ECSE 6520: Estimation and Detection Theory

Bayes Estimators

Class Notes - 10

March 6th, 2014

Contents

1 Bayesian Statistical Modeling

- In the Bayesian theory of parameter estimation, the unknown parameter θ is treated as a realization of a *random* variable with its own distribution $f(\theta)$ is called the *prior distribution.*

- A statistical model is specified in terms of the conditional pdf/pmf $f(x|\theta)$ and the prior distribution $f(\theta)$ of θ.

- The prior model is specified by the investigator based on his/her prior knowledge on the uncertainty of θ.

- The idea is to combine the data likelihood function $f(x|\theta)$ with the prior knowledge $f(\theta)$ to convert prior distribution into a distribution informed by the data likelihood, i.e, the *posterior distribution* and use this distribution for inference.

- Using Bayes rule, we can express the posterior probability of θ as follows:

$$\begin{aligned} f(\theta|x) &= \frac{f(x|\theta)f(\theta)}{f(x)} \\ &= \frac{f(x|\theta)f(\theta)}{\int f(x|\acute{\theta})f(\acute{\theta})d\acute{\theta}} \end{aligned}$$

- The prior distribution represents the uncertainty in θ before x is observed and the posterior distribution reflects our uncertainty in θ after x is observed.

- **Conjugate Priors -** A class of prior probability distributions $f(\theta)$ is said to be conjugate to a class of likelihood functions $f(x|\theta)$ if the resulting posterior distributions $f(\theta|x)$ are in the same family as $f(\theta)$.

- **Sufficiency and Bayesian Inference -** If $T = \tau(X)$ is a sufficient statistic for θ and $\tau(x_1) = \tau(x_2)$ for the observations x_1 and x_2, then x_1 and x_2 lead to the same Bayesian inference for θ.

2 Bayesian Estimation

- The objective is to estimate a specific value of θ given a set of observations based on the a posteriori model $f(\theta|x)$.

- Main ingredient of the Bayesian estimation is the "cost", "risk" or "loss" function

  $$c(\hat{\theta}(x), \theta).$$

  The cost function represents the investigators view of "loss" when θ is declared as $\hat{\theta}(x)$ for a given $X = x$.

- The optimum estimator in the Bayesian sense is the one that minimizes the expected cost, known as the Bayes risk

  $$R(\hat{\theta}) := E[c(\hat{\theta}(X), \theta)].$$

- Note that the expectation is with respect to both X and θ

  $$\begin{aligned} R(\hat{\theta}) &= \int c(\hat{\theta}(X), \theta) f(x, \theta) dx d\theta \\ &= \int c(\hat{\theta}(X), \theta) f(x|\theta) f(\theta) dx d\theta. \end{aligned}$$

- Minimizing the Bayes risk gives the Bayesian estimator:

  $$\hat{\theta} = \arg\min_{\phi} R(\phi).$$

- The optimal estimator can be expressed solely by the cost function $c(\hat{\theta}(x), \theta)$ and the posterior probability $f(\theta|x)$.

  – Note that

  $$\begin{aligned} R(\hat{\theta}) &= E[c(\hat{\theta}(X), \theta)] \\ &= E_X[E_{\theta|X}[c(\hat{\theta}(X), \theta)|X = x]]. \end{aligned}$$

- Thus to minimize $R(\hat{\theta})$, $E_{\theta|X}[c(\hat{\theta}(X), \theta)|X = x]$ must be minimized.

- Thus, Bayesian estimator can be reexpressed as

$$\hat{\theta} = \arg\min_{\phi} E_{\theta|X}[c(\phi, \theta)|X = x].$$

- This is called posterior expected loss and it depends on only posterior density and the loss.

2.1 Bayesian Cost Functions

Some commonly used cost functions are

- Squared error:

$$c(\hat{\theta}, \theta) = (\hat{\theta} - \theta)^T(\hat{\theta} - \theta).$$

- Absolute error:

$$c(\hat{\theta}, \theta) = \|\hat{\theta} - \theta\|^2_{L^1-norm} = \sum_{i=1}^{p} |\hat{\theta}_i - \theta_i|.$$

  For scalar parameters:

$$c(\hat{\theta}, \theta) = |\hat{\theta} - \theta|.$$

- Uniform error:

$$\begin{aligned} c(\hat{\theta}, \theta) &= \mathcal{I}_{\{\|\hat{\theta}-\theta\| > \epsilon\}} \\ &= \begin{cases} 1 & if \quad |\hat{\theta} - \theta| > \epsilon \\ 0 & otherwise \end{cases} \end{aligned}$$

  where $\epsilon > 0$.

For all three cost function, we can compute the expected cost and determine the Bayes risk:

- Mean Square Error:

$$MSE(\hat{\theta}) = E[(\hat{\theta} - \theta)^T (\hat{\theta} - \theta)].$$

- Mean Absolute Error:

$$MAE(\hat{\theta}) = E[|\hat{\theta} - \theta|].$$

- Error Probability:

$$P_e(\hat{\theta}) = P(\|\hat{\theta} - \theta\| > \epsilon).$$

2.2 Minimum Mean Squared Error Estimation

- We define the Bayesian MSE to be the Bayes risk when the cost function is the squared error.

$$MSE(\hat{\theta}) = E_{\theta|X}[(\hat{\theta} - \theta)^T (\hat{\theta} - \theta)|X = x].$$

  The estimator that minimizes the $MSE(\hat{\theta})$ is called the minimum mean squared error estimator (MMSE).

- The MMSE estimator is give by the posterior mean

$$\hat{\theta}(x) = E[\theta|X].$$

2.3 Minimum Mean Absolute Error Estimation

Minimum mean absolute error (MMAE) estimator is the conditional median (posterior median) estimator:

$$\hat{\theta} = \texttt{median}_{\theta\in\Theta}\{f(\theta|X)\}.$$

$$\begin{aligned}
\texttt{median}_{\theta\in\Theta}\{f(\theta|X)\} &= \min\{u \ | \int_{-\infty}^{u} f(\theta|X) d\theta = 1/2\} \\
&= \min\{u \ | \int_{-\infty}^{u} f(X|\theta) f(\theta) d\theta = \int_{u}^{\infty} f(X|\theta) f(\theta) d\theta\}.
\end{aligned}$$

2.4 Minimum Mean Uniform Error Estimation

- Minimum mean uniform error (MMUE) estimation uses mean uniform error criterion which only penalizes those errors that exceed a tolerance level $\epsilon > 0$. This penalty is uniform.

- For small ϵ the optimal estimator is the maximum a posteriori (MAP) estimator, which is called the posterior mode estimator:

$$
\begin{aligned}
\hat{\theta} &= \arg\max_{\theta \in \Theta} \{f(\theta|X)\} \\
&= \arg\max_{\theta \in \Theta} \{\frac{f(X|\theta)f(\theta)}{f(X)}\} \\
&= \arg\max_{\theta \in \Theta} \{f(X|\theta)f(\theta)\}.
\end{aligned}
$$

Remarks -

- For all three estimators, the estimate depends on x through the posterior: posterior mean, posterior median and posterior mode.

- If the posterior is continuous, symmetric and unimodal, then the MMSE, MMAE and MMUE estimators are equal.

- The MMSE and MMAE estimators require integrating with respect to $f(\theta|X)$. Often the calculation is intractable. We have to use numerical techniques for integration.

- If the posterior mode can not be determined analytically, then many of the numerical techniques used for MLE can be applied.

3 The Multivariate Gaussian Linear Model

Consider the following model:

$$X = H\theta + \omega$$

where

θ is an unknown $p \times 1$ vector

H is known $N \times p$ matrix

$\theta \sim \mathcal{N}(\mu_\theta, R_\theta)$

$\omega \sim \mathcal{N}(0, R_\omega)$

θ and ω are independent

R_ω, R_θ, and μ_θ are known.

- Then the posterior distribution of $\theta|X$ is Gauss:

$$\theta|X \sim \mathcal{N}(\mu_{\theta|X}, R_{\theta|X})$$

where

$$\begin{aligned}
\mu_{\theta|X} &= \mu_\theta + R_\theta H^T (H R_\theta H^T + R_\omega)^{-1}(x - H\mu_\theta) \\
&= \mu_\theta + (H^T R_\omega^{-1} H + R_\theta^{-1})^{-1} H^T R_\omega^{-1}(x - H\mu_\theta)
\end{aligned}$$

and

$$\begin{aligned}
R_{\theta|X} &= R_\theta - R_\theta H^T (H R_\theta H^T + R_\omega)^{-1} H^T R_\theta \\
&= (H^T R_\omega^{-1} H + R_\theta^{-1})^{-1}.
\end{aligned}$$

3.1 Bayes Estimation

The posterior distribution is Gauss, symmetric and unimodal. Therefore, the optimal Bayes estimator is

$$\hat{\theta}(x) = \mu_{\theta|X} = \mu_\theta + R_\theta H^T (H R_\theta H^T + R_\omega)^{-1}(x - H\mu_\theta)$$

regardless of which optimality criterion we use. (Recall that MMSE, MMAE and MMUE estimators are equivalent in this case.)

Remarks -

- The optimal estimator $\hat{\theta}(x)$ is a linear function of the data x.
- Consider the case where $R_\theta = \sigma^2 I$ and let $\sigma^2 \to \infty$. Then $R_\theta^{-1} \to 0$ and

  $$\begin{aligned} \hat{\theta}(x) &= \mu_\theta + (H^T R_\omega^{-1} H)^{-1} H^T R_\omega^{-1} (x - H\mu_\theta) \\ &= (H^T R_\omega^{-1} H)^{-1} H^T R_\omega^{-1} x \end{aligned}$$

  Note that this is the same as the maximum likelihood estimator and the minimum variance unbiased estimator.

- It suffices to focus on the case where $\mu_\theta = 0$. Then the Bayesian estimator is

  $$\begin{aligned} \mu_{\theta|X} &= R_\theta H^T (H R_\theta H^T + R_\omega)^{-1} x \\ &= (H^T R_\omega^{-1} H + R_\theta^{-1})^{-1} H^T R_\omega^{-1} x \end{aligned}$$

  If $\mu_\theta \neq 0$, we can apply the above estimator to $x - H\mu_\theta$ and add μ_θ to the result.

3.2 Simultenously Diagonalizable Covariance Matrices

Consider the problem of estimating a signal in Gaussian noise

$$x = s + \omega$$

where

x : Observed noisy measurements

s : True signal

ω : Noise

- Thus, in the general linear model introduced above $H = I$ and $\theta = s$. Assuming that

$$s \sim \mathcal{N}(0, R_s)$$

and

$$\omega \sim \mathcal{N}(0, R_\omega)$$

and that s and ω are statistically independent, the Bayesian estimate of s is given by

$$\hat{s} = R_s(R_s + R_\omega)^{-1}x.$$

- Suppose, R_s and R_ω are simultaneously diagonalizable, meaning there is an orthogonal matrix U such that

$$R_s = U\Sigma_s U^T$$

and

$$R_\omega = U\Sigma_\omega U^T$$

with Σ_s and Σ_ω being diagonal.

- Then the estimator becomes

$$\begin{aligned} \hat{s} &= R_s(R_s + R_\omega)^{-1}x \\ &= U \underbrace{\left[\Sigma_s(\Sigma_s + \Sigma_\omega)^{-1}\right]}_{\Sigma} U^T x \end{aligned}$$

where

$$\Sigma = \begin{pmatrix} \frac{\lambda_1^s}{\lambda_1^s+\lambda_1^\omega} & \cdots & 0 \\ \vdots & \frac{\lambda_2^s}{\lambda_2^s+\lambda_2^\omega} & \vdots \\ 0 & \cdots & \frac{\lambda_N^s}{\lambda_N^s+\lambda_N^\omega} \end{pmatrix}$$

- **Observations -**
  - U : Rotation matrix that changes the basis
  - $y = U^T x$: Rotated x vector
  - $z = \Sigma_s y$: Rescaling of y.
  - $s = Uz$: Projection back into the signal space.
  - $U^T s \sim \mathcal{N}(0, U^T R_s U) = \mathcal{N}(0, \Sigma_s)$ and

    $U^T \omega \sim \mathcal{N}(0, U^T R_\omega U) = \mathcal{N}(0, \Sigma_\omega)$
  - $U = [u_1, ..., u_N]$, we have

    $$u_i^T s \sim \mathcal{N}(0, \lambda_i^s)$$

    and

    $$u_i^T \omega \sim \mathcal{N}(0, \lambda_i^\omega).$$
  - $\lambda_i = \frac{\lambda_i^s}{\lambda_i^s+\lambda_i^\omega}$: The proportion of the projection onto u_i that is due to the signal.