Machine Learning

Collinearity

Irving Gómez Méndez

UNIVERSIDAD
IBEROAMERICANA
CIUDAD DE MÉXICO ®

Introduction

Two variables are collinear if the data vectors representing them lie on the same line. More generally, k variables are collinear if the vectors that represent them lie in a subspace of dimension less than k, that is, if one of the vectors is a linear combination of the others. In practice, such "exact collinearity" rarely occurs, then, two variables are collinear if they lie almost on the same line.

This is equivalent to saying that they have a high correlation between them. We can readily generalize this notion to more than two variables by saying that collinearity exists if there is a high multiple correlation when one of the variables is regressed on the others.

Collinearity has to do with specific characteristics of the data matrix $\mathbf{X}$ and not the statistical aspects of the linear regression model $\mathbf{Y} = \mathbf{X}\beta + \boldsymbol{\varepsilon}$. That is, collinearity is a data problem, not a statistical problem.

Variance Inflation Factors

Analysis of the Correlation Matrix

Examining the correlation matrix of the explanatory variables is a commonly employed procedure. If we assume the $\mathbf{X}$ data **have been centered and scaled** to have unit length, the **correlation matrix** $\mathbf{R}$ is simply $\mathbf{X}^T\mathbf{X}$.

While a high correlation coefficient between two explanatory variables can indeed point to a possible collinearity problem, the absence of high correlations cannot be viewed as evidence of no problem. It is clearly possible for three or more variables to be collinear while no two of the varaibles taken alone are highly correlated. The correlation matrix is wholly incapable of diagnosing such a situation.

Variance Inflation Factor (VIF)

The diagonal elements of $\mathbf{R}^{-1}$, are called the variance inflation factors, VIF_i, and their diagnostic value follows from the relation

$$VIF_i = \frac{1}{1 - R_i^2}$$

where R_i^2 is the coefficient of determination of $\mathbf{X}^i$ regressed on the remaining explanatory variables. The term "variance inflation factor" derives from the fact that the variance of the ith regression coefficient $\sigma^2_{\hat{\beta}_i}$, obeys the relation

$$\sigma^2_{\hat{\beta}_i} = \sigma^2 VIF_i$$

where σ^2 is the variance of the regression disturbance term.

A rule of the thumb is that a VIF bigger than 5 might indicate collinearity.

Relation with Numerical Analysis

It has been suggested to use the eigenvalues of $\mathbf{X}^T\mathbf{X}$ as a key to the presence of collinearity : collinearity is indicated by the presence of a “small” eigenvalue, where small is understood as “small to the other eigenvalues”. This idea comes from the extremely rich literature in numerical analysis showing the relevance of the condition number of a matrix.

Numerical analysts are interested in the properties (conditioning) of a matrix $\mathbf{A}$ of a linear system of equations $\mathbf{Az} = \mathbf{c}$ that allow a solution for $\mathbf{z}$ to be obtained with numerical stability.

The relevance of this to collinearity in least squares is readily apparent, for the least-squares estimator is a solution to the linear system $(\mathbf{X}^T\mathbf{X})\hat{\beta} = \mathbf{X}^T\mathbf{Y}$.

Then, collinearity among the data series of $\mathbf{X}$ results in a matrix $\mathbf{A} = \mathbf{X}^T\mathbf{X}$ whose ill conditioning causes both the solution for b and its variance-covariance matrix to be numerically unstable.

Singular Value Decomposition (SVD)

Singular Value Decomposition (SVD)

Any $n \times p$ matrix $\mathbf{X}$, considered here to be a matrix of n observations on p variables, may be decomposed as

$$\mathbf{X} = \mathbf{UDV}^T,$$

where $\mathbf{U}^T\mathbf{U} = \mathbf{V}^T\mathbf{V} = I_p$ and $\mathbf{D}$ is diagonal with nonnegative diagonal elements μ_k, $k = 1, ..., p$, called the singular values of $\mathbf{X}$.

For the purposes of the collinearity diagnostics, it is always desirable to scale $\mathbf{X}$ to have equal (unit) column lengths. However, if the data are relevant to a model with a constant term, $\mathbf{X}$ should contain uncentered data along with a column of ones; indeed, the use of the centered data matrix $\mathbf{X}$ in this situation is to be avoided, since centering can mask the role of the constant in any underlying near dependencies and produce misleading diagnostic results.

Relation with eigenvalues of $\mathbf{X}^T\mathbf{X}$

Noting that $\mathbf{X}^T\mathbf{X} = \mathbf{V}\mathbf{D}^2\mathbf{V}^T$, we see that $\mathbf{V}$ is an orthogonal matrix that diagonalizes $\mathbf{X}^T\mathbf{X}$ and hence the diagonal elements of $\mathbf{D}^2$, the squares of the singular values, must be the eigenvalues of the real symmetric matrix $\mathbf{X}^T\mathbf{X}$.

As a practical matter, however, there are reasons for preferring the use of the singular-value decomposition.

1. It applies directly to the data matrix $\mathbf{X}$ that is the focus of our concern, and not to the cross-product matrix $\mathbf{X}^T\mathbf{X}$.
2. The notion of a condition number of $\mathbf{X}$ is properly defined in terms of the singular values of $\mathbf{X}$ and not the square roots of the eigenvalues of $\mathbf{X}^T\mathbf{X}$.

3. Whereas the eigensystem and the SVD of a given matrix are mathematically equivalent, computationally they are not.Algorithms exist that allow the singular-value decomposition of $\mathbf{X}$ to be computed with much greater numerical stability than is possible in computing the eigensystem of $\mathbf{X}^T\mathbf{X}$.

Then, the collinearity diagnostics we discuss should always be carried out using the stable algorithm for the singular-value decomposition of $\mathbf{X}$ rather than an algorithm for determining the eigenvalues and eigenvectors of $\mathbf{X}^T\mathbf{X}$.

Exact Linear Dependencies

Let us assume $\mathbf{X}$ has exact linear dependencies among its columns, a case rarely encountered in actual practice, so that $\text{rank}(\mathbf{X}) = r < p$. Since, in the SVD of $\mathbf{X}$, $\mathbf{U}$ and $\mathbf{V}$ are each orthogonal (and hence are necessarily of full rank), we must have $\text{rank}(\mathbf{X}) = \text{rank}(\mathbf{D})$. There will therefore be exactly as many zero elements along the diagonal of $\mathbf{D}$ as the nullity of $\mathbf{X}$, and the SVD in may be partitioned as

$$\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^T = \mathbf{U}\begin{bmatrix} \mathbf{D}_{11} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}\mathbf{V}^T$$

where $\mathbf{D}_{11}$ is $r \times r$ and nonsingular.

Postmultiplying by $\mathbf{V}$ and further partitioning, we obtain

$$\mathbf{X}\begin{bmatrix}\mathbf{V}_1 & \mathbf{V}_2\end{bmatrix} = \begin{bmatrix}\mathbf{U}_1 & \mathbf{U}_2\end{bmatrix}\begin{bmatrix}\mathbf{D}_{11} & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{bmatrix}$$

where $\mathbf{V}_1$ is $p \times r$, $\mathbf{U}_1$ is $n \times r$, $\mathbf{V}_2$ is $p \times (p-r)$, and $\mathbf{U}_2$ is $n \times (p-r)$. This results in the two matrix equations

$$\mathbf{X}\mathbf{V}_1 = \mathbf{U}_1\mathbf{D}_{11} \tag{1}$$

$$\mathbf{X}\mathbf{V}_2 = \mathbf{0} \tag{2}$$

Interest centers on Equation (2), for it displays all of the linear dependencies of $\mathbf{X}$. The $p \times (p-r)$ matrix $\mathbf{V}_2$, provides an orthonormal basis for the null space associated with the columns of $\mathbf{X}$.

Near Linear Dependencies

If ,then, $\mathbf{X}$ possessed $p - r$ exact linear relations among its columns (and computers possessed exact arithmetic), there would also be exactly $p - r$ zero singular values in $\mathbf{D}$, and the variables involved in each of these dependencies would be determined by the nonzero elements of $\mathbf{V}_2$.

In most statistical applications, the interrelations among the columns of X are not exact dependencies, and computers deal in finite, not exact, arithmetic. Nevertheless, each near linear dependence among the columns of X will manifest itself in a small singular value, a small μ. The question now is to determine what is small, we are greatly aided in answering this question by the notion of condition number of a matrix $\mathbf{X}$.

Condition Indexes

Condition Index

A means for defining the conditioning of a matrix that accords somewhat with intuition is afforded by the singular-value decomposition.

The familiar Euclidean norm of an n-vector $\mathbf{z}$, denoted $\|\mathbf{z}\|$, is defined as

$$\|\mathbf{z}\| = (\mathbf{z}^T\mathbf{z})^{1/2}.$$

An important generalization of the Euclidean norm to an $n \times n$ matrix $\mathbf{A}$ is the spectral norm, denoted $\mathbf{A}$ and defined as

$$\|\mathbf{A}\| = \sup_{\|\mathbf{z}\|=1} \|\mathbf{A}\mathbf{z}\|.$$

It is readily shown that $\|\mathbf{A}\| = \mu_{max}$, that is, the maximal singular value of $\mathbf{A}$. Similarly, if $\mathbf{A}$ is square, $\|\mathbf{A}^{-1}\| = 1/\mu_{min}$. Further, like the Euclidean norm, the spectral norm can be shown to be a true norm; that is, it possesses the following properties:

1. $\|\lambda\mathbf{A}\| = |\lambda| \, \|\mathbf{A}\|$ for all real λ and all $\mathbf{A}$.
2. $\|\mathbf{A}\| = 0$ if and only if $\mathbf{A} = \mathbf{0}$, the matrix of zeros.
3. $\|\mathbf{A} + \mathbf{B}\| \leq \|\mathbf{A}\| + \|\mathbf{B}\|$ for all $m \times n$ matrices $\mathbf{A}$ and $\mathbf{B}$.

And, in addition, the spectral norm obeys the following relations:

4. $\|\mathbf{Az}\| \leq \|\mathbf{A}\| \, \|\mathbf{z}\|$.
5. $\|\mathbf{AB}\| \leq \|\mathbf{A}\| \, \|\mathbf{B}\|$ for all commensurate $\mathbf{A}$ and $\mathbf{B}$.

Analysis of Linear Systems

We shall now see that the spectral norm is directly relevant to an analysis of the conditioning of a linear system of equations $\mathbf{Az} = \mathbf{c}$. In the event that $\mathbf{A}$ is fixed but $\mathbf{c}$ changes by $\delta\mathbf{c}$, we have $\delta\mathbf{z} = \mathbf{A}^{-1}\delta\mathbf{c}$, or

$$\left\|\delta\mathbf{z}\right\| \leq \left\|\mathbf{A}^{-1}\right\| \left\|\delta\mathbf{c}\right\| .$$

Further, employing property 4 above to the equation system, we have

$$\left\|\mathbf{c}\right\| \leq \left\|\mathbf{A}\right\| \left\|\mathbf{z}\right\| ;$$

and from multiplying these last two expressions we obtain

$$\frac{\left\|\delta\mathbf{z}\right\|}{\left\|\mathbf{z}\right\|} \leq \left\|\mathbf{A}\right\| \left\|\mathbf{A}^{-1}\right\| \frac{\left\|\delta\mathbf{c}\right\|}{\left\|\mathbf{c}\right\|}.$$

A similar result holds for perturbations in the elements of the matrix **A**. Here it can be shown that

$$\frac{\|\delta \mathbf{z}\|}{\|\mathbf{z} + \delta \mathbf{z}\|} \leq \|\mathbf{A}\| \left\|\mathbf{A}^{-1}\right\| \frac{\|\delta \mathbf{A}\|}{\|\mathbf{A}\|}.$$

The magnitude $\|\mathbf{A}\| \left\|\mathbf{A}^{-1}\right\|$ is defined to be the **condition number** of the nonsingular matrix **A** and is denoted as $\kappa(\mathbf{A})$.

The concept of condition index is readily extended to a rectangular matrix and can be calculated without recourse to an inverse.

From the SVD, $\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^T$, it is easily shown that the generalized inverse $\mathbf{X}^+$ of $\mathbf{X}$ is$\mathbf{V}\mathbf{D}^+\mathbf{U}^T$, where $\mathbf{D}^+$ is the generalized inverse of $\mathbf{D}$ and is simply $\mathbf{D}$ with its nonzero diagonal elements inverted. Hence the singular values of $\mathbf{X}^+$ are merely the reciprocals of those of $\mathbf{x}$, and the maximal singular value of $\mathbf{X}^+$ is the reciprocal of the minimum (nonzero) singular value of $\mathbf{x}$.

Thus for any $n \times p$ matrix $\mathbf{X}$ we may define the condition number to be

$$\kappa(\mathbf{X}) = \frac{\mu_{max}}{\mu_{min}} \geq 1.$$

kth Condition Index

Define

$$\eta_k \equiv \frac{\mu_{max}}{\mu_k}, k = 1, \ldots, p$$

to be the kth condition index of the $n \times p$ data matrix $\mathbf{X}$. The largest value for η_k is also the condition number of the given matrix. Therefore there are as many near dependencies among the columns of a data matrix $\mathbf{X}$ as there are high condition indexes.

Weak dependencies are associated with condition indexes around 5 to 10, whereas moderate to strong relations are associated with condition indexes of 30 to 100.

Variance Decomposition

Variance-Decomposition Proportion

Using the SVD, $\mathbf{X} = \mathbf{UDV}^T$, the variance of matrix of $\hat{\beta}$, $\mathbb{V}\left(\hat{\beta}\right)$, may be written as

$$\mathbb{V}\left(\hat{\beta}\right) = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1} = \sigma^2\mathbf{V}\mathbf{D}^{-2}\mathbf{V}^T,$$

or, for the kth component of $\hat{\beta}$,

$$\mathbb{V}\left(\hat{\beta}_k\right) = \sigma^2\sum_{j=1}^{p}\frac{v_{kj}^2}{\mu_j^2}.$$

Define the k, jth variance-decomposition proportion as the proportion of the variance of the kth regression coefficient associated with the jth component of its decomposition.

Calculate of the Variance-Decomposition Proportions

These proportions are readily calculated as follows. First let

$$\phi_{kj} \equiv \frac{v_{kj}^2}{\mu_j^2} \quad \text{and} \quad \phi_k \equiv \sum_{j=1}^{p} \phi_{kj} \quad k = 1, \ldots, p.$$

Then, the variance-decomposition proportions are

$$\pi_{jk} \equiv \frac{\phi_{kj}}{\phi_k}, \quad k, j = 1, \ldots, p.$$

It is usefull to present this analysis in a summery table, sometimes called the Π matrix.

Associated Singular Value	Proportions of				Condition Index
	$\mathbb{V}\left(\hat{\beta}_1\right)$	$\mathbb{V}\left(\hat{\beta}_1\right)$	$\cdots$	$\mathbb{V}\left(\hat{\beta}_p\right)$	
μ_1	π_{11}	π_{12}	$\cdots$	π_{1p}	η_1
μ_2	π_{21}	π_{22}	$\cdots$	π_{2p}	η_2
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
μ_p	π_{p1}	π_{p2}	$\cdots$	π_{pp}	η_p

Near Orthogonality

Note that $v_kj = 0$ when the columns k and j of $\mathbf{X}$ belong to mutually orthogonal partitions. Thus we see that the bad effect of collinearity, resulting in relatively small μ's, may be mitigated for some coefficients by near orthogonality, resulting in small v_{kj}'s.

At Least Two Variables Must Be Involved

Since two or more variables are required to create a near dependency, it must be that two or more variables are adversely affected by high variance-decomposition proportions associated with a single singular value.

To illuminate this, consider a data matrix **X** with mutuallly orthogonal columns. Hence the associated Π matrix of variance-decomposition proportions must take the following form:

Associated Singular Value	Proportions of				Condition Index
	$\mathbb{V}\left(\hat{\beta}_1\right)$	$\mathbb{V}\left(\hat{\beta}_1\right)$	$\cdots$	$\mathbb{V}\left(\hat{\beta}_p\right)$	
μ_1	1	0	$\cdots$	0	η_1
μ_2	0	1	$\cdots$	0	η_2
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
μ_p	0	0	$\cdots$	1	η_p