

CSI771 Final  
May 14, 2007  
Closed book, computer, notes

Name (print) \_\_\_\_\_ G number \_\_\_\_\_

Work on this exam is to be done on your own. Please observe the Honor Code and certify that you have done so by signing the pledge.

Signature \_\_\_\_\_

Please put your answers on these pages. Write very carefully and legibly. *Show all work.*

1. Estimation of functions.

We will assume that a functional relationship exists between a variable of interest,  $y$ , and an observable variable,  $x$ . There are various cases of interest.

- (a) First, assume the functional relationship represents a mean value of all  $y$  at that particular value of  $x$ , and the relationship is of the form  $y = f(x) + \epsilon$ , where  $\epsilon$  is an unobservable random variable. Assume further that the functional form  $y = f(x)$  is known; that is, assume we know the function  $f$ , except possibly for some unknown parameters (that is,  $y = f(x; \theta)$ , and we know all about  $f$  except for  $\theta$ ).

Assume that we have a set of pairs of observations  $(y_1, x_1), \dots, (y_n, x_n)$ , where  $y_i$  is just one realization of  $y$  associated with the value  $x_i$ .

Describe how you would estimate  $f$ .

To estimate  $f$  in this case is just to estimate  $\theta$ . A simple (and good) way of estimating  $\theta$  is as

$$\hat{\theta} = \arg \min_t \|y - \underline{f}(x; t)\|,$$

where  $y$  and  $x$  are, respectively, the vectors  $(y_1, \dots, y_n)$  and  $(x_1, \dots, x_n)$ ,  $\underline{f}(x, t)$  is the vector  $(f(x_1; t), \dots, f(x_n; t))$ , and  $\|\cdot\|$  is some norm. If  $\|\cdot\|$  is  $L_2$  norm, for example, this is the “least squares” estimator of  $\theta$  and of the function  $f(x; \theta)$ .

- (b) Now, assume the functional form  $y = f(x)$  is **not** known, and again assume that we have a set of pairs of observations  $(y_1, x_1), \dots, (y_n, x_n)$ , where the  $y_i$  are measured with some error, or are realizations of a random variable with a mean that depends on  $x_i$ . Describe how you would estimate  $f$ . (This answer must be different from your answer to the previous question.) Be very specific. It is not sufficient merely to say “I would use \_\_\_.” If there are choices you need to make, be clear how you would make those choices.

There are several possibilities. One possibility is simply to smooth the points. To do this, we first order the data with respect to  $x$ ; that is, we rearrange the data as

$$(y^{(1)}, x_{(1)}), (y^{(2)}, x_{(2)}), \dots, (y^{(n)}, x_{(n)}),$$

where  $x_{(i)}$  is the  $i^{\text{th}}$  order statistic of  $x$ , and  $y^{(i)}$  is the corresponding value of  $y$ . We then construct a curve through the points in that order. There are many types of curves. The simplest would just be straight line segments.

A more complicated curve could be constructed using splines. (We would first do the ordering, as indicated above.)

Another possibility is to use orthogonal polynomials as a set of basis functions. We would choose the orthogonal polynomials based on the support of the function  $f$ . (This would be decided based on assumptions or known facts about  $f$ .) We then would have to estimate the  $c_j$ 's. We could do this by minimizing a norm as in the first part of this question.

- (c) Again, assume the functional form  $y = f(x)$  is not known. Also, assume  $f$  is a PDF; that is,  $y$  represents a probability density of the random variable  $X$  of which  $x$  is a realization. In this case, obviously we cannot observe a value for  $y$  even with error. Now our set of observations consists of the singletons  $x_1, \dots, x_n$ . We must estimate  $y$  indirectly.

Briefly describe three ways of doing this. (Just give names and formulas. Although you do not need to be very detailed, I have provided an additional page you can use.)

Three ways are histogram, kernel, and orthogonal series. See the text for a description of each.

2. Operations in linear spaces.

Let  $\mathcal{X}$  be a set that is closed under an addition operation (that is, for  $x, y \in \mathcal{X}$ ,  $x + y \in \mathcal{X}$ ) and that is closed under multiplication by a real number (that is,  $x \in \mathcal{X}$ , and  $t \in \mathbb{R}$  implies  $tx \in \mathcal{X}$ ). (Think of  $\mathcal{X}$  as a vector space or a space of functions.)

Now, let  $\phi$  be a linear function that takes  $\mathcal{X} \times \mathcal{X}$  into the reals; that is, for  $x, y \in \mathcal{X}$ ,  $\phi(x, y)$  is a real number, and for  $x, y, z \in \mathcal{X}$  and a real number  $t$ ,

$$\phi(tx + y, z) = t\phi(x, z) + \phi(y, z);$$

and that has the additional property that  $\phi(x, x) \geq 0$  for any  $x \in \mathcal{X}$ .

(Think of an inner product.)

- (a) Show that  $\phi(x, y)^2 \leq \phi(x, x)\phi(y, y)$ . *Hint:* consider  $\phi(tx + y, tx + y)$ .

This is the Cauchy-Schwarz inequality. See page 130.

- (b) Two different objects  $x$  and  $y$  in  $\mathcal{X}$  are said to be *orthogonal* if  $\phi(x, y) = 0$ . Assume that  $w$  and  $z$  in  $\mathcal{X}$  are such that  $\phi(w, z)^2 < \phi(w, w)\phi(z, z)$  ( $w$  and  $z$  are “linearly independent”), and find  $x$  and  $y$  in terms of  $w$  and  $z$  such that  $x$  and  $y$  are orthogonal.

This is the Gram-Schmidt orthogonalization.

Let

$$x = w / \sqrt{\phi(w, w)}$$

and

$$y = (z - \phi(x, z)x) / \sqrt{\phi(z - \phi(x, z)x, z - \phi(x, z)x)}$$

- (c) What is a norm?

A norm is a mapping from a linear space to the nonnegative reals. If  $\mathcal{X}$  is a linear space and  $\nu : \mathcal{X} \mapsto \mathbb{R} - \mathbb{R}_-$ ,  $\nu$  is a norm if and only if

- $\nu(x) = 0$  if and only if  $x = 0$ , where  $0$  is the additive identity in  $\mathcal{X}$ .
- $\nu(ax) = a\nu(x)$  for any  $x$  in  $\mathcal{X}$  and any  $a$  in  $\mathbb{R}$ .
- $\nu(x + y) \leq \nu(x) + \nu(y)$  for any  $x$  and  $y$  in  $\mathcal{X}$ .

How could you define a norm on  $\mathcal{X}$  in terms of  $\phi(\cdot, \cdot)$ ?

For  $x \in \mathcal{X}$ , let  $\nu(x) = \sqrt{\phi(x, x)}$ . Then  $\nu$  is a norm.

### 3. Integrated measures.

Consider the  $U(0, \theta)$  distribution, with  $\theta$  unknown. The true probability density is  $p(x) = 1/\theta$  over  $(0, \theta)$  and 0 elsewhere. Suppose we have a sample of size  $n$  and we estimate the density as  $\hat{p}(x) = 1/x_{(n)}$  over  $(0, x_{(n)})$  and 0 elsewhere, where  $x_{(n)}$  is the maximum order statistic. The density of the distribution of  $X_{(n)}$  is  $nx_{(n)}^{n-1}\theta^{-n}$  over  $(0, \theta)$  and 0 elsewhere.

- (a) Determine (that is, write an explicit expression for) the integrated squared bias, ISB, of  $\hat{p}(x)$ .

We first work out the bias, for which we need  $E(1/X_{(n)})$ :

$$\begin{aligned} E(1/X_{(n)}) &= \int_0^\theta nt^{n-2}\theta^{-n}dt \\ &= \frac{n}{n-1} \frac{1}{\theta}. \end{aligned}$$

Then we have

$$\begin{aligned} \text{ISB}(\hat{p}(x)) &= \int_0^\theta \left( \left( \frac{n}{n-1} - 1 \right) \frac{1}{\theta} \right)^2 dt \\ &= \frac{1}{(n-1)^2} \frac{1}{\theta} \end{aligned}$$

- (b) Determine the integrated squared error, ISE, of  $\hat{p}(x)$ .

$$\begin{aligned} \text{ISE}(\hat{p}(x)) &= \int_0^{x_{(n)}} \left( \frac{1}{x_{(n)}} - \frac{1}{\theta} \right)^2 dt + \int_{x_{(n)}}^\theta \left( \frac{1}{\theta} \right)^2 dt \\ &= \frac{1}{x_{(n)}} - \frac{1}{\theta}. \end{aligned}$$

- (c) Determine the mean integrated squared error, MISE, of  $\hat{p}(x)$ .

Using  $E(1/X_{(n)})$  from above, we have

$$\begin{aligned} \text{MISE}(\hat{p}(x)) &= E(\text{ISE}(\hat{p}(x))) \\ &= E\left( \frac{1}{X_{(n)}} - \frac{1}{\theta} \right) \\ &= \frac{1}{n-1} \frac{1}{\theta}. \end{aligned}$$

- (d) Determine the asymptotic (as  $n \rightarrow \infty$ ) mean integrated squared error, AMISE, of  $\hat{p}(x)$ .

The AMISE as a truncation of a Taylor series of the MISE, which we often use to determine the order of convergence, is exactly the same as the MISE.

From the MISE, as  $n \rightarrow \infty$ , we have  $\text{AMISE}(\hat{p}(x)) \rightarrow 0$ ; that is,  $\hat{p}(x)$  is *AMISE consistent* for  $p(x)$ .

- (e) The questions above assume we know  $p(x)$ . How would you use the bootstrap to estimate the bias in  $X_{(n)}$  for  $\theta$  (knowing only that  $\theta$  is the upper bound on the range, but not knowing the distribution is uniform)?

We would resample in the usual way and compute  $b_1$  as in equation (4.3).

Is the bootstrap very reliable in this case? Why or why not?

No. The bootstrap is not very reliable for estimating the expectation of an order statistic, especially an extreme order statistic.

4. Short answers.

- (a) Suppose a dataset is sphered (what does that mean?), and then a principal components analysis is performed. What can you say about the principal components for these sphered data?

The principal components are indeterminate (because the variance is the same in all directions).

- (b) How does K-means clustering differ from hierarchical clustering? Briefly describe each method.

Both methods require some measure of distance (or similarity).

In K-means, first, we must specify  $K$ . Then the method finds  $K$  clusters such that the total distances of all points in each cluster from the centroid of the cluster are minimized.

Hierarchical clustering can be either agglomerative or divisive. In either case there is a possibility of identifying any number of clusters from 1 to  $n$ . Agglomerative clustering begins by identifying the two most similar points. Those two form the first cluster. Successive clusters are formed by identifying the two most similar points or clusters or points and clusters.

- (c) Explain why an acceptance/rejection method for generating multivariate random variates is generally not very efficient in higher dimensions. Be specific. An example will help (even one in one or two dimensions).

The reason is that the rejection region becomes proportionally larger as the dimensionality increases. See Exercise 10.10.

- (d) Describe the basic ideas in projection pursuit.

The basic idea is that we look for one-dimensional projections that are most different from a sample from a normal distribution.

- (e) Describe two ways of ranking multivariate data. Assume we have  $d$ -variate data, with observations  $x_1, x_2, \dots, x_n$ , and we want to determine a meaningful ranking

$$x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}.$$

One way is by using an MST (see page 256 and Figure 10.13), and another way is convex hull peeling (see page 258 and Figure 10.14).