Lecture 16. Lehmann-Scheffé Theorem

16.1 Definition

Suppose that Y is a sufficient statistic for θ. We say that Y is *complete* if there are no nontrivial unbiased unbiased estimates of 0 based on Y, i.e., if $E[g(Y)] = 0$ for all θ, then $P_\theta\{g(Y) = 0\} = 1$ for all θ. Thus if we have two unbiased estimates of θ based on Y, say $\varphi(Y)$ and $\psi(Y)$, then $E_\theta[\varphi(Y) - \psi(Y)] = 0$ for all θ, so that regardless of θ, $\varphi(Y)$ and $\psi(Y)$ coincide (with probability 1). So if we find one unbiased estimate of θ based on Y, we have essentially found all of them.

16.2 Theorem (Lehmann-Scheffé)

Suppose that $Y_1 = u_1(X_1, \ldots, X_n)$ is a complete sufficient statistic for θ. If $\varphi(Y_1)$ is an unbiased estimate of θ based on Y_1, then among all possible unbiased estimates of θ (whether based on Y_1 or not), $\varphi(Y_1)$ has minimum variance. We say that $\varphi(Y_1)$ is a *uniformly minimum variance unbiased estimate* of θ, abbreviated UMVUE. The term "uniformly" is used because the result holds for *all possible values* of θ.

Proof. By Rao-Blackwell, if Y_2 is any unbiased estimate of θ, then $E[Y_2|Y_1]$ is an unbiased estimate of θ with $\mathrm{Var}[E(Y_2|Y_1)] \leq \mathrm{Var}\, Y_2$. But $E(Y_2|Y_1)$ is a function of Y_1, so by completeness it must coincide with $\varphi(Y_1)$. Thus regardless of the particular value of θ, $\mathrm{Var}_\theta[\varphi(Y_1)] \leq \mathrm{Var}_\theta(Y_2)$. ♣.

Note that just as in the Rao-Blackwell theorem, the Lehmann-Scheffé result holds equally well if we are seeking a UMVUE of a function of θ. Thus we look for an unbiassed estimate of $r(\theta)$ based on the complete sufficient statistic Y_1.

16.3 Definition and Remarks

There are many situations in which complete sufficient statistics can be found quickly. The *exponential class* (or *exponential family*) consists of densities of the form

$$f_\theta(x) = a(\theta)b(x)\exp\big[\sum_{j=1}^{m} p_j(\theta)K_j(x)\big]$$

where $a(\theta) > 0, b(x) > 0, \alpha < x < \beta, \theta = (\theta_1, \ldots, \theta_k)$ with $\gamma_j < \theta_j < \delta_j, 1 \leq j \leq k$ ($\alpha, \beta, \gamma_j, \delta_j$ are constants).

There are certain regularity conditions that are assumed, but they will always be satisfied in the examples we consider, so we will omit the details. In all our examples, k and m will be equal. This is needed in the proof of completeness of the statistic to be discussed in Lecture 17. (It is not needed for sufficiency.)

16.4 Examples

1. Binomial(n, θ) where n is known. We have $f_\theta(x) = \binom{n}{x}\theta^x(1-\theta)^{n-x}, x = 0, 1, \ldots, n$, where $0 < \theta < 1$. Take $a(\theta) = (1-\theta)^n, \quad b(x) = \binom{n}{x}, \quad p_1(\theta) = \ln\theta - \ln(1-\theta), K_1(x) = x$. Note that $k = m = 1$.

2. Poisson(θ). The probability function is $f_\theta(x) = e^{-\theta}\theta^x/x!, x = 0, 1, \ldots$, where $\theta > 0$. We can take $a(\theta) = e^{-\theta}, \quad b(x) = 1/x!, \quad p_1(\theta) = \ln\theta, \quad K_1(x) = x$, and $k = m = 1$.

3. Normal(μ, σ^2). The density is

$$f_\theta(x) = \frac{1}{\sqrt{2\pi}\sigma}\exp[-(x-\mu)^2/2\sigma^2], \quad -\infty < x < \infty, \quad \theta = (\mu, \sigma^2).$$

Take $a(\theta) = [1/\sqrt{2\pi}\sigma]\exp[-\mu^2/2\sigma^2], \quad b(x) = 1, \quad p_1(\theta) = -1/2\sigma^2, \quad K_1(x) = x^2, \quad p_2(\theta) = \mu/\sigma^2, \quad K_2(x) = x$, and $k = m = 2$.

4. Gamma(α, β). The density is $x^{\alpha-1}e^{-x/\beta}/[\Gamma(\alpha)\beta^\alpha], \quad x > 0, \quad \theta = (\alpha, \beta)$. Take $a(\theta) = 1/[\Gamma(\alpha)\beta^\alpha], \quad b(x) = 1, \quad p_1(\theta) = \alpha - 1, \quad K_1(x) = \ln x, \quad p_2(\theta) = -1/\beta, \quad K_2(x) = x$, and $k = m = 2$.

5. Beta(a, b). The density is $[\Gamma(a+b)/\Gamma(a)\Gamma(b)]x^{a-1}(1-x)^{b-1}, \quad 0 < x < 1, \quad \theta = (a, b)$. Take $a(\theta) = [\Gamma(a+b)/\Gamma(a)\Gamma(b)], \quad b(x) = 1, \quad p_1(\theta) = a - 1, \quad K_1(x) = \ln x, \quad p_2(\theta) = b - 1, \quad K_2(x) = \ln(1-x)$, and $k = m = 2$.

6. Negative Binomial

First we derive some properties of this distribution. In a possibly infinite sequence of Bernoulli trials, let Y_r be the number of trials required to obtain the r-th success (assume r is a known positive integer). Then $P\{Y_1 = k\}$ is the probability of $k-1$ failures followed by a success, which is $q^{k-1}p$ where $q = 1 - p$ and $k = 1, 2, \ldots$. The moment-generating function of Y_1 is

$$M_{Y_1}(t) = E[e^{tY_1}] = \sum_{k=1}^{\infty} q^{k-1}pe^{tk}.$$

Write e^{tk} as $e^{t(k-1)}e^t$. We get

$$M_{Y_1}(t) = pe^t(1 + qe^t + (qe^t)^2 + \cdots) = \frac{pe^t}{1 - qe^t}, \quad |qe^t| < 1.$$

The random variable Y_1 is said to have the *geometric* distribution. (The slightly different random variable appearing in Problem 3 of Lecture 14 is also frequently referred to as geometric.) Now Y_r (the negative binomial random variable) is the sum of r independent random variables, each geometric, so

$$M_{Y_r}(t) = \left(\frac{pe^t}{1 - qe^t}\right)^r.$$

The event $\{Y_r = k\}$ occurs iff there are $r - 1$ successes in the first $k - 1$ trials, followed by a success on trial k. Therefore

$$P\{Y_r = k\} = \binom{k-1}{r-1}p^{r-1}q^{k-r}p, \quad x = r, r+1, r+2, \ldots .$$

We can calculate the mean and variance of Y_r from the moment-generating function, but the differentiation is not quite as messy if we introduce another random variable. Let X_r be the number of failures preceding the r-th success. Then X_r plus the number of successes preceding the r-th success is the total number of trials preceding the r-th success. Thus

$$X_r + (r-1) = Y_r - 1, \quad \text{so} \quad X_r = Y_r - r$$

and

$$M_{X_r}(t) = e^{-rt} M_{Y_r}(t) = \left(\frac{p}{1-qe^t}\right)^r.$$

When $r = 1$ we have

$$M_{X_1}(t) = \frac{p}{1-qe^t}, \quad E(X_1) = \left.\frac{pqe^t}{(1-qe^t)^2}\right|_{t=0} = \frac{q}{p}.$$

Since $Y_1 = X_1 + 1$ we have $E(Y_1) = 1 + (q/p) = 1/p$ and $E(Y_r) = r/p$. Differentiating the moment-generating function of X_1 again, we find that

$$E(X_1^2) = \frac{(1-q)^2 pq + pq^2 2(1-q)}{(1-q)^4} = \frac{pq(1-q)[1-q+2q]}{(1-q)^4} = \frac{pq(1+q)}{p^3} = \frac{q(1+q)}{p^2}.$$

Thus $\operatorname{Var} X_1 = \operatorname{Var} Y_1 = [q(1+q)/p^2] - [q^2/p^2] = q/p^2$, hence $\operatorname{Var} Y_r = rq/p^2$.

Now to show that the negative binomial distribution belongs to the exponential class:

$$P\{Y_r = x\} = \binom{x-1}{r-1}\theta^r(1-\theta)^{x-r}, \quad x = r, r+1, r+2, \ldots, \theta = p.$$

Take

$$a(\theta) = \left(\frac{\theta}{1-\theta}\right)^r, \quad b(x) = \binom{x-1}{r-1}, \quad p_1(\theta) = \ln(1-\theta), \quad K_1(x) = x, \quad k = m = 1.$$

Here is the reason for the terminology "negative binomial":

$$M_{Y_r}(t) = \left(\frac{pe^t}{1-qe^t}\right)^r = p^r e^{rt}(1-qe^t)^{-r}.$$

To expand the moment-generating function, we use the binomial theorem with a negative exponent:

$$(1+x)^{-r} = \sum_{k=0}^{\infty} \binom{-r}{k} x^k$$

where

$$\binom{-r}{k} = \frac{-r(-r-1\cdots(-r-k+1)}{k!}.$$

Problems are deferred to Lecture 17.

Lecture 17. Complete Sufficient Statistics For The Exponential Class

17.1 Deriving the Complete Sufficient Statistic

The density of a member of the exponential class is

$$f_\theta(x) = a(\theta)b(x)\exp\big[\sum_{j=1}^{m} p_j(\theta)K_j(x)\big]$$

so the joint density of n independent observations is

$$f_\theta(x_1,\ldots,x_n) = (a(\theta))^n \prod_{i=1}^{n} b(x_i)\exp\big[\sum_{j=1}^{m} p_j(\theta)K_j(x_1)\big]\cdots\exp\big[\sum_{j=1}^{m} p_j(\theta)K_j(x_n)\big].$$

Since $e^r e^s e^t = e^{r+s+t}$, it follows that $p_j(\theta)$ appears in the exponent multiplied by the factor $K_j(x_1) + K_j(x_2) + \cdots + K_j(x_n)$, so by the factorization theorem,

$$\big(\sum_{i=1}^{n} K_1(x_i),\ldots,\sum_{i=1}^{n} K_m(x_i)\big)$$

is sufficient for θ. This statistic is also complete. First consider $m = 1$:

$$f_\theta(x_1,\ldots,x_n) = \big(a(\theta)\big)^n \prod_{i=1}^{n} b(x_i)\exp\big[p(\theta)\sum_{i=1}^{n} K(x_i)\big].$$

Let $Y_1 = \sum_{i=1}^{n} K(X_i)$; then $E_\theta[g(Y_1)]$ is given by

$$\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} g\big(\sum_{i=1}^{n} K(x_i)\big) f_\theta(x_1,\ldots,x_n)\,dx_1\cdots dx_n.$$

If $E_\theta[g(Y_1)] = 0$ for all θ, then for all θ, $g(Y_1) = 0$ with probability 1.

What we have here is analogous to a result from Laplace or Fourier transform theory: If for all t between a and b we have

$$\int_{-\infty}^{\infty} g(y)e^{ty}\,dy = 0$$

then $g = 0$. It is also analogous to the result that the moment-generating function determines the density uniquely.

When $m > 1$, the exponent in the formula for $f_\theta(x_1,\ldots,x_n)$ becomes

$$p_1(\theta)\sum_{i=1}^{n} K_1(x_i) + \cdots + p_m(\theta)\sum_{i=1}^{n} K_m(x_i)$$

and the argument is essentially the same as in the one-dimensional case. The transform result is as follows. If

$$\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \exp[t_1 y_1 + \cdots + t_m y_m] g(y_1, \ldots, y_n)\, dy_1 \cdots dy_m = 0$$

when $a_i < t_i < b_i, i = 1, \ldots, m$, then $g = 0$. The above integral defines a joint moment-generating function, which will appear again in connection with the multivariate normal distribution.

17.2 Example

Let $X_1, \ldots, X_n$ be iid, each normal(θ, σ^2) where σ^2 is known. The normal distribution belongs to the exponential class (see (16.4), Example 3), but in this case the term $\exp[-x^2/2\sigma^2]$ can be absorbed in $b(x)$, so only $K_2(x) = x$ is relevant. Thus $\sum_{i=1}^n X_i$, equivalently $\overline{X}$, is sufficient (as found in Lecture 14) and complete. Since $E(\overline{X}) = \theta$, it follows that $\overline{X}$ is a UMVUE of θ.

Let's find a UNVUE of θ^2. The natural conjecture that it is $(\overline{X})^2$ is not quite correct. Since $\overline{X} = n^{-1} \sum_{i=1}^n X_i$, we have $\operatorname{Var} \overline{X} = \sigma^2/n$. Thus

$$\frac{\sigma^2}{n} = E[(\overline{X})^2] - (E\overline{X})^2 = E[(\overline{X})^2] - \theta^2,$$

hence

$$E\big[(\overline{X})^2 - \frac{\sigma^2}{n}\big] = \theta^2$$

and we have an unbiased estimate of θ^2 based on the complete sufficient statistic $\overline{X}$. Therefore $(\overline{X})^2 - [\sigma^2/n]$ is a UMVUE of θ^2.

17.3 A Cautionary Tale

Restricting to unbiased estimates is not always a good idea. Let X be Poisson(θ), and take $n = 1$, i.e., only one observation is made. From (16.4), Example 2, X is a complete sufficient statistic for θ. Now

$$E[(-1)^X] = \sum_{k=0}^{\infty} (-1)^k \frac{e^{-\theta}\theta^k}{k!} = e^{-\theta} \sum_{k=0}^{\infty} \frac{(-\theta)^k}{k!} = e^{-\theta} e^{-\theta} = e^{-2\theta}.$$

thus $(-1)^X$ is a UMVUE of $e^{-2\theta}$. But $Y \equiv 1$ is certainly a better estimate, since 1 is closer to $e^{-2\theta}$ than is -1. Estimating a positive quantity $e^{-2\theta}$ by a random variable which can be negative is not sensible.

Note also that the maximum likelihood estimate of θ is X (Lecture 9, Problem 1a), so the MLE of $e^{-2\theta}$ is e^{-2X}, which looks better than Y.

Problems

1. Let X be a random variable that has zero mean for all possible values of θ. For example, X can be uniformly distributed between $-\theta$ and θ, or normal with mean 0 and variance θ. Give an example of a sufficient statistic for θ that is not complete.

2. Let $f_\theta(x) = \exp[-(x-\theta)], \theta < x < \infty$, and 0 elsewhere. Show that the first order statistic $Y_1 = \min X_i$ is a complete sufficient statistic for θ, and find a UMVUE of θ.

3. Let $f_\theta(x) = \theta x^{\theta-1}, 0 < x < 1$, where $\theta > 0$. Show that $u(X_1, \ldots, X_n) = \left[\prod_{i=1}^n X_i\right]^{1/n}$ is a complete sufficient statistic for θ, and that the maximum likelihood estimate $\hat{\theta}$ is a function of $u(X_1, \ldots, X_n)$.

4. The density $f_\theta(x) = \theta^2 x \exp[-\theta x], x > 0$, where $\theta > 0$, belongs to the exponential class, and $Y = \sum_{i=1}^n X_i$ is a complete sufficient statistic for θ. Compute the expectation of $1/Y$ under θ, and from the result find the UMVUE of θ.

5. Let Y_1 be binomial (n, θ), so that $Y_1 = \sum_{i=1}^n X_i$, where X_i is the indicator of a success on trial i. [Thus each X_i is binomial $(1, \theta)$.] By Example 1 of (16.4), the X_i, as well as Y_1, belong to the exponential class, and Y_1 is a complete sufficient statistic for θ. Since $E(Y_1) = n\theta$, Y_1/n is a UMVUE of θ.

   Let $Y_2 = (X_1 + X_2)/2$. In an effortless manner, find $E(Y_2|Y_1)$.

6. Let X be normal with mean 0 and variance θ, so that by Example 3 of (16.4), $Y = \sum_{i=1}^n X_i^2$ is a complete sufficient statistic for θ. Find the distribution of Y/θ, and from this find the UMVUE of θ^2.

7. Let $X_1, \ldots, X_n$ be iid, each Poisson (θ), where $\theta > 0$. (Then $Y = \sum_{i=1}^n X_i$ is a complete sufficient statistic for θ.) Let I be the indicator of $\{X_1 \le 1\}$.
   (a) Show that $E(I|Y)$ is the UMVUE of $P\{X_1 \le 1\} = (1+\theta)\exp(-\theta)$. Thus we need to evaluate $P\{X_1 = 0|Y = y\} + P\{X_1 = 1|Y = y\}$. When $y = 0$, the first term is 1 and the second term is 0.
   (b) Show that if $y > 0$, the conditional distribution of X_1 (or equally well, of any X_i) is binomial $(y, 1/n)$.
   (c) Show that

$$E(I|Y) = \left(\frac{n-1}{n}\right)^Y \left[1 + \frac{Y}{n-1}\right]$$

8. Let $\theta = (\theta_1, \theta_2)$ and $f_\theta(x) = (1/\theta_2)\exp[(x-\theta_1)/\theta_2], x > \theta_1$ (and 0 elsewhere) where θ_1 is an arbitrary real number and $\theta_2 > 0$. Show that the statistic $(\min_i X_i, \sum_{i=1}^n X_i)$ is sufficient for θ.

Lecture 18. Bayes Estimates

18.1 Basic Assumptions

Suppose we are trying to estimate the state of nature θ. We observe $X = x$, where X has density $f_\theta(x)$, and make decision $\delta(x) =$ our estimate of θ when x is observed. We incur a loss $L(\theta, \delta(x))$, assumed nonnegative. We now *assume* that θ is random with density $h(\theta)$. The Bayes solution minimizes the *Bayes risk* or *average loss*

$$B(\delta) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} h(\theta) f_\theta(x) L(\theta, \delta(x))\, d\theta\, dx.$$

Note that $h(\theta)f_\theta(x) = h(\theta)f(x|\theta)$ is the joint density of θ and x, which can also be expressed as $f(x)f(\theta|x)$. Thus

$$B(\delta) = \int_{-\infty}^{\infty} f(x) \left[\int_{-\infty}^{\infty} L(\theta, \delta(x)) f(\theta|x)\, d\theta \right] dx.$$

Since $f(x)$ is nonnegative, it is sufficient to minimize $\int_{-\infty}^{\infty} L(\theta, \delta(x)) f(\theta|x)\, d\theta$ for each x. The resulting δ is called the *Bayes estimate* of θ. Similarly, to estimate a function of θ, say $\gamma(\theta)$, we minimize $\int_{-\infty}^{\infty} L(\gamma(\theta), \delta(x)) f(\theta|x)\, d\theta$.

We can jettison a lot of terminology by recognizing that our problem is to observe a random variable X and estimate a random variable Y by $g(X)$. We must minimize $E[L(Y, g(X)]$.

18.2 Quadratic Loss Function

We now assume that $L(Y, g(X)) = (Y - g(X))^2$. By the theorem of total expectation,

$$E(Y - g(X))^2 = \int_{-\infty}^{\infty} E[(Y - g(X))^2 | X = x] f(x)\, dx$$

and as above, it suffices to minimize the quantity in brackets for each x. If we let $z = g(x)$, we are minimizing $z^2 - 2E(Y|X = x)z + E(Y^2|X = x)$ by choice of z. Now $Az^2 - 2Bz + C$ is a minimum when $z = B/A = E(Y|X = x)/1$, and we conclude that

$$\boxed{E[(Y - g(X))^2] \quad \text{is minimized when} \quad g(x) = E(Y|X = x).}$$

What we are doing here is minimizing $E[(W - c)^2] = c^2 - 2E(W)c + E(W^2)$ by our choice of c, and the minimum occurs when $c = E(W)$.

18.3 A Different Loss Function

Suppose that we want to minimize $E(|W - c|)$. We have

$$E(|W - c|) = \int_{-\infty}^{c} (c - w) f(w)\, dw + \int_{c}^{\infty} (w - c) f(w)\, dw$$

$$= c\int_{-\infty}^{c} f(w)\,dw - \int_{-\infty}^{c} wf(w)\,dw + \int_{c}^{\infty} wf(w)\,dw - c\int_{c}^{\infty} f(w)\,dw.$$

Differentiating with respect to c, we get

$$cf(c) + \int_{-\infty}^{c} f(w)\,dw - cf(c) - cf(c) + cf(c) - \int_{c}^{\infty} f(w)\,dw$$

which is 0 when $\int_{-\infty}^{c} f(w)\,dw = \int_{c}^{\infty} f(w)\,dw$, in other words when C is a *median* of W. Thus $E(|Y - g(X)|)$ is minimized when $g(x)$ is a median of the conditional distribution of Y given $X = x$.

18.4 Back To Quadratic Loss

In the statistical decision problem with quadratic loss, the Bayes estimate is

$$\delta(x) = E[\theta|X = x] = \int_{-\infty}^{\infty} \theta f(\theta|x)\,d\theta$$

and

$$f(\theta|x) = \frac{f(\theta, x)}{f(x)} = \frac{h(\theta)f(x|\theta)}{f(x)}.$$

Thus

$$\boxed{\delta(x) = \frac{\int_{-\infty}^{\infty} \theta h(\theta) f_\theta(x)\,d\theta}{\int_{-\infty}^{\infty} h(\theta) f_\theta(x)\,d\theta}}$$

If we are estimating a function of θ, say $\gamma(\theta)$, replace θ by $\gamma(\theta)$ in the integral in the numerator.

Problems

1. Let X be binomial(n, θ), and let the density of θ be

$$h(\theta) = \frac{\theta^{r-1}(1-\theta)^{s-1}}{\beta(r, s)} \quad [\text{beta}(r, s)].$$

   Show that the Bayes estimate with quadratic loss is

$$\delta(x) = \frac{r + x}{r + s + n}, \quad x = 0, 1, \ldots, n.$$

2. For this estimate, show that the *risk function* $R_\delta(\theta)$, defined as the average loss using δ when the parameter is θ, is

$$\frac{1}{(r + s + n)^2}[((r + s)^2 - n)\theta^2 + (n - 2r(r + s))\theta + r^2]$$

3. Show that if $r = s = \sqrt{n}/2$, then $R_\delta(\theta)$ is a constant, independent of n.
4. Show that a Bayes estimate δ with constant risk (as in Problem 3) is *minimax*, that is, δ minimizes $\max_\theta R_\delta(\theta)$.

Lecture 19. Linear Algebra Review

19.1 Introduction

We will assume for the moment that matrices have complex numbers as entries, but the complex numbers will soon disappear. If A is a matrix, the conjugate transpose of A will be denoted by A^*. Thus if

$$\mathbb{A} = \begin{bmatrix} a+bi & c+di \\ e+fi & g+hi \end{bmatrix} \quad \text{then} \quad A^* = \begin{bmatrix} a-bi & e-fi \\ c-di & g-hi \end{bmatrix}.$$

The transpose is

$$A' = \begin{bmatrix} a+bi & e+fi \\ c+di & g+hi \end{bmatrix}.$$

Vectors X, Y, etc., will be regarded as column vectors. The *inner product (dot product)* of n-vectors X and Y is

$$< X, Y >= x_1\overline{y}_1 + \cdots + x_n\overline{y}_n$$

where the overbar indicates complex conjugate. Thus $< X, Y >= Y^*X$. If c is any complex number, then $< cX, Y >= c < X, Y >$ and $< X, cY >= \overline{c} < X, Y >$. The vectors X and Y are said to be *orthogonal (perpendicular)* if $< X, Y >= 0$. For an arbitrary n by n matrix B,

$$< BX, Y >=< X, B^*Y >$$

because $< X, B^*Y >= (B^*Y)^*X = Y^*B^{**}X = Y^*BX =< BX, Y >$.

Our interest is in *real symmetric matrices*, and "symmetric" will always mean "real symmetric". If A is symmetric then

$$< AX, Y >=< X, A^*Y >=< X, AY > .$$

The *eigenvalue problem* is $AX = \lambda X$, or $(A - \lambda I)X = 0$, where I is the identity matrix, i.e., the matrix with 1's down the main diagonal and 0's elsewhere. A nontrivial solution ($X \neq 0$) exists iff $\det(A - \lambda I) = 0$. In this case, λ is called an *eigenvalue* of A and a nonzero solution is called an *eigenvector*.

19.2 Theorem

If A is symmetric then A has real eigenvalues.

Proof. Suppose $AX = \lambda X$ with $X \neq 0$. then $< AX, Y >=< X, AY >$ with $Y = X$ gives $< \lambda X, X >=< X, \lambda X >$, so $(\lambda - \overline{\lambda}) < X, X >= 0$. But $< X, X >= \sum_{i=1}^{n} |x_i|^2 \neq 0$, and therefore $\lambda = \overline{\lambda}$, so λ is real. ♣

The important conclusion is that for a symmetric matrix, the eigenvalue problem can be solved using only real numbers.

19.3 Theorem

If A is symmetric, then eigenvectors of distinct eigenvalues are orthogonal.

Proof. Suppose $AX_1 = \lambda_1 X_1$ and $AX_2 = \lambda_2 X_2$. Then $< AX_1, X_2 >=< X_1, AX_2 >$, so $< \lambda_1 X_1, X_2 >=< X_1, \lambda_2 X_2 >$. Since λ_2 is real we have $(\lambda_1 - \lambda_2) < X_1, X_2 >= 0$. But we are assuming that we have two distinct eigenvalues, so that $\lambda_1 \neq \lambda_2$. Therefore we must have $< X_1, X_2 >= 0$. ♣

19.4 Orthogonal Decomposition Of Symmetric Matrices

Assume A symmetric with distinct eigenvalues $\lambda_1, \ldots, \lambda_n$. The assumption that the λ_i are distinct means that the equation $\det(A - \lambda I) = 0$, a polynomial equation in λ of degree n, has no repeated roots. This assumption is actually unnecessary, but it makes the analysis much easier.

Let $AX_i = \lambda_i X_i$ with $X_i \neq 0, i = 1, \ldots, n$. Normalize the eigenvectors so that $\|X_i\|$, the length of X_i, is 1 for all i. (The length of the vector $x = (x_1, \ldots, x_n)$ is

$$\|x\| = \Big(\sum_{i=1}^{n} |x_i|^2\Big)^{1/2}$$

hence $\|x\|^2 =< x, x >$.) Thus we have $AL = LD$, where

$$L = [X_1|X_2|\cdots|X_n] \quad \text{and} \quad D = \begin{bmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_n \end{bmatrix}.$$

To verify this, note that multiplying L on the right by a diagonal matrix with entries $\lambda_1, \ldots, \lambda_n$ multiplies column i of L (namely X_i) by λ_i. (Multiplying on the left by D would multiply row i by λ_i.) Therefore

$$LD = [\lambda_1 X_1|\lambda_2 X_2|\cdots|\lambda_n X_n|] = AL.$$

The columns of the square matrix L are mutually perpendicular unit vectors; such a matrix is said to be *orthogonal*. The transpose of L can be pictured as follows:

$$L' = \begin{bmatrix} X_1' \\ X_2' \\ \vdots \\ X_n' \end{bmatrix}$$

Consequently $L'L = I$. Since L is nonsingular ($\det I = 1 = \det L' \det L$), L has an inverse, which must be L'. to see this, multiply the equation $L'L = I$ on the right by L^{-1} to get $L'I = L^{-1}$, i.e., $L' = L^{-1}$. Thus $LL' = I$.

Since a matrix and its transpose have the same determinant, $(\det L)^2 = 1$, so the determinant of L is ± 1.

Finally, from $AL = LD$ we get

$$\boxed{L'AL = D}$$

We have shown that *every symmetric matrix (with distinct eigenvalues) can be orthogonally diagonalized*.

19.5 Application To Quadratic Forms

Consider a *quadratic form*

$$X'AX = \sum_{i,j=1}^{n} a_{i,j}x_i x_j.$$

If we change variables by $X = LY$, then

$$X'AX = Y'L'ALY = Y'DY = \sum_{i=1}^{n} \lambda_i y_i^2.$$

The symmetric matrix A is said to be *nonnegative definite* if $X'AX \geq 0$ for all X. Equivalently, $\sum_{i=1}^{n} \lambda_i y_i^2 \geq 0$ for all Y. Set $y_i = 1, y_j = 0$ for all $j \neq i$ to conclude that A is nonnegative definite if and only if *all eigenvalues of* A are nonnegative. The symmetric matrix is said to be *positive definite* if $X'AX > 0$ except when all $x_i = 0$. Equivalently, *all eigenvalues of* A are strictly positive.

19.6 Example

Consider the quadratic form

$$q = 3x^2 + 2xy + 3y^2 = (x, y)\begin{bmatrix} 3 & 1 \\ 1 & 3 \end{bmatrix}\begin{pmatrix} x \\ y \end{pmatrix}$$

Then

$$A = \begin{bmatrix} 3 & 1 \\ 1 & 3 \end{bmatrix}, \quad \det(A - \lambda I) = \begin{vmatrix} 3-\lambda & 1 \\ 1 & 3-\lambda \end{vmatrix} = \lambda^2 - 6\lambda + 8 = 0$$

and the eigenvalues are $\lambda = 2$ and $\lambda = 4$. When $\lambda = 2$, the equation $A(x, y)' = \lambda(x, y)'$ reduces to $x + y = 0$. Thus $(1, -1)$ is an eigenvector. Normalize it to get $(1/\sqrt{2}, -1/\sqrt{2})'$. When $\lambda = 4$ we get $-x + y = 0$ and the normalized eigenvector is $(1/\sqrt{2}, 1/\sqrt{2})'$. Consequently,

$$L = \begin{bmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ -1/\sqrt{2} & 1/\sqrt{2} \end{bmatrix}$$

and a direct matrix computation yields

$$L'AL = \begin{bmatrix} 2 & 0 \\ 0 & 4 \end{bmatrix} = D$$

as expected. If $(x, y)' = L(v, w)'$, i.e., $x = (1/\sqrt{2})v + (1/\sqrt{2})w, \quad y = (-1/\sqrt{2})v + (1/\sqrt{2})w$, then

$$q = 3\left[\frac{v^2}{2} + \frac{w^2}{2} + vw\right] + 2\left[-\frac{v^2}{2} + \frac{w^2}{2}\right] + 3\left[\frac{v^2}{2} + \frac{w^2}{2} - vw\right].$$

Thus $q = 2v^2 + 4w^2 = (v, w)D(v, w)'$, as expected.

Lecture 20. Correlation

20.1 Definitions and Comments

Let X and Y be random variables with finite mean and variance. Denote the mean of X by μ_1 and the mean of Y by μ_2, and let $\sigma_1^2 = \operatorname{Var} X$ and $\sigma_2^2 = \operatorname{Var} Y$. Note that $E(XY)$ must be finite also, because $-X^2 - Y^2 \le 2XY \le X^2 + Y^2$. The *covariance* of X and Y is defined by

$$\operatorname{Cov}(X, Y) = E[(X - \mu_1)(Y - \mu_2)]$$

and it follows that

$$\operatorname{Cov}(X, Y) = E(XY) - \mu_1 E(Y) - \mu_2 E(X) + \mu_1\mu_2 = E(XY) - E(X)E(Y).$$

Thus $\operatorname{Cov}(X, Y) = \operatorname{Cov}(Y, X)$. Since expectation is linear, we have $\operatorname{Cov}(aX, bY) = ab \operatorname{Cov}(X, Y)$, $\operatorname{Cov}(X, Y + Z) = \operatorname{Cov}(X, Y) + \operatorname{Cov}(X, Z)$, $\operatorname{Cov}(X + Y, Z) = \operatorname{Cov}(X, Z) + \operatorname{Cov}(Y, Z)$, and $\operatorname{Cov}(X + a, Y + b) = \operatorname{Cov}(X, Y)$. Also, $\operatorname{Cov}(X, X) = E(X^2) - (EX)^2 = \operatorname{Var} X$.

The *correlation coefficient* is a normalized covariance:

$$\rho = \frac{\operatorname{Cov}(X, Y)}{\sigma_1\sigma_2}.$$

The correlation coefficient is a measure of *linear dependence* between X and Y. To see this, estimate Y by $AX + b$, equivalently (to simplify the calculation) estimate $Y - \mu_2$ by $c(X - \mu_1) + d$, choosing c and d to minimize

$$E[(Y - \mu_2 - (c(X - \mu_1) + d))^2] = \sigma_2^2 - 2c \operatorname{Cov}(X, Y) + c^2\sigma_1^2 + d^2.$$

Note that $E[2cd(X - \mu_1)] = 0$ since $E(X) = \mu_1$, and similarly $E[2d(Y - \mu_2)] = 0$. We can't do any better than to take $d = 0$, so we need to minimize $\sigma_2^2 - 2c\rho\sigma_1\sigma_2 + c^2\sigma_1^2$ by choice of c. Differentiating with respect to c, we have $-2\rho\sigma_1\sigma_2 + 2c\sigma_1^2$, hence

$$\boxed{c = \rho\frac{\sigma_2}{\sigma_1}}$$

The minimum expectation is

$$\sigma_2^2 - 2\rho\frac{\sigma_2}{\sigma_1}\rho\sigma_1\sigma_2 + \rho^2\frac{\sigma_2^2}{\sigma_1^2}\sigma_1^2 = \boxed{\sigma_2^2(1 - \rho^2)}$$

The expectation of a nonnegative random variable is nonnegative, so

$$\boxed{-1 \le \rho \le 1}$$

For a fixed σ_2, the closer $|\rho|$ is to 1, the better the estimate of Y by $aX + b$. If $|\rho| = 1$ then the minimum expectation is 0, so (with probability 1)

$$Y - \mu_2 = c(X - \mu_1) = \rho\frac{\sigma_2}{\sigma_1}(X - \mu_1) \quad \text{with} \quad \rho = \pm 1.$$

20.2 Theorem

If X and Y are independent then X and Y are uncorrelated ($\rho = 0$) but not conversely.

Proof. Assume X and Y are independent. Then

$$E[(X - \mu_1)(Y - \mu_2)] = E(X - \mu_1)E(Y - \mu_2) = 0.$$

For the counterexample to the converse, let $X = \cos\theta, Y = \sin\theta$, where θ is uniformly distributed on $(0, 2\pi)$. Then

$$E(X) = \frac{1}{2\pi}\int_0^{2\pi} \cos\theta\, d\theta = 0, \quad E(Y) = \frac{1}{2\pi}\int_0^{2\pi} \sin\theta\, d\theta = 0,$$

and

$$E(XY) = E[(1/2)\sin 2\theta] = \frac{1}{4\pi}\int_0^{2\pi} \sin 2\theta\, d\theta = 0,$$

so $\rho = 0$. But $X^2 + Y^2 = 1$, so X and Y are not independent. ♣

20.3 The Cauchy-Schwarz Inequality

This result, namely

$$|E(XY)|^2 \le E(X^2)E(Y^2)$$

is closely related to $-1 \le \rho \le 1$. Indeed, if we replace X by $X - \mu_1$ and Y by $Y - \mu_2$, the inequality says that $[\mathrm{Cov}(X, Y)]^2 \le \sigma_1^2\sigma_2^2$, i.e., $(\rho\sigma_1\sigma_2)^2 \le \sigma_1^2\sigma_2^2$, which gives $\rho^2 \le 1$. Thus Cauchy-Schwarz implies $-1 \le \rho \le 1$.

Proof. Let $h(\lambda) = E[(\lambda X + Y)^2] = \lambda^2 E(X^2) + 2\lambda E(XY) + E(Y^2)$. Since $h(\lambda) \ge 0$ for all λ, the quadratic equation $h(\lambda) = 0$ has no real roots or at worst a real repeated root. Therefore the discriminant is negative or at worst 0. Thus $[2E(XY)]^2 - 4E(X^2)E(Y^2) \le 0$, and the result follows. ♣

As a special case, let $P\{X = x_i\} = 1/n, 1 \le i \le n$. If $X = x_i$, take $Y = y_i$. (The x_i and y_i are arbitrary real numbers.) Then the Cauchy-Schwarz inequality becomes

$$\left(\sum_{i=1}^n x_i y_i\right)^2 \le \left(\sum_{i=1}^n x_i^2\right)\left(\sum_{i=1}^n y_i^2\right).$$

(There will be a factor of $1/n^2$ on each side of the inequality, which will cancel.) This is the result originally proved by Cauchy. Schwarz proved the analogous formula for integrals:

$$\left(\int_a^b f(x)g(x)\, dx\right)^2 \le \int_a^b [f(x)]^2\, dx \int_a^b [g(x)]^2\, dx.$$

Since an integral can be regarded as the limit of a sum, the integral result can be proved from the result for sums.

We know that if $X_1, \ldots, X_n$ are independent, then the variance of the sum of the X_i is the sum of the variances. If we drop the assumption of independence, we can still say something.

20.4 Theorem

Let $X_1, \ldots, X_n$ be arbitrary random variables (with finite mean and variance). Then

$$\operatorname{Var}(X_1 + \cdots + X_n) = \sum_{i=1}^{n} \operatorname{Var} X_i + 2 \sum_{\substack{i,j=1 \\ i<j}}^{n} \operatorname{Cov}(X_i, X_j).$$

For example, the variance of $X_1 + X_2 + X_3 + X_4$ is

$$\sum_{i=1}^{4} \operatorname{Var} X_i + 2[\operatorname{Cov}(X_1, X_2) + \operatorname{Cov}(X_1, X_3) + \operatorname{Cov}(X_1, X_4)$$

$$+ \operatorname{Cov}(X_2, X_3) + \operatorname{Cov}(X_2, X_4) + \operatorname{Cov}(X_3, X_4)].$$

Proof. We have

$$E[(X_1 - \mu_1) + \cdots + (X_n - \mu_n)]^2 = E\left[\sum_{i=1}^{n} (X_i - \mu_i)^2\right]$$

$$+2E\left[\sum_{i<j} (X_i - \mu_i)(X_j - \mu_j)\right]$$

as asserted. ♣

The reason for the $i < j$ restriction in the summation can be seen from an expansion such as

$$(x + y + z)^2 = x^2 + y^2 + z^2 + 2xy + 2xz + 2yz.$$

It is correct, although a bit inefficient, to replace $i < j$ by $i \neq j$ and drop the factor of 2. This amounts to writing $2xy$ as $xy + yx$.

20.5 Least Squares

Let $(x_1, y_1), \ldots, (x_n, y_n)$ be points in the plane. The problem is to find the line $y = ax + b$ that minimizes $\sum_{i=1}^{n} [y_i - (ax_i + b)]^2$. (The numbers a and b are to be determined.)

Consider the following random experiment. Choose X with $P\{X = x_i\} = 1/n$ for $i = 1, \ldots, x_n$. If $X = x_i$, set $Y = y_i$. [This is the same setup as in the special case of the Cauchy-Schwarz inequality in (20.3).] Then

$$E[(Y - (aX + b))^2] = \frac{1}{n} \sum_{i=1}^{n} [y_i - (ax_i + b)]^2$$

so the least squares problem is equivalent to finding the best estimate of Y of the form $aX + b$, where "best" means that the mean square error is to be minimized. This is the problem that we solved in (20.1). The least squares line is

$$\boxed{y - \mu_Y = \rho \frac{\sigma_Y}{\sigma_X}(x - \mu_X)}$$

To evaluate $\mu_X, \mu_Y, \sigma_X, \sigma_Y, \rho$:

$$\mu_X = \frac{1}{n}\sum_{i=1}^{n} x_i = \overline{x}, \quad \mu_Y = \frac{1}{n}\sum_{i=1}^{n} y_i = \overline{y},$$

$$\sigma_X^2 = E[(X-\mu_X)^2] = \frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})^2 = s_x^2, \quad \sigma_Y^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i-\overline{y})^2 = s_y^2,$$

$$\rho = \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X\sigma_Y}, \quad \rho\frac{\sigma_Y}{\sigma_X} = \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X^2}.$$

The last entry is the slope of the least squares line, which after cancellation of $1/n$ in numerator and denominator, becomes

$$\frac{\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})}{\sum_{i=1}^{n}(x_i-\overline{x})^2}.$$

If $\rho > 0$, then the least squares line has positive slope, and y tends to increase with x. If $\rho < 0$, then the least squares line has negative slope and y tends to decrease as x increases.

Problems

In Problems 1-5, assume that X and Y are independent random variables, and that we know $\mu_X = E(X), \mu_Y = E(Y), \sigma_X^2 = \operatorname{Var} X$, and $\sigma_Y^2 = \operatorname{Var} Y$. In Problem 2, we also know ρ, the correlation coefficient between X and Y.

1. Find the variance of XY.
2. Find the variance of $aX + bY$, where a and b are arbitrary real numbers.
3. FInd the covariance of X and $X + Y$.
4. FInd the correlation coefficient between X and $X + Y$.
5. FInd the covariance of XY and X.
6. Under what conditions will there be equality in the Cauchy-Schwarz inequality?