4/29/2017

MEEN 357 Honors Project

Jacob Hartzer Section 201

INTRODUCTION

As means of production have become faster and more advanced, industry has shifted focus towards maximizing efficiencies of production to maximize profit. To do this, an accurate model must be created to describe the performance based on a set of descriptor characteristics and then determine a desired range of operation for peak performance. Every manufacturing process contains errors, and it is therefore important to constantly be checking these errors for accuracy within a given degree and to ensure there is no trend in the errors of the measured quantity. One way to determine this is measuring deviations in the data across a spanning dataset. Data points with the largest deviations would be related to conditions that deviate the most from desired operating range and therefore should be investigated further. This generally keeps processes, fast, efficient, and working properly, thus maximizing the efficiency of the process.

In this paper, the process to be analyzed is a multivariate model describing the conversion of natural gas to energy in the form of steam or electricity. For this model, a series of possibly interrelated variables are considered to influence the energy output of the steam. To detect unusual occurrences in the power output of the natural gas, it is important that only independent variables are considered. Therefore, redundant variables will be removed and the remaining data will be analyzed for unusual occurrences by using statistical distances.

METHOD

To determine 'usualness' of any occurrence, one must first determine some method of comparing any one given datum to the set as a whole. A very common method is to take the Euclidean distance between the data point and the mean of the set. This distance for a single point is defined as:

$$D_e = \sqrt{\sum_{i=1}^{n} (x_i - \bar{x}_i)^2}$$

where x_i is an individual instance of each variable, and $\bar{x}_i$ is the corresponding mean of each individual variable.

The Euclidean distance, while very easy to compute, unfortunately does not consider the possible covariance between each of the variables. This is quite a common issue and is often difficult to intuitively detect. Instead, this paper utilizes Hotelling's T^2 distance, which considers the possible covariance between variables. Hotelling's T^2 is given by:

$$T^2 = (X - \mu)' \, \Sigma^{-1} \, (X - \mu)$$

where $X = (x_1; x_2; \dots; x_n)$, μ is the mean, and Σ is the variance-covariance matrix of the sample.

To first ensure that the data being analyzed is a span, redundant variables must be removed. To do this, we calculate the square root of the ratio of maximum eigenvalue to each of the eigenvalues and define these numbers as condition indices. If, for any variable, the condition index is greater than 30, this implies that there is severe collinearity among variables. To determine which variables are involved in collinearity, the associated eigenvector is analyzed and if two values are sufficiently large and close to equal, then the use of both variables is redundant and one can be removed.

Once all redundant variables are removed, Hotelling's T^2 can be calculated for each time step to create a model for determining when data points are outside desired operating conditions.

RESULTS

The MatLab code in Appendix A was initially used to find the Euclidean distances of each time step to the mean across the sample distribution. The results of these calculations are summarized in *Figure 1*.

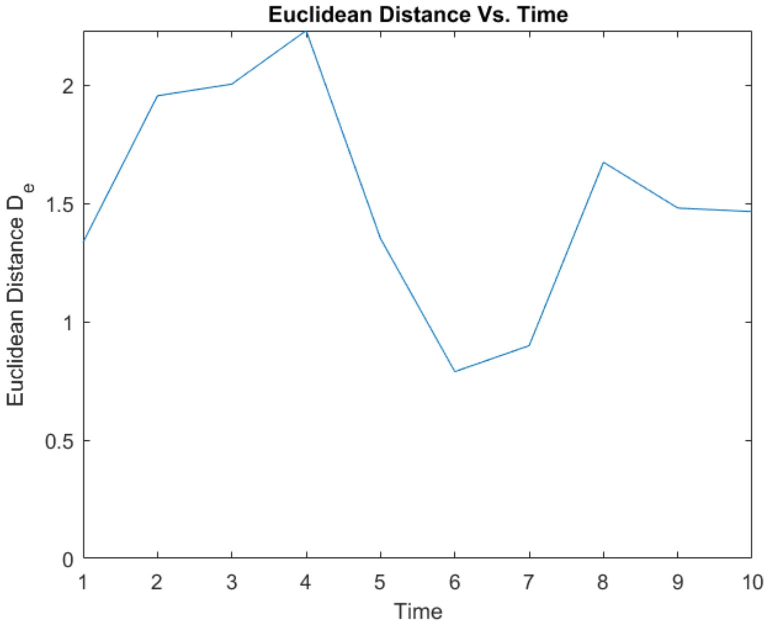

Figure 1. Euclidean distance from the mean at each time step

The code was also used to calculate the covariance matrix of the data and its corresponding eigenvectors and eigenvalues. By comparing each of the eigenvalues to the maximum and removing variables that had similar loadings in the vectors with the highest condition indices, it was determined variables 1, 2, & 3 were redundant as was the stated method for reduction of redundant variables. Once these variables were removed, the largest condition index was 16.6092. This shows that the variables are now sufficiently independent from each other and represent a span without significant correlation between variables.

With the redundant variables removed, the T^2 values could be computed at each time step. *Figure 2* summarizes T^2 distances at each time step.

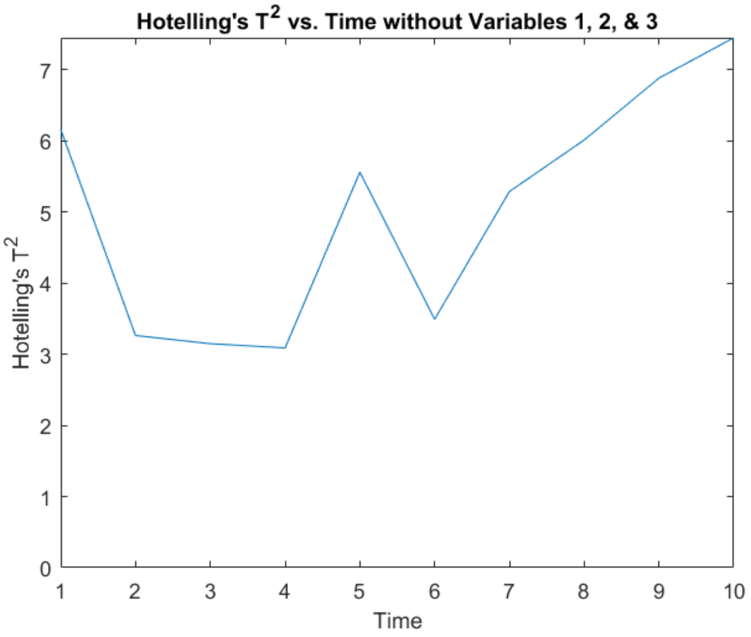

Figure 2. Hotelling's T^2 distance with the removal of redundant variables 1, 2, and 3 at each time step.

DISCUSSION

From looking at the Euclidean distance, it would initially appear that times 2, 3, 4, and 10 deviate the most from the mean of the sample. However, from knowing that it is common to have interdependency between variables, this must be checked. Using the prescribed method, it was found that variables 1, 2, and 3 were all redundant. This implies that most of the variance in power production, our desired measured quantity, could be explained by variables 4 through 7. Therefore, by removing variables 1 through 3, it is possible to still holistically describe the dataset, without putting unequal weights on variables by counting them twice.

Therefore, with the redundant variables removed, the Hotelling's T^2 distance at each time step should represent a more accurate deviation from the mean. Comparing the new T^2 distance to the Euclidean distance shows that times 2, 3, and 4, which would have previously thought to have been the data points that deviated most from the mean using just the Euclidean distance, are some of the lowest. Additionally, times 9 and 10 have the largest deviations from the mean using the T^2 distance, compared to being moderately low in Euclidean distance. This implies that the redundant variables were compounding at these points and would've given a false analysis, further stressing the need to remove dependent variables.

Because times 8, 9, and 10 have the largest deviation from the mean as described by the T^2 distance, it is recommended that further analysis is done to determine why these times produced the largest deviations and under what conditions. This could lead to a better understanding of the system and an overall more efficient and controllable process.

A final note would be a recommendation to analyze the system for time dependency as the largest T^2 values were computed at sequential time steps.

CONCLUSION

In conclusion, it was determined that the data set contained near linearly dependent variables. Through analysis on the eigenvectors and eigenvalues, it was determined that removing variables 1 through 3 would remove any redundancy between variables and create a nearly linearly independent set. With these variables removed, Hotelling's T^2 distance was computed at each time step, showing that times 8, 9, and 10 deviated most from the mean. It is recommended that further analysis is done into the variables and conditions at these times to best determine why the power production had the largest deviation at these times. Finally, further analysis could be done to determine if there is any time dependency in the system.

APPENDICES

A. Code

```
clc; clear; close all;

data = csvread('Honors_data.csv',1,1);
data = data(:,[4,6,7]);

Mins = zeros([1,length(data(1,:))]);
Maxes = zeros([1,length(data(1,:))]);
Normalizeddata = zeros(size(data));

for i = 1:length(data(1,:))
    Mins(i) = min(data(:,i));
    Maxes(i) = max(data(:,i));
    Normalizeddata(:,i) = (data(:,i) - Mins(i))/(Maxes(i)-Mins(i));
end

De_H = zeros([length(data(:,1))],1);

for i = 1:length(data(:,1))
    De_H(i,1) = sqrt(sum(Normalizeddata(i,:).^2));
end

sigma = cov(data);

[eig_vect,D] = eig(sigma);
eig_val = diag(D);

disp(sqrt(abs((max(eig_val)./eig_val)))>30);


sigma_inv = 0;
for i = 1:length(eig_val)
    sigma_inv = sigma_inv +
1/eig_val(i)*eig_vect(:,i)*eig_vect(:,i)';
end

means = zeros(1,size(data,2));
for i = 1:size(data,2)
    means(i) = mean(data(:,i));
end

T_squared = zeros(size(data,1),1);
for i = 1:size(data,1)
    T_squared(i) = (data(i,:) - means)*sigma_inv*(data(i,:) -
means)';
end

figure
subplot(1,2,1)
```

```
plot(De_H);
xlabel('Time');
ylabel('Euclidean Distance D_e');
title('Euclidean Distance Vs. Time');

subplot(1,2,2)
plot(T_squared);
xlabel('Time');
ylabel('Hotelling''s T^2');
title('Hotelling''s T^2 vs. Time without Variables 1, 2, 3, & 5')
```