9 Bayes Estimation of Random Parameters

Bayes' Estimation for Random Parameters

- In the parameter estimation problem, we have a parameterized statistic model $\{P_\theta\}_{\theta\in\Theta}$, which gives us a conditional probability mass/density function $f(x\mid\theta)$ for each $\theta\in\Theta$.
- In some applications, we not only just know the parameter space Θ, but also the a **prior information** about the distribution of θ specified by PDF/PMF $f(\theta)$. Then by making use of this information, we can compute the **posterior** distribution of θ when observing $x\in\mathcal{X}$ by the Bayes' rule

$$f(\theta\mid x)=\frac{f(x,\theta)}{f(x)}=\frac{f(x\mid\theta)f(\theta)}{f(x)},$$

where $f(x)$ is the marginal PDF determined by marginalization over θ, i.e.,

$$f(x)=\int_\Theta f(x,\theta)d\theta=\int_\Theta f(x\mid\theta)f(\theta)d\theta$$

- Then for an estimator $\hat{\theta}:\mathcal{X}\to\Theta$, the estimate is $\hat{\theta}(x)$ when x is observed. Suppose that the true value of the parameter is θ, then an **estimation cost** $\mathsf{cost}(\hat{\theta}(x),\theta)$ occurs, where cost is a function

$$\mathsf{cost}:\Theta\times\Theta\to[0,\infty)$$

Note that both θ and x are random and therefore, cost is actually a random variable, which is a function of Θ and X. By knowing $f(\theta)$ and $f(x)$, we can computed the expected cost, called the **Bayes' average cost/risk**, associated with an estimator $\hat{\theta}$, which is

$$E(\mathsf{cost}(\hat{\theta},\theta))=\int_\Theta\int_{\mathcal{X}}\mathsf{cost}(\hat{\theta}(x),\theta)f(x,\theta)dxd\theta=\int_\Theta\int_{\mathcal{X}}\mathsf{cost}(\hat{\theta}(x),\theta)f(x\mid\theta)f(\theta)dxd\theta$$

Then the **optimal Bayes estimator** is defined by

$$\hat{\theta}=\arg\min_{\hat{\theta}\in\Theta}E(\mathsf{cost}(\hat{\theta},\theta))$$

- Clearly, the optimal Bayes estimator depends on the specific form of the cost function. Here, we will investigate three most widely used cost functions
  - **squared error:** $\mathsf{cost}(\hat{\theta},\theta)=|\hat{\theta}-\theta|^2$
  - **absolute error:** $\mathsf{cost}(\hat{\theta},\theta)=|\hat{\theta}-\theta|$
  - **uniform error:** $\mathsf{cost}(\hat{\theta},\theta)=I(|\hat{\theta}-\theta|>\epsilon)$

  For each of the above cost functions, its expectation is called
  - **mean squared error (MSE):** $\mathsf{MSE}(\hat{\theta})=E(|\hat{\theta}-\theta|^2)$
  - **mean absolute error (MAE):** $\mathsf{MAE}(\hat{\theta})=E(|\hat{\theta}-\theta|)$
  - **ϵ-error probability:** $P_e(\hat{\theta})=P(|\hat{\theta}-\theta|>\epsilon)$

Bayes' Estimator for Squared Error

- The MSE is the most widespread estimation criterion and arguably the one with the longest history. The optimal minimum mean squared error estimator (MMSEE) is the **conditional mean estimator** (CME) defined as

$$\hat{\theta}_{\text{CME}} = E(\theta \mid X) = \int_{\Theta} \theta f(\theta \mid X) d\theta$$

- The CME has an intuitive mechanical interpretation as the center of mass (1st moment of inertia) of the mass density $f(\theta \mid x)$. The CME corresponds to the posterior average value of the parameter after you have observed the data sample. As we have already seen in the part of conditional expectation, the CME satisfies an orthogonality condition: the Bayes estimator error is orthogonal to any (linear or non-linear) function of the data, i.e.,

$$E((\hat{\theta}_{\text{CME}} - \theta)g(X)) = 0$$

Therefore, we have

$$\begin{aligned} E((\hat{\theta} - \theta)^2) =& E(((\hat{\theta} - \hat{\theta}_{\text{CME}}) - (\theta - \hat{\theta}_{\text{CME}}))^2) \\ =& E((\hat{\theta} - \hat{\theta}_{\text{CME}})^2) + E((\theta - \hat{\theta}_{\text{CME}})^2) - 2E((\hat{\theta} - \hat{\theta}_{\text{CME}})(\theta - \hat{\theta}_{\text{CME}})) \\ =& E((\hat{\theta} - \hat{\theta}_{\text{CME}})^2) + E((\theta - \hat{\theta}_{\text{CME}})^2) \end{aligned}$$

Therefore, $\hat{\theta}_{\text{CME}} = E(\theta \mid X)$ attains the minimum.

Bayes' Estimator for Absolution Error

- We assume that the CDF $F(\theta \mid x)$ is always continuous. Then the minimal mean absolute error estimator (MMAEE) is the **conditional median estimator** (CmE) defined as

$$\hat{\theta}_{\text{CmE}} = \underset{\theta \in \Theta}{\text{median}}\, f(\theta \mid X) \quad \text{such that} \quad \int_{-\infty}^{\hat{\theta}_{\text{CmE}}(x)} f(\theta \mid x) dx = \int_{\hat{\theta}_{\text{CmE}}(x)}^{\infty} f(\theta \mid x) dx = \frac{1}{2}$$

- The median of a density separates the density into two halves of equal mass. The correctness follows from the fact that the CmE also satisfies an orthogonality condition:

$$\begin{aligned} E(\text{sgn}(\hat{\theta}_{\text{CmE}} - \theta)g(X)) &= \int_{\mathcal{X}} \int_{\Theta} \text{sgn}(\hat{\theta}_{\text{CmE}}(x) - \theta)g(x)f(\theta, x)d\theta dx \\ &= \int_{\mathcal{X}} \int_{\Theta} \text{sgn}(\hat{\theta}_{\text{CmE}}(x) - \theta)g(x)f(\theta \mid x)f(x)d\theta dx \\ &= \int_{\mathcal{X}} \underbrace{E(\text{sgn}(\hat{\theta}_{\text{CmE}}(x) - \theta)g(x)f(x) \mid X = x)}_{=0} dx = 0 \end{aligned}$$

Then we have

$$\begin{aligned} \mathsf{MAE}(\hat{\theta}) =& E(|\hat{\theta} - \theta|) = E(|\underbrace{(\theta - \hat{\theta}_{\text{CmE}})}_{a} + \underbrace{(\hat{\theta}_{\text{CmE}} - \hat{\theta})}_{\Delta}|) \\ =& E(|\hat{\theta} - \hat{\theta}_{\text{CmE}}|) + \underbrace{E(\text{sgn}(\theta - \hat{\theta}_{\text{CmE}})\Delta)}_{=0} + E(\text{sgn}(a + \Delta) - \text{sgn}(a))(a + \Delta)) \end{aligned}$$

Therefore, the minimum is attained when $\Delta = 0$, i.e., $\theta = \hat{\theta}_{\text{CmE}}$.

Bayes' Estimator for Uniform Error

- Unlike the MSE or MAE, the MUE penalizes only those errors that exceed a tolerance level $\epsilon > 0$ and this penalty is uniform. For small ϵ the optimal estimator is the maximum a posterior (MAP) estimator, which is also called the posterior mode estimator defined as

$$\hat{\theta}_{\text{MAP}} = \arg\max_{\theta\in\Theta}\{f(\theta \mid X)\} = \arg\max_{\theta\in\Theta}\left\{\frac{f(X \mid \theta)f(\theta)}{f(X)}\right\} = \arg\max_{\theta\in\Theta}\{f(X \mid \theta)f(\theta)\}$$

- The proof is very simple. Assume that ϵ is a small and positive number. The probability that the magnitude estimator error exceeds ϵ is simply expressed as

$$\begin{aligned} P_e(\hat{\theta}) &= 1 - P(|\theta - \hat{\theta}| \leq \epsilon) \\ &= 1 - \int_{\mathcal{X}}\int_{\Theta} \mathbf{1}_{\{(x,\theta):|\theta-\hat{\theta}(x)|\leq\epsilon\}} f(\theta, x)d\theta dx \\ &= 1 - \int_{\mathcal{X}}\int_{\{\theta:|\theta-\hat{\theta}(x)|\leq\epsilon\}} f(\theta \mid x)d\theta f(x)dx \end{aligned}$$

Consider the inner integral (over θ) in the above expression. This is an integral over θ within a window, which we call the length 2ϵ window, centered at $\hat{\theta}$. It should be evident that, if ϵ is sufficiently small, this integral will be maximized by centering the length 2ϵ window at the value of θ that maximizes the integrated $f(\theta \mid x)$. This value is of course the definition of the MAP estimate $\hat{\theta}$.

Comments for Different Bayes' Estimators

- The CmE may not exist for discrete Θ since the median may not be well defined.

- Only the CME requires (often difficult) computation of the normalization factor $f(x)$ in the posterior $f(\theta \mid x) = f(x \mid \theta)f(\theta)/f(x)$.

- When the posterior is continuous, unimodal, and symmetric then each of the above estimators are identical.

- Each of these estimators depends on X only through posterior $f(\theta \mid X)$.

- If $T = T(X)$ is a sufficient statistic, then the posterior depends on x only through T. Suppose $f(X \mid \theta) = g(T(X), \theta)h(X)$. We have

$$f(\theta \mid X) = \frac{f(X \mid \theta)f(\theta)}{\int_{\Theta} f(X \mid \theta)f(\theta)d\theta} = \frac{g(T, \theta)f(\theta)}{\int_{\Theta} g(T, \theta)f(\theta)d\theta} = f(\theta \mid T)$$

- The CME has the following linearity property. For any random parameter variables θ_1 and θ_2, we have

$$\hat{\theta}_{1+2} = E(\theta_1 + \theta_2 \mid X) = E(\theta_1 \mid X) + E(\theta_2 \mid X) = \hat{\theta}_1 + \hat{\theta}_2$$

This property is not shared by the CmE or the MAP estimator. (find counter example for each as a homework)

Example: Web Link Latency Estimation

- A networked computer terminal takes a random amount of time to connect to another terminal after sending a connection request at time $t = 0$. You, the user, wish to schedule a transaction with a potential client as soon as possible after sending the request. However, if your machine does not connect within the scheduled time then your client will go elsewhere. If one assumes that the connection delay is a random variable X that is uniformly distributed over the time interval $[0, \theta]$ you can assure your client that the delay will not exceed θ. The problem is that you do not know θ so it must be estimated from past experience, e.g., the sequence of previously observed connection delays $X_1, \ldots, X_n$. By assuming a prior distribution on θ an optimal estimate can be obtained using the theory developed above.

- Now let's formulate this in our language of estimation theory. We assume that, given θ, observations $X_1, \ldots, X_n$ are i.i.d. uniform samples each with conditional density

$$f(x_i \mid \theta) = \frac{1}{\theta}\mathbf{1}_{[0,\theta]}(x_i)$$

Let's say that based on your experience with lots of different clients you determine that a reasonable prior on θ is

$$f(\theta) = \theta e^{-\theta}, \theta > 0$$

We will derive the CME, CmE, and MAP estimators of θ as follows.

- First, we find the posterior

$$f(\theta \mid x) = \frac{f(x \mid \theta) f(\theta)}{f(x)}$$

Specifically, we have

$$f(x \mid \theta) f(\theta) = \left(\prod_{i=1}^{n} \frac{1}{\theta}\mathbf{1}_{[x_i,\infty)}(\theta)\right)\left(\theta e^{-\theta}\right) = \frac{e^{-\theta}}{\theta^{n-1}}\mathbf{1}_{[x_{(1)},\infty)}(\theta)$$

where $x_{(1)} = \max\{x_i\}$ and function $\frac{e^{-\theta}}{\theta^{n-1}}$ is monotone decreasing. Also, we have

$$f(x) = \int_0^\infty f(x \mid \theta) f(\theta) d\theta = \int_0^\infty e^{-\theta}\theta^{-n+1}\mathbf{1}_{[x_{(1)},\infty)}(\theta) d\theta = \int_{x_{(1)}}^\infty e^{-\theta}\theta^{-n+1} d\theta = q_{-n+1}(x_{(1)})$$

where $q_n(x) := \int_x^\infty \theta^n e^{-\theta} d\theta$ is the incomplete Euler function, which is monotone decreasing and has a recursive formula $q_{-n-1}(x) = \frac{1}{n}(\frac{1}{x^n}e^{-x} - q_{-n}(x)), n = 0, -1, -2, \ldots$.

- Then we find the optimal estimator functions

$$\hat{\theta}_{\text{MAP}} = X_{(1)}$$
$$\hat{\theta}_{\text{CME}} = \frac{q_{-n+2}(X_{(1)})}{q_{-n+1}(X_{(1)})}$$
$$\hat{\theta}_{\text{CmE}} = q_{-n+1}^{-1}(0.5 q_{-n+1}(X_{(1)}))$$

Note that only the MAP estimator is a simple function of X while the two others require more difficult computation of integrals q_n and/or an inverse function q_n^{-1}.

Example: Estimation of Gaussian Amplitude

- A very common assumption arising in many signal extraction problems is the assumption of a Gaussian distributed signal observed in additive Gaussian noise. For example, a radar target acquisition system might transmit a pulse to probe for possible targets in a cell located at a particular point in space. If a strong reflecting target is present at that point then it reflects some of the energy in the radar pulse back to the radar, resulting in a high energy signal, called a radar return, at the radar receiver. The amplitude of this signal might contain useful information about the identity of the target. Estimation of the radar return is complicated by the presence of ambient noise generated in the radar receiver (thermal noise) or by interference from other sources (clutter) in the cell.

- Formally, we consider two jointly Gaussian variables S and X with

$$f(s,x)=\frac{1}{2\pi\sigma_S\sigma_X\sqrt{1-\rho^2}}e^{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{s-\mu_S}{\sigma_S}\right)^2-2\rho\left(\frac{s-\mu_S}{\sigma_S}\right)\left(\frac{x-\mu_X}{\sigma_X}\right)+\left(\frac{x-\mu_X}{\sigma_X}\right)^2\right]}$$

  i.e., $E(S)=\mu_S, E(X)=\mu_X, var(S)=\sigma_S^2, var(X)=\sigma_X^2$ and $cov(S,X)=\rho\sigma_S\sigma_X$. Specifically, S will play the role of the signal and X will be the measurement. The objective is to find an optimal estimator of S given measured X.

- First, we need to find the posterior density. A fundamental fact about jointly Gaussian random variables is that if you condition on one of the variables then the other variable is also Gaussian, but with different mean and variance equal to its conditional mean and variance. In particular, the conditional density of S given $X=x$ is Gaussian with mean parameter

$$\mu_{S|X}(x)=E(S\mid X=x)=\mu_S+\rho\frac{\sigma_S}{\sigma_X}(x-\mu_X)$$

  and variance parameter

$$\sigma_{S|X}^2=E((S-E(S\mid X))^2\mid X=x)=(1-\rho^2)\sigma_S^2$$

  Therefore, the conditional desity takes the form

$$f_{S|X}(s\mid x)=\frac{f_{X|S}(x\mid s)f_S(s)}{f_X(x)}=\frac{1}{\sqrt{2\pi}\sigma_{S|X}}\exp\left(-\frac{(s-\mu_{S|X}(x))^2}{2\sigma_{S|X}^2}\right)$$

- We immediately note that, as the posterior is continuous, symmetric and unimodal, the MAP, CME, and CmE estimators are of identical form. Bringing out the explicit dependency of the estimator $\hat{S}$ on the observed realization x we have:

$$\hat{S}(x)=\mu_{S|X}(x)$$

- An interesting special case, relevant to the radar example discussed above, is the independent additive noise model where $X=S+V$, where V is an independent Gaussian. For this case, we have $\sigma_X^2=\sigma_S^2+\sigma_V^2, \rho^2=\frac{\sigma_S^2}{\sigma_S^2+\sigma_V^2}$ and therefore

$$\hat{S}(x)=\mu_S+\frac{\sigma_S^2}{\sigma_S^2+\sigma_V^2}(x-\mu_X)$$

Supplementary Materials for Jointly Gaussian

- A random vector $X = (X_1, X_2, \dots, X_n)^{\mathrm{T}}$ is said to be **Gaussian** if any linear combination of itself components of the form $\sum_{i=1}^{n} c_i X_i$ is a Gaussian random variable, where c_i are real numbers. Or equivalently, we call $X_1, X_2 \dots, X_n$ **jointly Gaussian**.

- One can show (by using characteristic functions) that when $X_1, X_2, \dots, X_n$ are jointly Gaussian, their joint PDF is

$$f_{X_1,\dots,X_n}(x_1,\dots,x_n) = \frac{1}{(2\pi)^{\frac{n}{2}}\sqrt{\det(\Sigma_X)}} \exp\left(-\frac{1}{2}(X - E(X))^{\mathrm{T}} C_X^{-1} (X - E(X))\right)$$

where Σ_X is the covariance matrix of $X = (X_1, X_2, \dots, X_n)$, i.e.,

$$\Sigma_X(i,j) = E((X_i - E(X_i))(X_j - E(X_j)))$$

- One important property of jointly Gaussian is that **uncorrelated implies independent**, which is not true is general. Specifically, suppose that $X_1, X_2, \dots, X_n$ are jointly Gaussian and are uncorrelated, i.e.,

$$\Sigma_X = \begin{pmatrix} \sigma_{X_1}^2 & 0 & \cdots & 0 \\ 0 & \sigma_{X_2}^2 & & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_{X_n}^2 \end{pmatrix} \quad \text{and} \quad \Sigma_X^{-1} = \begin{pmatrix} (\sigma_{X_1}^2)^{-1} & 0 & \cdots & 0 \\ 0 & (\sigma_{X_2}^2)^{-1} & & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & (\sigma_{X_n}^2)^{-1} \end{pmatrix}$$

Therefore, we have

$$f_{X_1,\dots,X_n}(x_1,\dots,x_n) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma_{X_i}} \exp\left[-\frac{1}{2}\left(\frac{x_i - E(X_i)}{\sigma_{X_i}}\right)^2\right] = \prod_{i=1}^{n} f_{X_i}(x_i)$$

which means that $X_1, X_2, \dots, X_n$ are actually independent!

- Another important property of jointly Gaussian is that the conditional distribution is still Gaussian. Specifically, if $X_1, X_2, \dots, X_n$ are jointly Gaussian, then

$$f_{X_n|X_1,\dots,X_{n-1}}(X_n \mid X_1, \dots, X_{n-1}) = \frac{f_{X_1,\dots,X_n}(x_1,\dots,x_n)}{f_{X_1,\dots,X_{n-1}}(x_1,\dots,x_{n-1})} \sim N(\hat{X}, \hat{\Sigma})$$

where we have

$$\begin{aligned} \hat{X} =& E(X_n \mid X_1, \dots, X_{n-1}) = E(X_n) + \Sigma_{X_n,\underline{X_{n-1}}} \Sigma_{\underline{X_{n-1}}}^{-1} (\underline{X_{n-1}} - E(\underline{X_{n-1}})) \\ \hat{\Sigma} =& \sigma_{X_n}^2 - \Sigma_{X_n,\underline{X_{n-1}}} \Sigma_{\underline{X_{n-1}}}^{-1} \Sigma_{\underline{X_{n-1}},X_n} \end{aligned}$$

Note that $\underline{X_{n-1}}$ is the data you observed and X_n is the value you want to estimate. The above actually tells us the following important properties for jointly Gaussian

  - $E(X \mid X_1, \dots, X_n)$ is a linear function of the data; and
  - Σ is actually data independent.

Note that both are not correct for the general case. (try to find counter-examples as a homework)