KTH ROYAL INSTITUTE OF TECHNOLOGY

Estimation theory – Lecture 5

- Ch. 10 The Bayesian Philosophy
- Ch. 11 General Bayesian Estimators

Magnus Jansson

The Geneal Bayesian Philosophy

Assume we have prior knowledge about θ, e.g. we know $0 < \theta < 10[V]$.

MLE:

$$\hat{\theta} = \arg \max_{0<\theta<10} \ln p(\mathbf{x};\theta)$$

Alt: Consider θ as a random variable with PDF $p(\theta)$, e.g., $\theta \sim \mathcal{U}(0,10)$

The problem is then decribed by the joint PDF $p(\mathbf{x},\theta)$

Reconsider MMSE estimation

Bayesian MSE:
$\mathrm{Bmse}(\hat{\theta}) = \mathrm{E}(\theta-\hat{\theta})^2 = \int\int(\theta-\hat{\theta})^2 p(\mathbf{x},\theta)d\mathbf{x}d\theta$

Cf. "classical" MSE:
$\mathrm{mse}(\hat{\theta}) = \mathrm{E}_{\mathbf{x}}(\theta-\hat{\theta})^2 = \int(\theta-\hat{\theta})^2 p(\mathbf{x};\theta)d\mathbf{x}$

MMSE cont'd

We have $p(\mathbf{x},\theta) = p(\theta|\mathbf{x})p(\mathbf{x})$ and

$$\mathrm{Bmse}(\hat{\theta}) = \int \left[\underbrace{\int (\theta-\hat{\theta})^2 p(\theta|\mathbf{x})d\theta}_{=:I(\hat{\theta},\mathbf{x})} \right] p(\mathbf{x})d\mathbf{x}$$

Try to minimize inner integral for fixed $\mathbf{x}$:

$$\frac{\partial I(\hat{\theta},\mathbf{x})}{\partial\hat{\theta}} = \int -2(\theta-\hat{\theta})p(\theta|\mathbf{x})d\theta$$
$$= -2\int \theta p(\theta|\mathbf{x})d\theta + 2\hat{\theta}\int p(\theta|\mathbf{x})d\theta$$

MMSE cont'd

This equals zero when

$$\hat{\theta} = \int \theta p(\theta|\mathbf{x}) d\theta = \mathrm{E}(\theta|\mathbf{x})$$

the conditional mean or the mean of the *posterior* PDF $p(\theta|\mathbf{x})$.

The BMMSE (or simply the MMSE) estimate will not depend on the true θ as it generally does in the classical setting. It depends on the prior PDF instead.

Need to find $p(\theta|\mathbf{x})$. Using Bayes rule

$$p(\theta|\mathbf{x}) = \frac{p(\mathbf{x}|\theta)p(\theta)}{\int p(\mathbf{x}|\theta)p(\theta)d\theta}$$

Requires multiple integrations in general.

Multivariate Gaussian

Assume $\mathbf{x} \in \mathbb{R}^k, \ \mathbf{y} \in \mathbb{R}^l$ and:

$$\begin{bmatrix}\mathbf{x}\\ \mathbf{y}\end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix}\mathbf{m}_x\\ \mathbf{m}_y\end{bmatrix}, \begin{bmatrix}\mathbf{C}_{xx} & \mathbf{C}_{xy}\\ \mathbf{C}_{yx} & \mathbf{C}_{yy}\end{bmatrix}\right)$$

Then

$$p(\mathbf{x},\mathbf{y}) = \frac{1}{(2\pi)^{\frac{k+l}{2}}\sqrt{\det(\mathbf{C})}} \exp\left\{-\frac{1}{2}\begin{bmatrix}\mathbf{x}-\mathbf{m}_x\\ \mathbf{y}-\mathbf{m}_y\end{bmatrix}^T \mathbf{C}^{-1}\begin{bmatrix}\mathbf{x}-\mathbf{m}_x\\ \mathbf{y}-\mathbf{m}_y\end{bmatrix}\right\}$$

The random variable $\mathbf{y}|\mathbf{x}$ is also Gaussian with

$$\mathrm{E}(\mathbf{y}|\mathbf{x}) = \mathbf{m}_y + \mathbf{C}_{yx}\mathbf{C}_{xx}^{-1}[\mathbf{x}-\mathbf{m}_x]$$
$$\mathbf{C}_{\mathbf{y}|\mathbf{x}} = \mathbf{C}_{yy} - \mathbf{C}_{yx}\mathbf{C}_{xx}^{-1}\mathbf{C}_{xy}$$

Bayesian general linear model

$$\mathbf{x} = \mathbf{H}\boldsymbol{\theta} + \mathbf{w}$$
$$\boldsymbol{\theta} \sim \mathcal{N}(\boldsymbol{\mu}_\theta, \mathbf{C}_\theta)$$
$$\mathbf{w} \sim \mathcal{N}(0, \mathbf{C}_w); \quad \text{independent of } \boldsymbol{\theta}$$

Then, $\mathbf{x}$ and $\boldsymbol{\theta}$ are jointly Gaussian with

$$\hat{\boldsymbol{\theta}} = \mathrm{E}(\boldsymbol{\theta}|\mathbf{x}) = \boldsymbol{\mu}_\theta + \mathbf{C}_\theta\mathbf{H}^T[\mathbf{H}\mathbf{C}_\theta\mathbf{H}^T + \mathbf{C}_w]^{-1}(\mathbf{x} - \mathbf{H}\boldsymbol{\mu}_\theta)$$
$$\mathbf{C}_{\theta|\mathbf{x}} = \mathbf{C}_\theta - \mathbf{C}_\theta\mathbf{H}^T[\mathbf{H}\mathbf{C}_\theta\mathbf{H}^T + \mathbf{C}_w]^{-1}\mathbf{H}\mathbf{C}_\theta$$

Alternative expressions

$$\hat{\boldsymbol{\theta}} = \mathrm{E}(\boldsymbol{\theta}|\mathbf{x}) = \boldsymbol{\mu}_\theta + [\mathbf{C}_\theta^{-1} + \mathbf{H}^T\mathbf{C}_w^{-1}\mathbf{H}]^{-1}\mathbf{H}^T\mathbf{C}_w^{-1}(\mathbf{x} - \mathbf{H}\boldsymbol{\mu}_\theta)$$
$$\mathbf{C}_{\theta|\mathbf{x}} = [\mathbf{C}_\theta^{-1} + \mathbf{H}^T\mathbf{C}_w^{-1}\mathbf{H}]^{-1}$$

or

$$\mathbf{C}_{\theta|\mathbf{x}}^{-1} = \mathbf{C}_\theta^{-1} + \mathbf{H}^T\mathbf{C}_w^{-1}\mathbf{H}$$

Note that $\mathbf{H}^T\mathbf{C}_w^{-1}\mathbf{H}$ is the inverse of the covariance matrix of $\boldsymbol{\theta}$ for the classical linear model. Cf. additivity of FIM.

Evaluation of estimators

Note that in the Bayesian setting, $\boldsymbol{\theta}$ takes different values from $p(\boldsymbol{\theta})$ in each experiment similar to the noise $\mathbf{w}$. That is, in our Monte Carlo simulations we should average over both $\boldsymbol{\theta}$ and the noise $\mathbf{w}$.

Nuisance parameters

Assume $\boldsymbol{\theta} = [\boldsymbol{\alpha}^T \ \boldsymbol{\beta}^T]^T$ and we are only interested in estimating $\boldsymbol{\alpha}$. Compute

$$p(\boldsymbol{\alpha}|\mathsf{x}) = \int p(\boldsymbol{\alpha}, \boldsymbol{\beta}|\mathsf{x}) d\boldsymbol{\beta}$$

(marginalization over $\boldsymbol{\beta}$). Alternatively,

$$p(\boldsymbol{\alpha}|\mathsf{x}) = \frac{p(\mathsf{x}|\boldsymbol{\alpha})p(\boldsymbol{\alpha})}{\int p(\mathsf{x}|\boldsymbol{\alpha})p(\boldsymbol{\alpha})d\boldsymbol{\alpha}}$$

where

$$p(\mathsf{x}|\boldsymbol{\alpha}) = \int p(\mathsf{x}|\boldsymbol{\alpha}, \boldsymbol{\beta})p(\boldsymbol{\beta}|\boldsymbol{\alpha})d\boldsymbol{\beta}$$

$$= \text{if } \boldsymbol{\beta}, \boldsymbol{\alpha} \text{ indep.} = \int p(\mathsf{x}|\boldsymbol{\alpha}, \boldsymbol{\beta})p(\boldsymbol{\beta})d\boldsymbol{\beta}$$

Ch.11 General Bayesian Estimators

Define the cost function

$$C(\varepsilon) = \varepsilon^2 = (\theta - \hat{\theta})^2$$

for each realization of θ and $\mathbf{x}$.

Then $\text{Bmse}(\hat{\theta}) = \text{E}[C(\varepsilon)]$ and minimizing Bmse gave the MMSE estimator.

We could consider other cost functions!

Cost functions

Consider e.g.

$$C(\varepsilon) = |\varepsilon|$$

$$C(\varepsilon) = \begin{cases} 1 & |\varepsilon| > \delta \\ 0 & |\varepsilon| \leq \delta \end{cases}; \delta \to 0$$

The last cost is called the "hit-or-miss" function.

In general $R = \text{E}[C(\varepsilon)]$ is termed the Bayes risk.

Bayesian Estimators

The cost function $C(\varepsilon) = |\varepsilon|$ leads to the estimator:

$$\hat{\theta} \text{ is the } \textit{median} \text{ such that } \Pr(\theta \le \hat{\theta}|\mathbf{x}) = 1/2$$

The hit-or-miss cost function leads to

$$\hat{\theta} = \arg\max_{\theta} p(\theta|\mathbf{x})$$

The maximum a posteriori (MAP) estimator.

MAP estimators

Note that

$$p(\theta|\mathbf{x}) = \frac{p(\mathbf{x}|\theta)p(\theta)}{p(\mathbf{x})}$$

and, hence,

$$\hat{\theta}_{\text{MAP}} = \arg\max_{\theta} p(\theta|\mathbf{x}) = \arg\max_{\theta} p(\mathbf{x}|\theta)p(\theta)$$
$$= \arg\max_{\theta} [\ln p(\mathbf{x}|\theta) + \ln p(\theta)]$$

Cf. MLE when $p(\mathbf{x}|\theta) = p(\mathbf{x};\theta)$ and $p(\theta)$ is flat over the support of $p(\mathbf{x};\theta)$ (non-informative prior).

MAP estimators cont'd

No integration is needed in the scalar case. However, in the vector case

$$\hat{\theta}_i = \arg\max_{\theta_i} p(\theta_i|\mathbf{x}); \qquad i = 1, 2, \ldots, p$$

where $p(\theta_i|\mathbf{x})$ is obtained by marginalizing (integrating) $p(\boldsymbol{\theta}|\mathbf{x})$ over the other parameters in $\boldsymbol{\theta}$.

$\hat{\theta}_i$ minimizes the risk $R_i = \mathrm{E}_{\mathbf{x},\theta_i}[C(\theta_i - \hat{\theta}_i)]$ for the hit-or-miss cost.

Vector MAP

Use "vector MAP" instead:

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} p(\boldsymbol{\theta}|\mathbf{x})$$

It minimizes the risk $R = \mathrm{E}[C(\boldsymbol{\varepsilon})]$ with $\boldsymbol{\varepsilon} = \boldsymbol{\theta} - \hat{\boldsymbol{\theta}}$ and

$$C(\boldsymbol{\varepsilon}) = \begin{cases} 1 & \|\boldsymbol{\varepsilon}\| > \delta \\ 0 & \|\boldsymbol{\varepsilon}\| \le \delta \end{cases}; \delta \to 0$$

Not the same as above but usually the one referred to as MAP.

MAP Example

$$
\begin{aligned}
\mathbf{x} &= \mathbf{H}\boldsymbol{\theta} + \mathbf{w} \\
\mathbf{w} &\sim \mathcal{N}(0, \sigma^2 \mathbf{I}); \quad \text{independent of } \boldsymbol{\theta} \\
p(\boldsymbol{\theta}) &= \prod_{i=1}^{p} \frac{1}{2b} \exp(-|\theta_i|/b)
\end{aligned}
$$

That is, $\boldsymbol{\theta}$ has a Laplace prior distribution. We have

$$
p(\mathbf{x}|\boldsymbol{\theta}) = \frac{1}{(2\pi\sigma^2)^{\frac{N}{2}}} \exp\left\{ -\frac{1}{2\sigma^2} (\mathbf{x} - \mathbf{H}\boldsymbol{\theta})^T (\mathbf{x} - \mathbf{H}\boldsymbol{\theta}) \right\}
$$

MAP Example cont'd

MAP estimator:

$$
\begin{aligned}
\hat{\boldsymbol{\theta}}_{\text{MAP}} &= \arg\max_{\boldsymbol{\theta}} \{\ln p(\mathbf{x}|\boldsymbol{\theta}) + \ln p(\boldsymbol{\theta})\} \\
&= \arg\min_{\boldsymbol{\theta}} \left\{ \|\mathbf{x} - \mathbf{H}\boldsymbol{\theta}\|_2^2 + \lambda \sum_{i=1}^{p} |\theta_i| \right\} \\
&= \arg\min_{\boldsymbol{\theta}} \left\{ \|\mathbf{x} - \mathbf{H}\boldsymbol{\theta}\|_2^2 + \lambda \|\boldsymbol{\theta}\|_1 \right\}
\end{aligned}
$$

for some constant λ.