# A Squared Correlation Coefficient of the Correlation Matrix


Rong Fan*

Southern Illinois University


August 25, 2016


**Abstract**

Multivariate linear correlation analysis is important in statistical analysis and applications. This paper defines a one number summary $\gamma^2$ of the population correlation matrix that behaves like a squared correlation. The squared Pearson's correlation coefficient is a special case of $\gamma^2$ for two variables. Unlike the coefficient of multiple determination, also known as the multiple correlation coefficient, $\gamma^2$ does not depend on the choice of the dependent variable.



**KEY WORDS:** Bootstrap, Confidence Interval, Correlation Matrix.


## 1 INTRODUCTION

This paper defines a one number summary $\gamma^2$ of the population correlation matrix that acts like a squared correlation. The following notation will be useful. Let $\boldsymbol{x} = (X_1 ..., X_p)^T$ and $\text{Cov}(\boldsymbol{x}) = \boldsymbol{\Sigma}_{\boldsymbol{x}} = (\sigma_{ij})$ be the covariance matrix of $\boldsymbol{x}$ where $\sigma_{ij} = \text{Cov}(X_i, X_j)$ is the covariance of $X_i$ and $X_j$. Let $\text{Cor}(\boldsymbol{x}) = \boldsymbol{\rho}_{\boldsymbol{x}} = (\rho_{ij})$ be the correlation matrix of $\boldsymbol{x}$ where

$$\rho_{ij} = \frac{\text{Cov}(X_i, X_j)}{\sigma_i \sigma_j}$$

is the correlation of $X_i$ and $X_j$, and $\sigma_i^2 = \sigma_{ii}$ is the variance of $X_i$. Let $det(\boldsymbol{\rho}_{\boldsymbol{x}})$ be the determinant of $\text{Cor}(\boldsymbol{x})$, and let $\|\boldsymbol{A}\|_F$ denote the Frobenius norm of a matrix $\boldsymbol{A}$. Let $\boldsymbol{I}_p$ be the $p \times p$ identity matrix.

Let the $p \times p$ population standard deviation matrix

$$\boldsymbol{\Delta} = \text{diag}(\sqrt{\sigma_{11}}, ..., \sqrt{\sigma_{\text{pp}}}).$$

Then

$$\boldsymbol{\Sigma}_{\boldsymbol{x}} = \boldsymbol{\Delta} \boldsymbol{\rho}_{\boldsymbol{x}} \boldsymbol{\Delta}, \tag{1.1}$$

*Rong Fan is Ph.D. Student, Department of Mathematics, Southern Illinois University, Carbondale, IL 62901, USA.

and

$$\boldsymbol{\rho}_{\boldsymbol{x}} = \boldsymbol{\Delta}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{\Delta}^{-1}. \tag{1.2}$$

The correlation or Pearson correlation coefficient $\rho_{12}$ is used to measure the linear relationship between two variables $X_1$ and $X_2$, and we define a one number summary of the correlations between $p$ random variables $X_1, ..., X_p$ that acts like a squared correlation. When there are $p = 2$ random variables, the squared Pearson correlation coefficient is the special case of our squared coefficient.

The *squared correlation coefficient of the correlation matrix*

$$\gamma^2 = \frac{\|\boldsymbol{\rho}_{\boldsymbol{x}} - \boldsymbol{I}_p\|_F^2}{2det(\boldsymbol{\rho}_{\boldsymbol{x}}) + \|\boldsymbol{\rho}_{\boldsymbol{x}} - \boldsymbol{I}_p\|_F^2} = \frac{\sum_{i<j} \rho_{ij}^2}{det(\boldsymbol{\rho}_{\boldsymbol{x}}) + \sum_{i<j} \rho_{ij}^2}. \tag{1.3}$$

Here

$$\|\boldsymbol{\rho}_{\boldsymbol{x}} - \boldsymbol{I}_p\|_F^2 = 2\sum_{i<j} \rho_{ij}^2 = \text{trace}[(\boldsymbol{\rho}_{\boldsymbol{x}} - \boldsymbol{I}_{\text{p}})^2]$$

is used to measure the discrepancy between $\boldsymbol{\rho}_{\boldsymbol{x}}$ and $\boldsymbol{I}_p$.

# 2 PROPERTIES OF $\gamma^2$

The following theorem gives properties of $\gamma^2$. The random variable $X_k$ is a linear combination of the other $p-1$ random variables if

$$X_k = a_0 + a_1X_1 + \cdots + a_{k-1}X_{k-1} + a_{k+1}X_{k+1} + \cdots + a_pX_p = a_0 + \sum_{j\neq k} a_jX_j$$

where the constants $a_1, ..., a_{k-1}, a_{k+1}, ..., a_p$ are not all zero.

**Theorem 1.** Assume all pairwise covariances and correlations exist.
i) $0 \leq \gamma^2 \leq 1$.
ii) If $X_1, ..., X_p$ are pairwise uncorrelated ($\rho_{ij} = 0$ for $i \neq j$), then $\gamma^2 = 0$.
iii) If $X_k$ is a linear combination of the other $p-1$ random variables, then $\gamma^2 = 1$.
iv) Conversely, if $\gamma^2 = 1$, then with probability 1, $X_k$ is a linear combination of the other $p-1$ random variables for some $k$.
v) If $n = 2$, then $\gamma^2 = \rho_{12}^2$, the squared Pearson correlation coefficient.
vi) If $n = 3$, then $\gamma^2 = \rho_{13}^2 + \rho_{23}^2$ if $\rho_{12} = 0$, $\gamma^2 = \rho_{12}^2 + \rho_{23}^2$ if $\rho_{13} = 0$, $\gamma^2 = \rho_{12}^2 + \rho_{13}^2$ if $\rho_{23} = 0$, $\gamma^2 = \rho_{12}^2$ if $\rho_{12} = \rho_{23} = 0$, $\gamma^2 = \rho_{13}^2$ if $\rho_{12} = \rho_{23} = 0$, and $\gamma^2 = \rho_{23}^2$ if $\rho_{12} = \rho_{13} = 0$.
vii) If $\rho_{ij} = 1$, for some $i \neq j$, then $\gamma^2 = 1$ with probability 1.

**Proof.** i) First, $det(\boldsymbol{\rho}_{\boldsymbol{x}}) + \|\boldsymbol{\rho}_{\boldsymbol{x}} - \boldsymbol{I}_p\|_F^2 \geq 0$ and $2det(\boldsymbol{\rho}_{\boldsymbol{x}}) + \|\boldsymbol{\rho}_{\boldsymbol{x}} - \boldsymbol{I}_p\|_F^2 \geq 0$, since the squared norm $\geq 0$ and the symmetric covariance and correlation matrices of $\boldsymbol{x}$ are positive-semidefinite, i.e. $det(\text{Cor}(\boldsymbol{x})) = det(\boldsymbol{\rho}_{\boldsymbol{x}}) \geq 0$ and $det(\text{Cov}(\boldsymbol{x})) \geq 0$.

Second, let's show $2det(\boldsymbol{\rho}_{\boldsymbol{x}}) + \|\boldsymbol{\rho}_{\boldsymbol{x}} - \boldsymbol{I}_p\|_F^2 \neq 0$. This result is true unless both $det(\boldsymbol{\rho}_{\boldsymbol{x}}) = 0$ and $\|\boldsymbol{\rho}_{\boldsymbol{x}} - \boldsymbol{I}_p\|_F^2 = 0$. If $det(\boldsymbol{\rho}_{\boldsymbol{x}}) = 0$, then $det(\boldsymbol{\rho}_{\boldsymbol{x}} - \boldsymbol{I}_p) \neq 0$, and the squared norm $\|\boldsymbol{\rho}_{\boldsymbol{x}} - \boldsymbol{I}_p\|_F^2 \neq 0$. If $\|\boldsymbol{\rho}_{\boldsymbol{x}} - \boldsymbol{I}_p\|_F^2 = 0$ then $\boldsymbol{\rho}_{\boldsymbol{x}} = \boldsymbol{I}_p$, and $det(\boldsymbol{\rho}_{\boldsymbol{x}}) = 1 \neq 0$.

Thus $\gamma^2$ is well defined and

$$0 \le \frac{\|\boldsymbol{\rho}_{\boldsymbol{x}} - \boldsymbol{I}_p\|_F^2}{2det(\boldsymbol{\rho}_{\boldsymbol{x}}) + \|\boldsymbol{\rho}_{\boldsymbol{x}} - \boldsymbol{I}_p\|_F^2} \le 1.$$

ii) If $X_1, ..., X_p$ are pairwise uncorrelated, then $\boldsymbol{\rho}_{\boldsymbol{x}} = \boldsymbol{I}_p$, and $\gamma^2 = 0$.

iii) Without loss of generality, suppose there exist constants $a_1, ..., a_{p-1}$ not all zero, and constant $a_0$ such that $X_p = a_0 + a_1 X_1 + \cdots + a_{p-1} X_{p-1}$. Then $\text{Cov}(X_i, X_p) = \text{Cov}(X_i, \sum_{j=1}^{p-1} a_j X_j) = \sum_{j=1}^{p} a_j \text{Cov}(X_i, X_j)$. This implies that the last column of $\text{Cov}(\boldsymbol{x})$ is a linear combination of the first $p-1$ columns. Thus $det(\text{Cov}(\boldsymbol{x})) = 0$. Since the pairwise correlations exist, each $\sigma_i^2 > 0$, and we have

$$det(\boldsymbol{\rho}_{\boldsymbol{x}}) = det(\boldsymbol{\Delta}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{\Delta}^{-1}) = det(\boldsymbol{\Delta}^{-1})det(\boldsymbol{\Sigma}_{\boldsymbol{x}})det(\boldsymbol{\Delta}^{-1}) = 0.$$

Thus

$$\gamma^2 = \frac{\|\boldsymbol{\rho}_{\boldsymbol{x}} - \boldsymbol{I}_p\|_F^2}{2det(\boldsymbol{\rho}_{\boldsymbol{x}}) + \|\boldsymbol{\rho}_{\boldsymbol{x}} - \boldsymbol{I}_p\|_F^2} = 1.$$

iv) If $\gamma^2 = 1$, then $det(\boldsymbol{\rho}_{\boldsymbol{x}}) = 0$ and there exists $\boldsymbol{a} = (a_1, ..., a_p)^T \neq \boldsymbol{0}$, the zero vector, such that $\boldsymbol{a}^T \text{Cov}(\boldsymbol{x})\boldsymbol{a} = 0$. Let $Y = \boldsymbol{a}^T \boldsymbol{x}$. Then $V(Y) = E(Y - E(Y))^2 = E[(\boldsymbol{a}^T(\boldsymbol{x} - E(\boldsymbol{x})))^2] = E[\boldsymbol{a}^T(\boldsymbol{x} - E(\boldsymbol{x}))(\boldsymbol{a}^T(\boldsymbol{x} - E(\boldsymbol{x}))^T] = \boldsymbol{a}^T E[(\boldsymbol{x} - E(\boldsymbol{x}))(\boldsymbol{x} - E(\boldsymbol{x}))^T]\boldsymbol{a} = \boldsymbol{a}^T \text{Cov}(\boldsymbol{x})\boldsymbol{a} = 0$. $V(Y) = 0$ implies that $P(Y - E(Y) = 0) = 1$, i.e. $1 = P(Y = E(Y)) = P(a_1 X_1 + \cdots + a_p X_p = E(Y)) = 1$. Hence $X_k$ is a linear combination of the other $p-1$ random variables for some $k$.

v) If $\boldsymbol{x} = (X_1, X_2)^T$, then

$$det(\text{Cor}(\boldsymbol{x})) = \begin{vmatrix} 1 & \rho_{12} \\ \rho_{12} & 1 \end{vmatrix} = 1 - \rho_{12}^2.$$

Thus

$$\gamma^2 = \frac{\rho_{12}^2}{1 + \rho_{12}^2 - \rho_{12}^2} = \rho_{12}^2.$$

vi) If $\boldsymbol{x} = (X_1, X_2, X_3)^T$, then

$$det(\text{Cor}(\boldsymbol{x})) = \begin{vmatrix} 1 & \rho_{12} & \rho_{13} \\ \rho_{12} & 1 & \rho_{23} \\ \rho_{13} & \rho_{23} & 1 \end{vmatrix} = 1 + 2\rho_{12}\rho_{13}\rho_{23} - (\rho_{12}^2 + \rho_{13}^2 + \rho_{23}^2).$$

Thus

$$\gamma^2 = \frac{\rho_{12}^2 + \rho_{13}^2 + \rho_{23}^2}{1 + 2\rho_{12}\rho_{13}\rho_{23}},$$

and the result follows.

vii) Without loss of generality, suppose $\rho_{12} = 1$. Then $X_1 = aX_2 + b$, with probability 1, where $a > 0$. Then $V(X_1) = a^2 V(X_2)$ and $\text{Cov}(X_1, X_i) = \text{Cov}(aX_2 + b, X_i) = a\text{Cov}(X_2, X_i)$ for $i = 1, ..., p$. Thus $\rho_{1i} = \rho_{2i}$ for $i = 1, ..., p$. Then $det(\boldsymbol{\rho}_{\boldsymbol{x}}) = det(\text{Cor}(\boldsymbol{x})) = 0$, and $\gamma^2 = 1$ with probability 1. $\square$

Let $\boldsymbol{R} = \boldsymbol{R_x} = (r_{ij})$ be the sample correlation matrix where $r_{ij}$ is the sample correlation of $X_i$ and $X_j$. Then the *sample squared correlation coefficient of the correlation matrix*

$$\hat{\gamma}^2 = \frac{\|\boldsymbol{R} - \boldsymbol{I}_p\|_F^2}{2det(\boldsymbol{R}) + \|\boldsymbol{R} - \boldsymbol{I}_p\|_F^2} = \frac{\sum_{i<j} r_{ij}^2}{det(\boldsymbol{R}) + \sum_{i<j} r_{ij}^2},$$

and $\hat{\gamma}^2$ has the same properties as $\gamma^2$ except $\hat{\gamma}^2 > 0$ unless $\boldsymbol{R} = \boldsymbol{I}_p$. If $\boldsymbol{R}$ is a consistent estimator of $\boldsymbol{\rho_x}$, then $\hat{\gamma}^2$ is a consistent estimator of $\gamma^2$ since $\gamma^2$ is a continuous function of $\boldsymbol{\rho_x}$.

# 3 EXAMPLE AND SIMULATIONS

The nonparametric bootstrap takes a bootstrap sample of size $n$ with replacement from the $n$ cases $\boldsymbol{x}_i$. Compute the sample correlation matrix from the bootstrap sample and obtain $\hat{\gamma}_1^{2*}$. Then repeat $B$ times to get $\hat{\gamma}_1^{2*}, ..., \hat{\gamma}_B^{2*}$.

The shorth confidence intervals (CIs) are useful. Let $T$ be a statistic and $T^*$ the statistic computed from a bootstrap sample. Let $T^*_{(1)}, ..., T^*_{(B)}$ be the order statistics of $T^*_1, ..., T^*_B$. Consider intervals that contain $c$ cases: $(T^*_{(1)}, T^*_{(c)}), (T^*_{(2)}, T^*_{(c+1)}), ..., (T^*_{(B-c+1)}, T^*_{(B)})$. Compute $T^*_{(c)} - T^*_{(1)}, T^*_{(c+1)} - T^*_{(2)}, ..., T^*_{(B)} - T^*_{(B-c+1)}$. Then let the shortest closed interval containing at least $c$ of the $T^*_i$ be

$$\mathrm{shorth(c)} = [\mathrm{T}^*_{(\mathrm{s})}, \mathrm{T}^*_{(\mathrm{s+c-1})}]. \tag{2.1}$$

Let

$$k_B = \lceil B(1-\delta) \rceil \tag{2.2}$$

where $\lceil x \rceil$ is the smallest integer $\geq x$, e.g., $\lceil 7.7 \rceil = 8$. Frey (2013) showed that for large $B\delta$ and iid data, the shorth($k_B$) prediction interval (PI) has undercoverage that depends on the distribution of the data with maximum undercoverage $\approx 1.12\sqrt{\delta/B}$, and used the shorth($c$) estimator as the large sample $100(1-\delta)\%$ PI where

$$c = \min(B, \lceil B[1 - \delta + 1.12\sqrt{\delta/B}\ ]\ \rceil). \tag{2.3}$$

Olive (2014, p. 283) suggested using the shorth as a bootstrap confidence interval. Hall (1988) discussed the shortest bootstrap interval based on all bootstrap samples. We will use the shorth($c$) interval applied to the bootstrap sample $T_i^* = \hat{\gamma}_i^{2*}$ as a large sample confidence interval for $\gamma^2 \in (0,1)$. One sided confidence intervals are used when a lower or upper bound on the parameter $\gamma^2$ is desired, and can be useful if $\gamma^2 = 0$ or 1 is on the boundary of [0,1] of the parameter space of $\gamma^2$. The large sample $100(1-\delta)\%$ lower CI for $\gamma^2$ is $[0, \hat{\gamma}^{2*}_{(c)}]$, while the large sample $100(1-\delta)\%$ upper CI for $\gamma^2$ is $[\hat{\gamma}^{2*}_{(B-c+1)}, 1]$.

For the simulation, suppose that $\boldsymbol{\rho_x} = (\rho_{ij})$ where $\rho_{ij} = \rho$ for $i \neq j$ and $\rho_{ij} = 1$ for $i = j$. Then $det(\boldsymbol{\rho_x}) = (1-\rho)^{p-1}[1 + (p-1)\rho]$. See Graybill (1969, p. 204). Hence

$$\gamma^2 = \frac{p(p-1)\rho^2}{2(1-\rho)^{p-1}[1 + (p-1)\rho] + p(p-1)\rho^2}. \tag{2.4}$$

The simulation simulated iid data $\boldsymbol{w}$ with $\boldsymbol{x} = \boldsymbol{A}\boldsymbol{w}$ and $\boldsymbol{A}_{ij} = \psi$ for $i \neq j$ and $\boldsymbol{A}_{ii} = 1$. Hence $\mathrm{cor}(X_i, X_j) = \rho = [2\psi + (p-2)\psi^2]/[1 + (p-1)\psi^2]$. We used $\boldsymbol{w} \sim N_p(\boldsymbol{0}, \boldsymbol{I}_p)$,

$\boldsymbol{w} \sim (1-\tau)N_m(\boldsymbol{0}, \boldsymbol{I}) + \tau N_m(\boldsymbol{0}, 25\boldsymbol{I})$ with $0 < \tau < 1$ and $\tau = 0.25$ in the simulation, $\boldsymbol{w} \sim LN(\boldsymbol{0}, \boldsymbol{I}_p)$ where the marginals are iid lognormal(0,1), or $\boldsymbol{w} \sim MVT_p(d)$, a multivariate $t$ distribution with $d = 7$ degree of freedom.

**Example 1.** This example used the SASHELP.CARS data set available from the SASHELP library, and this example is useful for illustrating that $\hat{\gamma}^2$ can be used to quickly determine that the variables have a strong linear relationship. This data set has 428 observations and 15 variables. It contains data for make and models of cars, such as miles per gallon, number of cylinders, cost, etc. Four variables, number of cylinders of the car (Cylinders), weight (Weight), length (Length) and gas mileage in the city (MPG_City), are used to check for the possible relationships among them. The squared correlation coefficient $\hat{\gamma}^2 = 0.9640$ suggests that the four variables are highly correlated in some way. The coefficient of determination $R^2 = 0.5893$ implies a moderate linear relationship among the variables when gas mileage in the city is the response variable.

Then we checked the linear association between 7 variable invoice, Horsepower, number of cylinders, gas per mile on highway (MPG_Highway), Weight, Wheelbase, and Engine Size. Then $\hat{\gamma}^2 = 0.9999$. If we took invoice, the number of cylinders, engine size, horsepower, weight, gas per mile on highway (MPG_Highway), wheelbase, as the dependent variable respectively, then we got corresponding coefficients of determination $R^2 = 0.7394$, $R^2 = 0.8828$, $R^2 = 0.8517$, $R^2 = 0.848$, $R^2 = 0.8332$, $R^2 = 0.6881$, and $R^2 = 0.6566$. Note that the $R^2$ varied widely based on the different choice of dependent variable.

# 4 CONCLUSIONS

This paper has given a one number summary of the correlation matrix that acts like a squared correlation. Olive (2016abc) showed that applying certain prediction regions to a bootstrap sample results in confidence regions. Note that $H_0 : \gamma^2 = 0$ is equivalent to $H_0 : \rho_{ij} = 0$ for all $i < j$. There are $r = p(p-1)/2$ such correlations. Olive (2016a, section 5.3.5; 2016c) gives a bootstrap test for $H_0$, but the test needs $n \geq 50r$ and is computationally intensive.

Olive (2016a, Remark 5.8) suggests the following graphical diagnostic. Let $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$ be iid with sample covariance matrix $\boldsymbol{S}_{\boldsymbol{x}}$ where $n \geq 10p$. Let $\boldsymbol{z}_1, ..., \boldsymbol{z}_n$ be the standardized data that has sample mean $\overline{\boldsymbol{z}} = \boldsymbol{0}$ and sample covariance matrix equal to the sample correlation matrix of the $\boldsymbol{x}$: $\boldsymbol{S}_{\boldsymbol{z}} = \boldsymbol{R}_{\boldsymbol{x}}$. Let the squared sample Mahalanobis distance $D_i^2(\boldsymbol{0}, \boldsymbol{A}) = (\boldsymbol{z}_i - \overline{\boldsymbol{z}})^T \boldsymbol{A}^{-1} (\boldsymbol{z}_i - \overline{\boldsymbol{z}}) = \boldsymbol{z}_i^T \boldsymbol{A}^{-1} \boldsymbol{z}_i$ for $i = 1, ..., n$. Plot $D_i(\boldsymbol{0}, \boldsymbol{I}_p)$ on the horizontal axis versus $D_i(\boldsymbol{0}, \boldsymbol{R}_{\boldsymbol{x}})$ on the vertical axis. Add the identity line with unit slope and zero intercept as a visual aid. If $H_0 : \boldsymbol{\rho}_{\boldsymbol{x}} = \boldsymbol{I}_p$ is true, then as $n \to \infty$, the plotted points should cluster tightly about the identity line.

Some tests for independence when $p/n \to \theta$ as $n \to \infty$ are given in Mao (2014), Schott (2005), Srivastava (2005), and Srivastava, Kollo, and von Rosen (2005).

Simulations were done in *R*. See R Core Team (2016). Functions for the simulation are in the collection of functions *mpack.txt* available from (http://lagrange.math.siu.edu/Olive /mpack.txt). The function `shorthLU` gets the shorth($c$) CI, the lower CI, and the upper CI. The function `gsqboot` bootstraps $\hat{\lambda}^2$, while the function `gsqbootsim` does the simulation.

**Acknowledgments**

The author thanks Dr. Olive for suggesting the bootstrap confidence intervals.

# 5 References

Frey, J. (2013), "Data-Driven Nonparametric Prediction Intervals," *Journal of Statistical Planning and Inference*, 143, 1039-1048.

Graybill, F.A. (1969), *Introduction to Matrices with Applications in Statistics*, Wadsworth, Belmont, CA.

Hall, P. (1988), "Theoretical Comparisons of Bootstrap Confidence Intervals," (with discussion), *The Annals of Statistics*, 16, 927-985.

Mao, G. (2014), "A New Test of Independence for High-Dimensional Data," *Statistics & Probability Letters*, 93, 14-18.

Olive, D.J. (2014), *Statistical Theory and Inference*, Springer, New York, NY.

Olive, D.J. (2016a), *Robust Multivariate Analysis*, Springer, New York, NY, to appear.

Olive, D.J. (2016b), "Applications of Hyperellipsoidal Prediction Regions," *Statistical Papers*, to appear.

Olive, D.J. (2016bc), "Bootstrapping Hypothesis Tests and Confidence Regions," preprint, see (http://lagrange.math.siu.edu/Olive/ppvselboot.pdf).

R Core Team (2016), "R: a Language and Environment for Statistical Computing," R Foundation for Statistical Computing, Vienna, Austria, (www.R-project.org).

Schott, J.R. (2005), "Testing for Complete Independence in High Dimensions," *Biometrika*, 92, 951-956.

Srivastava, M.S. (2005), " Some Tests Concerning the Covariance Matrix in High Dimensional Data," *Journal of the Japan Statistical Society*, 35, 251-272.

Srivastava, M.S., Kollo, T., and von Rosen, D. (2005), "Some Tests for the Covariance Matrix with Fewer Observations than the Dimension Under Non-normality," *Journal of Multivariate Analysis*, 102, 1090-1103.