Distribution of R^2 for Single Regression of Uncorrelated Gaussian Random Variables

Sahand Rabbani

Problem Statement

Consider an i.i.d. sequence of uncorrelated jointly Gaussian random variables X_i and Y_i, $i \in \{1, 2, \ldots, n\}$. We seek the distribution of the coefficient of determination R^2 obtained by regressing the sequences X_i on Y_i:

$$R^2 = \frac{\left(\sum_{i=1}^n X_i Y_i\right)^2}{\sum_{i=1}^n X_i^2 \sum_{i=1}^n Y_i^2}$$

Solution

First, consider the vectors $\mathbf{X}$ and $\mathbf{Y}$ in n-dimensional space defined by the sequences X_i and Y_i respectively. The coefficient of determination can be written as

$$R^2 = \left(\frac{\mathbf{X}^T\mathbf{Y}}{\|\mathbf{X}\|\,\|\mathbf{Y}\|}\right)^2 = \cos^2\theta$$

where $\theta \in [0, \pi]$ is the angle formed between the two vectors in n-dimensional space. We first derive the cumulative density function of the random variable θ.

This probability is given by the integral of the n-dimensional Gaussian density function over the infinite hypercone defined by the locus of points within an angle θ of the vector $\mathbf{Y}$, which we call $\mathcal{R}(\theta, \mathbf{Y})$. By total probability, we have

$$F_\Theta(\theta) = \int_{\mathbb{R}^n} \int_{\mathcal{R}(\theta,\mathbf{Y})} f_{\mathbf{X}|\mathbf{Y}}(\mathbf{x}|\mathbf{y}) f_{\mathbf{Y}}(\mathbf{y})\, d\mathbf{x}\, d\mathbf{y}$$

However, because $\mathbf{x}$ and $\mathbf{y}$ are jointly Gaussian and uncorrelated, they are independent and $f_{\mathbf{X}|\mathbf{Y}}(\mathbf{x}|\mathbf{y}) = f_{\mathbf{X}}(\mathbf{x})$. Due to the radial symmetry of the marginal distribution of $\mathbf{X}$, its integral over the infinite conical region $\mathcal{R}(\theta, \mathbf{y})$ is equal for all $\mathbf{y}$, allowing us to choose $\mathbf{y} = e_1$, where e_1 is a vector of all zeros except for its first element, which is unity. The integral becomes

$$F_\Theta(\theta) = \int_{\mathbb{R}^n} \int_{\mathcal{R}(\theta,e_1)} f_{\mathbf{X}}(\mathbf{x}) f_{\mathbf{Y}}(\mathbf{y})\, d\mathbf{x}\, d\mathbf{y} = \int_{\mathbb{R}^n} f_{\mathbf{Y}}(\mathbf{y})\, d\mathbf{y} \int_{\mathcal{R}(\theta,e_1)} f_{\mathbf{X}}(\mathbf{x})\, d\mathbf{x} = \int_{\mathcal{R}(\theta,e_1)} f_{\mathbf{X}}(\mathbf{x})\, d\mathbf{x}$$

Because of the radial symmetry of the Gaussian distribution, this probability is equal to the ratio of the finite conical section of a hypersphere of arbitrary radius to the volume of this hypersphere.

Consider the volume of the hypersphere with unit radius:

$$V_s = \int_{r=0}^{1} \int_{\phi_1=0}^{\pi} \cdots \int_{\phi_{n-2}=0}^{\pi} \int_{\phi_{n-1}=0}^{2\pi} r^{n-1} \sin^{n-2}\phi_1 \sin^{n-3}\phi_2 \cdots \sin\phi_{n-2}\, dr\, d\phi_1\, d\phi_2\, \cdots\, d\phi_{n-1}$$

The volume of the hyperconical section of this sphere with angular span θ is similarly given by

$$V_c = \int_{r=0}^{1} \int_{\phi_1=0}^{\theta} \cdots \int_{\phi_{n-2}=0}^{\pi} \int_{\phi_{n-1}=0}^{2\pi} r^{n-1} \sin^{n-2}\phi_1 \sin^{n-3}\phi_2 \cdots \sin\phi_{n-2}\, dr\, d\phi_1\, d\phi_2\, \cdots\, d\phi_{n-1}$$

The multiple integrals devolve into a product of single integrals. The ratio of these quantities admits the simple expression

$$F_\Theta(\theta) = \frac{\int_0^\theta \sin^{n-2}\phi \, d\phi}{\int_0^\pi \sin^{n-2}\phi \, d\phi}$$

The probability density function is the derivative of this expression:

$$f_\Theta(\theta) = \frac{dF_\Theta(\theta)}{d\theta} = \frac{\sin^{n-2}\theta}{\int_0^\pi \sin^{n-2}\phi \, d\phi}$$

The integral in the denominator has a closed-form solution in terms of the gamma function:

$$\int_0^\pi \sin^{n-2}\phi \, d\phi = \sqrt{\pi}\,\frac{\Gamma\left(\frac{n-1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)}$$

This gives

$$f_\Theta(\theta) = \frac{\Gamma\left(\frac{n}{2}\right)}{\sqrt{\pi}\,\Gamma\left(\frac{n-1}{2}\right)} \sin^{n-2}\theta$$

To find the probability density function of R^2, we employ a simple transformation of variables given by $R^2 = \cos^2\Theta$. That is,

$$f_{R^2}(r^2) = \frac{2 f_\Theta(\cos^{-1}\sqrt{r^2})}{|dr^2/d\theta|}$$

The factor of two accounts for the squaring, which maps all negative values of $\cos\theta$ to positive values of r^2. Noting that $\sin\cos^{-1}\sqrt{r^2} = \sqrt{1-r^2}$ and $|dr^2/d\theta| = 2\sqrt{r^2}\sqrt{1-r^2}$, we have the final solution:

$$f_{R^2}(r^2) = \frac{\Gamma\left(\frac{n}{2}\right)}{\sqrt{\pi r^2}\,\Gamma\left(\frac{n-1}{2}\right)} (1-r^2)^{\frac{n-3}{2}}$$