



## International Journal of Research Publications

# Golden Batch Identification using Statistical Tools

Modhuli D. Goswami<sup>a</sup>

<sup>a</sup>Dept. Of Instrumentation and Control Institute of Technology Nirma University, Ahmedabad 380015, India

---

### Abstract

This paper discusses use of data analysis using statistical methods for prediction of the Golden Batch in a sample industrial dataset (Wine manufacturing). The Golden Batch identification is extremely important as it allows us to reach, a plant process operating curve, which delivers the most optimal product mix, along with enhanced quality of final product. PCA (Principle Component Analysis) and PLS (Partial Least Square Regression) analysis are the tools used for analyzing the data set. These Data sets are the telemeter and archived process/plant parameters from a large number of such product runs. Principal Components Analysis (PCA) is used for dimension reduction and Partial Least Square Regression (PLS) is used for predictive analysis, RMSE is used to get the error between actual and predicted dataset. Business Managers, Process and control engineers can use this method to detect batch to batch error and recipe of the Golden batch which when implemented in other batches, results in optimal plant and product output, thus increasing efficiency of the manufacturing and increasing profits.

© 2018 Published by IJRP.ORG. Selection and/or peer-review under responsibility of International Journal of Research Publications (IJRP.ORG)

Keywords: Golden Batch; Data Set; Matlab; IIOT; PCA; PLS; Asset Performance Management

---

## 1. Introduction

Analysis of telemetered industrial process measurement data, which is collected at a central data repository, especially as a part of Industrial Internet of Things (IIOT), can give us insights into the optimal process parameters, by subjecting the data sets to Statistical Analysis and Machine Learning. The optimization can result in significant improvement in costs, energy, time and the quality of the outcome from the industry or plant.

Batch processes are generally found in industries manufacturing small lots of material which are produced through chemical, electro-chemical or biological reactions. The controlling of batch process is more complex than continuous process as multiple products are produced using the same equipment. The slow response of the batch processes, and the fact that all the input lots undergo the “batch process” in one go, makes it difficult for the operator/controllers to identify abnormal conditions which affect the final product quality. [1] A range of process, environment, plant, inputs, quality parameters (Data Set) can be measured in order to maintain the quality of the final product.

Many key performance indicators (KPI's) can be used to measure the effectiveness of batch production. Ability to repeatedly produce on-specification and saleable product is absolutely required, but once this is accomplished, many opportunities remain to improve production and KPI's for cost savings and capacity increases without large capital expenditures. The quality of the final product differs in each batch as the recipe differs, when we get the best quality product, it defined by a standard as an exception or ‘golden batch’. [2]

The KPI's and inputs of the golden batch are considered as the master recipe and used for determination of a batch as good or bad. A golden batch is defined as the time- based profile of the measurement values that were recorder for a particular batch that met produced quality targets; once the golden batch is identified the other batches are judged by how closely are they to the golden batch. [2]

The concept of ‘Golden Batch’ addresses such optimal performance point, which can be derived from the existing Plant or Process performance data sets.

## 2. Approach for the analysis of Batch Process dataset

One difficult aspect of using multiple process parameters is that they usually do not have equal importance with regard to the overall quality and hence considering individual parameter can give conflicting results, thus making it difficult to improve the quality. Hence to get proper batch process parameters into consideration we have used Principle Component Analysis (PCA) which does dimension reduction in order to take into consideration the most important parameters which have maximum effect on the Batch and quality of the final product.[3] After getting the principle components, these components are used to create a predictive model of the system for accurate process/quality prediction, this is done using Partial Least Square Regression. Finally the error between the predicted and the actual model is found using Root Mean Square Error (RMSE). As a data-driven model and statistical method, PCA (Principal Component Analysis) and PLS (Partial Least Squares) are widely used in engineering and science applications in general, and in particular for quality monitoring.

### 2.1 Dataset used for testing

The data taken into consideration is a wine-quality dataset having 11 variables: fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates,

alcohol and 1 output: quality of the wine (median of at least 3 evaluations made by wine experts), each expert graded the wine quality between 0 (very bad) and 10 (very excellent). with 4500 samples. The batches are divided into 45 samples each, having total of 100 batches.[4]

### 3. Finding Golden Batch Benchmarks

#### 3.1 Software

The software used for analysis and testing is MATLAB version 2013, along with in-line MATLAB code for data set handling, plots, formatted Input/ Output.

#### 3.2 Normalization of Data

The raw data cannot be directly used for PCA as the range of data for each input parameters is varied, hence data has been normalized in order to convert all the data of the input parameters in the scale of 1-100. Figure1 shows the boxplot of all the 11 input parameters.

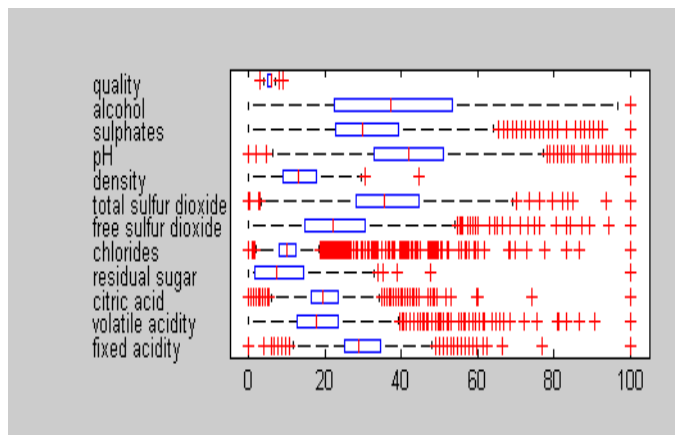


Figure 1.

#### 3.3 Results from Principal Component Analysis (PCA)

Principal component analysis enables identification and evaluation of product and process variables that may be critical to the product quality and performance of the process. It also helps to understand the relationship between the process measurement and analysis of the final product. It is a statistical procedure that uses an orthogonal transformation to convert a set of observations of possibly correlated variable into a set of values that are linearly uncorrelated variables called principle components.[5,6] The first principle component accounts for as much of the variability in the data as possible and each of the succeeding component accounts for as much of the remaining variability as possible. It is a dimension reduction tool that can be used to reduce a large set of variable to a small set that still contains the most important information of the large dataset.

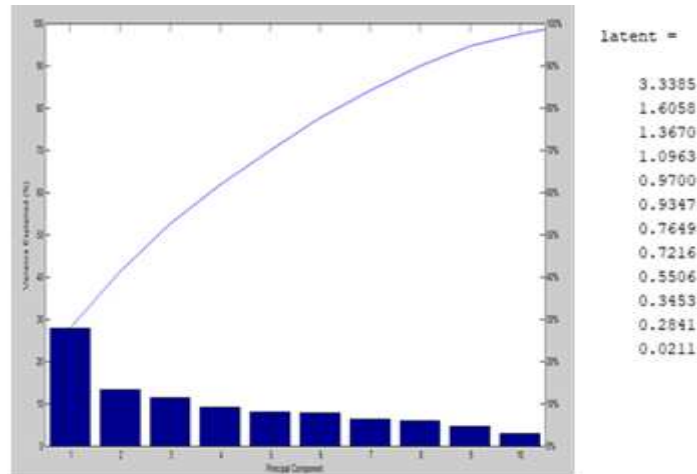


Figure. 2

The PCA has created 3 principal components, thus reducing the parameters for consideration from 11 to 3, Figure.2 shows the latent matrix which gives the variance due to each principal component in the final quality, and the pareto plot (figure. 2) shows the same in the form of percentage.

Figure.3 is the biplot of the first two principal components, the distance of each variable from the center gives the statistical measure of that variable, for eg, the input parameter having maximum distance from the center along x- axis has the maximum effect on principle component 1. The same thing can be found analytically by finding the maximum correlation of each principal component with the parameters. The results obtained from the biplot are: principle component 1 is maximum affected by alcohol, hence it being the first most important variable for consideration the 2<sup>nd</sup> principle component is maximum affected by the pH and 3<sup>rd</sup> principle component is affected by sulphates. To get better results, further analysis is done to find the relation these variables posses between each other, and the parameter(s) having maximum positive or negative correlation with these three principal parameter are also considered for quality prediction. Figure.4 shows the relationships.

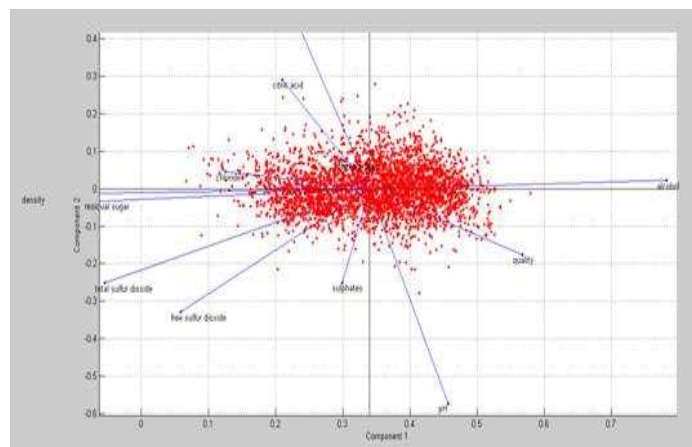


Figure. 3

```

'density'
is inversely related with
'alcohol'

'fixed acidity'
is inversely related with
'pH'

'chlorides'
is inversely related with
'sulphates'

```

Figure. 4

### 3.4 Results from Partial Least Square Regression (PLS)

In this method, PLS is used for analyzing the impact of processing conditions on the final product and also continuous prediction of the end of the batch quality parameters. PLS is a statistical method that bears some relation to principle component regression. It finds a linear regression model by projecting the predicted variables and the observable variables to a new space. A linear model specifies the (linear) relationship between a dependent (response) variable Y, and a set of predictor variables, the X's. The PCA is done in the model to get the variables responsible for maximum variance in the sample data. These variables are used as independent variables (x) and a predicted model (y) is expressed as linear regression of the independent variables. [7,8]

PLS gives a predictive model for wine quality prediction using 4500 data samples of telemetered data. The number of components for the PLS is determined by the previous PCA analysis, where we got 3 principal components and 4 component are highly correlated to them either directly or indirectly hence 7 variables are chosen and the predictors data is taken from the standardized dataset and observed values are the observed quality values in the actual data. The PLS gives beta which when multiplied by predictor (parameters) and the intercept (constant) gives yfit matrix, giving the predicted quality values. Figure.5 shows the graph comparing the actual and predicted quality of the dataset for every 9<sup>th</sup> data. The correlation obtained between both the models is 0.9391. The error obtained is 11.80 %

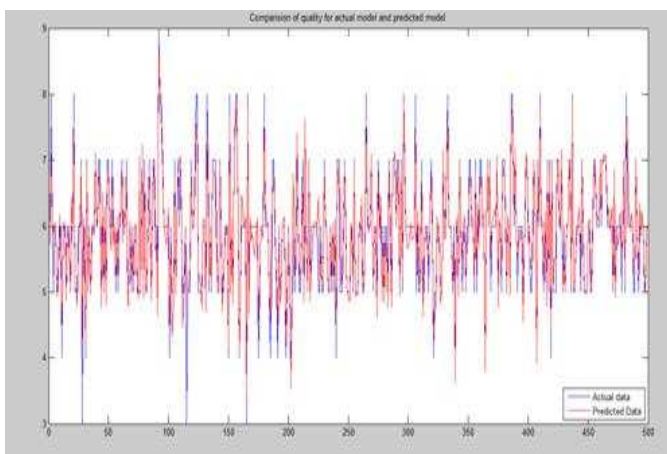


Figure 5

### 3.5 Golden Batch identification

The entire 4500 dataset is divided into 45 batches of 100 sample data each. A for loop is run (as part of the processing program) for each batch and a predetermined good quality number (mean =5.8), 6 (greater than mean) is chosen and now each batch sample is compared with this number. If a batch of 100 sample data has at least 63 samples greater than this quality then it is considered for batch analysis. After running this “for loop” through entire dataset we get the golden batch which is found to be 34 for this dataset.

Figure 6 shows the actual and predicted data for the golden batch where horizontal values represent the observation number and vertical axis represents the quality (0-10). Now we find a mathematical equation relating the quality and predictor variables (3 Principal Components), figure.7 shows the matrix representing the regression coefficient with the 1<sup>st</sup> element being intercept and the following numbers are the coefficient for the corresponding Principal Component.

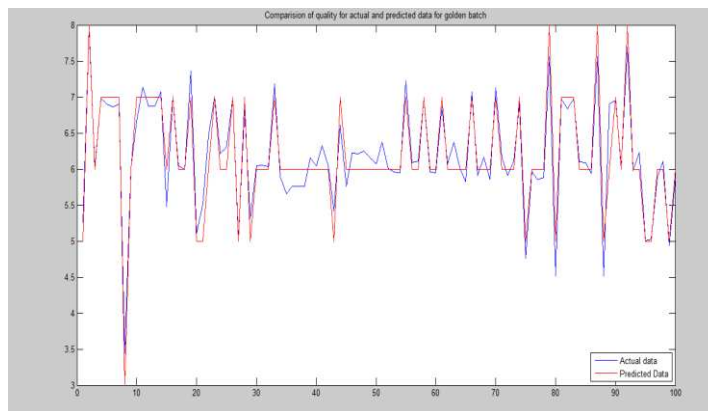


Figure.6

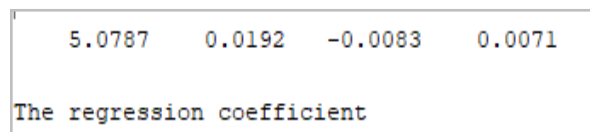


Figure.7

This model depicts the recipe of the golden batch, which when implemented in other batches is expected to give results like the golden batch.

### 3.6 Comparison between batches

RMSE is used to find the error between a batch and the golden batch, now this is done so that particular batch can be improved by improving proper principal component. It measures the differences between values (sample or population values) predicted by a model or an estimator and the values observed. Root mean square error is the standard deviation of the residual (predicted data). RMSE is the measure of how spread the residuals are and are different from the actual data (line fit).

Figure. 8 shows the error in the 3 principle components in all the 45 batches with respect to the golden batch, we can observe that the least error is in the 32<sup>nd</sup> batch and maximum error is in the 9<sup>th</sup> batch. Hence this graph can be used to identify which batch had least error compared to the golden batch and which batch has maximum faults and is far away from the standard of the golden batch.

Figure.9 and figure.10 shows comparison of a single batch (here batch 29, user choice batch number) with the golden batch. The box plot shows the medians and quartiles for both the batches quality, where as the next graph shows the error in three principle components in both the batches, here the error is maximum in sulphates followed by alcohol and least in pH. The errors in each Principle component is also shown analytically. (Figure.11)

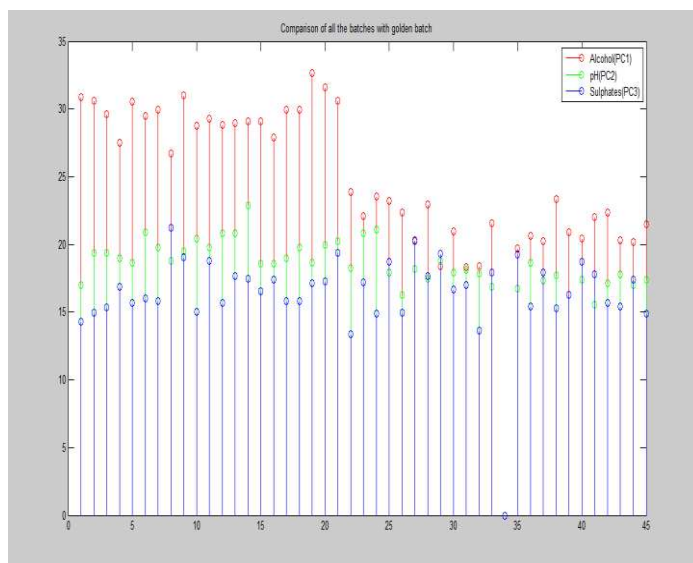


Figure. 8

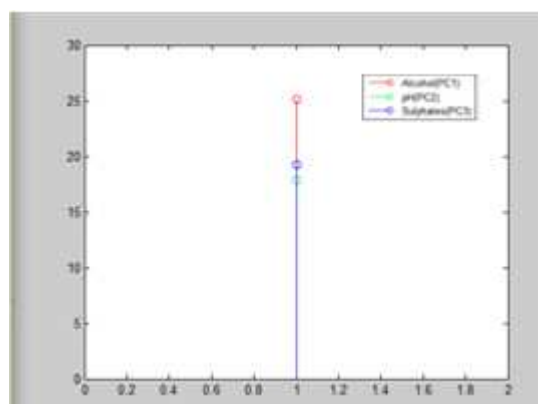


Figure. 9

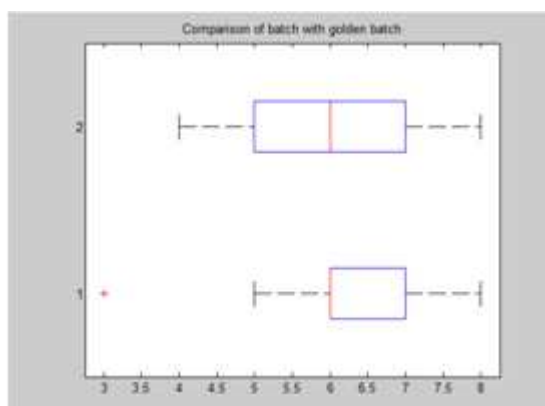


Figure. 10

```

1- for single batch comparison 2- all batch comparison
Enter the batch to be compared with the golden batch: 29
RMSE in Alcohol (PC1) is
    25.1690

RMSE in pH (PC2) is
    17.8180

RMSE in Sulphates (PC3) is
    19.2263

```

Figure. 11

#### 4. Conclusion and Summary

This MATLAB code can be used for any dataset by changing few parameters and the program will provide graphically and analytically the following:

- The main process variables of the data
- Their relation with other variables
- A predictive model
- Identify the golden batch
- Compare it with either the entire dataset or particular batch.

The presented algorithm, method and code was tested using an actual industrial batch data set of Wine manufacturing, and the results are presented. The method can be used for very large or smaller batch sets as well, without any changes. The output plots, Figure. 6 and 8, show that the method provides identification of the ‘Golden Batch’, with minimum RMSE, and also helps to indicate to plant managers, the process/plant parameter errors for other batches of the product.

IIOT has become an important part of today’s industries. Predictive maintenance, digital twin, predictive model, process optimization are now a requirement for the meeting ever growing competition and quality requirements as well as to meet the expectations of the consumers.

Asset Performance Management (APM) is one of the main products for converting a simple manufacturing plant into an IOT enabled plant. It not only helps to reduce downtime, and provide optimized solutions it also reduces the revenue and risks associated with the assets.

‘Golden batch identification’ as a part of Asset Performance Management helps to identify the best recipe that can be implemented in other batches to meet the golden batch standards. Advantages of this are:

- Optimization of batch process: implementing golden batch benchmarks and accordingly implementing control strategies
- Economic benefits: increasing the batch quality is done by changing the recipe of already used input parameters hence there are no added costs and the quality of the product is also improved.
- Use of this statistical approach will allow user to identify the relationship between process variables and their importance in affecting the product quality parameters. It also provides additional information to help the process control engineer to pinpoint which and where the process needs improvement. [3]

#### References

- [1] Data Analytics in Batch Operations by Robert Wojewodka and Terry Blevins
- [2] <https://www.yokogawa.com/library/resources/media-publications/match-the-golden-batch/>
- [3] Data Mining Applications for Finding Golden Batch Benchmarks and Optimizing Batch Process Control by Yuelong Su and Fengqin Yu (2016 12th World Congress on Intelligent Control and Automation (WCICA) June 12-15, 2016, Guilin, China)
- [4] <https://archive.ics.uci.edu/ml/datasets/wine+quality>
- [5] [https://en.wikipedia.org/wiki/Principal\\_component\\_analysis](https://en.wikipedia.org/wiki/Principal_component_analysis)
- [6] <https://in.mathworks.com/help/stats/pca.html>
- [7] [https://en.wikipedia.org/wiki/Partial\\_least\\_squares\\_regression](https://en.wikipedia.org/wiki/Partial_least_squares_regression)
- [8] <https://www.mathworks.com/help/stats/plsregress.html>