

Multivariate Regression, leading up to Component Regression/Analysis to Partial Least Squares
-- an introductory tutorial to some of the most important ideas in multivariate regression.

Instructor: Nam Sun Wang

...continued...

It would be nice if we can somehow decide which component to take away first so that most of the variation in X can be captured with just a few components. With the salad analogy, if a chicken salad contains mostly lettuce and chicken meat, it does not make sense to try to describe the salad with components it has very little of or, worse, with components it does not have. A good description of the salad will probably start with its lettuce and chicken content. A significant problem with the normal equation is that when the number of independent variable is large and highly correlated, the inverse of $X^T X$, i.e., $(X^T X)^{-1}$, becomes highly unstable or nonexistent. Output from array sensors, including spectral data, are highly correlated among neighboring array elements. The salary of a football player is determined largely by his performance, which, in turn, are based on a set of highly correlated numbers that define his performance. The consumer appeal of a gourmet dish (Y) depends on hundreds of individual chemical ingredients, e.g., level of sugar, msg, salt, vinegar, pepper, etc. (X). However, these independent variables in a given set of samples are usually correlated. The dish's seasoning may come from a combination of several basic, commercially readily available ingredients (e.g., mayonnaise, ketchup, soup stock, mustard, Tabasco sauce, salad dressing, etc.). We try to reduce the number of variables by choosing a good combination of the independent variables. In another words, we want to choose a set of basis vectors/functions wisely. Instead of describing the given data in terms of the original set of independent variables, we try to describe the same system with a (smaller) set of basis vectors/functions. In the the gourmet dish example, we try to describe the make up of a dish not necessarily by its chemical composition, but perhaps in terms of the amount of a few major sauces that the chef has access to. Mathematically, if we know very little about what to choose as the basis vectors/functions, we can let the computer decide which combination explains away most of the variations present in the given set of independent vectors. These vectors are called **principal components** because they represent the major components that make up the system. These vectors are also called **latent variables** because they tells us the hidden underlying structure in a given data set. This is accomplished by **principal component analysis (PCA)**, and we refer to regression based on principal components as **principal component regression (PCR)**. These mathematically derived variables do not necessarily match the physical reality. For example, given the chemical content of a set of dishes, we may analyze the given data mathematically and arrive at several principal components. However, these mathematical components do not necessarily reflect the actual composition of the different sauces used to prepare the dishes. Nevertheless, principal component regression is valuable when the independent variables are not completely independent.

Principal Component Regression (PCR).

Instead of performing linear regression with the straight original set of independent variables $x^{<0>}$, $x^{<1>}$, and $x^{<2>}$, we try to find a vector that is a linear combination of the original independent variables. This 0th principal vector is the eigenvector that corresponds to the largest eigenvalue of the covariance matrix $X^T \cdot X$. It explains the most amount of variation contained in the data matrix X . In summary, the steps for principal component regression are:

- Step 0. Provide X and Y vectors.
- Step 1. Mean-center X and Y . If necessary, normalize the columns of X and Y with the respective covariance (variance-scaling).
- Step 2. Find the basis vectors v by examining X . These are the eigenvectors of $X^T \cdot X$.
- Step 3. Find the portion of the X matrix that lies along the basis vectors v . $\text{score} = X \cdot v$
Apply the scalar normal equation to find regression coefficient a :
 $a = (\text{score}^T \text{score})^{-1} \text{score}^T Y$.
- Step 4. Compute the residual of X and Y .
- Step 5. The regression equation is: $y_{\text{regress}} = \text{score} \cdot a = X \cdot v \cdot a$.
(Take care of mean-centering and variance-scaling as needed.)
- Step 6. Steps 2-5 are performed with one basis vector at a time.
Repeat Steps 2-5 for each additional basis vector.

Below, we will numerically demonstrate these steps.

Step 0. Generate X and Y Data. The first two independent variables $x^{<0>}$ and $x^{<1>}$ are mostly dependent, $x^{<0>}$ being 10 times of $x^{<1>}$.

Number of points: $N := 50$ $i := 0..N$

Dimension: $m := 2$ $j := 0..m$

$$X^{<i>} := (\text{rnd}(1) - 0.5) \cdot \begin{pmatrix} 10 \\ 1 \\ 0 \end{pmatrix} + (\text{rnd}(1) - 0.5) \cdot \begin{pmatrix} 0 \\ 0.001 \\ 0 \end{pmatrix} + (\text{rnd}(1) - 0.5) \cdot \begin{pmatrix} 0 \\ 0 \\ 0.1 \end{pmatrix} \quad X := X^T$$

↑ Without this small noise term, $x^{<0>}$ and $x^{<1>}$ are completely dependent and $X^T \cdot X$ is singular.

$$Y_i := X_{i,0} + 3 \cdot X_{i,1} + (\text{rnd}(0.1) - 0.05) \quad y(x) := x \cdot \begin{pmatrix} 1 \\ 3 \\ 0 \end{pmatrix}$$

Step 1. Mean-Centering: $X_{\text{save}} := X$ $Y_{\text{save}} := Y$ (save them for use later.)

$$x_{\text{mean},j} := \text{mean}(X^{<j>}) \quad x_{\text{mean}} := x_{\text{mean}}^T \quad X^{<j>} := X^{<j>} - x_{\text{mean},j}$$

$$x_{\text{mean}} = (-0.318 \quad -0.032 \quad 0.003)$$

$$y_{\text{mean}} := \text{mean}(Y) \quad Y := Y - y_{\text{mean}} \quad y_{\text{mean}} = -0.42$$

Step 2. Find the eigenvalues and eigenvectors of the covariance matrix.

Covariance matrix:

$$\mathbf{X}^T \cdot \mathbf{X} = \begin{pmatrix} 403.701 & 40.366 & 1.346 \\ 40.366 & 4.036 & 0.135 \\ 1.346 & 0.135 & 0.039 \end{pmatrix} \quad \left| \mathbf{X}^T \cdot \mathbf{X} \right| = 4.369 \cdot 10^{-5} \quad \leftarrow \text{Almost singular.}$$

$$\text{conde}(\mathbf{X}^T \cdot \mathbf{X}) = 1.313 \cdot 10^8$$

Eigenvalues/eigenvectors:

$$\lambda := \text{reverse}(\text{sort}(\text{eigenvals}(\mathbf{X}^T \cdot \mathbf{X}))) \quad \lambda^T = (407.741 \quad 0.034 \quad 3.106 \cdot 10^{-6})$$

↑ One of the eigenvalues is almost 0, again, indicating that $\mathbf{X}^T \cdot \mathbf{X}$ is almost singular.

$$v^{<j>} := \text{eigenvec}(\mathbf{X}^T \cdot \mathbf{X}, \lambda_j) \quad v = \begin{pmatrix} 0.995 & -0.003 & -0.1 \\ 0.099 & -0.002 & 0.995 \\ 0.003 & 1 & 0.002 \end{pmatrix}$$

The 0th eigenvector $v^{<0>}$ successfully recovers the vector (1 0.1 0) originally used to generate X, and the 1st eigenvector $v^{<1>}$ captures the vector (0 0 1). The percentage of variation in the X matrix captured with the 0th eigenvector alone is:

$$\frac{\lambda_0}{\sum \lambda} = 99.992 \cdot \% \quad \dots \text{ in terms of squared (variance) quantities because the eigenvalues are calculated from the covariance matrix } \mathbf{X}^T \cdot \mathbf{X}.$$

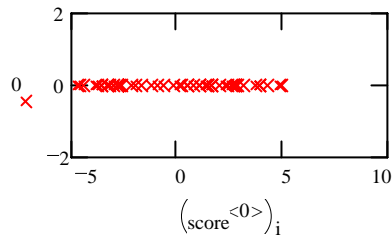
$$\sqrt{\frac{\lambda_0}{\sum \lambda}} = 99.996 \cdot \% \quad \dots \text{ in terms of the un-squared (standard deviation) quantities.}$$

Check that eigenvectors are orthonormal. Because the covariance matrix $\mathbf{X}^T \cdot \mathbf{X}$ is real and symmetric, we have a complete set of the eigenvectors even when eigenvalues are not distinct, and these eigenvectors must be mutually orthogonal. Furthermore, because the covariance matrix $\mathbf{X}^T \cdot \mathbf{X}$ is positive semi-definite, all the eigenvalues must be non-negative. Because Mathcad normalizes the eigenvector to be a unit length, the eigenvectors calculated by Mathcad are mutually orthonormal.

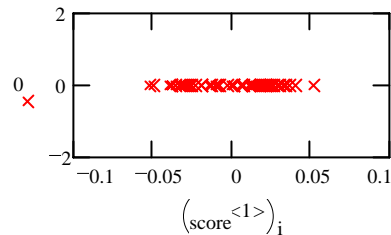
$$v \cdot v^T = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \quad v^T \cdot v = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

Component of X in the $v^{<0>}$ direction is X projected along the $v^{<0>}$ direction, i.e., $X \cdot v^{<0>}$. Likewise, component of X in the $v^{<1>}$ direction is X projected along the $v^{<1>}$ direction. The directional vector is called **loading**, and the projection of X in that direction (i.e., the length) is called **score**. Remember that a vector has both direction (loading) and length (score).

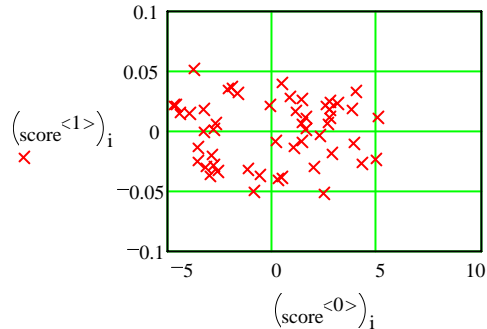
$$\text{score: } \text{score}^{<0>} := X \cdot v^{<0>}$$



$$\text{score}^{<1>} := X \cdot v^{<1>}$$



Scatter plot in the $v^{<0>}$ and $v^{<1>}$ plane:



Check: $X^T \cdot \text{score}^{<0>} = X^T \cdot X \cdot v^{<0>} = \lambda_0 \cdot v^{<0>}$

$$\left| X^T \cdot \text{score}^{<0>} \right| = 407.741 \quad \leftarrow \text{Compare} \rightarrow \quad \lambda_0 = 407.741$$

Check: Conversely, the loading v is the projection of X onto the score vector.

$$\frac{X^T \cdot \text{score}^{<0>}}{\left| X^T \cdot \text{score}^{<0>} \right|} = \begin{pmatrix} 0.995 \\ 0.099 \\ 0.003 \end{pmatrix} \quad \leftarrow \text{Compare} \rightarrow \quad v^{<0>} = \begin{pmatrix} 0.995 \\ 0.099 \\ 0.003 \end{pmatrix}$$

Step 3. Regression of the dependent variable against the 0th principal component score. Note that we are calculating only one (scalar) regression coefficient. Because we break the original matrix X into principal components, regression is performed one component at a time. If we employ the normal formula for linear regression, there is no inverse of a matrix, only the inverse of a scalar. Thus, we avoid the problem that arises from having to invert a nearly singular matrix with a large condition number.

$$a_0 := \text{slope}(\text{score}^{<0>}, Y) \quad a_0 = 1.295$$

Check: $\text{intercept}(\text{score}^{<0>}, Y) = 0$ \leftarrow This is expected because we have already mean-centered X and Y .

Step 4. Residual matrices. Take away the 0th principal component from X; E contains the residual after the 0th principal component has been subtracted.

$$E := X - \text{score}^{<0>} \cdot v^{<0>T} \quad \text{Note that the X vector associated with score}^{<0>} \text{ is: } \text{score}^{<0>} \cdot v^{<0>T}$$

Take away the part of Y that has already been captured by the 0th principal component.

$$\text{sse}_{\text{old}} := Y \cdot Y \quad \text{sse}_{\text{old}} = 684.34$$

$$F := Y - \text{score}^{<0>} \cdot a_0 \quad \text{Note that the Y vector captured by score}^{<0>} \text{ is: } \text{score}^{<0>} \cdot a_0$$

$$\text{sse} := F \cdot F \quad \text{sse} = 0.042 \quad \text{Compared to the original sse before taking away the 0th principal component, there is not much left. A drastic decrease in the sse tells us that most of the variation in Y has been captured.}$$

We can also look at the correlation coefficient or the r^2 value to see how much information has been captured.

$$r^2 := \frac{\text{sse}_{\text{old}} - \text{sse}}{\text{sse}_{\text{old}}} \quad r^2 = 99.994\%$$

$$r := \sqrt{r^2} \quad r = 99.997\%$$

Note that these numbers are different from those calculated from the eigenvalues ($\lambda_0/\Sigma\lambda$). The former is a measure of the amount of information in Y captured by the regression with the 0th principal component; whereas, the latter indicates the amount of variation in X captured by the 0th principal component. We can probably stop at this point with just one component.

Step 5. Regression Model. The regression model is:

$$y_{\text{regress}}(x) := (x - x_{\text{mean}}) \cdot v^{<0>} \cdot a_0 + y_{\text{mean}}$$

Equivalently,

$$y_{\text{regress}}(x) := x \cdot v^{<0>} \cdot a_0 + (y_{\text{mean}} - x_{\text{mean}} \cdot v^{<0>} \cdot a_0)$$

Check: Let us examine the slope and intercept.

$$\text{Intercept: } y_{\text{mean}} - x_{\text{mean}} \cdot v^{<0>} \cdot a_0 = -0.006 \quad \text{which is practically 0; thus, } y=x \cdot \text{slope}$$

$$\text{slope: } v^{<0>} \cdot a_0 = \begin{pmatrix} 1.289 \\ 0.129 \\ 0.004 \end{pmatrix} \quad \leftarrow \text{compare} \rightarrow \quad y(x) := x \cdot \begin{pmatrix} 1 \\ 3 \\ 0 \end{pmatrix}$$

Comparing the regression slope to that in the original generating function $y(x)$, which is reproduced above, we see that there is apparently a difference in the formula. However, if we consider that the calibration data have correlated $x^{<0>}$ and $x^{<1>}$ (i.e., $x^{<0>} = 10 \cdot x^{<1>}$), the slope calculated from the regression coefficient $v^{<0>} \cdot a_0$ essentially gives:

$$y = (1.287 + 0.1 \cdot 0.129) \cdot x^{<0>} = 1.300 \cdot x^{<0>}$$

which is exactly the same as in the generating equation $y(x)$. Thus, we can claim that we have successfully captured the underlying structure.

Examples:

$$y_{\text{regress}}((5 \ 0.5 \ 0.05)) = 6.504 \quad y((5 \ 0.5 \ 0.05)) = 6.5 \quad \leftarrow \text{O.K.}$$

$$y_{\text{regress}}((5 \ -0.5 \ 0.05)) = 6.375 \quad y((5 \ -0.5 \ 0.05)) = 3.5 \quad \leftarrow \text{significantly off -- extrapolation.}$$

The regression function is valid for only a one-point argument. We probably have to utilize the programming feature in Mathcad Version 6.0 to allow multiple number of points. For now, we first create a one-matrix I that has the same number of rows as the input to match the size between X_{test} and x_{mean} . Mathcad will complain about "array size mismatch" if we do not take care of it.

$$X_{\text{test}} := \begin{pmatrix} 5 & 0.5 & 0.05 \\ 5 & -0.5 & 0.05 \end{pmatrix} \quad \text{index} := 0 \dots \text{rows}(X_{\text{test}}) - 1 \quad I_{\text{index}} := 1 \quad I = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

$$y_{\text{regress}}(x) := (x - I \cdot x_{\text{mean}}) \cdot v^{<0>} \cdot a_0 + y_{\text{mean}}$$

$$y_{\text{regress}}(X_{\text{test}}) = \begin{pmatrix} 6.504 \\ 6.375 \end{pmatrix} \quad \leftarrow \text{compare} \rightarrow \quad y(X_{\text{test}}) = \begin{pmatrix} 6.5 \\ 3.5 \end{pmatrix}$$

Example -- Similarity Transform. Here we treat principal component analysis as a form of similarity transformation. We transform the original given independent covariance matrix $X^T \cdot X$ by the nonsingular transformation matrix v such that:

$$(X^T \cdot X) \cdot v = v \cdot (Z^T \cdot Z) \quad Z^T \cdot Z = v^{-1} \cdot (X^T \cdot X) \cdot v$$

However, this does not solve the problem of having to invert $X^T \cdot X$ in multiple linear regression, because similar matrices have the same eigenvalues. That is $Z^T \cdot Z$ has exactly the same set of eigenvalues as $X^T \cdot X$, which is close to being singular.

$$\text{Check: } \lambda_z := \text{reverse} \left[\text{sort} \left[\text{eigenvals} \left[v^{-1} \cdot (X^T \cdot X) \cdot v \right] \right] \right]$$

$$\lambda_z = \begin{bmatrix} 407.741 \\ 0.034 \\ 3.106 \cdot 10^{-6} \end{bmatrix} \quad \text{eigenvalues of } Z^T \cdot Z \leftrightarrow \text{eigenvalues of } X^T \cdot X \quad \lambda = \begin{bmatrix} 407.741 \\ 0.034 \\ 3.106 \cdot 10^{-6} \end{bmatrix}$$

Perform linear regression on transformed independent variable Z. This is another way to view principal component regression. This is also the same as change of basis vectors.

Form Z:

$$Z := X \cdot v$$

Linear Regression between Z and Y: $a = 1.295$... the value from principal component regression based on X.

$$a := (Z^T \cdot Z)^{-1} \cdot Z^T \cdot Y \quad a = \begin{pmatrix} 1.295 \\ -0.087 \\ 5.058 \end{pmatrix} \quad \dots \text{the value from regression with transformed Z.}$$

\leftarrow compare the 0th element of a to the value above.

Regression Model:

$$y_{\text{regress}}(x) := x \cdot v \cdot a$$

An example:

$$y_{\text{regress}}(X_{\text{test}}) = \begin{pmatrix} 6.508 \\ 1.345 \end{pmatrix} \quad \leftarrow \text{compare} \rightarrow \quad y(X_{\text{test}}) = \begin{pmatrix} 6.5 \\ 3.5 \end{pmatrix}$$

To avoid singularity problem, we can perform regression by increasing the number of basis vectors one at a time until most variations in Y have been captured by the regression model. This is similar to fitting given pairs of (x,y) points with power series (1, x, x², x³, ...) by adding one term at a time until the the coefficient of correlation is close to 100%.

Regression for the given number of terms. -- **All terms are calculated simultaneously with matrix inverse.**

$$a := 0 \quad a^{<j>} := \left(\text{submatrix}(Z, 0, N, 0, j)^T \cdot \text{submatrix}(Z, 0, N, 0, j) \right)^{-1} \cdot \text{submatrix}(Z, 0, N, 0, j)^T \cdot Y$$

$$a = \begin{pmatrix} 1.295 & 1.295 & 1.295 \\ 0 & -0.087 & -0.087 \\ 0 & 0 & 5.058 \end{pmatrix} \quad \leftarrow \text{Note that since the basis vectors, which are based on orthogonal eigenvectors, are orthogonal, the regression coefficient } a \text{ does not change as we add more terms.}$$

Regression model.

$$y_{\text{regress}}(x, \text{order}) := \sum_{j=0}^{\text{order}} x \cdot v^{<j>} \cdot (a^{<\text{order}>})_j$$

Goodness of fit.

$$\text{sse}(\text{order}) := (Y - y_{\text{regress}}(X, \text{order})) \cdot (Y - y_{\text{regress}}(X, \text{order}))$$

$$\text{sse}(0) = 0.042 \quad \leftarrow \text{There is no more change in sse after the 0th term. Thus, only the 0th term is needed.}$$

$$\text{sse}(1) = 0.042$$

$$\text{sse}(2) = 0.042$$

Because the transformed independent variable is now expressed with orthogonal basis vectors, we can perform regression based on the transformed Z one column at a time.

Regression for the given number of terms. -- **Each term is calculated separately with only scalar inverse.** The ability to do this is the major difference between the plain multiple linear regression and principal component regression.

$$a := 0 \quad a_j := \frac{1}{Z^{<j>} \cdot Z^{<j>}} \cdot Z^{<j>} \cdot Y \quad a = \begin{pmatrix} 1.295 \\ -0.087 \\ 5.058 \end{pmatrix} \quad \leftarrow \text{The coefficients are identical to the case when the matrix inverse is involved.}$$

Regression model.

$$y_{\text{regress}}(x, \text{order}) := \sum_{j=0}^{\text{order}} x \cdot v^{<j>} \cdot a_j$$

Goodness of fit.

$$\text{sse}(\text{order}) := (Y - y_{\text{regress}}(X, \text{order})) \cdot (Y - y_{\text{regress}}(X, \text{order}))$$

$$\text{sse}(0) = 0.042$$

$$\text{sse}(1) = 0.042$$

Example. Here is an example where the dependence of Y on X is a bit different from the last example. Since the X variable is unchanged from the last example, we do not need to recalculate the eigenvalues and eigenvectors of $X^T \cdot X$.

$$Y_i := X_{\text{save}_{i,0}} + 3 \cdot X_{\text{save}_{i,1}} + 100 \cdot X_{\text{save}_{i,2}} + (\text{rnd}(0.1) - 0.05)$$

↑ We added a nontrivial degree of dependence on $x_{<2>}$.

$$y(x) := x \cdot \begin{pmatrix} 1 \\ 3 \\ 100 \end{pmatrix}$$

Mean-centering: $y_{\text{mean}} := \text{mean}(Y) \quad y_{\text{mean}} = -0.1$

$$Y := Y - y_{\text{mean}}$$

Linear regression. $a := 0 \quad a_0 := \text{slope}(\text{score}^{<0>}, Y) \quad a_0 = 1.627$

Check: $\text{intercept}(\text{score}^{<0>}, Y) = 0$

Residual matrices: $E := X - \text{score}^{<0>} \cdot v^{<0>T}$

$$\text{sse}_{\text{old}} := Y \cdot Y \quad \text{sse}_{\text{old}} = 1.425 \cdot 10^3$$

$$F := Y - \text{score}^{<0>} \cdot a_0$$

$$\text{sse} := F \cdot F \quad \text{sse} = 346.037$$

Regression coefficient of correlation:

$$r2 := \frac{\text{sse}_{\text{old}} - \text{sse}}{\text{sse}_{\text{old}}} \quad r2 = 75.719 \cdot \%$$

$$r := \sqrt{r2} \quad r = 87.016 \cdot \%$$

It is clear from the r^2 value that there is still a significant amount of information (25%) in Y that remains to be explained. We repeat the regression procedure with the next principal component. In this problem, we have already calculated all the three eigenvalues and eigenvectors earlier. We wish to point out that this is not always desirable, because the dimension may be very large in certain applications (e.g., spectral data analysis). Furthermore, an impetus for resorting to principal component regression is precisely that we want to reduce the dimension of a given problem, in which case we do not even want to compute the un-needed eigenvalues/eigenvectors. Thus, it is common to pursue a sequential algorithm where we compute only one eigenvalue and one eigenvector at each regression step. In a sequential procedure, we calculate the eigenvalue and eigenvector based on the residual matrix E at each step, and these should be identical to the ones calculated based on the original matrix X. Because the eigenvectors are orthogonal, we can take away one component at a time from X without affecting what has already been done up to that point. We will take a detour to demonstrate this point. Had the eigenvectors not been orthogonal, we would have to calculate all the eigenvalues and eigenvectors at the beginning of the regression procedure, and a major advantage of the principal component regression would be lost.

The next eigenvalue and eigenvector calculated based on X are:

$$\lambda_1 = 0.034 \quad v^{<1>} = \begin{pmatrix} -0.003 \\ -0.002 \\ 1 \end{pmatrix} \quad \leftarrow \text{Compare this to the 0th column of } v_E.$$

These are the same as the largest eigenvalue and the corresponding eigenvector calculated based on the residual E.

$$\lambda_E := \text{reverse}(\text{sort}(\text{eigenvals}(E^T \cdot E))) \quad \lambda_E^T = (0.034 \quad 3.106 \cdot 10^{-6} \quad 0)$$

$$v_E^{<j>} := \text{eigenvec}(E^T \cdot E, \lambda_{E_j})$$

$$v_E = \begin{pmatrix} -0.003 & -0.1 & 0.995 \\ -0.002 & 0.995 & 0.099 \\ 1 & 0.002 & 0.003 \end{pmatrix} \quad \leftarrow \text{compare the last column of } v_E \text{ to the 0th principal component } v^{<0>} \rightarrow v^{<0>} = \begin{pmatrix} 0.995 \\ 0.099 \\ 0.003 \end{pmatrix}$$

Note that the last eigenvalue of the residual E is 0 because that component, having already been taken away, is absent in E. And the eigenvector that corresponds to the eigenvalue of 0 (i.e., the last column of v_E) is identical to the 0th principal component $v^{<0>}$.

We repeat the regression. $a_1 := \text{slope}(\text{score}^{<1>}, F)$ $a_1 = 100.152$

$$\text{Check: } \text{intercept}(\text{score}^{<1>}, F) = 0$$

Residual matrices: $E := E - \text{score}^{<1>} \cdot v^{<1>T}$

$$F := F - \text{score}^{<1>} \cdot a_1$$

$$\text{sse} := F \cdot F \quad \text{sse} = 0.047$$

Regression coefficient of correlation:

$$r2 := \frac{\text{sse}_{\text{old}} - \text{sse}}{\text{sse}_{\text{old}}} \quad r2 = 99.997\%$$

$$r := \sqrt{r2} \quad r = 99.998\%$$

We see that we have captured almost all the information in Y with two components. If this were not the case, we would iterate the regression process. The regression equation is:

$$V^{<0>} := v^{<0>} \quad V^{<1>} := v^{<1>} \quad \leftarrow \text{This step would not be needed in Mathcad, had we calculated only those eigenvectors that were needed.}$$

$$y_{\text{regress}}(x) := (x - x_{\text{mean}}) \cdot V \cdot a + y_{\text{mean}}$$

Check: Let us examine the slope and intercept.

Intercept: $y_{\text{mean}} - x_{\text{mean}} \cdot V \cdot a = -0.007$ which is practically 0; thus, $y=x \cdot \text{slope}$

slope: $V \cdot a = \begin{pmatrix} 1.309 \\ -0.079 \\ 100.156 \end{pmatrix} \leftarrow \text{compare} \rightarrow y(x) := x \cdot \begin{pmatrix} 1 \\ 3 \\ 100 \end{pmatrix}$

Considering that the calibration data have correlated $x^{<0>}$ and $x^{<1>}$, we conclude that the slope calculated from the regression coefficient $V \cdot a$ essentially matches that from $y(x)$.

Examples:

$y_{\text{regress}}((5 \ 0.5 \ 0.05)) = 11.506 \leftrightarrow y((5 \ 0.5 \ 0.05)) = 11.5 \leftarrow \text{O.K.}$

$y_{\text{regress}}((5 \ -0.5 \ 0.05)) = 11.585 \leftrightarrow y((5 \ -0.5 \ 0.05)) = 8.5 \leftarrow \text{significantly off}$
 -- extrapolation.

Example. Here, we shall consider yet another example where the dependent variable Y depends most heavily on $x^{<2>}$ and to a minor extent on $x^{<1>}$. Again, since the X variable is unchanged from the last example, we do not need to recalculate the eigenvalues and eigenvectors of $X^T \cdot X$.

$$Y_i := 0.1 \cdot X_{\text{save}_{i,1}} + 100 \cdot X_{\text{save}_{i,2}} + (\text{rnd}(0.1) - 0.05)$$

$$y(x) := x \cdot \begin{pmatrix} 0 \\ 0.1 \\ 100 \end{pmatrix}$$

↑ Y depend mostly on $x^{<2>}$.

Mean-centering: $y_{\text{mean}} := \text{mean}(Y)$ $y_{\text{mean}} = 0.315$

$$Y := Y - y_{\text{mean}}$$

Linear regression. $a_0 := \text{slope}(\text{score}^{<0>}, Y)$ $a_0 = 0.34$

Check: $\text{intercept}(\text{score}^{<0>}, Y) = 0$

Residual matrices: $E := X - \text{score}^{<0>} \cdot v^{<0>T}$

$$\text{sse}_{\text{old}} := Y \cdot Y$$

$$\text{sse}_{\text{old}} = 391.291$$

$$F := Y - \text{score}^{<0>} \cdot a_0$$

$$\text{sse} := F \cdot F$$

$$\text{sse} = 344.028$$

Regression coefficient of correlation:

$$r^2 := \frac{\text{sse}_{\text{old}} - \text{sse}}{\text{sse}_{\text{old}}}$$

$$r^2 = 12.079 \cdot \%$$

$$r := \sqrt{r^2}$$

$$r = 34.755 \cdot \%$$

With a minuscule r^2 value, the 0th principal component captured hardly any of the information in Y . Remember that we calculate the principal components solely based on X without ever considering the sensitivity of Y to different components of X .

We repeat the regression. $a_1 := \text{slope}(\text{score}^{<1>}, F)$ $a_1 = 99.862$

Check: $\text{intercept}(\text{score}^{<1>}, F) = 0$

Residual matrices: $E := E - \text{score}^{<1>} \cdot v^{<1>T}$

$$F := F - \text{score}^{<1>} \cdot a_1$$

$$\text{sse} := F \cdot F$$

$$\text{sse} = 0.038$$

Regression coefficient of correlation:

$$r^2 := \frac{\text{sse}_{\text{old}} - \text{sse}}{\text{sse}_{\text{old}}}$$

$$r^2 = 99.99 \cdot \%$$

$$r := \sqrt{r^2}$$

$$r = 99.995 \cdot \%$$

Thus, a subsequent component captured most of the information in Y . The regression equation is:

$$V^{<0>} := v^{<0>} \quad V^{<1>} := v^{<1>}$$

$$y_{\text{regress}}(x) := (x - x_{\text{mean}}) \cdot V \cdot a + y_{\text{mean}}$$

Check: Let us examine the slope and intercept.

Intercept: $y_{\text{mean}} - x_{\text{mean}} \cdot V \cdot a = -0.003$ which is practically 0; thus, $y = x \cdot \text{slope}$

slope: $V \cdot a = \begin{pmatrix} 0.03 \\ -0.206 \\ 99.862 \end{pmatrix} \leftarrow \text{compare} \rightarrow y(x) := x \cdot \begin{pmatrix} 0 \\ 0.1 \\ 100 \end{pmatrix}$

Examples:

$y_{\text{regress}}((5 \ 0.5 \ 0.05)) = 5.036 \leftrightarrow y((5 \ 0.5 \ 0.05)) = 5.05 \leftarrow \text{O.K.}$

$y_{\text{regress}}((5 \ -0.5 \ 0.05)) = 5.243 \leftrightarrow y((5 \ -0.5 \ 0.05)) = 4.95 \leftarrow \text{O.K. only because } y \text{ is not a function of } x^{<1>}$.

Example -- Variance-Scaling. Here, we shall add the variance-scaling step to the data from the last example. In that example, recall that the relative magnitudes of the three variables were dissimilar. Because independent variables may have different physical units or completely different magnitudes, the process of variance-scaling puts all variables on a comparable footing.

Step 1. Variance-Scaling (Note that we have already mean-centered the X matrix at the beginning of this worksheet.)

$$x_{\text{stdev}} := \text{stdev}(X^{<j>}) \quad x_{\text{stdev}} := x_{\text{stdev}}^T \quad X^{<j>} := \frac{X^{<j>}}{x_{\text{stdev}}}$$

$$x_{\text{stdev}} = (2.813 \quad 0.281 \quad 0.028)$$

Step 2. Find the eigenvalues and eigenvectors of the mean-centered and variance-scaled covariance matrix.

Covariance matrix:

$$X^T \cdot X = \begin{pmatrix} 51 & 51 & 17.304 \\ 51 & 51 & 17.295 \\ 17.304 & 17.295 & 51 \end{pmatrix} \quad |X^T \cdot X| = 0.091 \quad \leftarrow \text{Almost singular.}$$

$$\text{conde}(X^T \cdot X) = 6.013 \cdot 10^6 \quad \leftarrow \text{Very large.}$$

Eigenvalues/eigenvectors:

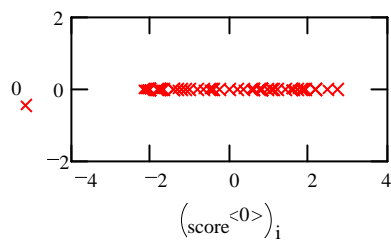
$$\lambda := \text{reverse}(\text{sort}(\text{eigenvals}(X^T \cdot X))) \quad \lambda^T = (111.838 \quad 41.162 \quad 1.982 \cdot 10^{-5})$$

$$v^{<j>} := \text{eigenvec}(X^T \cdot X, \lambda_j) \quad v = \begin{bmatrix} 0.656 & -0.264 & -0.707 \\ 0.656 & -0.264 & 0.707 \\ 0.373 & 0.928 & 1.457 \cdot 10^{-4} \end{bmatrix}$$

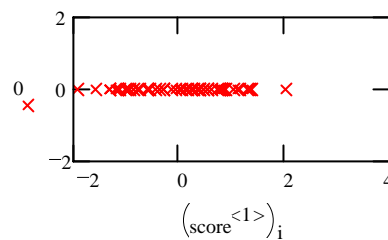
Check for orthogonality.

$$v^T \cdot v = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \quad v \cdot v^T = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

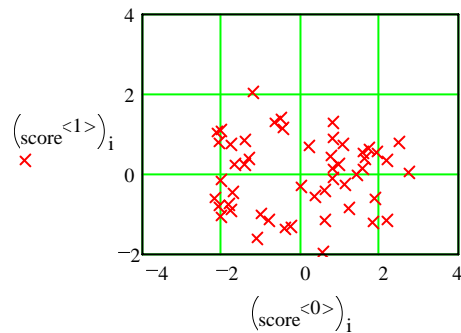
score: $\text{score}^{<0>} := X \cdot v^{<0>}$



score: $\text{score}^{<1>} := X \cdot v^{<1>}$



Scatter plot in the $v^{<0>}$ and $v^{<1>}$ plane:



Step 3.1. Regression -- 1st principal component.

Linear regression. $a_0 := \text{slope}(\text{score}^{<0>}, Y)$ $a_0 = 1.047$

Check: $\text{intercept}(\text{score}^{<0>}, Y) = 0$

Residual matrices: $E := X - \text{score}^{<0>} \cdot v^{<0>T}$
 $\text{sse}_{\text{old}} := Y \cdot Y$ $\text{sse}_{\text{old}} = 391.291$
 $F := Y - \text{score}^{<0>} \cdot a_0$
 $\text{sse} := F \cdot F$ $\text{sse} = 268.686$

Regression coefficient of correlation:

$$r2 := \frac{\text{sse}_{\text{old}} - \text{sse}}{\text{sse}_{\text{old}}} \quad r2 = 31.334\%$$

Step 3.2. Regression -- 2nd principal component.

Linear regression. $a_1 := \text{slope}(\text{score}^{<1>}, F)$ $a_1 = 2.555$

Check: $\text{intercept}(\text{score}^{<1>}, F) = 0$

Residual matrices: $E := E - \text{score}^{<1>} \cdot v^{<1>T}$
 $F := F - \text{score}^{<1>} \cdot a_1$
 $\text{sse} := F \cdot F$ $\text{sse} = 0.038$

Regression coefficient of correlation:

$$r2 := \frac{\text{sse}_{\text{old}} - \text{sse}}{\text{sse}_{\text{old}}} \quad r2 = 99.99\%$$

$$r := \sqrt{r2} \quad r = 99.995\%$$

Step 4. Regression Equation. (Be sure to take care of both mean-centering and variance-scaling)

$$V^{<0>} := v^{<0>} \quad V^{<1>} := v^{<1>} \quad x_{\text{stdev_inv},j,j} := \frac{1}{x_{\text{stdev}_{0,j}}}$$

$$y_{\text{regress}}(x) := (x - x_{\text{mean}}) \cdot x_{\text{stdev_inv}} \cdot V \cdot a + y_{\text{mean}}$$

Check: Let us examine the slope and intercept.

Intercept: $y_{\text{mean}} - x_{\text{mean}} \cdot x_{\text{stdev_inv}} \cdot V \cdot a = -0.003$ which is practically 0; thus, $y=x$ -slope

slope:

$$x_{\text{stdev_inv}} \cdot V \cdot a = \begin{pmatrix} 0.005 \\ 0.045 \\ 99.863 \end{pmatrix} \leftarrow \text{compare} \rightarrow y(x) := x \cdot \begin{pmatrix} 0 \\ 0.1 \\ 100 \end{pmatrix}$$

Examples:

$$y_{\text{regress}}((5 \ 0.5 \ 0.05)) = 5.036 \quad \longleftrightarrow \quad y((5 \ 0.5 \ 0.05)) = 5.05 \quad \leftarrow \text{O.K.}$$

$$y_{\text{regress}}((5 \ -0.5 \ 0.05)) = 4.991 \quad \longleftrightarrow \quad y((5 \ -0.5 \ 0.05)) = 4.95 \quad \leftarrow \text{O.K. only because } y \text{ is not a function of } x^{<1>}.$$