Ongoing Notes - 546

Definitions

Chapter 1

Properties of a CDF:

(1): F_X is monotonically nondecreasing: if $x \leq y$, then $F_X(x) \leq F_X(y)$

(2): $F_X(x)$ tends to 0 as $x \to -\infty$ and to 1 as $x \to \infty$

(3): $F_X(x)$ is a continuous function of x.

A Key Property We Will Use Time and Time Again:

$$0 \leq Var[X] = E[X^2] - (E[X])^2$$

Thm. 1.1:

If the covariance matrix of $\mathbf{Y}$ is Σ_{YY}, then the covariance matrix of $\mathbf{Z} = \mathbf{c} + \mathbf{AY}$ is

$$\Sigma_{ZZ} = \mathbf{A\Sigma_{YY}A}^\top$$

Thm. 1.2:

Let $\mathbf{X}$ be a random n vector with mean μ and covariance Σ and let $\mathbf{A}$ be a fixed matrix. Then:

$$\mathbb{E}[\mathbf{X}^\top \mathbf{AX}] = \mathbf{tr}(\mathbf{A\Sigma}) + \mu^\top \mathbf{A}\mu$$

Thm. 1.3:

Let $\mathbf{X}$ be a random vector with covariance matrix $\Sigma_X X$.

If: $\mathbf{Y} = \mathbf{A_{p\times n}X}$ and $\mathbf{Z} = \mathbf{B_{m\times n}X}$, where $\mathbf{A}$ and $\mathbf{B}$ are fixed matrices. Then, the cross-covariance matrix of $\mathbf{Y}$ and $\mathbf{Z}$ is:

$$\Sigma_{YZ} = \mathbf{A\Sigma_{XX}B}^\top$$

Limiting Behavior of Functions

Big O Let f and g be two functions defined on some subset of the real numbers. One writes:

$$f(x) = O(g(x)) \text{ as x} \to \infty$$

iff $\exists M$ (some positive constant) such that for all sufficienctly large values of x, f(x) is at most M multiplied by the absolute value of g(x).

Alternative formulation:

$$f(x) = O(g(x)) \iff \exists M \in \mathbb{R}^+ \text{ and } \exists x_0 \in \mathbb{R} \text{ such that } |f(x)| \leq M|g(x)| \quad \forall x \geq x_0$$

Note: Typically this course with use $n \to \infty$

little o Description: This means that g(x) grows **much faster** than f(x).

$$f(x) = o(g(x)) \text{ as } x \to \infty$$

Means: for every positive constant ϵ, there exists a constant N such that:

$$|f(n)| \leq \epsilon |g(n)| \quad \forall n \geq N$$

Note: If something is little o, then it is also Big O; the reverse is not true.

Also: If g(x) is nonzero, or at least becomes nonzero beyond a certain point, the relation $f(x) = o(g(x))$ is equivalent to:

$$\lim_{x \to \infty} \frac{f(x)}{g(x)} = 0$$

Limiting Behavior of Random Variables:

When X is a R.V., then:

(Big O): $X_n = O_p(a_n)$, means that the set of values X_n/a_n is stochastically bounded. That is, for any $\epsilon > 0$, there exists a finite M > 0 such that:

$$P[|X_n/a_n \geq M] < \epsilon \quad \forall n$$

(little o):

$X_n = o_p(a_n)$ means that the set of values X_n/a_n converges to zero in probability as n approaches an appropriate limit. Equivalently,

$X_n = o_p(a_n)$ can be written as $X_n/a_n = o_p(1)$, where $X_n = o_p(1)$ is defined as:

$$\lim_{n \to \infty} P(|X_n| \geq \epsilon) = 0$$

Note: $o_p(1)$ is short for a sequence of random vectors that converges to zero in probability.

7 Most Used Proof Rules (Using Big O and little o formulas)

(1): $o_p(1) + o_p(1) = o_p(1)$

(2): $o_p(1) + O_p(1) = O_p(1)$

(3): $O_p(1)o_p(1) = o_p(1)$

(4): $1 + o_p(1))^{-1} = O_p(1)$

(5): $o_p(R_n) = R_n o_p(1)$

(6): $O_p(R_n) = R_n O_p(1)$

(7): $o_p(O_p(1)) = o_p(1)$

Chapter 2

2.1

To evaluate the *global performance* of a density estimate, the most intuitively appealing global criterion is the L_∞ norm

$$\sup_x \left|\hat{f}_X(x) - f_X(x)\right|.$$

(max deviation)

At the other end of the spectrum is the L_1 norm

$$\int \left|\hat{f}_X(x) - f_X(x)\right| dx.$$

The L_1 nor the L_∞ criterion is as easily manipulated as the L_2 norm, which is referred to as **integrated squared error (ISE)**

$$\text{ISE}(\hat{f}_X) \;=\; \int \left(\hat{f}_X(x) - f_X(x)\right)^2 dx,$$

Note: ISE is a R.V.

However, the MISE is not a R.V. because of expectation.

$$E[ISE(\hat{f_X})] = IMSE(\hat{f}_X)$$

2.2

Histogram Estimator Choose an origin t_0 and a bin width $h > 0$, where the bin width is the width of the classes (i.e., bins).

The kth bin is given by

$$B_k = [t_k, t_{k+1}], \quad k \in \mathbb{Z}$$

with

$$t_{k+1} = t_k + h, \quad k \in \mathbb{Z}.$$

Denote by ν_k the *bin count* of the kth bin, i.e., the number of sample points falling in the bin B_k.
The histogram estimator is defined as

$$\hat{f}_X(x) = \frac{\nu_k}{nh} = \frac{1}{nh}\sum_{i=1}^{n} \mathbf{1}_{[t_k, t_{k+1})}(X_i), \quad \text{for } x \in B_k \tag{2.1}$$

The histogram estimator is a very elementary estimator, but it can give the first good idea about the underlying unknown density function.

But if one wants to work further with the density estimate (discriminant analysis, hazard function estimation, ...) then a more accurate estimator is needed.

The histogram is a discontinuous function (a step function), and hence the density is estimated with a step function.

However, there are two unknown quantities in (2.1), i.e.,

- the bin width h
- the origin t_0 (position of the edges of the bins).

2.3

Reading Notes

Chapter 1

1.2: Smoothing: general concepts

Two main types of problems we'll study.

Density Estimation: Want to estimate the pdf f_X when we have a random sample from a distribution.

Regression: $Y_i = m(X_i) + \epsilon_i$, where m is the regression function (what we estimate!) and our **key assumption** $E[e|X] = 0$

Throughout the course, we DO NOT REQUIRE Normality assumptions (but we do require uncorrelated errors!).

For convenience though, we will also assume X's are independent

There is no "gold standard" for non-parametric estimation; it is best treated on a case-by-case basis.

1.3: Some concepts on continuous random variables

We know that a CDF always exists; the issue is that sometimes the pdf does not exist (or at least, does not exist in an easy closed form)

Big O and little o: Descriptions of the limiting behavior of a function when the argument tends towards a particular value or infinity, usually in terms of simpler functions, e.g. x, x^2, etc.

Big O convergence: Is like convergence in Probability

Little o convergence: Like Markov, Chebychev inequalities

Chapter 2

2.1

L_1 vs. L_2

- The L_1 criterion $\int|\hat{f}_X(x)-f_X(x)|dx$ puts more emphasis on the tails of a density than the L_2 criterion. The latter de-emphasizes the relatively small density values by squaring.
- Note that

$$\int|\hat{f}_X(x)-f_X(x)|dx \;\leq\; \int|\hat{f}_X(x)|dx+\int|f_X(x)|dx \;\leq\; 2$$

if $\hat{f}_X$ is a density. Hence, it follows that

$$0 \;\leq\; \int|\hat{f}_X(x)-f_X(x)|dx \;\leq\; 2.$$

- For the L_2 criterion, we have that

$$0 \;\leq\; \int\big(\hat{f}_X(x)-f_X(x)\big)^2dx \;\leq+\infty.$$

- In practical situations, the estimators that optimize these criteria are similar.
- The analytical simplicity of squared error and its adequacy in practical applications makes the L_2 criterion often the criterion of choice.
- L_1 error is invariant under any smooth monotone transformation.

Scheffé's Lemma

(Scheffé, 1947; Devroye and Györfi, 1985). For all densities f and g on $\mathbb{R}^d$:

$$\int|f(x)-g(x)|dx \;=\; 2\,TV(f,g) \;=\; 2\int\max(f(x)-g(x),0)\,dx \;=\; 2\int\max(g(x)-f(x),0)\,dx$$

- The result in Scheffé's lemma provides a connection with statistical classification.

2.2

Histogram Estimator The analysis of the histogram random variable $\hat{f}_X$ is quite simple once one recognizes that the bin counts are Binomial random variables.
For the bin count of the kth bin

$$\nu_k\sim\text{Bin}(n,p_k)\quad\text{where}\quad p_k=\int_{B_k}f(t)\,dt.$$

Hence, we have

$$\mathbb{E}[\nu_k] = n \cdot p_k \quad \text{and} \quad \text{Var}[\nu_k] = n \cdot p_k \cdot (1 - p_k).$$

and therefore, for $x \in B_k$ (see also Figure 2.4)

$$\mathbb{E}[\hat{f}_X(x)] = \frac{p_k}{h} \quad \text{and} \quad \text{Var}[\hat{f}_X(x)] = \frac{p_k \cdot (1 - p_k)}{nh^2}.$$

The exact bias of the histogram estimator is

$$\mathbb{E}[\hat{f}_X(x)] - f_X(x) = \frac{p_k}{h} - f_X(x).$$

Denote the optimal bin width by $h_{n,\text{MISE}}$. Then, an approximation for $h_{n,\text{MISE}}$ is obtained by minimizing the asymptotic expression for $\text{MISE}(\hat{f}_X)$ given in (2.3):

$$h_{n,\text{AMISE}} = \arg\min_h \text{AMISE}(\hat{f}_X)$$

$$= \arg\min_h \left[\frac{1}{nh} + \frac{1}{12} h^2 \int f_X''(x)^2 \, dx \right]$$

$$= \left[\frac{6}{\int f_X''(x)^2 \, dx} \right]^{1/3} n^{-1/3}. \tag{2.4}$$

However, (2.4) is not useful in practice since it depends on the true unknown density f_X.

A quick and simple bin width selection rule is obtained by referring to a normal density.

If $f_X = N(\mu, \sigma^2)$, then

$$\int f_X''(x)^2 \, dx = \frac{1}{4\sqrt{\pi}\, \sigma^3}$$

and consequently

$$\hat{h}_{n,\text{AMISE}} = \left[\frac{24\sqrt{\pi}\, \hat{\sigma}^3}{n} \right]^{1/3} \approx 3.5 \hat{\sigma} n^{-1/3},$$

where $\hat{\sigma}$ is a consistent estimator of σ, e.g.,

$$\hat{\sigma} = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2} \quad \text{with} \quad \bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i.$$

This bin width selector is the so-called **rule-of-thumb bin width selector**.

2.3