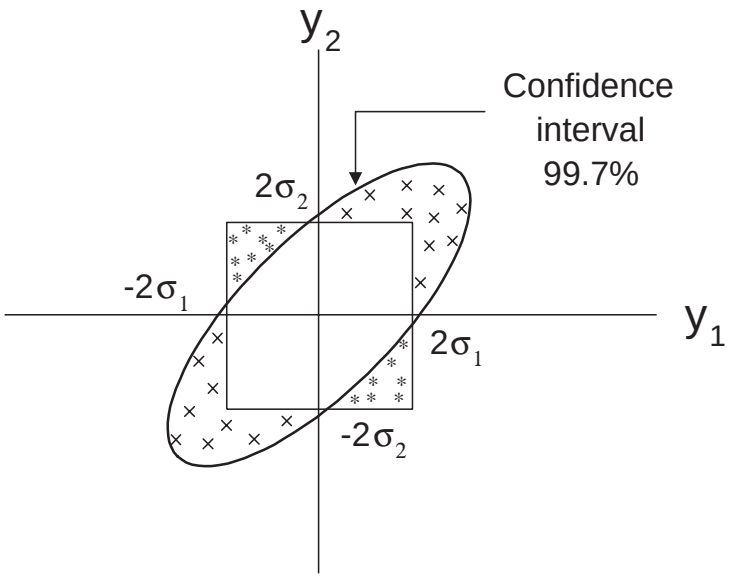
- simultaneous analysis of several quality variables
- including other (more easily and efficiently measured) process variables into the monitoring.

**Very Important Point**

These measurements are often not independent, but carry significant correlation. It is important to account for the existing correlation through the use of multivariate statistics.

**Motivating Example**

Let us demonstrate the importance of considering the correlation through the following simple example. Assume that we are monitoring two outputs $y_1$ and $y_2$ and their underlying probability distribution is jointly normal. If the correlation is strong, the data distribution and joint-confidence interval looks as below:




Considering the two measurements to be independent results in the conclusion that the probability of being outside the box is approximately $(1 - 0.95)^2 \approx 0.03$. The problems are:

- There are points (marked with $*$ in the above) that are outside the probability level $\gamma$, but fall well within the two $\sigma$s on both univariate charts. This means missed faults.
- There are points (marked with $\times$) that are inside the joint confidence interval of 99.7% probability level, but are outside the box. This means false alarms.

**Conclusions:**

- The most effective thing to do is to establish an elliptical confidence interval corresponding to a desired probability level $\gamma$ and see if the measurement falls outside the interval.
- *q-in-a-row* concept can be utilized as before, if desired.
- On the other hand, as the dimension of the output rises, graphical inspection is clearly out of question. It is desirable to reduce the variables into one variable that can be used for a monitoring purpose.

### 1.3.2 BASICS OF MULTIVARIABLE STATISTICS AND CHI-SQUARE MONITORING

**Computation of Sample Mean and Covariance**

Let $y$ be a vector containing $n$ variables:

$$y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} \tag{1.7}$$

Then the sample mean and covariance can be defined as before:

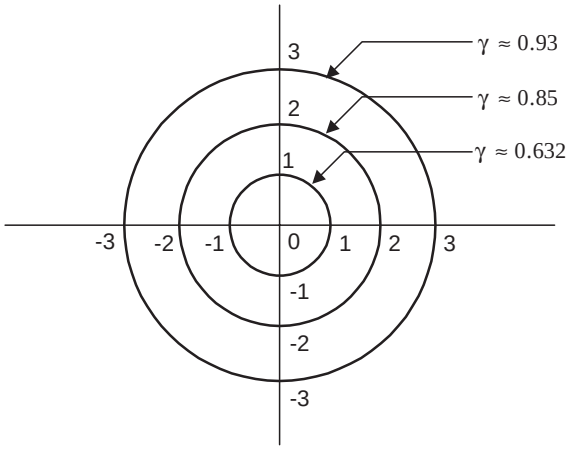
$$\bar{y} = \begin{bmatrix} \bar{y}_1 \\ \vdots \\ \bar{y}_n \end{bmatrix} = \frac{1}{N} \sum_{i=1}^{N} \begin{bmatrix} y_1(i) \\ \vdots \\ y_n(i) \end{bmatrix} \tag{1.8}$$

$$R_y = \frac{1}{N} \sum_{i=1}^{N} \left\{ \left( \begin{bmatrix} y_1(i) \\ \vdots \\ y_n(i) \end{bmatrix} - \begin{bmatrix} \bar{y}_1 \\ \vdots \\ \bar{y}_n \end{bmatrix} \right) \left( \begin{bmatrix} y_1(i) \\ \vdots \\ y_n(i) \end{bmatrix} - \begin{bmatrix} \bar{y}_1 \\ \vdots \\ \bar{y}_n \end{bmatrix} \right)^T \right\} \tag{1.9}$$

As $N \to \infty$, the above should approach the mean and covariance (assuming stationarity). Hence, $N$ should be fairly large for the above to be meaningful.

**Decorrelation & Normalization: For Normally Distributed Variables**

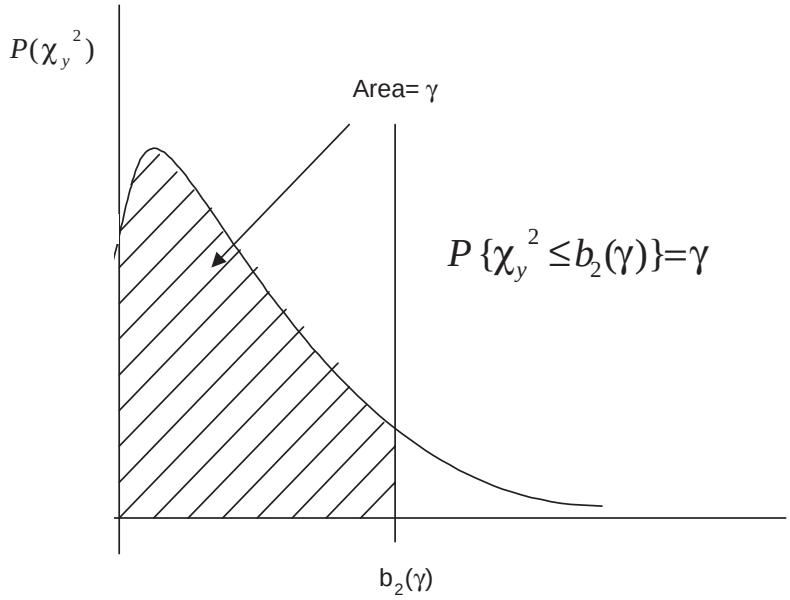
Assuming the underlying distribution is *normal*, the distribution of $z \triangleq R_y^{-1/2}(y - \bar{y})$ is normal with zero mean and identity covariance matrix. Hence, $R_y^{-1/2}$ can be interpreted as a transformation performing both decorrelation and normalization.The distribution for the two-dimensional case looks as below:



**Chi-Square Distribution**

Hence, the following quantity takes on the chi-square distribution of degree-of-freedom $n$:

$$\chi_y^2 \triangleq z^T z = (y - \bar{y})^T R_y^{-1} (y - \bar{y}) \tag{1.10}$$




Distribution of $\chi_y^2 = z^T z = z_1^2 + z_2^2$

For any given probability level $\gamma$, one can establish the elliptical confidence interval

$$(y - \bar{y})^T R_y^{-1} (y - \bar{y}) \leq b_n(\gamma) \tag{1.11}$$

simply by reading off the values $b_n(\gamma)$ from a chi-square value table.

**Chi-square percen** $\chi_u^2(n)$

| n \ u | 0.005 | 0.01 | 0.025 | 0.05 | 0.1 | 0.9 | 0.95 | 0.975 | 0.99 | 0.995 |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0 | 0 | 0 | 0 | 0.02 | 2.71 | 3.84 | 5.02 | 6.63 | 7.88 |
| 2 | 0.01 | 0.02 | 0.05 | 0.1 | 0.21 | 4.61 | 5.99 | 7.38 | 9.21 | 10.6 |
| 3 | 0.07 | 0.11 | 0.22 | 0.35 | 0.58 | 6.25 | 7.81 | 9.35 | 11.34 | 12.84 |
| 4 | 0.21 | 0.3 | 0.48 | 0.71 | 1.06 | 7.78 | 9.49 | 11.14 | 13.28 | 14.86 |
| 5 | 0.41 | 0.55 | 0.83 | 1.15 | 1.61 | 9.24 | 11.07 | 12.83 | 15.09 | 16.75 |
| 6 | 0.68 | 0.87 | 1.24 | 1.64 | 2.2 | 10.64 | 12.59 | 14.45 | 16.81 | 18.55 |
| 7 | 0.99 | 1.24 | 1.69 | 2.17 | 2.83 | 12.02 | 14.07 | 16.01 | 18.48 | 20.28 |
| 8 | 1.34 | 1.65 | 2.18 | 2.73 | 3.49 | 13.36 | 15.51 | 17.53 | 20.09 | 21.96 |
| 9 | 1.73 | 2.09 | 2.7 | 3.33 | 4.17 | 14.68 | 16.92 | 19.02 | 21.67 | 23.59 |
| 10 | 2.16 | 2.56 | 3.25 | 3.94 | 4.87 | 15.99 | 18.31 | 20.48 | 23.21 | 25.19 |
| 11 | 2.6 | 3.05 | 3.82 | 4.57 | 5.58 | 17.28 | 19.68 | 21.92 | 24.73 | 26.76 |
| 12 | 3.07 | 3.57 | 4.4 | 5.23 | 6.3 | 18.55 | 21.03 | 23.34 | 26.22 | 28.3 |
| 13 | 3.57 | 4.11 | 5.01 | 5.89 | 7.04 | 19.81 | 22.36 | 24.74 | 27.69 | 29.82 |
| 14 | 4.07 | 4.66 | 5.63 | 6.57 | 7.79 | 21.06 | 23.68 | 26.12 | 29.14 | 31.32 |
| 15 | 4.6 | 5.23 | 6.26 | 7.26 | 8.55 | 22.31 | 25 | 27.49 | 30.58 | 32.8 |
| 16 | 5.14 | 5.81 | 6.91 | 7.69 | 9.31 | 23.54 | 26.3 | 28.85 | 32 | 34.27 |
| 17 | 5.7 | 6.41 | 7.56 | 8.67 | 10.09 | 24.77 | 27.59 | 30.19 | 33.41 | 35.72 |
| 18 | 6.26 | 7.01 | 8.23 | 9.39 | 10.86 | 25.99 | 28.87 | 31.53 | 34.81 | 37.16 |
| 19 | 6.84 | 7.63 | 8.91 | 10.12 | 11.65 | 27.2 | 30.14 | 32.85 | 36.91 | 38.58 |
| 20 | 7.42 | 8.26 | 9.59 | 10.85 | 12.44 | 28.41 | 31.41 | 34.17 | 37.57 | 40 |
| 22 | 8.6 | 9.5 | 11 | 12.3 | 14 | 30.8 | 33.9 | 36.8 | 40.3 | 42.8 |
| 24 | 9.9 | 10.9 | 12.4 | 13.8 | 15.7 | 33.2 | 36.4 | 39.4 | 43 | 45.6 |
| 26 | 11.2 | 12.2 | 13.8 | 15.4 | 17.3 | 35.6 | 38.9 | 41.9 | 45.6 | 48.3 |
| 28 | 12.5 | 13.6 | 15.3 | 16.9 | 18.9 | 37.9 | 41.3 | 44.5 | 48.3 | 51 |
| 30 | 13.8 | 15 | 16.8 | 18.5 | 20.6 | 40.3 | 43.8 | 47 | 50.9 | 53.7 |
| 40 | 20.7 | 22.2 | 24.4 | 26.5 | 29.1 | 51.8 | 55.8 | 59.3 | 63.7 | 66.8 |
| 50 | 28 | 29.7 | 32.4 | 34.8 | 37.7 | 63.2 | 67.5 | 71.4 | 76.2 | 79.5 |

For $n > 50$: $\chi_u^2(n) = \frac{1}{2}(z_u + \sqrt{2n-1})^2$

Now one can simply monitor $\chi_y^2(k)$ against the established bound.

## Limitations of Chi-Square Test

The chi-square monitoring method that we discussed has two drawbacks.

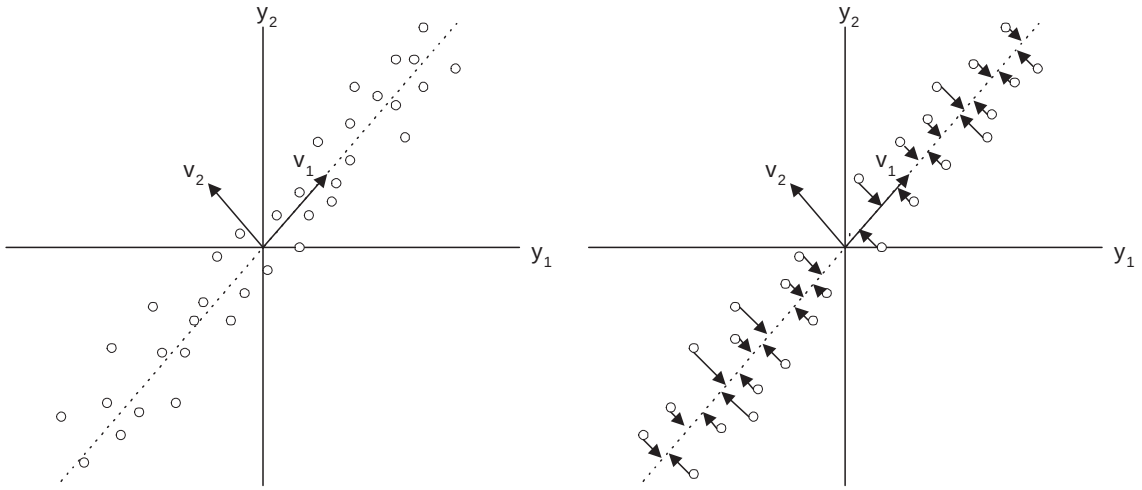
- *No Useful Insight for Diagnosis*
  Although the test suggests that there may be an abnormality in the operation, it does not provide any more insight. One can store all the output variables and analyze their behavior whenever an abnormality is indicated by the chi-square test. However, this requires a large storage space and analysis based on a large correlated data set is anything but a difficult, cumbersome task.

- *Sensitivity to Outliers and Noise*
  Note that the variables are normalized through $R_y^{-1/2}$. For an ill-conditioned $R_y$, gains of very different magnitudes are applied to different combinations of the $y$ elements in the normalization process. This can cause extreme sensitivity to noise, outliers, etc.

### 1.3.3 PRINCIPAL COMPONENT ANALYSIS

A solution to the both problem is to monitor and store only the *principal components* of the output vector.

**What's The Idea?**
Consider the following two-dimensional case:




It is clear that, through an appropriate coordinate transformation, one can explain most of the variation with a single variable.

The SVD of the covariance matrix provides a useful insight for doing this. For the above case, the SVD looks like

$$R_y = \begin{bmatrix} v_1 & v_2 \end{bmatrix} \begin{bmatrix} \sigma_1 & 0 \\ 0 & \sigma_2 \end{bmatrix} \begin{bmatrix} v_1^T \\ v_2^t \end{bmatrix}, \qquad \sigma_1 \gg \sigma_2$$

**Computing the Princial Components: Using SVD**

The principal components may be computed using the singular value decomposition of $R_y$ as follows:

$$R_y = \left[\begin{array}{ccc|ccc} v_1 & \cdots & v_m & v_{m+1} & \cdots & v_n \end{array}\right] \left[\begin{array}{ccc|ccc} \sigma_1 & & & & & \\ & \ddots & & & & \\ & & \sigma_m & & & \\ \hline & & & \sigma_{m+1} & & \\ & & & & \ddots & \\ & & & & & \sigma_n \end{array}\right] \left[\begin{array}{c} v_1^T \\ \vdots \\ v_m^T \\ \hline v_{m+1}^T \\ \vdots \\ v_n^T \end{array}\right] \tag{1.12}$$

One can, for instance, choose $m$ such that

$$\frac{\sum_{i=1}^{m} \sigma_i}{\sum_{i=1}^{n} \sigma_i} \geq \gamma \tag{1.13}$$

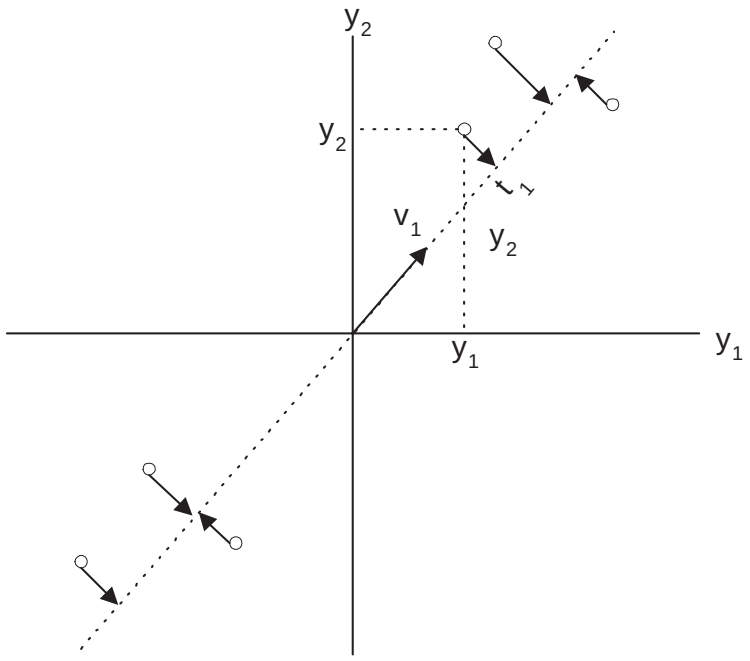
where $\gamma$ is the tolerance parameter close to 1 (say .99), or such that

$$\sigma_m \gg \sigma_{m+1} \tag{1.14}$$

Usually, $m \ll n$.

$v_1, \cdots, v_m$ are called *principal component directions*. Define the *score variables* for the principal component directions as

$$t_i = v_i^T y, \qquad i = 1, \cdots, m \tag{1.15}$$




These score variables are independent of one another since

$$
\begin{aligned}
E\left\{ \begin{bmatrix} t_1 \\ \vdots \\ t_m \end{bmatrix} \begin{bmatrix} t_1 \\ \vdots \\ t_m \end{bmatrix}^T \right\} &= E\left\{ \begin{bmatrix} v_1^T \\ \vdots \\ v_m^T \end{bmatrix} (y(k)-\bar{y})(y(k)-\bar{y})^T \begin{bmatrix} v_1 & \cdots & v_m \end{bmatrix} \right\} \\
&= \begin{bmatrix} v_1^T \\ \vdots \\ v_m^T \end{bmatrix} R_y \begin{bmatrix} v_1 & \cdots & v_m \end{bmatrix} \\
&= \begin{bmatrix} \sigma_1^2 & & \\ & \ddots & \\ & & \sigma_m^2 \end{bmatrix} \qquad (1.16)
\end{aligned}
$$

**Example**

Show a 4-dimensinal case, perform SVD and explain what it means.
Actually generate a large set of data and show projection to each principal component direction.

- Generate 1000 data points (from the normal distribution).
- Plot each $y$ (time vs. value plot for each variable).
- Compute the sample mean and covariance.
- Perform SVD of the sample covariance.
- Compute principal components.
- Plot each variable along with its prediction from the two principal components ($\hat{y} = t_1 v_1 + t_2 v_2$)

**Monitoring Based on Principal Component Analysis**

- $t_i$'s are perfect candidates for monitoring since they are: (1) independent of one another, and (2) relatively low in dimension.
- The residual vector can be computed as

$$r(k) = y(k) - \sum_{i=1}^{m} \underbrace{(v_i^T y(k))}_{t_i(k)} v_i \tag{1.17}$$

  The above residual vector represents the contribution of the parts that were thrown out because their variations were judged to be insignificant from the normal operating data. The size of the residual vector should be monitored in addition, since a significant growth in its size can indicate an abnormal (*out-of-control*) situation.

**Advantages**

The advantage of the two-tier approach is that one can gain much more information from the monitoring. Often times, when the monitoring test indicates a problem, useful additional insights can be gained by examining

- the direction of the principal component(s) which has violated the bound
- the residual vector if its size has gone over the tolerance level.

### 1.3.4 EXAMPLE: MULTIVARIATE ANALYSIS VS. SINGLE VARIATE ANALYSIS

Compare Univariate vs. Multivariate. Compare chi-square test vs. PC monitoring.

## 1.4 TIME SERIES MODELING

### 1.4.1 LIMITATIONS OF THE TRADITIONAL SPC METHODS

In the process industries, there are two major sources for quality variances:

- Equipment / instrumentation malfunctioning.
- feed variations and other disturbances.

Usually, for the latter, the dividing line between normal and abnormal are not as clear-cut since

- they occur very often.
- they tend to fluctuate quite a bit from one time to another (but with strong temporal correlations).
- they often cannot be eliminated at source.

Because of the frequency and nature of these disturbances, they cannot be classified as *Pareto's glitches* and *normal* periods (*in-control* epochs) must be defined to include their effects.

The implications are