Topics: Interpolation, decimation, and optimum linear filtering

Spring 2010 Final: Problem 4 (M-estimators and nonlinear filters)

Consider a sequence of N i.i.d. random variables, X_n, each with density

$$p(x|\mu) = \frac{1}{z} \exp\left(-\rho(x-\mu)\right)$$

where μ is a parameter of the distribution, and z is a normalizing constant given by

$$z = \int_{\Re} \exp\left(-\rho(x)\right) dx \ .$$

Then the maximum likelihood estimate of μ is defined as

$$\begin{aligned} \hat{\mu} &= \arg\max_{\mu} \{p(x_1, \cdots, x_N|\mu)\} \\ &= \arg\max_{\mu} \{\log p(x_1, \cdots, x_N|\mu)\} \end{aligned}$$

where $p(x_1, \cdots, x_N|\mu)$ is the joint density for the sequence of random variables $(X_1, \cdots, X_N)$.

a) Derive an expressions for the joint density, $p(x_1, \cdots, x_N|\mu)$, and $\log p(x_1, \cdots, x_N|\mu)$.

Solution:

$$\begin{aligned} p(x_1, \ldots, x_n|\mu) &= \frac{1}{Z^N} \prod_{i=1}^{N} \exp\left\{-\rho(x_i-\mu)\right\} \\ &= \frac{1}{Z^N} \exp\left\{ -\sum_{i=1}^{N} \rho(x_i-\mu) \right\} \end{aligned}$$

$$\log p(x_1, \ldots, x_N|\mu) = -\sum_{i=1}^{N} \rho(x_i-\mu) - N\log(Z)$$

b) Derive a general expression for the maximum likelihood estimate of μ.

Solution:

$$\hat{\mu} = \arg\min_{\mu} \sum_{i=1}^{N} \rho(x_i-\mu)$$

c) Calculate the maximum likelihood estimate of μ when $\rho(x) = x^2$.

Solution:

$$\hat{\mu} = \arg\min_{\mu} \sum_{i=1}^{N} (x_i - \mu)^2$$

$$\frac{d}{d\mu} \sum_{i-1}^{N} (x_i - \mu)^2 = 0$$

$$\sum_{i=1}^{N} 2(x_i - \hat{\mu})(-1) = 0$$

$$\sum_{i=1}^{N} x_i - N\hat{\mu} = 0 \Rightarrow \hat{\mu} = \frac{1}{N} \sum_{i=1}^{N} x_i$$

d) Calculate the maximum likelihood estimate of μ when $\rho(x) = |x|$.

Solution:

$$\frac{d}{d\mu} \sum_{i=1}^{N} |x_i - \mu| = 0$$

$$\sum_{i=1}^{N} \text{sign}(x_i - \hat{\mu}) = 0 \Rightarrow (\text{number of } x_i > \hat{\mu}) = (\text{number of } x_i < \hat{\mu})$$

$$\hat{\mu} = \text{median}(x_1, \ldots, x_N)$$

e) What is the advantage of using $\rho(x) = |x|$ rather than $\rho(x) = x^2$?

Solution:

The function $\rho(x) = |x|$ results in an ML estimate that is less sensitive to outliers, or equivalently more robust.

Spring 2009 Final: Problem 4 (Training for MMSE and LS estimation)

Consider a non-linear prediction problem for which we are trying to predict the value of a scalar Y_n from a vector of observations Z_n. Our assumption is that we can estimate Y_n using the non-linear predictor given by

$$\hat{Y}_n = f(Z_n, \theta)$$

where $\theta \in \Re^p$ is a p dimensional parameter vector that controls the behavior of the nonlinear predictor.

Fortunately, we are given some training data pairs with the form (Y_n, Z_n).[1] The data is partitioned into two sets. The first set, $n \in S_1$, contains $N = |S_1|$ pairs, and is used for training purposes. The second set, $n \in S_2$, contains $M = |S_2|$ pairs, and is used for testing purposes.

Using these data, we can define the training MSE as

$$MSE_1(\theta) = \frac{1}{N} \sum_{n \in S_1} \| Y_n - f(Z_n, \theta) \|^2 \ ,$$

the testing MSE as

$$MSE_2(\theta) = \frac{1}{M} \sum_{n \in S_2} \| Y_n - f(Z_n, \theta) \|^2 \ ,$$

and the expected MSE as

$$MSE_3(\theta) = E\left[\| Y_n - f(Z_n, \theta) \|^2 \right] \ .$$

Based on these error measures, we can define the following two estimates for the parameter vector.

$$\hat{\theta} = \arg\min_{\theta} MSE_1(\theta)$$

$$\theta^* = \arg\min_{\theta} MSE_3(\theta)$$

[1] Assume that each training data pair is independent, and each pairs has the same distribution.

a) Which of the two quantities would you expect to be smaller, $MSE_2(\hat{\theta})$ or $MSE_2(\theta^*)$? Why?

Solution:

$$MSE_2(\theta^*) \text{ is expected to be smaller.}$$

$$E\left[||Y_n - f(Z_n, \theta^*)||^2\right] \leq E\left[Y_n - f(Z_n, \hat{\theta})||^2\right]$$

$$\begin{aligned}
E\left[MSE(\theta^*)\right] &= E\left[\frac{1}{M}\sum_{n\in S_2}||Y_n - f(Z_n, \theta^*)||^2\right] \\
&= \frac{1}{M}\sum_{n\in S_2} E\left[||Y_n - f(Z_n, \theta^*)||^2\right] \\
&= E\left[||Y_n - f(Z_n, \theta^8)||^2\right] \\
&\leq E\left[||Y_n - f(Z_n, \hat{\theta})||^2\right] \\
&= \frac{1}{M}\sum_{n\in S_2} E\left[||Y_n - f(Z_n, \hat{\theta})||^2\right] \\
&= E\left[MSE_2(\hat{\theta})\right]
\end{aligned}$$

$$\text{Therefore, } E\left[MSE_2(\theta^*)\right] \leq E\left[MSE_2(\hat{\theta})\right]$$

b) What is the disadvantage of using $MSE_2(\theta^*)$?

Solution:

In order to use θ^*, we have to know the distribution of Y_n and Z_n, which is usually not available.

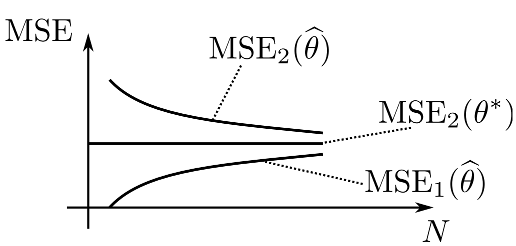
c) Approximately how large should N be in order for $\hat{\theta}$ to be useful?

Solution:

N should be at least larger than p, but the larger, the better.

d) Sketch the plots of $MSE_1(\hat{\theta})$, $MSE_2(\hat{\theta})$, and $MSE_2(\theta^*)$ as a function of the amount of training data N.

Solution:



e) Which value would you expect to be smaller, $MSE_1(\hat{\theta})$ or $MSE_2(\hat{\theta})$. Why?

Solution:

$MSE_1(\hat{\theta})$ is smaller because it is testing on the same dataset as the training data, whereas $MSE_2(\hat{\theta})$ is testing on the testing data using the estimate from the training data.

f) If you are reporting results of your experiment, which value should you report, $MSE_1(\hat{\theta})$ or $MSE_2(\hat{\theta})$. Why?

Solution:

Report $MSE_2(\hat{\theta})$, because in real cases it's meaningless to test on training data. Our goal is to use the estimated $\hat{\theta}$ along with new data to estimate Y_n. The new data will not be the same as the training data.

Spring 2009 Final: Problem 3 (MMSE prediction)

Let $Y \in \Re^N$ be a vector containing the pixels in an image window. We can model Y as

$$Y = tS + W$$

where $t \in \Re^N$ is a deterministic column vector of length N, S is scalar valued Gaussian random variable with mean 0 and variance σ^2, and W is a independent Gaussian random vector of correlated noise with distribution $N(0, R_w)$ where R_w is an $N \times N$ positive definite covariance matrix.

Intuitively, Y is composed of a signal tS obscured by noise W. Our objective is to estimate the value of S from the observations Y. To do this, we will form a MMSE linear estimator for S given by

$$\hat{S} = Y^t \theta$$

where $\theta \in \Re^N$ is a vector of coefficients.

Furthermore, define the covariance matrix of Y given by

$$R_y = E\left[YY^t\right] \ ,$$

and the cross-covariance column vector of Y and S given by

$$b = E\left[YS\right] \ .$$

a) Calculate an expression for the MSE given by $E\left[||S - \hat{S}||^2\right]$ in terms of R_y, b, σ^2, and θ.

Solution:

$$\begin{aligned}
E\left[||S - \hat{S}||^2\right] &= E\left[(S - Y^t\theta)^2\right] \\
&= E\left[S^2 - 2\theta^t Y S + \theta^t Y Y^t \theta\right] \\
&= E\left[S^2\right] - 2\theta^t E\left[YS\right] + \theta^t E\left[YY^t\right]\theta \\
&= \sigma^2 - 2\theta^t b + \theta^t R_y \theta
\end{aligned}$$

b) Use the expression from part a) to compute the value of θ that produces the MMSE estimate of S.

Solution:

$$\frac{\partial}{\partial\theta} E\left[||S - \hat{S}||^2\right] = 2R_y\theta - 2b = 0 \quad \Rightarrow \ \theta = R_y^{-1} b$$

c) Calculate R_y in terms of t, σ^2, and R_w.

Solution:

$$\begin{aligned} R_y &= E\left[YY^t\right] \\ &= E\left[(tS+W)(tS+W)^t\right] \\ &= E\left[S^2tt^t + StW^tSWt^t + WW^t\right] \end{aligned}$$

Since S and W are independent, $E\left[SWt^t\right] = E\left[S\right]E\left[W\right]t^t = 0$ and $E\left[StW^t\right] = tE\left[S\right]E\left[W^t\right] = 0$.

$$R_y = \sigma^2 tt^t + R_w$$

d) Calculate b in terms of t, σ^2, and R_w.

Solution:

$$\begin{aligned} b &= E\left[YS\right] \\ &= E\left[(tS+W)S\right] \\ &= E\left[tS^2 + WS\right] \\ &= tE\left[S^2\right] + E\left[W\right]E\left[S\right] \\ &= \sigma^2 t \end{aligned}$$

e) Use the above results to calculate a closed form expression for $\hat{S}$.

Solution:

$$\hat{S} = Y^t\theta = Y^tR_y^{-1}b = Y^t(\sigma^2tt^t + R_w)^{-1}\sigma^2 t$$

Note that we can show that R_y is positive definite. Remember $R_y = \sigma^2tt^t + R_w$. Since tt^t is positive semi-definite, and R_w is positive definite, we have that R_y is positive definite.

This is helpful in explaining why $\theta = R_y^{-1}b$ makes $E\left[||S-\hat{S}||^2\right]$ minimum.

($\frac{\partial^2}{\partial\theta^2}E\left[||S-\hat{S}||^2\right] = 2R_y$, R_y is positive definite.)

Spring 2002 Final: Problem 3 (2D interpolation)

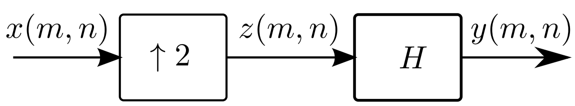
Let the image $y(m,n)$ be formed by applying 2-D interpolation by a factor of $L = 2$ to the signal $x(m,n)$ with an interpolation filter of the form

$$
\begin{aligned}
h(m,n) \quad = \quad & 0.25\delta(m-1,n-1) + 0.5\delta(m,n-1) + 0.25\delta(m+1,n-1) \\
& +0.5\delta(m-1,n) + \delta(m,n) + 0.5\delta(m+1,n) \\
& +0.25\delta(m-1,n+1) + 0.5\delta(m,n+1) + 0.25\delta(m+1,n+1)
\end{aligned}
$$

a) Use a free boundary condition to compute $y(m,n)$ for the input $x(m,n)$ given by

$$
\begin{array}{ccc}
0 & 1 & 1 \\
0 & 1 & 1 \\
0 & 0 & 0
\end{array}
$$

Solution:



$$
z(m,n) = \begin{array}{ccccc}
0 & 0 & 1 & 0 & 1 \\
0 & 0 & 0 & 0 & 0 \\
0 & 0 & 1 & 0 & 1 \\
0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0
\end{array}
$$

$$
y(m,n) = \begin{array}{ccccc}
0 & \frac{1}{2} & 1 & 1 & 1 \\
0 & \frac{1}{2} & 1 & 1 & 1 \\
0 & \frac{1}{2} & 1 & 1 & 1 \\
0 & \frac{1}{4} & \frac{1}{2} & \frac{1}{2} & \frac{1}{2} \\
0 & 0 & 0 & 0 & 0
\end{array}
$$

b) Compute $H(e^{j\mu}, e^{j\nu})$ the DSFT of the filter $h(m,n)$.

Solution:

$$h(m,n) = f(m)f(n), \text{ where } f(m) = 0.5\delta(m-1) + \delta(m) + 0.5\delta(m+1)$$

$$\begin{aligned} F\left(e^{j\omega}\right) &= DTFT\left\{f(m)\right\} \\ &= 1 + 0.5e^{j\omega} + 0.5e^{-j\omega} \\ &= 1 + \frac{e^{j\omega} + e^{-j\omega}}{2} = 1 + \cos(\omega) \end{aligned}$$

$$\begin{aligned} H\left(e^{j\mu}, e^{j\nu}\right) &= DSFT\left\{h(m,n)\right\} \\ &= (1 + cos\mu)(1 + cos\nu) \end{aligned}$$

c) Write an expression for $Y(e^{j\mu}, e^{j\nu})$ in terms of $X(e^{j\mu}, e^{j\nu})$ and $H(e^{j\mu}, e^{j\nu})$.

Solution:

$$Y\left(e^{j\mu}, e^{j\nu}\right) = \sum_{k=0}^{1}\sum_{l=0}^{1} H\left(e^{j(\mu-2\pi k)/2}, e^{j(\nu-2\pi l)/2}\right) X\left(e^{j(\mu-2\pi k)/2}, e^{j(\nu-2\pi l)/2}\right)$$

d) What are the advantages and disadvantages of this interpolation method?

Solution:

Advantages: easy to compute, reduces aliasing compared to pixel replications

Disadvantages: softens image due to attenuation in passband, allows some aliased frequency energy