EE126: Probability and Random Processes

Lecture 24: Bayesian Estimation II

Abhay Parekh

UC Berkeley

April 21, 2011

Logistics

- Take-home handed back after class.
- Error in grading problem 2. Please get in touch with Arash if this applies to you.
- HW due Friday; next HW due Wed.

Bayesian Estimation

If we are trying to estimate a parameter Θ and have observations $X = (X_1, X_2, ..., X_n)$:

- The distribution of p_θ (f_Θ) is assumed to be known. This is called the **Prior Distribution**.
- A complete solution to the estimation problem is provided by $p_{\theta|X}$ (or $f_{\Theta|X}$). This is called the **Posterior Distribution**.

Bayesian Estimation of Gaussians

We make n measurements of an unknown quantity θ and make n independent observations:

$$X_i = \theta + W_i, i = 1, 2, ..., n$$

where W_i is noise distributed $N(0, \sigma^2)$ and independent of θ. We start with the prior $\theta \sim N(x_0, \sigma^2)$.
Then $\theta | X_1, ..., X_n \sim N(m, v)$ where

$$m = \frac{x_0 + ... + x_n}{n+1}, \quad v = \frac{\sigma^2}{n+1}$$

In general, if the prior is distributed $N(x_0, \sigma_0^2)$ and $W_i \sim N(0, \sigma_i^2)$ then:
$\theta | X_1, ..., X_n \sim N(m, v)$

$$m = \frac{\sum_{i=0}^{n} s_i x_i}{\sum_{i=0}^{n} s_i}, \quad v = \frac{1}{\sum_{i=0}^{n} s_i}$$

where $s_i = \frac{1}{\sigma_i^2}$:
So the posterior has the same kind of distribution as the prior!

Maximum a Posteriori (MAP) Rule

Suppose we are forced to pick a point estimate instead of a distribution.
Then as we have seen, the Bayesian is going to choose the rule:

$$\hat{\theta} = \text{argmax}_\theta p_{\Theta|x}(\theta|x) \quad (\Theta \text{ discrete})$$

$$\hat{\theta} = \text{argmax}_\theta f_{\Theta|x}(\theta|x) \quad (\Theta \text{ continuous})$$

Example: Suppose there are two possible values: θ_1 and θ_2. Then MAP picks θ_1 iff $p_{\Theta|X)}(\theta_1|x) \geq p_{\Theta|X}(\theta_1|x)$, I.e.,

$$\frac{P(X = x|\theta = \theta_1)}{P(X = x|\theta = \theta_2} \geq \frac{P(\theta_2)}{P(\theta_1)}$$

Notice that $P(X = x)$ always cancels out...

How good is the MAP Rule?

The MAP rule can be pretty good esp. when θ is discrete. Let $g_M(X)$ be the MAP rule.
Suppose there is another procedure that produces an estimate $g_o(X)$ given observations X. Let I_o be a boolean r.v. that is 1 when the procedure is correctly estimates θ and zero o.w.

$$E[I_o|X] = P(g(X) = \theta|X) \leq P(\Theta = \hat{\theta}|X) = E[I_m|X]$$

Iterated expectations:

$$E[I_o] \leq E[I_m]$$

So the MAP rule maximizes the probability of estimating, over all rules. All bets are off when θ is continuous...

Issue with Point Estimators such as MAP

1. Sometimes the maximum of a distribution is not typical.
2. Example: Higher dimensional data: Say θ is a point in $\Re^n$ and the posterior is a joint gaussian of n independent standard normals. Maximum of the distribution is at the origin (the mean). For $n = 1000$, 90% of the probability is in a shell of radius 31.6 and thinkness 2.8.
3. So use with care.

What about picking $E[\Theta|X]$ as a point estimator?
The mean of the distribution might be more typical then the point of maximum density.

$E[\Theta|X]$ minimizes the conditional MSE

How to find the $\hat{\Theta}$ which minimizes $E[(\Theta - \hat{\theta})^2]$, i.e. the mean square error?

$$\begin{aligned} E[(\Theta - \hat{\theta})^2] &= var(\Theta - \hat{\theta}) + (E[\Theta - \hat{\theta}])^2 \\ &= var(\Theta) + (E[\Theta] - \hat{\theta})^2 \end{aligned}$$

So pick $\hat{\theta} = E[\Theta]$ Now suppose we make an observation for random variable Θ, i.e. $X = x$. Then what should our estimate be? Again, we want to minimize mean square error (given $X = x$) so:

$$E[(\Theta - \hat{\theta})^2|X = x] \text{ is minimized at } \hat{\theta} = E[\Theta|X = x]$$

General Result

Suppose $g(X)$ is another decision rule for arriving at some estimate of Θ. Then

$$E[(\Theta - E[\Theta|X = x])^2|X = x] \leq E[(\Theta - g(x))^2|X = x]$$

Since this holds for every value $X = x$:

$$E[(\Theta - E[\Theta|X])^2|X] \leq E[(\Theta - g(x))^2|X]$$

Iterated Expectations:

$$E[(\Theta - E[\Theta|X])^2] \leq E[(\Theta - g(x))^2]$$

So $E[\Theta|X]$ is the estimator with the lowest unconditional mean-square error.

Some other Properties of $E[\Theta|X]$

The estimation error is $\bar{\Theta} = \hat{\theta} - \theta$. Let $\hat{\theta} = E[\Theta|X]$:

1. $E[\bar{\theta}] = 0$ Error is unbiased
2. $\mathrm{cov}(\hat{\theta}, \bar{\theta}) = 0$: error is uncorrelated with the estimate.
3. $var(\Theta) = var(\hat{\theta}) + \mathrm{var}(\bar{\theta})$

Example

$X = \Theta + W$ where W is distributed uniformly on $[-1, 1]$ and Θ is uniform on $[4, 10]$. Then $X|\Theta = \theta$ is uniform on $[\theta - 1, \theta + 1]$. Then, $f_{\Theta,X}(\theta, x) = \frac{1}{2}\frac{1}{6} = \frac{1}{12}$. Find $E[\Theta|X = x]$.

Example

$X = \Theta + W$ where W is distributed uniformly on $[-1, 1]$ and Θ is uniform on $[4, 10]$. Then $X|\Theta = \theta$ is uniform on $[\theta - 1, \theta + 1]$. Then, $f_{\Theta,X}(\theta, x) = \frac{1}{2}\frac{1}{6} = \frac{1}{12}$. Find MSE $= var(\Theta|X = x)$.

Alice and Bob revisited

Recall: Bob is late by an amount randomly distributed on $[0, \Theta]$, Θ uniform on $[0, 1]$. If Bob is late by x, we saw that

$$f_{\Theta|X}(\theta|x) = \frac{c}{\theta}$$

for $\theta \in [x, 1]$ and zero otherwise. So the MAP rule picks $\theta_M = x$. The LMS estimate is

$$E[\Theta|X = x] = \int_x^1 \theta \frac{c}{\theta} d\theta = c(1 - x)$$

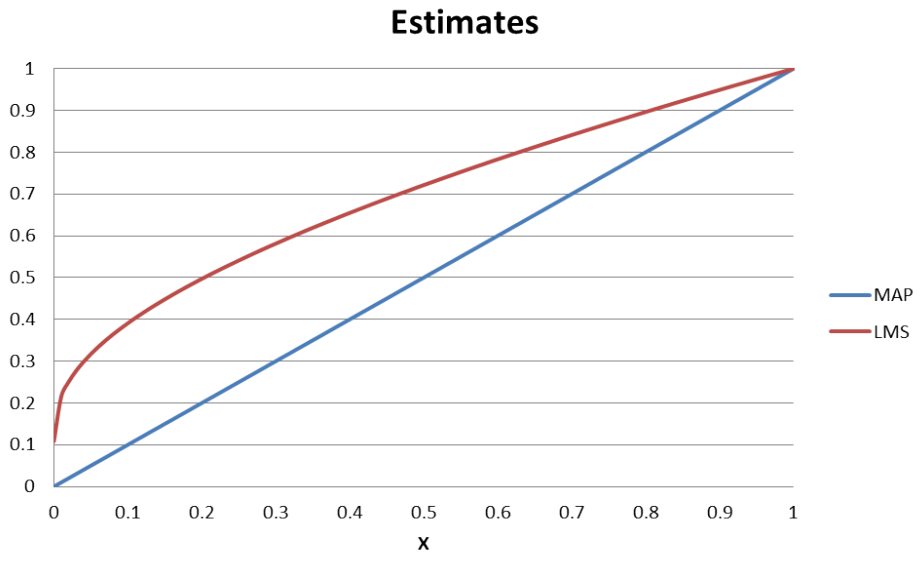
Alice and Bob revisited

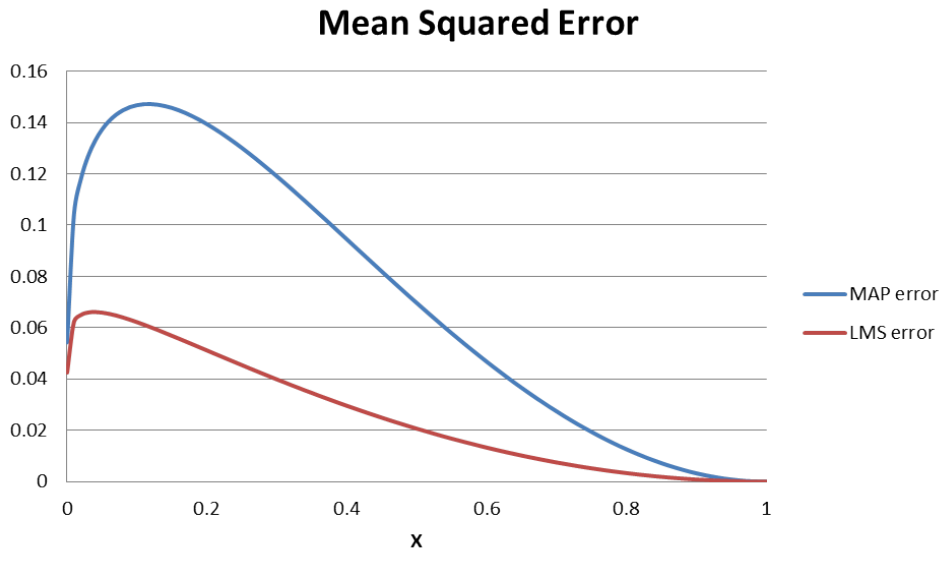
What is the MSE? If the estimate is $\hat{\theta}$:

$$\begin{aligned} E[(\hat{\theta}-\theta)^2|X=x] &= \int_x^1 (\hat{\theta}-\theta)^2 \frac{c}{\theta} d\theta \\ &= \int_x^1 (\hat{\theta}^2 - 2\hat{\theta}\theta + \theta^2)\frac{c}{\theta} d\theta \\ &= \hat{\theta}c\int_x^1 \frac{1}{\theta} d\theta - 2\hat{\theta}c\int_x^1 2d\theta + c\int_x^1 \theta d\theta \\ &= \hat{\theta}c|lnx| - 2c\hat{\theta}(1-x) + \frac{c}{2}(1-x^2) \end{aligned}$$

Substitute $\hat{\theta} = x$ for MAP and $\hat{\theta} = c(1-x)$ in the expression. What do you get?

Comparison between MAP and LMS





Estimating Multiple Parameters

We need to estimate Θ_i, $i = 1, 2, ..., k$. What to do?
Minimize the sum of the individual MSE's:

$$E[\Theta_1 - \hat{\theta}_1] + \cdots + E[\Theta_k - \hat{\theta}_k]$$

Solve k decoupled MSE problems: $\hat{\theta}_i = E[\Theta_i|X]$.

MSE and multiple observations

Observations are $X_1, ..., X_n$: MSE is minimized by $E[\Theta|X_1, ..., X_n]$.
Calculations are frequently very cumbersome.

1. Need joint distribution $f_{\Theta,X_1,....,X_n}$.
2. Expectation integral can be quite tricky

Need a computationally easier method that is still pretty good.

Bayes Linear Least Squares Estimate

Find the estimator that minimizes MSE but that is of the form

$$\hat{\theta} = a_1 X_1 + a_2 X_2 + \cdots + a_n X_n + b$$

This estimate is linear in the observations and is called the Bayesian Linear Least Squared Estimate.
We want to choose $a_1, a_2, .., a_n, b$ which minimize:

$$E[(\Theta - (\sum_i a_i X_i) - b)^2]$$

Calculations involve covariances between pairs of random variables and are manageable.

Bayes Linear LSE: One observation

Choose a, b which minimize: $E[(\Theta - aX - b)^2]$.
Let's focus on b first. For any given value of a, The MSE of the r.v. $\Theta - aX$ is minimized by $E[\theta - aX]$. Thus set

$$b = E[\theta - aX]$$

.Substitute this value of b to find the best a:

$$E[(\Theta - a_1X - b)^2] = E[(\Theta - a_1X - E[\theta - aX])^2] = var(\Theta - a_1X)$$

$$var(\Theta - a_1X) = \sigma_\Theta^2 + a_1^2\sigma_X^2 + 2cov(\Theta, -a_1X) = \sigma_\Theta^2 + a_1^2\sigma_X^2 - 2acov(\Theta, X)$$

Take derivative wrt a and set to zero.

$$2a\sigma_X^2 = 2cov(\Theta, X) \Rightarrow a = \frac{cov(\Theta, X)}{\sigma_X^2}.$$

Rewriting using $cov(\Theta, X) = \rho\sigma_\Theta\sigma_X$:

$$a = \rho\frac{\sigma_\theta}{\sigma_X}$$

Bayes Linear LSE: One observation

What about the MSE?
Since $b = E[\theta - aX]$,

$$E[\Theta - \hat{\theta}] = E[\Theta] - aE[X] - E[\Theta] + aE[X] = 0.$$

So MSE is the same as $var(\theta - \hat{\theta})$.

$$var(\Theta - \hat{\Theta}) = \sigma_\Theta^2 a^2 \sigma_X^2 - 2acov(\Theta, X)$$

Substituting: MSE $= (1 - \rho^2)\sigma_\Theta^2$.

Bayes LLSE with one observation

- Given an observation X:

$$\hat{\Theta} = E[\Theta] + \frac{cov(X,\Theta)}{\sigma_X^2}(X - E[X]) = E[\Theta] + \rho\frac{\sigma_\Theta}{\sigma_X}(X - E[X])$$

where

$$\rho_{\Theta,X} = \frac{cov(\Theta,X)}{\sigma_X\sigma_\Theta}$$

- The MSE is

$$(1-\rho^2)\sigma_\Theta^2.$$

Alice and Bob Revisited

Recall: Bob is late by an amount randomly distributed on $[0, \Theta]$, Θ uniform on $[0, 1]$. What is the Bayes LLSE for X, the amount he is late? Need to find: σ_X, σ_Θ, $cov(X, \Theta)$.
Deal with X first:
Since $E[X|\Theta] = \Theta/2$, by iterated expectations: $E[X] = \frac{1}{4}$.
From the Law of Total Variance:

$$var(X) = E[var(X|\Theta)] + var(E[X|\Theta)$$

Also, $var(X|\Theta) = \frac{\Theta^2}{12}$.

$$E[var(X|\Theta)] = \int_0^1 \frac{\theta^2}{12} d\theta = \frac{1}{36}$$

and

$$var(E[X|\Theta]) = var(\frac{\Theta}{2}) = \frac{1}{4}\frac{1}{12} = \frac{1}{48}$$

$$var(X) = \frac{7}{144}$$

Alice and Bob Revisited

Recall: Bob is late by an amount randomly distributed on $[0, \Theta]$, Θ uniform on $[0, 1]$. What is the Bayes LLSE for X, the amount he is late?

Next deal with Θ: $E[\Theta] = 1/2, E[\Theta^2] = 1/12 + 1/4 = 1/4$Finally, $cov(\Theta, X) = E[\Theta X] - E[X]E[\Theta]$:

$$E[\Theta X] = E[E[\Theta X]|\Theta] = E[\frac{\Theta^2}{2}] = \frac{1}{6}.$$

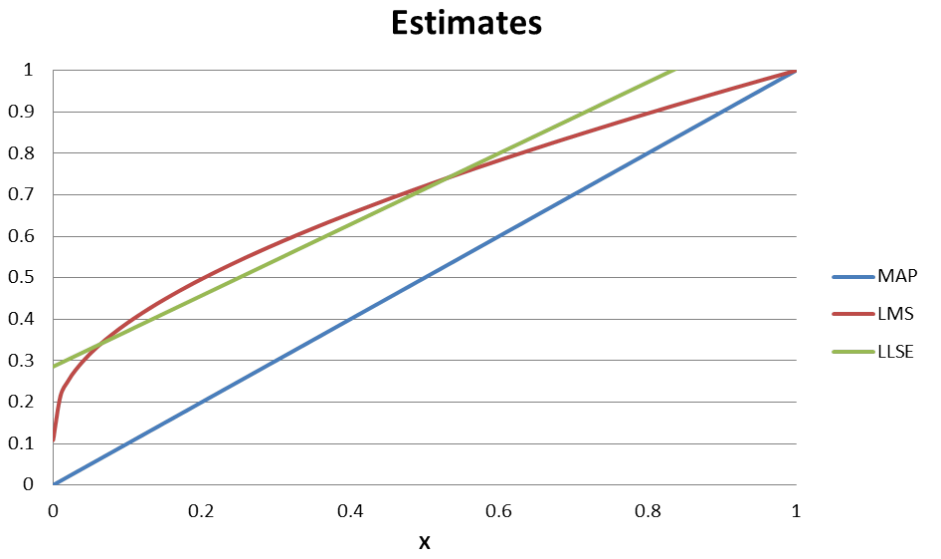
$$cov(X, \Theta) = \frac{1}{6} - \frac{1}{4}\frac{1}{2} = \frac{1}{24}$$

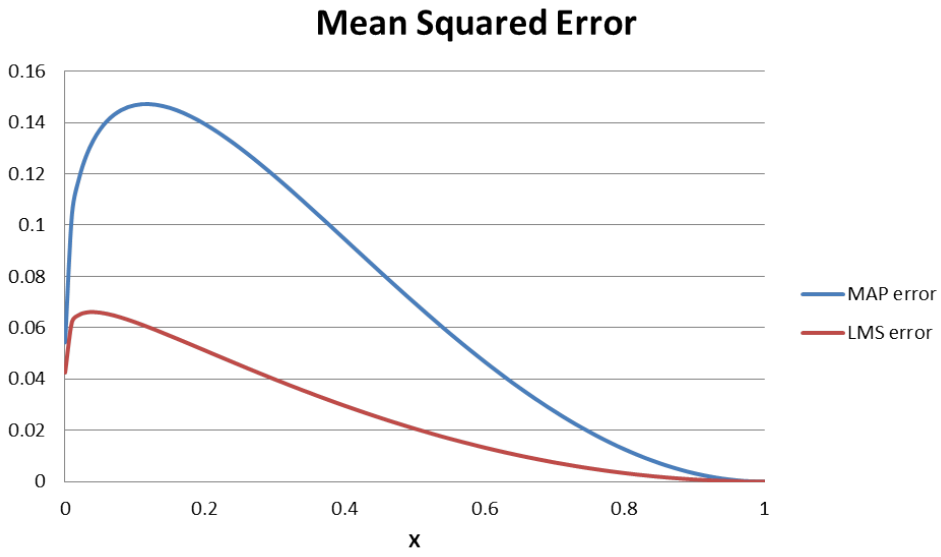
So:

$$\hat{\Theta} = E[\Theta] + \frac{cov(X, \Theta)}{\sigma_X^2}(X - E[X]) = \frac{1}{2} + \frac{1/24}{7/144}(X - \frac{1}{4})$$

$$\hat{\Theta} = \frac{6}{7}X + \frac{2}{7}$$

Comparing Estimates and MSE





Multiple Observations

Suppose we want to estimate Θ and observe $X_1, ..., X_m$:

$$X_i = \Theta + W_i$$

where the W_i and Θ are uncorrelated. The prior for Θ has mean μ and variance σ_0^2. Then we want to minimize:

$$E[(\Theta - (\sum_i a_i X_i) - b)^2]$$

By taking partial derivatives, setting to zero and solving for unknowns:

$$\hat{\Theta} = \frac{\mu/\sigma_0^2 + \sum_{i=1}^m X_i/\sigma_i^2}{\sum_{i=0}^m 1/\sigma_i^2}$$

If all the $\sigma = \sigma_i$ for $i = 0, 1, 2, ..., m$:

$$\hat{\Theta} = \frac{\mu + x_1 + \cdots + x_m}{m+1}$$

Gaussian Interpretation of Bayes LLSE

The LLSE is expressed in terms of means and (co) variances. So is a Gaussian rv.

We saw in the last lecture that for the problem of estimating Θ where

$$X_i = \Theta + W_i$$

X_i independent of each other Θ; X_i Gaussian, the posterior is Gaussian if the prior for Θ is Gaussian.

In fact: LMS $=$ LLSE!

One way of thinking about LLSE is that it is the LMS estimate when we assume that all the distributions are Gaussian.

Summary

- Bayesian Estimation makes a lot of sense if you know the priors.
- Even if you do not, you can learn them given a sufficient number of observations.
- Finding the posterior distribution is computationally difficult
- The MAP rule works well for discrete parameters
- The LMS estimator is generally preferred for continuous distributions, but it is computationally hard to calculate
- Linear LSE estimate is easier – assume that everything is Gaussian.