THU-70250043, Pattern Recognition (Spring 2016) **Homework: 2**

Parameter Estimation Method

Lecturer: Changshui Zhang zcs@mail.tsinghua.edu.cn

Student: Qingfu Wen (2015213495) qingfu.wen@gmail.com

1.

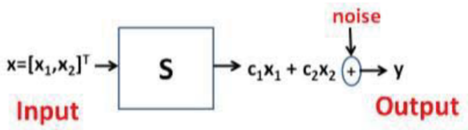

Figure 1: System S

Figure 1 shows a system S which takes two inputs x_1, x_2(which are deterministic) and outputs a linear combination of those two inputs, $c_1x_1 + c_2x_2$, introduces an additive error ϵ which is a random variable following some distribution. Thus the output y that you observe is given by equation (1). Assume that you have $n > 2$ instances $< x_{j1}, x_{j2}, y_j >_{j=1,...,n}$

$$y = c_1x_1 + c_2x_2 + \epsilon \qquad (1)$$

In other words having n equations in your hand is equivalent to having n equations of the following form:$y_j = c_1x_{j1} + c_2x_{j2} + \epsilon_j, j = 1, ..., n$ The goal is to estimate c_1, c_2 from those measurements by maximizing conditional log-likelihood given the input, under different assumptions for the noise. Specifically:

1) Assume that the ϵ_i for $i = 1, ..., n$ are iid Gaussian random variables with zero mean and variance σ^2.

(a) Find the conditional distribution of each y_i given the inputs

SOLUTION:

Since $y_j = c_1x_{j1} + c_2x_{j2} + \epsilon_j, j = 1, ..., n$, $y_j - c_1x_{j1} - c_2x_{j2} \sim N(0, \sigma^2)$.

Thus, $y_j \sim N(c_1x_{j1} + c_2x_{j2}, \sigma^2)$.

(b) Compute the log-likelihood of y given the inputs

SOLUTION:

The PDF of y_j is

$$f(y_j) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y_j - c_1x_{j1} - c_2x_{j2})^2}{2\sigma^2}\right)$$

the log-likelihood function is

$$\begin{aligned}
l(c_1, c_2) &= \ln L(c_1, c_2) \\
&= \ln \prod_{j=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y_j - c_1 x_{j1} - c_2 x_{j2})^2}{2\sigma^2}\right) \\
&= \frac{n \ln(\sqrt{2\pi}\sigma)}{2\sigma^2} \sum_{j=1}^{n} (y_j - c_1 x_{j1} - c_2 x_{j2})^2
\end{aligned}$$

(c) Maximize the likelihood above to get c_{ls}

SOLUTION:

$$c_{ls} = \underset{c}{\operatorname{argmin}} \sum_{j=1}^{n} (y_j - c_1 x_{j1} - c_2 x_{j2})^2$$

for c_1, we get

$$\begin{aligned}
\frac{\partial \sum_{j=1}^{n} (y_j - c_1 x_{j1} - c_2 x_{j2})^2}{\partial c_1} &= 0 \\
2c_1 \sum_{j=1}^{n} x_{j1}^2 - 2 \sum_{j=1}^{n} x_{j1} y_j &= 0 \\
c_1 &= \frac{\sum_{j=1}^{n} x_{j1} y_j}{\sum_{j=1}^{n} x_{j1}^2}
\end{aligned}$$

similarly, we get

$$c_1 = \frac{\sum_{j=1}^{n} x_{j2} y_j}{\sum_{j=1}^{n} x_{j2}^2}$$

Let $y = [y_1, y_2, \cdots, y_n]^T$, X be a $n * 2$ matrix that $X_{ij} = x_{ij}$, $c = [c_1, c_2]^T$. Then

$$c = (X^T X)^{-1} X^T y$$

2) Assume that the ϵ_i for $i = 1, ..., n$ are independent Gaussian random variables with zero mean and variance $Var(\epsilon_i) = \sigma_i$.

(a) Find the conditional distribution of each y_i given the inputs

SOLUTION:

$y_j \sim N(c_1 x_{j1} + c_2 x_{j2}, \sigma_i^2)$.

(b) Compute the log-likelihood of y given the inputs

SOLUTION:

The PDF of y_j is

$$f(y_j) = \frac{1}{\sqrt{2\pi}\sigma_j} \exp\left(-\frac{(y_j - c_1 x_{j1} - c_2 x_{j2})^2}{2\sigma_j^2}\right)$$

the log-likelihood function is

$$
\begin{aligned}
l(c_1, c_2) &= \ln L(c_1, c_2) \\
&= \ln \prod_{j=1}^{n} \frac{1}{\sqrt{2\pi}\sigma_j} \exp\left(-\frac{(y_j - c_1 x_{j1} - c_2 x_{j2})^2}{2\sigma_j^2}\right) \\
&= \sum_{j=1}^{n} \ln(\sqrt{2\pi}\sigma_j) \sum_{j=1}^{n} \frac{(y_j - c_1 x_{j1} - c_2 x_{j2})^2}{2\sigma_j^2}
\end{aligned}
$$

(c) Maximize the likelihood above to get c_{wls}

SOLUTION:

we need to minimize $||W(y - Xc)||$ that W is a diagonal matrix, $W_{ii} = \frac{1}{\sigma_i}$. Similarly,$c_{wls} = (X^T W^T W X)^{-1} X^T W^T W$

3) Assume that the ϵ_i for $i = 1, ..., n$ has density $f_{\epsilon_i}(x) = f(x) = \frac{1}{2b} exp(-\frac{|x|}{b})$. In other words our noise is iid following Laplace distribution with location parameter $\mu = 0$ and scale parameter b.

(a) Find the conditional distribution of each y_i given the inputs

SOLUTION:

Since $y_j = c_1 x_{j1} + c_2 x_{j2} + \epsilon_j, j = 1, ..., n$, $\epsilon_j = y_j - c_1 x_{j1} - c_2 x_{j2}$.

Thus, the density function of y_j is $f(y_j) = \frac{1}{2b} exp(-\frac{|y_j - c_1 x_{j1} - c_2 x_{j2}|}{b})$. y_j is a Laplace distribution with $\mu = c_1 x_{j1} + c_2 x_{j2}$ and scale parameter b.

(b) Compute the log-likelihood of y given the inputs

SOLUTION:

The PDF of y_j is

$$
f(y_j) = \frac{1}{2b} exp(-\frac{|y_j - c_1 x_{j1} - c_2 x_{j2}|}{b})
$$

the log-likelihood function is

$$
\begin{aligned}
l(c_1, c_2) &= \ln L(c_1, c_2) \\
&= \ln \prod_{j=1}^{n} \frac{1}{2b} exp(-\frac{|y_j - c_1 x_{j1} - c_2 x_{j2}|}{b}) \\
&= \frac{n \ln(2b)}{b} \sum_{j=1}^{n} |y_j - c_1 x_{j1} - c_2 x_{j2}|
\end{aligned}
$$

(c) Comment on why this model leads to more robust solution.

2. Consider a normal $p(x) \sim N(\mu, \sigma^2)$ and Parzen-window function $\phi(x) \sim N(0, 1)$ Show that the Parzen-window estimate

$$
p_n(x) = \frac{1}{n h_n} \sum_{i=1}^{n} \phi(\frac{x - x_i}{h_n})
$$

has the following properties:

(a)$\overline{p}_n(x) \sim N(\mu, \sigma^2 + h_n^2)$

SOLUTION:

$$\begin{aligned}
\overline{p}_n(x) &= E[p_n(x)] \\
&= E[\frac{1}{n}\sum_{i=1}^{n}\frac{1}{h_n}\phi(\frac{x-x_i}{h_n})] \\
&= \frac{1}{h_n}E[\phi(\frac{x-x_i}{h_n})] \\
&= \frac{1}{h_n}\int \phi(\frac{x-v}{h_n})p(v)dv \\
&= \frac{1}{h_n}\int h_n\frac{1}{\sqrt{2\pi}h_n}\exp(-\frac{(x-v)^2}{h_n^2}) * \frac{1}{\sqrt{2\pi}\sigma}\exp(-\frac{(v-\mu)^2}{\sigma^2})dv \\
&= N(x;\mu,\sigma^2+h_n^2)
\end{aligned}$$

(b)$Var[p_n(x)] \simeq \frac{1}{2nh_n\sqrt{\pi}}p(x)$

SOLUTION:

$$\begin{aligned}
Var[p_n(x)] &= \frac{1}{n^2}\sum_{i=1}^{n}\left(\frac{1}{h_n^2}E(\phi^2(\frac{x-x_i}{h_n})) - E^2(p(x))\right) \\
&= \frac{1}{n}\left(\int N^2(x;v,h_n^2)N(v;\mu,\sigma^2)dv - E^2(p(x))\right) \\
&= \frac{1}{n}\left(\frac{1}{2h_n\sqrt{\pi}}N(x;\mu,\sigma+\frac{h_n^2}{2}) - \frac{1}{2\sqrt{(\sigma^2+h_n^2)\pi}}N(x;\mu,\frac{\sigma^2+h_n^2}{2})\right)
\end{aligned}$$

When h_n gets small,

$$Var[p_n(x)] \simeq \frac{1}{2nh_n\sqrt{\pi}}p(x)$$

(c)$p(x) - \overline{p}_n(x) \simeq \frac{1}{2}(\frac{h_n}{\sigma})^2[1-(\frac{x-\mu}{\sigma})^2]p(x)$ for small h_n

(Note: if $h_n = \frac{h_1}{\sqrt{n}}$, this show that the error due to bias goes to zero as $1/n$, whereas the standard deviation of the noise only goes to zero as $\sqrt[4]{n}$.)

SOLUTION:

$$\begin{aligned}
p(x) - \overline{p}_n(x) &= N(x;\mu,\sigma^2) - N(x;\mu,sigma^2+h_n^2) \\
&= (1 - \frac{N(x;\mu,\sigma^2)}{N(x;\mu,sigma^2+h_n^2)})N(x;\mu,h_n^2) \\
&= \left(1-\sqrt{\frac{\sigma^2}{\sigma^2+h_n^2}}\exp(\frac{h_n^2(x-\mu)^2}{2(\sigma^2+h_n^2)\sigma^2})\right)p(x)
\end{aligned}$$

From Taylor series, when h_n is very small,

$$\sqrt{\frac{\sigma^2}{\sigma^2+h_n^2}} = \sqrt{1-\frac{h_n^2}{\sigma^2+h_n^2}} \approx 1 - \frac{h_n^2}{2(\sigma^2+h_n^2)}$$

$$\exp(\frac{h_n^2(x-\mu)^2}{2(\sigma^2+h_n^2)\sigma^2}) \approx 1 + \frac{h_n^2(x-\mu)^2}{2(\sigma^2+h_n^2)\sigma^2}$$

Thus

$$
\begin{aligned}
p(x) - \overline{p}_n(x) &\approx \left(1 - (1 - \frac{h_n^2}{2(\sigma^2 + h_n^2)})(1 + \frac{h_n^2(x-\mu)^2}{2(\sigma^2 + h_n^2)\sigma^2})\right) p(x) \\
&\approx \frac{h_n^2}{2\sigma^2}(1 - \frac{(x-\mu)^2}{\mu^2})p(x)
\end{aligned}
$$

3. One measure of the difference between two distributions in the same space is the Kullback-Leibler divergence of Kullback-Leibler "distance":

$$
D_{KL}(p_1(x), p_2(x)) = \int p_1(x) ln \frac{p_1(x)}{p_2(x)} dx
$$

(This "distance" does not obey the requisite symmetry and triangle inequalities for a metric.)Suppose we seek to approximate an arbitrary distribution $p_2(x)$ by a normal $p_1(x) \sim N(\mu, \Sigma)$. Show that the values that lead to the smallest Kullback-Leibler divergence are the obvious ones:

$$
\begin{aligned}
\mu &= \epsilon_2[x] \\
\Sigma &= \epsilon_2[(x-\mu)(x-\mu)^T]
\end{aligned}
$$

where the expectation ϵ_2 taken is over the density $p_2(x)$.

SOLUTION:

$p_1(x) \sim N(\mu, \Sigma)$, $p_2(x)$ is an arbitrary distribution. From $p_2(x)$, we can get KL divergence

$$
\begin{aligned}
D_{KL}(p_2(x), p_1(x)) &= \int p_2(x) ln \frac{p_2(x)}{p_1(x)} dx \\
&= \int p_2(x) ln p_2(x) dx + \frac{1}{2} \int p_2(x)[\ln(2\pi) + \ln \Sigma + (x-\mu)^T \Sigma^{-1} (x-\mu)] dx
\end{aligned}
$$

To minimize $D_{KL}(p_2(x), p_1(x))$ with parameter μ, Σ, set

$$
\frac{\partial D_{KL}(p_2(x), p_1(x))}{\partial \mu} = -\int p_2(x) \Sigma^{-1}(x-\mu) dx = 0
$$

$$
\frac{\partial D_{KL}(p_2(x), p_1(x))}{\partial \Sigma} = \int p_2(x)[\Sigma^{-1} - (x-\mu)^T \Sigma^{-2}(x-\mu)] dx = 0
$$

Then we get

$$
\int p_2(x)(x-\mu) dx = 0 \Rightarrow \epsilon_2(x-\mu) = 0
$$

$$
\int p_2(x)[\Sigma - (x-\mu)(x-\mu)^T] dx = 0 \Rightarrow \epsilon_2(\Sigma - (x-\mu)(x-\mu)^T) = 0
$$

Thus

$$
\begin{aligned}
\mu &= \epsilon_2(x) \\
\Sigma &= \epsilon_2[(x-\mu)(x-\mu)^T]
\end{aligned}
$$

4. (Programming) Assume $p(x) \sim 0.1N(-1,1) + 0.9N(1,1)$. Draw n samples from $p(x)$, for example, $n = 5, 10, 50, 100, \cdots, 1000, \cdots, 10000$. Use Parzen-window method to estimate $p_n(x) \approx p(x)$ (Hint: use randn() function in matlab to draw samples)

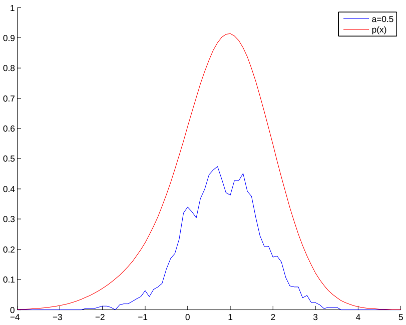

Figure 2: a=0.5

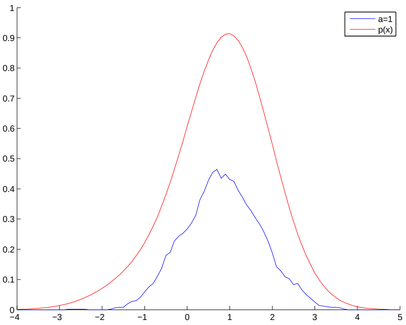

Figure 3: a=1

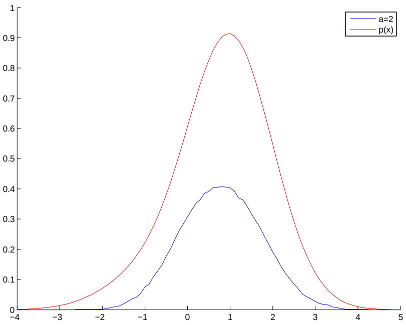

Figure 4: a=2

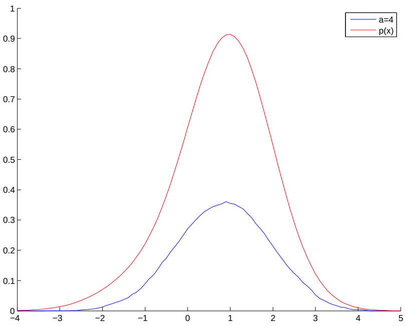

Figure 5: a=4

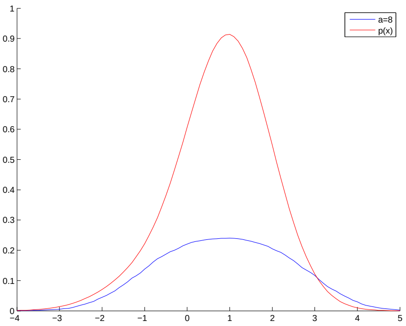

Figure 6: a=8

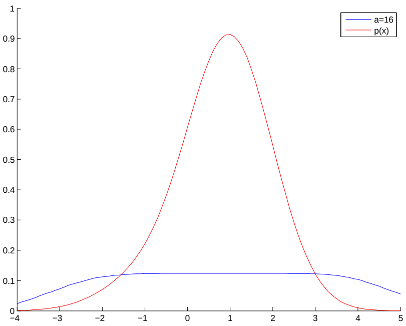

Figure 7: a=16

(a) Try window-function $P(x) = \begin{cases} \frac{1}{a}, -\frac{1}{2}a \leq x \leq \frac{1}{2}a \\ 0, otherwise. \end{cases}$. Estimate $p(x)$ with different window width a.

(b)Derive how to compute $\epsilon(p_n) = \int [p_n(x) - p(x)]^2 dx$ numerically.
SOLUTION:
we can compute $\epsilon(p_n)$ by $\epsilon(p_n) \approx \sum_{i=1}^{T} [p_n(x_i) - p(x_i)]^2 \Delta x_i$

(c)Demonstrate the expectation and variance of $\epsilon(p_n)$ w.r.t different n and a .

(d)With n given, how to choose optimal a from above the empirical experiences?

(e)Substitute $h(x)$ in (a) with Gaussian window. Repeat (a)-(e).

(g)Try different window functions and parameters as many as you can. Which window function/parameter is the best one? Demonstrate it numerically.
SOLUTION:
From Figure 11 to Figure 13, we can see that Gaussian window functions is the best with smallest square error $\epsilon(p_n)$. From Figure 14 to Figure 16, we can see that Gaussian windows functions with parameter $h_1 = 20$ is the best with smallest square error.

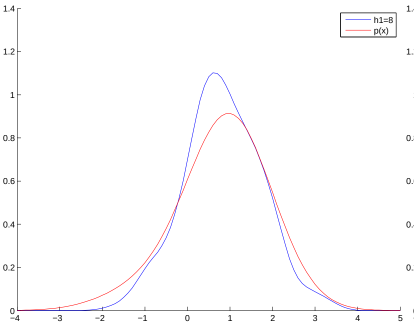

Figure 8: $h_1 = 8$

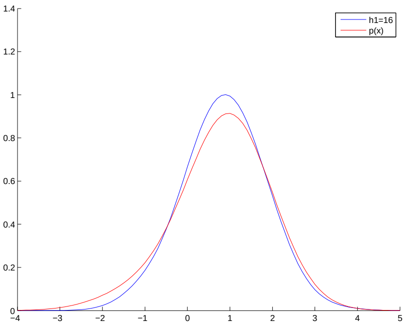

Figure 9: $h_1 = 16$

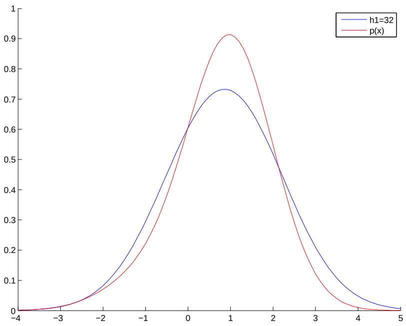

Figure 10: $h_1 = 32$

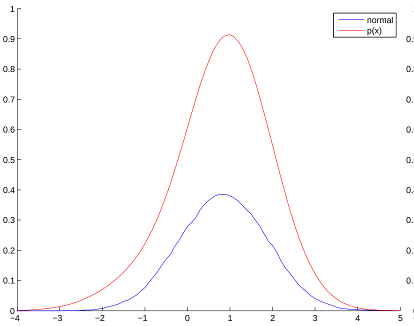

Figure 11: $error = 0.5367$

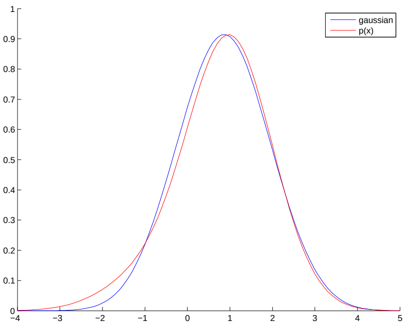

Figure 12: $error = 0.0073$

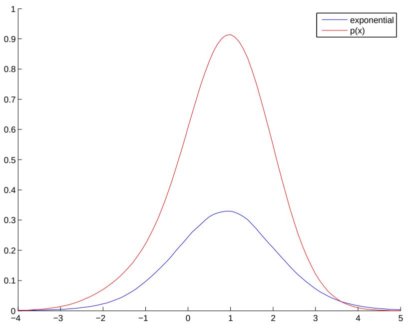

Figure 13: $error = 0.5887$

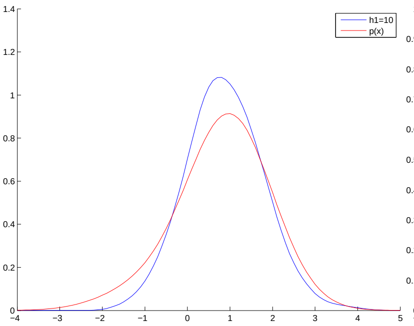

Figure 14: $error = 0.0561$

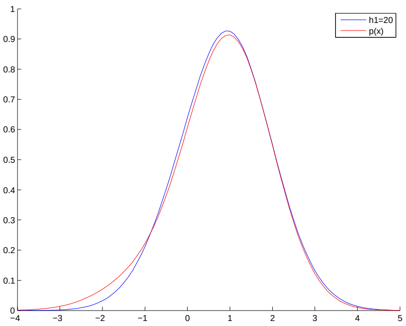

Figure 15: $error = 0.0057$

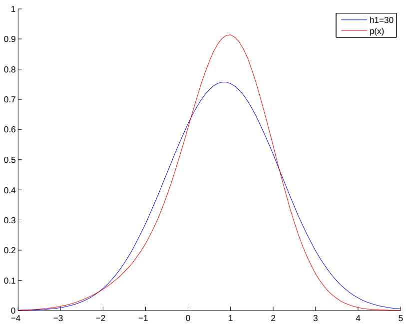

Figure 16: $error = 0.0381$