ECE 830 Fall 2011 Statistical Signal Processing

instructor: R. Nowak

Lecture 19: 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: $p(x|\theta)$ - distribution of x given θ viewed as function of 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 (e.g., $S(\theta) = H\theta$ with H a known $n \times k$ matrix). 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: $p(\theta|x) \propto p(x|\theta)p(\theta)$

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

A loss function is obtained 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.

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

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

$$\begin{aligned}
\hat{\theta}_{\text{MAP}} &= \arg\max_\theta p(\theta|x) \\
&= \arg\max_\theta p(x|\theta)p(\theta) \\
&= \arg\max_\theta \left(\log p(x|\theta) \;+\; \log p(\theta)\right)
\end{aligned}$$

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

1.2 Posterior Mean Estimator

The posterior mean estimator is defined as:

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

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 $\mathbb{E}\Big[\parallel \widehat{\theta} - \theta \parallel_2^2\Big]$. The estimator that minimizes this MSE is the posterior mean. To see this first note

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

Now differentiate with respect to $\widetilde{\theta}$ and set equal to zero to find the minimizer. $\widehat{\theta}$

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

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 is depicted in the Figure 1, below. The point estimators of θ are:*

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

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

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 $\widehat{\theta} = A^T X$ where A is a $k \times n$ matrix. Let's find the A to minimize the Bayesian MSE:

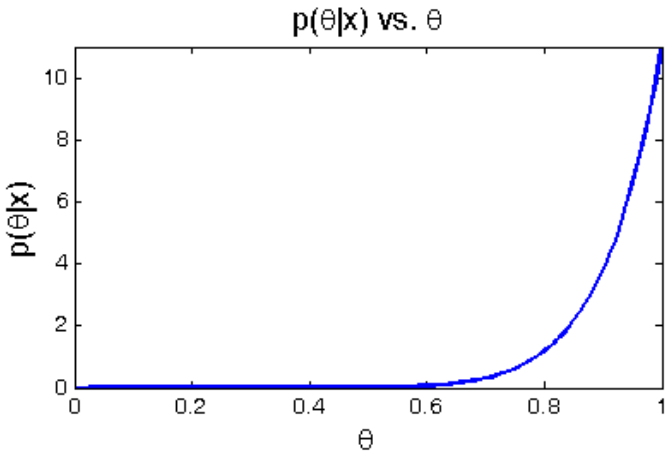

Figure 1: Posterior Probability Function from Example 2.

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

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

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

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

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

The resulting LMMSE estimator is

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

and the matrix $\Sigma_{x\theta}\Sigma_{xx}^{-1}$ is often called 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:*

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

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

The LMMSE is $\widehat{A}^T X$, where

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

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

Note that as SNR increases the LMMSE tends to the MLE. That is,

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

To see this, observe that we can assume wlog that the columns of H are orthonormal. Then we can form a basis for $\mathbb{R}^n$ using these columns and $n-k$ additional orthornomal vectors. Let $U := \left[h_1\ h_2\ \ldots\ h_k\ \widetilde{h}_{k+1}\ \ldots\ \widetilde{h}_n\right]$ The LMMSE can be expressed as

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

where D^{-1} is a diagonal matrix with

$$diag(D^{-1}) = \left[\underbrace{\frac{1}{(\sigma_\theta^2+\sigma_W^2)},\cdots,\frac{1}{(\sigma_\theta^2+\sigma_W^2)}}_{k-times}\ ,\ \underbrace{\frac{1}{\sigma_W^2},\cdots,\frac{1}{\sigma_W^2}}_{(n-k)-times}\right]\ .$$

Note that

$$H^TUD^{-1}U^T \;=\; \frac{1}{\sigma_\theta^2+\sigma_W^2}\left[I_{k\times k}\quad 0_{k\times n}\right]U^T \;=\; \frac{1}{\sigma_\theta^2+\sigma_W^2}H^T\ .$$

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 $H^TH = I$.

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