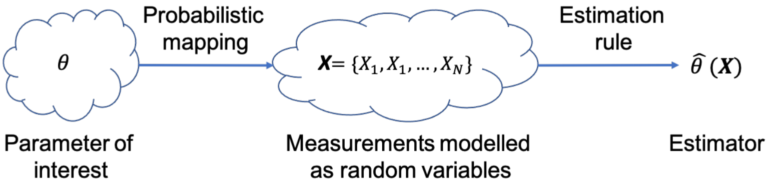
Bayesian estimators

Raj Thilak Rajan

Overview

Estimation Philosophy



- Let $X = \{X_1, X_2, \ldots, X_N\}$ be a set of random samples drawn from probability distributions $f_{X_n}(x_n; \boldsymbol{\theta})\ \forall\ 1 \leq n \leq N$, where $\boldsymbol{\theta}$ is the parameter of interest
- We aim to
  (a) recover the unknown $\boldsymbol{\theta}$ from the measurements X, and
  (b) provide a performance measure of the estimated $\boldsymbol{\theta}$
- Bayesian philosophy : θ is a random variable and the statistics of θ is known.

Bayesian mean square error (Bmse)

- θ is viewed as a random variable
- We would like to minimize the MSE

$$\text{Bmse}(\hat{\theta}) = \mathbb{E}[(\hat{\theta} - \theta)^2]$$

where both $\mathbf{x}$ and θ are random, and the statistics of $\hat{\theta}$ depend on the statistics of both $\mathbf{x}$ and θ.

- Note the difference between these two MSEs:

$$\mathrm{mse}(\hat{\theta}) = \mathbb{E}[(\hat{\theta} - \theta)^2] = \int (\hat{\theta} - \theta)^2 p(\mathbf{x}; \theta) d\mathbf{x}$$
$$\text{Bmse}(\hat{\theta}) = \mathbb{E}[(\hat{\theta} - \theta)^2] = \int \int (\hat{\theta} - \theta)^2 p(\mathbf{x}, \theta) d\mathbf{x} d\theta$$

- Note that *mse* depends on θ, but *Bmse* does not, only on its statistics.

MMSE estimator: Gaussian prior

Consider the estimation of A

$$x[n] = A + w[n], \quad n = 0, \cdots, N-1, \quad w[n] \sim \mathcal{N}(0, \sigma^2) \quad A \sim \mathcal{N}(\mu_A, \sigma_A^2)$$

MMSE estimator:

$$\hat{A} = \mathbb{E}(A|\mathbf{x}) = \mu_{A|x} = \frac{\frac{N}{\sigma^2}\bar{x} + \frac{\mu_A}{\sigma_A^2}}{\frac{N}{\sigma^2} + \frac{1}{\sigma_A^2}} = \frac{\sigma_A^2 \bar{x} + \frac{\sigma^2}{N}\mu_A}{\frac{\sigma^2}{N} + \sigma_A^2} = \alpha\bar{x} + (1-\alpha)\mu_A \tag{1}$$

where $\alpha = \frac{\sigma_A^2}{\sigma_A^2 + \frac{\sigma^2}{N}}$ and $0 \leq \alpha \leq 1$.

Remarks:

- α: the interplay between the prior knowledge (μ_A) and the data ($\bar{x}$).
- For small N or large σ^2: $\alpha \to 0$, $\sigma_A^2 << \sigma^2/N$ and $\hat{A} = \mu_A$.
- For larger N or small σ^2: $\alpha \approx 1$ and $\hat{A} = \bar{x}$.
- Note that the MMSE estimator always exists, given a prior $p(\theta)$.

Bivariate Gaussian process

If x and θ are jointly Gaussian, with joint mean and covariance matrix

$$\mathbb{E}\left(\begin{bmatrix} x \\ \theta \end{bmatrix}\right) = \begin{bmatrix} \mathbb{E}(x) \\ \mathbb{E}(\theta) \end{bmatrix}, \mathbf{C} = \begin{bmatrix} var(x) & cov(x,\theta) \\ cov(\theta,x) & var(\theta) \end{bmatrix}$$

such that

$$p(x,\theta) = \frac{1}{(2\pi)\sqrt{\det(\mathbf{C})}} \exp \mathbf{Q}$$

where

$$\mathbf{Q} = -\frac{1}{2}\left[\begin{bmatrix} x - \mathbb{E}(x) \\ \theta - \mathbb{E}(\theta) \end{bmatrix}^T \mathbf{C}^{-1} \begin{bmatrix} x - \mathbb{E}(x) \\ \theta - \mathbb{E}(\theta) \end{bmatrix}\right]$$

then the conditional PDF $p(\theta|x)$ is also Gaussian with mean and variance

$$\begin{aligned} \mathbb{E}(\theta|x) &= \mathbb{E}(\theta) + \frac{cov(\theta,x)}{var(x)}(x - \mathbb{E}(x)) \\ var(\theta|x) &= var(\theta) - \frac{cov(x,\theta)^2}{var(x)} = var(\theta)\left(1 - \frac{cov(x,\theta)^2}{var(x)var(\theta)}\right) \\ &= var(\theta)\left(1 - \rho^2\right) \end{aligned}$$

Multivariate Gaussian process

If $\mathbf{x}$ and $\boldsymbol{\theta}$ are jointly Gaussian, where $\mathbf{x}$ is $k \times 1$ and $\boldsymbol{\theta}$ is $l \times 1$, with joint mean and covariance matrix

$$\mathbb{E}\left(\begin{bmatrix}\mathbf{x}\\ \boldsymbol{\theta}\end{bmatrix}\right) = \begin{bmatrix}\mathbb{E}(\mathbf{x})\\ \mathbb{E}(\boldsymbol{\theta})\end{bmatrix}, \mathbf{C} = \begin{bmatrix}\mathbf{C}_{xx} & \mathbf{C}_{xy}\\ \mathbf{C}_{yx} & \mathbf{C}_{yy}\end{bmatrix}$$

such that

$$p(\mathbf{x}, \boldsymbol{\theta}) = \frac{1}{\sqrt{(2\pi)^{k+l}\det(\mathbf{C})}} \exp \mathbf{Q}$$

where

$$\mathbf{Q} = -\frac{1}{2}\left[\begin{bmatrix}\mathbf{x} - \mathbb{E}(\mathbf{x})\\ \boldsymbol{\theta} - \mathbb{E}(\boldsymbol{\theta})\end{bmatrix}^T \mathbf{C}^{-1} \begin{bmatrix}\mathbf{x} - \mathbb{E}(\mathbf{x})\\ \boldsymbol{\theta} - \mathbb{E}(\boldsymbol{\theta})\end{bmatrix}\right]$$

then the conditional PDF $p(\boldsymbol{\theta}|\mathbf{x})$ is also Gaussian with moments

$$\begin{aligned}\mathbb{E}(\boldsymbol{\theta}|\mathbf{x}) &= \mathbb{E}(\boldsymbol{\theta}) + \mathbf{C}_{\theta x}\mathbf{C}_{xx}^{-1}(\mathbf{x} - \mathbb{E}(\mathbf{x}))\\ \mathbf{C}_{\theta|x} &= \mathbf{C}_{\theta\theta} - \mathbf{C}_{\theta x}\mathbf{C}_{xx}^{-1}\mathbf{C}_{x\theta}\end{aligned}$$

MMSE estimator: Linear Gaussian model

- Consider the generalized linear Gaussian model:

$$\mathbf{x} = \mathbf{H}\boldsymbol{\theta} + \mathbf{w}, \quad \mathbf{w} \sim \mathcal{N}(\mathbf{0}, \mathbf{C})$$

  where $\boldsymbol{\theta}$ is a random vector with distribution $\mathcal{N}(\boldsymbol{\mu}_\theta, \mathbf{C}_\theta)$.
- Here, $p(\boldsymbol{\theta}|\mathbf{x})$ is also Gaussian with mean and covariance matrix

$$\begin{aligned}\mathbb{E}(\boldsymbol{\theta}|\mathbf{x}) &= \boldsymbol{\mu}_\theta + \mathbf{C}_\theta\mathbf{H}^T(\mathbf{H}\mathbf{C}_\theta\mathbf{H}^T + \mathbf{C})^{-1}(\mathbf{x} - \mathbf{H}\boldsymbol{\mu}_\theta) \\ \mathbf{C}_{\theta|x} &= \mathbf{C}_\theta - \mathbf{C}_\theta\mathbf{H}^T(\mathbf{H}\mathbf{C}_\theta\mathbf{H}^T + \mathbf{C})^{-1}\mathbf{H}\mathbf{C}_\theta\end{aligned}$$

- Alternative formulation using matrix inversion lemma:

$$\begin{aligned}\mathbb{E}(\boldsymbol{\theta}|\mathbf{x}) &= \boldsymbol{\mu}_\theta + (\mathbf{C}_\theta^{-1} + \mathbf{H}^T\mathbf{C}^{-1}\mathbf{H})^{-1}\mathbf{H}^T\mathbf{C}^{-1}(\mathbf{x} - \mathbf{H}\boldsymbol{\mu}_\theta) \\ \mathbf{C}_{\theta|x} &= (\mathbf{C}_\theta^{-1} + \mathbf{H}^T\mathbf{C}^{-1}\mathbf{H})^{-1}\end{aligned}$$

Bayes risk

- Bayesian MSE Bmse$(\hat{\theta})$

$$\mathbb{E}[\underbrace{(\hat{\theta}(\mathbf{x})-\theta)^2}_{\mathcal{C}(\epsilon)}] = \int\int \mathcal{C}(\epsilon)p(\mathbf{x},\theta)d\mathbf{x}d\theta, = \int\left[\int \mathcal{C}(\epsilon)p(\theta|\mathbf{x})d\theta\right]p(\mathbf{x})d\mathbf{x},$$

- We can more generally minimize the Bayes risk $\mathcal{R} = \mathbb{E}[\mathcal{C}(\epsilon)]$, where $\epsilon = \theta - \hat{\theta}$ and $\mathcal{C}$ is a cost function that can take many forms e.g.,

$$\mathcal{C}(\epsilon) = \epsilon^2, \qquad \mathcal{C}(\epsilon) = |\epsilon|, \qquad \mathcal{C}(\epsilon) = \begin{cases} 0 & |\epsilon| \leq \delta \\ 1 & |\epsilon| > \delta \end{cases}, \text{ with } \delta \to 0$$

- As for the MMSE, we now have to minimize (the inner integral of Bmse)

$$g(\hat{\theta}) = \int \mathcal{C}(\theta-\hat{\theta})p(\theta|\mathbf{x})d\theta.$$

- Recollect that for $\mathcal{C}(\epsilon) = \epsilon^2$, $\hat{\theta} = E[\theta|\mathbf{x}]$ i.e., the mean of the posterior.

MMSE estimator: "Absolute" error

- Consider the cost $\mathcal{C}(\epsilon) = |\epsilon|$:

$$\int |\theta - \hat{\theta}| p(\theta|\mathbf{x}) d\theta = \int_{-\infty}^{\hat{\theta}} (\hat{\theta} - \theta) p(\theta|\mathbf{x}) d\theta + \int_{\hat{\theta}}^{\infty} (\theta - \hat{\theta}) p(\theta|\mathbf{x}) d\theta.$$

- Differentiation with respect to $\hat{\theta}$, setting the result to zero we have

$$\int_{-\infty}^{\hat{\theta}} p(\theta|\mathbf{x}) d\theta = \int_{\hat{\theta}}^{\infty} p(\theta|\mathbf{x}) d\theta$$

- Hence, for $\mathcal{C}(\epsilon) = |\epsilon|$, the MMSE estimator is the median of the posterior.

Property: Leibniz rule for differentiation of integral:

$$\frac{\partial}{\partial u} \int_{\phi_1(u)}^{\phi_2(u)} h(u,v) dv = \int_{\phi_1(u)}^{\phi_2(u)} \frac{\partial}{\partial u} h(u,v) dv + \frac{d\phi_2(u)}{du} h(u, \phi_2(u)) - \frac{d\phi_1(u)}{du} h(u, \phi_1(u))$$

MMSE estimator: "Hit-or-miss" error

- Consider the "hit-or-miss" cost function:

$$\mathcal{C}(\epsilon) = \begin{cases} 0 & |\epsilon| \leq \delta \\ 1 & |\epsilon| > \delta \end{cases}, \text{ with } \delta \to 0$$

- Hence, we minimize

$$g(\hat{\theta}) = \int \mathcal{C}(\epsilon) p(\theta|\mathbf{x}) d\theta = \int_{-\infty}^{\hat{\theta}-\delta} 1 p(\theta|\mathbf{x}) d\theta + \int_{\hat{\theta}+\delta}^{\infty} 1 p(\theta|\mathbf{x}) d\theta.$$

- Alternatively, maximizing

$$\int_{\hat{\theta}-\delta}^{\hat{\theta}+\delta} p(\theta|\mathbf{x}) d\theta,$$

- For an arbitrarily small δ, this implies $\hat{\theta}$ corresponds to the location of the maximum of $p(\theta|\mathbf{x})$ i.e., the mode of the posterior.

Maximum a posteriori (MAP)

- The MAP estimator corresponds to

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

- Using Bayes' rule, this is thus identical to

$$\hat{\theta} = \arg\max_{\theta} p(\mathbf{x}|\theta)p(\theta) = \arg\max_{\theta} log\left(p(\mathbf{x}|\theta)\right) + log\left(p(\theta)\right).$$

- MAP is easier to calculate than the MMSE, since integration is avoided.

MAP estimator: Properties

- The MAP estimator corresponds to

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

- Note that if $p(\theta)$ is uniform and $p(\mathbf{x}|\theta)$ falls within this interval, then

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

  which is essentially the Bayesian MLE.
- If $N \to \infty$, the pdf $p(\mathbf{x}|\theta)$ becomes dominant over $p(\theta)$ and the MAP becomes thus identical to the Bayesian MLE.
- If the $\mathbf{x}$ and $\boldsymbol{\theta}$ are jointly Gaussian, then the MAP estimator is identical to the MMSE estimator.

Linear MMSE estimator

- Optimal Bayesian estimators:
  - In general, difficult to determine in closed form.
  - Easy to determine under jointly Gaussian assumptions.
  - MMSE estimator: Generally involves multidimensional integration.
  - MAP estimator: Generally involves multidimensional maximization.
- Proposition: Constrain the estimator to be linear i.e.,

$$\hat{\theta} = \sum_{n=0}^{N-1} a_n x[n] + a_N$$

and choose the weighting coefficients a_n's to minimize

$$\text{Bmse}(\hat{\theta}) = \mathbb{E}\left[(\theta - \hat{\theta})^2\right].$$

LMMSE estimator: Solution (1)

- Solve for a_N: Substituting for $\hat{\theta}$ in the Bmse expression and differentiating

$$\frac{\partial}{\partial a_N}\mathbb{E}\left[\left(\theta - \sum_{n=0}^{N-1} a_n x[n] - a_N\right)^2\right] = -2\mathbb{E}\left[\theta - \sum_{n=0}^{N-1} a_n x[n] - a_N\right]$$

which on setting to 0 yields

$$a_N = \mathbb{E}(\theta) - \sum_{n=0}^{N-1} a_n x[n]$$

- Subsequently,

$$\begin{aligned}
\text{Bmse}(\hat{\theta}) &= \mathbb{E}\left(\left[\sum_{n=0}^{N-1} a_n(x[n] - \mathbb{E}(\theta)) - (\theta - \mathbb{E}(\theta))\right]^2\right) \\
&= \mathbb{E}\left(\left[\mathbf{a}^T(\mathbf{x} - \mathbb{E}(\mathbf{x})) - (\theta - \mathbb{E}(\theta))\right]^2\right) \\
&= \mathbf{a}^T\mathbf{C}_{xx}\mathbf{a} - \mathbf{a}^T\mathbf{C}_{x\theta} - \mathbf{C}_{\theta x}\mathbf{a} + C_{\theta\theta}
\end{aligned}$$

where $\mathbf{a} = [a_0, a_1, \ldots, a_{N-1}]$ are the unknown parameters.

LMMSE estimator: Solution (2)

- Taking the partial derivative of Bmse,

$$\begin{aligned}\frac{\partial \text{Bmse}(\hat{\theta})}{\partial \mathbf{a}} &= \frac{\partial}{\partial \mathbf{a}}\left[\mathbf{a}^T\mathbf{C}_{xx}\mathbf{a} - \mathbf{a}^T\mathbf{C}_{x\theta} - \mathbf{C}_{\theta x}\mathbf{a} + C_{\theta\theta}\right] \\ &= 2\mathbf{C}_{xx}^{-1}\mathbf{a} - 2\mathbf{C}_{x\theta}\end{aligned}$$

and setting to zero, we have

$$\mathbf{a} = \mathbf{C}_{xx}^{-1}\mathbf{C}_{x\theta}$$

- Finally, the LMMSE estimator is

$$\begin{aligned}\hat{\theta} &= \mathbf{a}^T\mathbf{x} + a_N = (\mathbf{C}_{xx}^{-1}\mathbf{C}_{x\theta})^T\mathbf{x} + \mathbb{E}(\theta) - (\mathbf{C}_{xx}^{-1}\mathbf{C}_{x\theta})^T\mathbb{E}(\mathbf{x}) \\ &= \mathbb{E}(\theta) + \mathbf{C}_{\theta x}\mathbf{C}_{xx}^{-1}(\mathbf{x} - \mathbb{E}(\mathbf{x}))\end{aligned}$$

and the corresponding Bmse is

$$Bmse(\hat{\theta}) = C_{\theta\theta} - \mathbf{C}_{\theta x}\mathbf{C}_{xx}^{-1}\mathbf{C}_{x\theta}$$

LMMSE estimator: Example

- Consider the estimation of A

$$x[n] = A + w[n], \quad n = 0, \cdots, N-1, \quad w[n] \sim \mathcal{N}(0, \sigma^2) \quad A \sim U(-A_0, A_0)$$

- Recollect the expression for LMMSE estimator:

$$\begin{aligned} \hat{A} &= \mathbb{E}(A) + \mathbf{C}_{Ax}\mathbf{C}_{xx}^{-1}(\mathbf{x} - \mathbb{E}(\mathbf{x})) \\ &= \mathbf{C}_{Ax}\mathbf{C}_{xx}^{-1}\mathbf{x} \quad (\text{since } \mathbb{E}(\mathbf{x}) = \mathbb{E}(\mathbf{A}) = 0) \end{aligned}$$

where the covariance matrices are

$$\begin{aligned} \mathbf{C}_{xx} &= \mathbb{E}(\mathbf{x}\mathbf{x}^T) = \mathbb{E}(A^2)\mathbf{1}\mathbf{1}^{\mathbf{T}} + \sigma^{\mathbf{2}}\mathbf{I} = \sigma_{\mathbf{A}}^{\mathbf{2}}\mathbf{1}\mathbf{1}^{\mathbf{T}} + \sigma^{\mathbf{2}}\mathbf{I} \\ \mathbf{C}_{Ax} &= \mathbb{E}(A\mathbf{x}^T) = \mathbb{E}(A^2)\mathbf{1}^{\mathbf{T}} = \sigma_{\mathbf{A}}^{\mathbf{2}}\mathbf{1}^{\mathbf{T}} \end{aligned}$$

- Hence, we have

$$\hat{A} = \mathbf{C}_{Ax}\mathbf{C}_{xx}^{-1}\mathbf{x} = \sigma_A^2\mathbf{1}^{\mathbf{T}}(\sigma_{\mathbf{A}}^{\mathbf{2}}\mathbf{1}\mathbf{1}^{\mathbf{T}} + \sigma^{\mathbf{2}}\mathbf{I})^{-\mathbf{1}}\mathbf{x} = \frac{\sigma_{\mathbf{A}}^{\mathbf{2}}}{\sigma_{\mathbf{A}}^{\mathbf{2}} + \frac{\sigma^{\mathbf{2}}}{\mathbf{N}}}\bar{\mathbf{x}}$$

LMMSE estimator: Properties

- Bayesian Gauss-Markov model:

$$\mathbf{x} = \mathbf{H}\boldsymbol{\theta} + \mathbf{w}$$

with $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \mathbf{C}_w)$ and $\boldsymbol{\theta} \sim \mathcal{N}(\boldsymbol{\mu}_\theta, \mathbf{C}_\theta)$, the LMMSE estimator is

$$\begin{aligned} \hat{\boldsymbol{\theta}} &= \boldsymbol{\mu}_\theta + \mathbf{C}_\theta \mathbf{H}^T (\mathbf{H}\mathbf{C}_\theta \mathbf{H}^T + \mathbf{C}_w)^{-1} (\mathbf{x} - \mathbf{H}\boldsymbol{\mu}_\theta) \\ &= \boldsymbol{\mu}_\theta + (\mathbf{H}^T \mathbf{C}_w^{-1} \mathbf{H} + \mathbf{C}_\theta^{-1})^{-1} \mathbf{H}^T \mathbf{C}_w^{-1} (\mathbf{x} - \mathbf{H}\boldsymbol{\mu}_\theta) \end{aligned}$$

and for $\epsilon = \boldsymbol{\theta} - \hat{\boldsymbol{\theta}}$, the performance of the estimator is

$$\mathbf{C}_\epsilon = \mathbb{E}(\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T) = (\mathbf{C}_\theta^{-1} + \mathbf{H}^T \mathbf{C}_w^{-1} \mathbf{H})^{-1}$$

- LMMSE estimators are identical in form to the MMSE estimator for jointly Gaussian $\mathbf{x}$ and θ
- LMMSE estimators are commutative and additive for affine transformations
- A parameter uncorrelated with the data cannot be linearly estimated by an LMMSE estimator

Summary

Key points:

- MMSE estimator takes the form of the mean/median/mode of the posterior, when expectation of the cost function (Bayes risk) is quadratic, linear or 'hit-or-miss' respectively
- MAP estimator maximizes the aposteriori likelihood function
- MAP is identical to a Bayesian MLE as number of measurements increase
- LMMSE estimators constraint the estimates to be linear in data. They are commutative and additive for affine transformations
- For a Bayesian Gauss-Markov model MMSE, MAP and LMMSE estimators are identical

Next session:

- Wiener and Kalman filters

Assignments

Solve:

- Example 11.4, Problem 11.16, Problem 12.2, 12.3 12.19

Reading:

- Kay-I, Section 12.4: Geometrical interpretations of LMMSE
- Kay-I, Section 11.5: MAP for vector parameters