CORRELATION AND LINEAR LEAST SQUARES PREDICTION

MICHAEL J. SHARPE
MATHEMATICS DEPARTMENT, UCSD

1. Notation

Throughout this section, we assume that all random variables have a finite second moment. That is, we assume $EX^2 < \infty$ for every random variable X that will appear here. If we use the inequality

$$|x| \le 1 + x^2, \qquad x \in \mathbf{R}, \tag{1.1}$$

which follows from the trivial estimates $|x| \le x^2$ for $|x| > 1$ and $|x| \le 1$ for $|x| \le 1$, we find, taking expectation,

$$E|X| \le 1 + EX^2,$$

so that X also has a finite first moment. The last inequality is crude, and will be improved below. The second simple estimate we need is that

$$|EX| \le E|X|, \tag{1.2}$$

which follows by writing $X = X^+ - X^-$, the positive and negative parts of X, so that $|X| = X^+ + X^-$ and $|EX| = |EX^+ - EX^-| \le EX^+ + EX^- = E|X|$. (Note: for any real x, x^+ is defined to be x if $x \ge 0$, and equal to 0 otherwise; similarly, x^- is defined to be $-x$ if $x < 0$ and 0 otherwise.)

We shall use the notation μ_X sometimes in place of EX, and σ_X^2 for the variance of X, namely $\sigma_X^2 = E(X - \mu_X)^2$.

{prop:sumofl

Proposition 1.3. *If X and Y each have finite second moment, then so does $aX + bY$ for any scalars a, b.*

Proof. Just observe that the simple inequality $(u + v)^2 \le 2u^2 + 2v^2$ (which comes about from expanding the first square and using $2uv \le u^2 + v^2$) yields $E(aX + bY)^2 \le 2a^2EX^2 + 2b^2EY^2 < \infty$. □

1.1. **Cauchy-Schwarz inequality.**

Theorem 1.4. *(Cauchy-Schwarz) Let X and Y have finite second moment. Then $E|XY| < \infty$, and*

(1.5) {ineq:cs} $$|E(XY)| \le \sqrt{E(X^2)E(Y^2)},$$

and equality holds if and only if one of X, Y is a scalar multiple of the other.

Proof. For every real t, $X + tY$ has a finite second moment by Proposition ??, and so the function $g(t) := E(X + tY)^2$ is finite valued. Expanding the square gives

$$g(t) = EX^2 + 2tE(XY) + t^2EY^2.$$

That is, $g(t)$ is quadratic in t, and clearly $g(t) \ge 0$ for all t. If $EY^2 = 0$, then Y vanishes almost surely, so $E(XY) = 0$, and (??) is clearly satisfies, with Y a scalar multiple (0) of X. otherwise, if $EY^2 > 0$, we use the fact that the discriminant of the quadratic must be ≤ 0, which is to say $4(E(XY))^2 - 4EX^2EY^2 \le 0$. This clearly proves (??). If equality holds in (??), then the discriminant of $g(t)$ vanishes, hence $g(t)$ has a single real root, say at t_0. The fact that $g(t_0) = 0$ implies that the positive random variable $(X + t_0Y)^2$ has expectation 0, and so must vanish almost surely. Thus we find $X = -t_0Y$ almost surely. □

1.2. **Covariance and correlation.** Given X, Y with finite variances σ_X^2, σ_Y^2 and means μ_X, μ_Y, define the covariance $\mathrm{Cov}(X,Y)$ between X and Y by

$$\mathrm{Cov}(X,Y) := E\big((X-\mu_X)(Y-\mu_Y)\big) = E(XY) - \mu_X\mu_Y. \tag{1.6}$$ {eq:cov}

In view of the Cauchy-Schwarz inequality, we have

$$|\,\mathrm{Cov}(X,Y)| \le \sigma_X\sigma_Y \tag{1.7}$$

with equality if and only if one of $X-\mu_X$, $Y-\mu_Y$ is a scalar multiple of the other, which is to say that either $Y = aX+b$ for some scalars a, b, or vise-versa. To put this another way, let's assume that $\sigma_X > 0$ and $\sigma_Y > 0$, and then define the correlation coefficient $\rho_{X,Y}$ between X and Y by

$$\rho_{X,Y} := \frac{\mathrm{Cov}(X,Y)}{\sigma_X\sigma_Y}.$$

Then one finds from (??) that

$$|\rho_{X,Y}| \le 1, \qquad \text{and } \rho_{X,Y} = \pm 1 \text{ if and only } Y = aX+b \text{ for some scalars } a, b. \tag{1.8}$$

(With a little additional work, one sees that $\rho_{X,Y} = 1$ implies $a > 0$, while $\rho_{X,Y} = -1$ implies $a < 0$.)

Note that covariance and correlation coefficient are insensitive to change of location: that is,

$$\mathrm{Cov}(X+\alpha, Y+\beta) = \mathrm{Cov}(X,Y); \qquad \rho_{X+\alpha,Y+\beta} = \rho_{X,Y}.$$

Their sensitivity to scale is also simple:

$$\mathrm{Cov}(aX, bY) = ab\,\mathrm{Cov}(X,Y); \qquad \rho_{aX,bY} = \rho_{X,Y}. \tag{1.9}$$

{prop:indep}

Proposition 1.10. *If X and Y are independent, then* $\mathrm{Cov}(X,Y) = \rho_{X,Y} = 0$. *(The converse is definitely untrue, without further strong hypotheses.)*

Proof. Because X and Y are independent, so are $X-\mu_X$ and $Y-\mu_Y$. using the fact that the expectation of a product of independent random variables is the product of the expectation, we find

$$\mathrm{Cov}(X,Y) = E\big((X-\mu_X)(Y-\mu_Y)\big) = E(X-\mu_X)E(Y-\mu_Y) = 0.$$

□

Covariance enters into computations of variances of sums, in the following way. (We omit the easily checked calculation.)

$$\mathrm{Var}(X+Y) = \mathrm{Var}(X) + 2\,\mathrm{Cov}(X,Y) + \mathrm{Var}(Y). \tag{1.11}$$

This is of course the same as writing $\sigma_{X+Y}^2 = \sigma_X^2 + 2\rho_{X,Y}\sigma_X\sigma_Y + \sigma_Y^2$.

1.3. **Linear Least Squares Prediction.** Once again, we assume X and Y have finite variances. The issue here is how to best predict Y using a linear function of X. That is, we wish to choose scalars u, v so that $uX+v$ is as good as possible a predictor of Y. As criterion for "best", we measure error of prediction by $E(Y-(uX+v))^2$. We shall let $Y^* := Y-\mu_Y$ and $X^* := X-\mu_X$. Then it suffices to solve the least squares predictor problem $E(Y^*-(sX^*+t))^2$. Expanding out the square and taking expectations yields

$$E(Y^*-(sX^*+t))^2 = E(Y^*)^2 + s^2E(X^*)^2 + t^2 - 2sE(X^*Y^*) - 2tE(Y^*) - 2stE(X^*).$$

Since $EX^* = EY^* = 0$ and $E(X^*)^2 = \sigma_X^2$, the right side reduces immediately to

$$\sigma_Y^2 + s^2\sigma_X^2 - 2s\,\mathrm{Cov}(X,Y) + t^2.$$

For every s, the minimum as a function of t occurs when $t=0$. The remaining term is a quadratic in s, complete the square, it becomes

$$\sigma_X^2(s^2 - 2\rho_{X,Y}\frac{\sigma_Y}{\Sigma_X} + \frac{\sigma_Y^2}{\sigma_X^2} = \sigma_X^2(s-\rho_{X,Y}\frac{\sigma_Y}{\sigma_X})^2 + \sigma_Y^2(1-\rho_{X,Y}^2),$$

which is clearly smallest when $s = \rho_{X,Y}\frac{\sigma_Y}{\sigma_X}$. From this we deduce that the best (least squares sense) linear predictor of Y^* given X^* is given by $\rho_{X,Y}\frac{\sigma_Y}{\sigma_X}X^*$, and consequently, the best (least squares sense) linear predictor of Y given X is

$$\hat{Y} := \mu_Y + \rho_{X,Y}\frac{\sigma_Y}{\sigma_X}(X - \mu_X). \tag{1.12}$$

{eq:predictor}

It may easily be checked that $\mathrm{Cov}(Y - \hat{Y}, \hat{Y}) = 0$, and hence that $\sigma_Y^2 = \sigma_{\hat{Y}}^2 + \sigma_{Y-\hat{Y}}^2$, where of course $\sigma_{\hat{Y}}^2 = \rho_{X,Y}^2\sigma_Y^2$.