obtained by averaging the squared-error loss over both the random measurements *and* parameter realizations. It is computable before the data has been collected, hence the name preposterior.

✻ COMMENTS:

- Classical MSE generally depends on the true value of the parameter θ. Classical minimum mean-square error (MMSE) "estimates" usually depend on θ and, therefore, do not exist.

- Since Θ is integrated out in (3), BMSE *does not* depend on θ; hence, Bayesian MMSE estimates exist.

  Which estimator $\widehat{\theta}(\boldsymbol{x})$ minimizes the BMSE? Since

$$\begin{aligned}\mathrm{BMSE}\{\widehat{\theta}(\boldsymbol{x})\} &= \mathrm{E}_{\boldsymbol{X},\Theta}\big([\widehat{\theta}(\boldsymbol{X}) - \Theta]^2\big) \\ &= \mathrm{E}_{\boldsymbol{X}}\Big\{\underbrace{\mathrm{E}_{\Theta|\boldsymbol{X}}\big([\widehat{\theta}(\boldsymbol{X}) - \Theta]^2 \,|\, \boldsymbol{X}\big)}_{\rho(\widehat{\theta}|\boldsymbol{X})}\Big\}\end{aligned} \tag{3}$$

and, for every given $\boldsymbol{x}$, we know that

$$\widehat{\theta}(\boldsymbol{x}) = \mathrm{E}_{\Theta \mid \boldsymbol{X}}(\Theta \mid \boldsymbol{x})$$

posterior mean of Θ

minimizes the posterior expected squared loss $\rho(\widehat{\theta} \mid \boldsymbol{x})$. Therefore, $\widehat{\theta}(\boldsymbol{x}) = \mathrm{E}_{\Theta \mid \boldsymbol{X}}(\Theta \mid \boldsymbol{x})$ minimizes the BMSE in (3).

Bayes Risk

DEFINE the estimation error

$$\begin{aligned}\epsilon &= \epsilon(\boldsymbol{x}, \theta)\\ &= \widehat{\theta}(\boldsymbol{x}) - \theta\end{aligned}$$

and assign a loss (cost) function $\mathbb{L}(\epsilon)$. We may choose $\widehat{\theta}(\boldsymbol{x})$ to minimize the *Bayes (preposterior) risk*:

$$\mathrm{E}_{\boldsymbol{X},\Theta}\big[\mathbb{L}(\epsilon)\big] = \mathrm{E}_{\boldsymbol{X},\Theta}\big[\mathbb{L}\big(\widehat{\theta}(\boldsymbol{x}) - \Theta\big)\big]$$

but this is equivalent to minimizing the *posterior expected loss*:

$$\begin{aligned}\rho(\widehat{\theta} \mid \boldsymbol{x}) &= \mathrm{E}_{\Theta \mid \boldsymbol{X}}\big[\mathbb{L}(\epsilon) \mid \boldsymbol{x}\big]\\ &= \int \mathbb{L}\big(\widehat{\theta}(\boldsymbol{x}) - \theta\big)\, f_{\Theta \mid \boldsymbol{X}}(\theta \mid \boldsymbol{x})\, \mathrm{d}\theta\end{aligned}$$

for each $\boldsymbol{X} = \boldsymbol{x}$, which is a Bayesian criterion. The proof is the same as for the squared-error loss:

$$\mathrm{E}_{\boldsymbol{X},\Theta}\big[\mathbb{L}\big(\widehat{\theta}(\boldsymbol{x}) - \Theta\big)\big] \overset{\text{iter. exp.}}{=} \mathrm{E}_{\boldsymbol{X}}\Big\{\underbrace{\mathrm{E}_{\Theta \mid \boldsymbol{X}}\big[\mathbb{L}\big(\widehat{\theta}(\boldsymbol{X}) - \Theta\big) \,\big|\, \boldsymbol{X}\big]}_{\rho(\widehat{\theta}(\boldsymbol{X}) \mid \boldsymbol{X})}\Big\}. \tag{4}$$

Loss functions

POPULAR choices:

1. $\mathbb{L}(\epsilon) = \epsilon^2$, squared-error loss (accurate, most popular),
2. $\mathbb{L}(\epsilon) = |\epsilon|$, absolute-error loss (robust to outliers),
3. 0-1 loss (tractable)

$$\begin{aligned} \mathbb{L}(\epsilon) &= \begin{cases} 0, & |\epsilon| \le \Delta/2 \\ 1, & |\epsilon| > \Delta/2 \end{cases} \\ &= 1 - \mathbb{1}_{[-\Delta/2, \Delta/2]}(\epsilon). \end{aligned}$$

Estimators $\widehat{\theta} = \widehat{\theta}(\boldsymbol{x})$ that minimize the corresponding Bayes risks:

1. MMSE estimator:

$$\widehat{\theta} = \widehat{\theta}(\boldsymbol{x}) = \mathrm{E}_{\Theta \mid \boldsymbol{X}}(\Theta \mid \boldsymbol{X} = \boldsymbol{x})$$

   the *posterior mean of θ given $\boldsymbol{X} = \boldsymbol{x}$*;

2. *Posterior median of θ given $\boldsymbol{X} = \boldsymbol{x}$*, i.e., the optimal $\widehat{\theta}$ satisfies:

$$\int_{-\infty}^{\widehat{\theta}} f_{\Theta \mid \boldsymbol{X}}(\theta \mid \boldsymbol{x}) \, \mathrm{d}\theta = \int_{\widehat{\theta}}^{+\infty} f_{\Theta \mid \boldsymbol{X}}(\theta \mid \boldsymbol{x}) \, \mathrm{d}\theta$$

   HW: check this.

3. maximum *a posteriori* (MAP) estimator:

$$\begin{aligned} \widehat{\theta} &= \widehat{\theta}_{\mathrm{MAP}}(\boldsymbol{x}) \\ &= \arg\max_{\theta} f_{\Theta \mid \boldsymbol{X}}(\theta \mid \boldsymbol{x}) \end{aligned} \tag{5}$$

   the *posterior mode of θ given $\boldsymbol{X} = \boldsymbol{x}$*.

   We now show the MAP estimator result in 3. Start from (4):

$$\mathrm{E}_X \left\{ \mathrm{E}_{\Theta \mid X}[\mathbb{L}(\epsilon) \mid X] \right\} = \mathrm{E}_X \left[1 - \int_{\widehat{\theta}-\Delta/2}^{\widehat{\theta}+\Delta/2} f_{\Theta \mid \boldsymbol{X}}(\theta \mid x) \, \mathrm{d}\theta \right].$$

To minimize this expression with respect to $\widehat{\theta}$, we maximize

$$\int_{\widehat{\theta}-\Delta/2}^{\widehat{\theta}+\Delta/2} f_{\Theta \mid \boldsymbol{X}}(\theta \mid x) \, \mathrm{d}\theta$$

with respect to $\widehat{\theta}$ which, for small Δ, reduces to maximizing

$$f_{\Theta \mid \boldsymbol{X}}(\widehat{\theta} \mid x)$$

with respect to $\widehat{\theta}$ and (5) follows. Since

$$f_{\Theta|\boldsymbol{X}}(\theta \mid \boldsymbol{x}) \propto f_{\boldsymbol{X}|\Theta}(\boldsymbol{x} \mid \theta)\, f_\Theta(\theta)$$

we have

$$\widehat{\theta}_{\text{MAP}}(\boldsymbol{x}) = \arg\max_\theta \big[\ln f_{\boldsymbol{X}|\Theta}(\boldsymbol{x} \mid \theta) + \ln_\Theta f(\theta)\big].$$

Equivalence of MAP and ML for flat priors

NOTE: $\ln f_{\boldsymbol{X}|\Theta}(\boldsymbol{x} \mid \theta)$ is the log-likelihood function of θ. Thus, for a *flat prior*,

$$f_\Theta(\theta) \propto 1$$

the MAP estimator of θ *coincides with* the corresponding classical maximum-likelihood (ML) estimator.

Acronyms

BMSE Bayesian mean-square error. 2, 3

MAP maximum *a posteriori*. 4, 5

ML maximum-likelihood. 5

MMSE minimum mean-square error. 2, 4

MSE mean-square error. 2

References

[Her15] A. O. Hero, *Statistical methods for signal processing*, Lecture notes, Univ. Michigan, Ann Arbor, MI, 2015.