ECE 830 Fall 2010 Statistical Signal Processing

instructor: R. Nowak , **scribe:** K. Surender

Lecture 18: Bayesian Signal Processing

Most signal processing algorithms are designed and based on prior knowledge of signal and noise characteristics. It is therefore natural (and useful) to view them as Bayesian inference strategies. Let's first review the Bayesian set-up:

Prior: $p(\theta)$ - Prior probability distribution on signal parameters

Likelihood: $p(x|\theta)$ - Conditional Distribution of observation x given θ

Posterior: $p(\theta|x) = \frac{p(x,\theta)}{p(x)} = \frac{p(x|\theta)p(\theta)}{\int p(x|\theta)p(\theta)\,d\theta}$ - Posterior probability distribution of θ given x

Example 1 *Let $S(\theta) \in \mathbb{R}^n, \theta \in \mathbb{R}^k$. S is a n-dimensional signal vector determined by $k \leq n$ parameters θ. We observe $X = S(\theta) + W$, $W \sim \mathcal{N}(0, \sigma^2 I)$*

Prior: $p(\theta)$

Likelihood: $X|\theta \sim \mathcal{N}(S(\theta), \sigma^2 I)$

Posterior: $p(\theta|x) \propto p(x|\theta)p(\theta)$

$$p(x|\theta)p(\theta) = \frac{1}{(2\pi\sigma^2)^{n/2}} exp\bigg(-\frac{1}{2\sigma^2}\Big(X - S(\theta)\Big)^T\Big(X - S(\theta)\Big)\bigg)p(\theta)$$

A loss function is formed by taking the negative log of the posterior probability. Note that this loss function is comprised of two θ dependent terms. One term is based on the observed data and the other term is based on the prior probability.

$$-log(p(\theta|x)) = \frac{\| X - S(\theta) \|_2^2}{2\sigma^2} - log(p(\theta)) + constants$$

1 Point Estimators

Usually we are interested in obtaining an estimator of θ given x. Here are the two most common Bayesian estimators.

1.1 Maximum A Posteriori (MAP) Estimator

$$\begin{aligned} \hat{\theta}_{MAP} &= arg\max_{\theta} p(\theta|x) \\ &= arg\max_{\theta} p(x|\theta)p(\theta) \end{aligned}$$

Note if $p(\theta) =$ constant then $\hat{\theta}_{MAP} = \hat{\theta}_{MLE}$, where $\hat{\theta}_{MLE}$ is the maximum likelihood estimate of θ.

1.2 Posterior Mean Estimator

The posterior mean estimator is defined as:

$$\hat{\theta}_{PM} = \mathbb{E}[\theta|x] = \int \theta\, p(\theta|x)\; d\theta$$

Note that $p(\theta|x)$ requires knowledge of the normalizing term $p(x)$. Specifically,

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

where the denominator represents the normalization.

1.3 Posterior Mean and Bayesian MSE

The Bayesian MSE is $\mathbb{E}\Big[\parallel \hat{\theta} - \theta \parallel_2^2\Big]$. The estimator that minimizes this MSE is the posterior mean:

$$\begin{aligned}
\hat{\theta} &= arg\min_{\theta} \mathbb{E}\Big[\parallel \hat{\theta} - \theta \parallel_2^2\Big] \\
&= \min_{\tilde{\theta}} \iint \parallel \tilde{\theta} - \theta \parallel_2^2 p(x,\theta)\, dx\, d\theta \\
&= arg\min_{\tilde{\theta}} \int \parallel \tilde{\theta} - \theta \parallel_2^2 p(\theta|x)\, d\theta
\end{aligned}$$

Now differentiate with respect to $\tilde{\theta}$ and set equal to zero to find the minimizer $\hat{\theta}$

$$\begin{aligned}
\frac{\partial}{\partial\tilde{\theta}} \int \parallel \tilde{\theta} - \theta \parallel_2^2 p(\theta|x)\, d\theta &= \int 2\,(\hat{\theta} - \theta)\, p(\theta|x)\, d\theta = 0 \\
\implies \hat{\theta} \int p(\theta|x)\, d\theta &= \int \theta\, p(\theta|x)\, d\theta \\
\int p(\theta|x)\, d\theta = 1 \text{ therefore: } \hat{\theta} &= \int \theta\, p(\theta|x)\, d\theta \\
&= \hat{\theta}_{PM}
\end{aligned}$$

Example 2 *Coin Tossing. Let $\theta \sim Uniform[0,1]$ be a parameter describing the probability that a coin toss lands heads. Additionally assume we observe $x = 10$ heads in 10 flips. Then the posterior probability is given as $p(\theta|x) = 11\theta^{10}$ and the point estimators of θ are:*

$$\hat{\theta}_{MAP} = 1$$

$$\hat{\theta}_{PM} = \int_0^1 \theta\, p(\theta|x)\, d\theta = 11\int_0^1 \theta^{11}\, d\theta = 11\frac{\theta^{12}}{12}|_0^1 = \frac{11}{12}$$

In the above example it is seen that the MAP estimator is more aggresive than the posterior mean estimator.

2 Linear Minimum MSE (LMMSE) Estimators

Suppose $p(\theta)$ is such that $\int \theta\, p(\theta)\, d\theta = 0$. Also, assume that $\mathbb{E}[x] = 0$. Consider an estimator of the form $\hat{\theta} = A^T X$ where A is a (kxn) matrix. Let's find the A to minimize the Bayesian MSE:

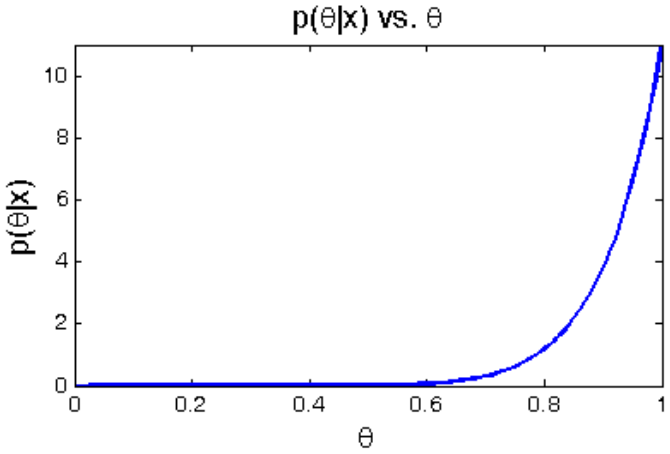

Figure 1: Posterior Probability Function from Example 2.

$$\begin{aligned} MSE(A) &= \mathbb{E}\Big[\| \theta - A^T X \|_2^2\Big] \\ &= \mathbb{E}\Big[tr\Big((\theta - A^T X)(\theta - A^T X)^T\Big)\Big] \\ &= tr\Big(\mathbb{E}\Big[(\theta - A^T X)(\theta - A^T X)^T\Big]\Big) \\ &= tr\Big(\mathbb{E}[\theta\theta^T] - A^T\mathbb{E}[X\theta^T] - \mathbb{E}[\theta X^T]A + A^T\mathbb{E}[XX^T]A\Big) \\ &= tr\Big(\Sigma_{\theta\theta} - A^T\Sigma_{x\theta} - \Sigma_{\theta x}A + A^T\Sigma_{xx}A\Big) \end{aligned}$$

Now differentiate with respect to A and set equal to zero to find the minimizer $\hat{A}$

$$\frac{\partial}{\partial A}MSE(A) = -2\Sigma_{x\theta} + 2\Sigma_{xx}\hat{A} = 0$$

The following equation is called the Wiener Hopf Equation and provides a solution for $\hat{A}$.

$$\begin{aligned} \Sigma_{x\theta} &= \Sigma_{xx}\hat{A} \\ \implies \hat{A} &= \Sigma_{xx}^{-1}\Sigma_{x\theta} \end{aligned}$$

Therefore the LMMSE Estimator is given by the Wiener Filter and the observed data X

$$\hat{\theta}_{LMMSE} = \Sigma_{x\theta}\Sigma_{xx}^{-1}X$$

$\Sigma_{x\theta}\Sigma_{xx}^{-1}$ is the Wiener Filter

Example 3 *Let $X = H\theta + W$ where $\theta \sim \mathcal{N}(0, \sigma_\theta^2 I)$ and $W \sim \mathcal{N}(0, \sigma_W^2 I)$ are independent. Therefore $X \sim \mathcal{N}(0, \sigma_\theta^2 HH^T + \sigma_W^2 I)$. To find the LMMSE define the covariance matrices:*

$$\Sigma_{xx} = \sigma_\theta^2 HH^T + \sigma_W^2 I$$

$$\Sigma_{x\theta} = \mathbb{E}[X\theta^T] = \mathbb{E}[(H\theta + W)\theta^T] = H\Sigma_{\theta\theta} = \sigma_\theta^2 H$$

Now the LMMSE can be found as $\hat{A}^T X$.

$$\hat{A} = (\sigma_\theta^2 HH^T + \sigma_W^2 I)^{-1}\sigma_\theta^2 H$$

$$\hat{\theta}_{LMMSE} = \sigma_\theta^2 H^T(\sigma_\theta^2 HH^T + \sigma_W^2 I)^{-1}X = H^T(HH^T + \frac{\sigma_W^2}{\sigma_\theta^2}I)^{-1}X$$

Note that as SNR increases the LMMSE goes to the MLE. That is:

$$As\ \frac{\sigma_\theta^2}{\sigma_W^2} \to \infty \quad \hat{\theta}_{LMMSE} \to (H^TH)^{-1}H^TX = \hat{\theta}_{MLE}$$

We can rewrite the LMMSE by letting H be a matrix of k orthonormal column vectors $H = \begin{bmatrix} h_1 \mid h_2 \mid \cdots \mid h_k \end{bmatrix}$ *and U an orthonormal basis for* $\mathbb{R}^N$ $U = \begin{bmatrix} h_1 \mid h_2 \mid \cdots \mid h_k \mid \tilde{h}_{k+1} \mid \cdots \mid \tilde{h}_n \end{bmatrix}$.

$$\begin{aligned} \hat{\theta}_{LMMSE} &= \sigma_\theta^2 H^T(\sigma_\theta^2 U \begin{bmatrix} I_{k\times k} & 0_{k\times n-k} \\ 0_{n-k\times k} & 0_{n-k\times n-k} \end{bmatrix} U^T + \sigma_W^2 UU^T)^{-1}X \\ &= \sigma_\theta^2 H^T\Big(U(\sigma_\theta^2 \begin{bmatrix} I_{k\times k} & 0_{k\times n-k} \\ 0_{n-k\times k} & 0_{n-k\times n-k} \end{bmatrix} + \sigma_W^2)U^T\Big)^{-1}X = \sigma_\theta^2 H^T U D^{-1} U^T X \end{aligned}$$

Note that:

$$H^TU = \begin{bmatrix} I_{k\times k} & 0_{k\times(n-k)} \end{bmatrix} \quad D^{-1} = \begin{bmatrix} \frac{1}{(\sigma_\theta^2+\sigma_W^2)} & & & & & \\ & \ddots & & & & \\ & & \frac{1}{(\sigma_\theta^2+\sigma_W^2)} & & & \\ & & & \frac{1}{\sigma_W^2} & & \\ & & & & \ddots & \\ & & & & & \frac{1}{\sigma_W^2} \end{bmatrix}$$

Then:

$$H^TUD^{-1}U^T = \frac{1}{\sigma_\theta^2+\sigma_W^2}\begin{bmatrix} I_{k\times k} & 0_{k\times n} \end{bmatrix}U^T = \frac{1}{\sigma_\theta^2+\sigma_W^2}H^T$$

Now the LMMSE can be written as the MLE multiplied by a shrinkage term $\frac{\sigma_\theta^2}{\sigma_\theta^2+\sigma_W^2}$ *by noting that in this case* $H^TH = I$.

$$\hat{\theta}_{LMMSE} = \frac{\sigma_\theta^2}{\sigma_\theta^2+\sigma_W^2}(H^TH)^{-1}H^TX$$