Analysis to Few Days of Drilling Operation in the North Sea

Mohammad Alidoust*
Department of Physics, Norwegian University of Science and Technology, N-7491 Trondheim, Norway
(Dated: September 6, 2013)

In this report we analyze the given data set which is ascribed to five days of drilling works in the north sea. To analyze the data set, we briefly analytically discuss the most important 'latent variable' methods i.e. the principal component analysis (PCA) for unsupervised data sets. Since the magnitude order of reported variables' value are more or less very different, we centralize first and then perform autoscaling on the given data of each variable by its standard deviation. We calculate the explained variance of raw data set and determine the number of principal components (PCs) needed to describe validly variable correlations. Using the obtained PCs, we present the loading and score plots to visualize the data distribution and discuss appeared patterns. we analyze the data and discuss the important variables involved, their positive and negative correlations, to have a global picture of the drilling operation.

INTRODUCTION

Unsupervised analysis is a study of finding possible interesting and correlation relations among variables of a data set. If we project a set of available variables within a data set onto a lower dimensional space we arrive at a new set of alternative variables which may lead to much deeper insights to understand underlying phenomena in the data set. These new set of variables are so-called as 'latent variables' which are a linear combination of the original variables. There are several methods for visualizing higher dimensional spaces such as principal component analysis (PCA), factor analysis (FA), projection pursuit (PP) and multidimensional scaling (MDS). The most important method in visualizing higher dimensional space is the PCA which we use it in our analysis in this report. PCA is a popular method to recognize possible patterns in a data set. Depending on the field of application, PCA is named the eigenvalue decomposition in physics for instance. In this method, in effect, we determine new independent variables via variance of the data set. Each of these new coordinate variables (directions) in the data set are along the maximum variance which is so-called to principal component (PC). This way we find those variables which are highly correlated, introduce latent variables and conclude some varaibles explain the same underlying phenomenon. Consider a sample $\mathcal{X}$ where $\mathcal{T}$ has maximum variance via

$$\mathcal{T} = \mathcal{X}\mathcal{P}_1, \;\; \max(\mathcal{T}\mathcal{T}^T) = \max(\mathcal{P}_1^T \mathcal{X}^T \mathcal{X}\mathcal{P}_1), \qquad (1)$$

where $|\mathcal{P}_1| = \mathcal{P}_1^T\mathcal{P}_1 = 1$ is a constraint and $\mathcal{P}_1$ determines the first principal component. This first PC gives the major variation. The next PCs carry smaller variations and thus should have less importance in data analysis (it is worthwhile to be mentioned that the mentioned PCs are orthogonal). This is an optimization problem in the presence of constraints which, in effect, can be solved via the Lagrange multipliers λ. The next principal components ($\mathcal{P}_2$, $\mathcal{P}_3$, ...) can be determined via the Lagrange multipliers.

If we assume matric $\mathcal{X}$ has a dimension $M \times N$, we may define a 'score' matrix $\mathcal{T}$ which is $M \times K$ where K is the number of principal components, M is the number of objects, and N is the number of variables in the raw data set. The 'loading' matrix also contains the number of defining directions of the different principal components. By these definitions, $\mathcal{X}$ can be expressed by the multiplication of 'score' and 'loading' matrices:

$$\mathcal{X} = \mathcal{T}\mathcal{P}^T. \qquad (2)$$

All K components of $\mathcal{T}$ (all principal components) describe the system exactly. However, we need to reduce the dimension of the system. Thus we may introduce a residual matrix and just investigate J components of $\mathcal{T}$. Therefore we may write:

$$\mathcal{X} = \mathcal{T}_J\mathcal{P}_J^T + \mathcal{R}_J. \qquad (3)$$

The goodness of a PCA model can be expressed by how much variation has the residual matrix $\mathcal{R}$. If we consider a single object i for variable j the residual is defined by;

$$\mathcal{R}_{ij} = \mathcal{X}_{raw-i,j} - \mathcal{M}_{i,j} - \sum_{\ell=1}^{K} \mathcal{T}_{i,\ell}\mathcal{T}_{\ell,j}. \qquad (4)$$

$\mathcal{M}$ is the mean value of given objects for a variable. This way, we shift the data to origin $\mathcal{X} = \mathcal{X}_{raw} - \mathcal{M}^T = \mathcal{T}\mathcal{P}^T + \mathcal{R}$. $\mathcal{X}_{raw}$ denotes our raw data set.

To obtain the optimal number of PCs, we may use the residual matrix $\mathcal{R}$. The total residual variance can be calculated by: $\mathcal{R}^2_{tot}(K) = \sum_{i=1}^{K} \mathcal{R}^2_J$. We then can plot the residual variance in terms of zeroth iteration:

$$\frac{\mathcal{R}^2_{tot}(K)}{\mathcal{R}^2_{tot}(0)} \times 100\%. \qquad (5)$$

Before calculating the mentioned quantities above, we may need to preprocessing calculations due to irrelevancy of variables' amplitudes. Mean subtraction (centering) is important to perform PCA and make sure that the first PC describes the direction of maximum variations.

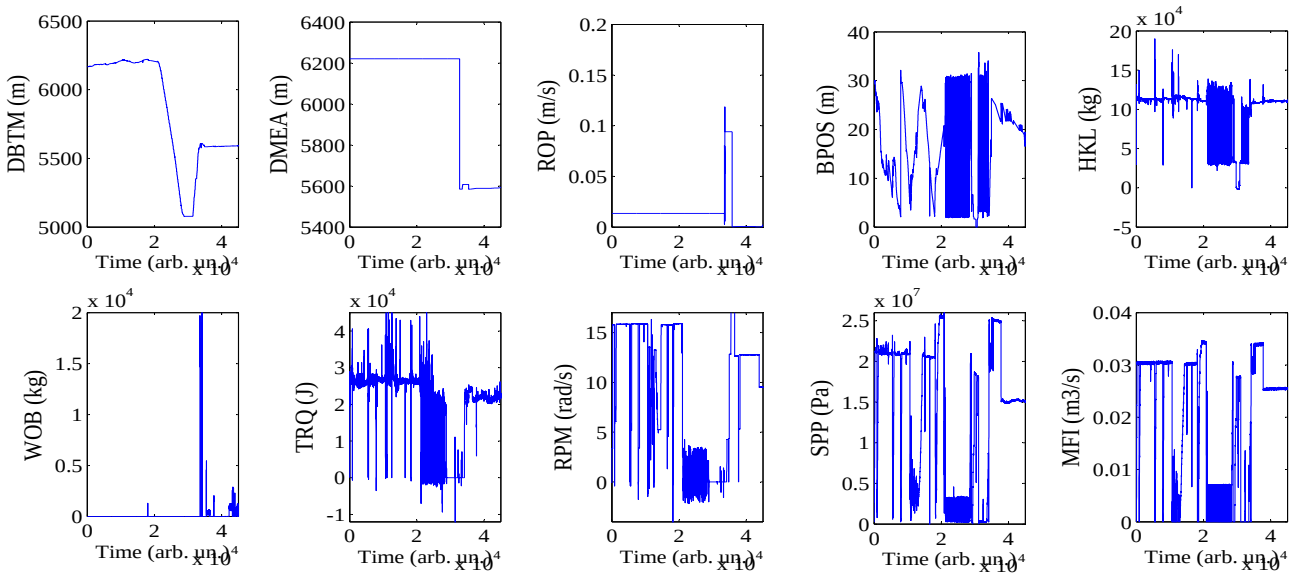

FIG. 1: (Color online) Raw data for ten variables involved the drilling operation. The time axis is considered for arbitrary units (arb. un.). There are approximately 45000 objects for each variable versus time. In these plots we have attributed a unit to each object.

This centering is also a preprocessing stage. Variables that have large overall amplitudes can change highly our analysis without being important. Therefore, to avoid such problems, it is necessary we autoscale our data set using the standard deviation of given objects of a specific variable. If we assume the standard deviation of jth variable can be given by Sd_j and mean value of the objects by M_j, the autoscaled data can be computed using the following formula:

$$\tilde{x}_j = \frac{x_{ij} - M_j}{Sd_j}, \tag{6}$$

where x_{ij} is an element in the original raw data set.

To have a good insight of our data analysis, we include the drilling variables in a table. The involved variable to the drilling operation are:

Abbreviation	Variable
DBTM	Bit Depth. The current depth of the drill bit, measured in meters below the drilling deck
DMEA	Hole Depth. The depth of the hole, measured in meters below the drilling deck
ROP	Rate of Penetration. Penetration speed while drilling in meters per hour
BPOS	Block Position. The position of the block, measured in meters above the drilling deck
HKL	Hook load. This is the weight of the drill pipe and bottom hole assembly in metric tonnes
WOB	Weigh on Bit. WOB is measured in metric tonnes
TRQ	Torque. The amount of torque in kilo Newton meters applied on the rotary engine to maintain the current RPM
RPM	Rounds per Minute.
SPP	Stand Pipe Pressure. The SPP is a measurement of the pressure of the fluid at the top of the drill pipe and is measured in bars
MFI	Flow rate of mud into the pipe. This rate is directly controlled by the driller and is measured in liters per minute

There are ten variables which an abbreviation is assigned to each variable. In what follows, we try to analyze the data set and find some useful relations among the variables to interpret the drilling operation.

PRINCIPAL COMPONENT ANALYSIS

To begin with, we plot the raw data set for each variable in Fig. 1. Since the scales of the variables are dif-

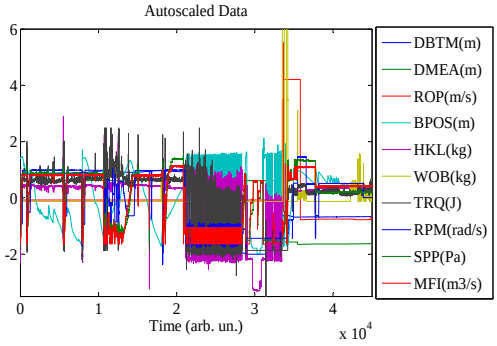

FIG. 2: (Color online) Autoscaled objects for all the involved variables as a function of time with arbitrary units.

ferent, we have presented them in ten different panels. Nonetheless, as seen, the plots carry much and fast oscillations that make them impossible to be interpreted and even make any conclusion. The only thing that may be concluded is that DBTM, DMEA, ROP, and WOB might have some kind of correlations together which is perhaps absent among the other variables since these mentioned variables have an abrupt change in a more or less identical time ($\sim 3.5 \times 10^4$). (We may also plot the variables against each other in the same plots to get more insights regarding their behavior. However, since there are 10 variables involved the operation, it is much more efficient to recourse to other methods like PCA.)

To overcome the mentioned issue and see all the objects in the same scales, we use the introduced auto-scaling method in the previous section. PCA is sensitive to the scaling of variables. This scaling makes the PCA robust in finding a pattern or model for a data set. However, this preprocessing compresses the oscillations and fluctuations in data set. Figure 2 exhibits the auto-scaled data set as a function of time with an arbitrary unit (arb. un.). Now, the new data set is easy and valid for interpretations.

Since there are ten variables involved the problem, we use the introduced PCA to reduce the dimension of variables. To do so, we resort to the plot of residual variance or equivalently the 'explained variance'. The results are shown in Table I. This Table might be interpreted as the percentage of data that can be described as a function of PCs used. The table I shows the exact percentages that can be described using a certain number of PCs. To have an exact interpretation one should use 10 PCs that is equal to the number of original variables involved to the problem (indeed, this is unnecessary to take all the ten PCs into account to have a good prediction of variables and their correlations). In fact, it is sufficient we just take two PCs and determine the positive and negative correlations among the variable. However, for completeness, we have considered four PCs in our plots. According to table I, these four PCs can explain $\sim 81\%$ of information altogether.

The PCA is sensitive to outliers in the data which produce large errors. However, in data with for example correlation clustering, the realization of cluster points and outliers can be difficult. This is a usual possibility that there happens some irrelevant and problematic data in a data set due to measurement of something else. That can have any source but the main responsibility is that we exclude them from our data set to be able to find correct patterns. This highly helps us to have the best predictions of our data set. These problematic objects are so-called as 'outliers'. To find the outliers, we should transforms variable values corresponding to a particular data point in a frame. this way, we may realize easily the outliers in our data set and then exclude them in our calculations.

The score plots of our drilling data set are shown in Fig. 3. We have considered four PCs in this figure. However, as mentioned above, we just restrict ourself to the first two PCs in our analysis and interpretation. As seen, there seems some clusters in the data set and it is difficult to find the outliers and problematic values in the data set. To find possibly outliers in the data set, we may have to resort to other methods which demands much more efforts and works. Therefore, One of the main assumptions we consider here is that there is no problematic outlier in the data set and consequently the PCA is robust and we can use the results of PCA to recognize a pattern in the data set. In effect, this plot shows the samples (which in our drilling operation the samples are different times in which the given variables are measured) in the variable space where now the ten-dimensional space is projected to a two-dimensional space defined with PC1 and PC2 (left side of Fig. 3). The samples are mainly distributed over the entire (positive and negative) values of PC1 which describes 40.30% of data variations. However, the samples are restricted approximately to positive

Number of PCs	Explained variance (%)
1	40.30%
2	57.05%
3	70.63%
4	81.62%
5	88.62%
6	94.50%
7	96.95%
8	98.55%
9	99.85%
10	100.00 %

TABLE I: Explained variance defined with Eq. 5 to determine the number of principal components (PCs) with the calculated percentages. First PC (PC1) describes 40.30% of the given information. Likewise, with PC2, PC3, and PC4, 57.05%, 70.63%, and 81.62% of objects can be described. To describe precisely the data (which is indeed unnecessary) one must consider 10 PCs (see text).

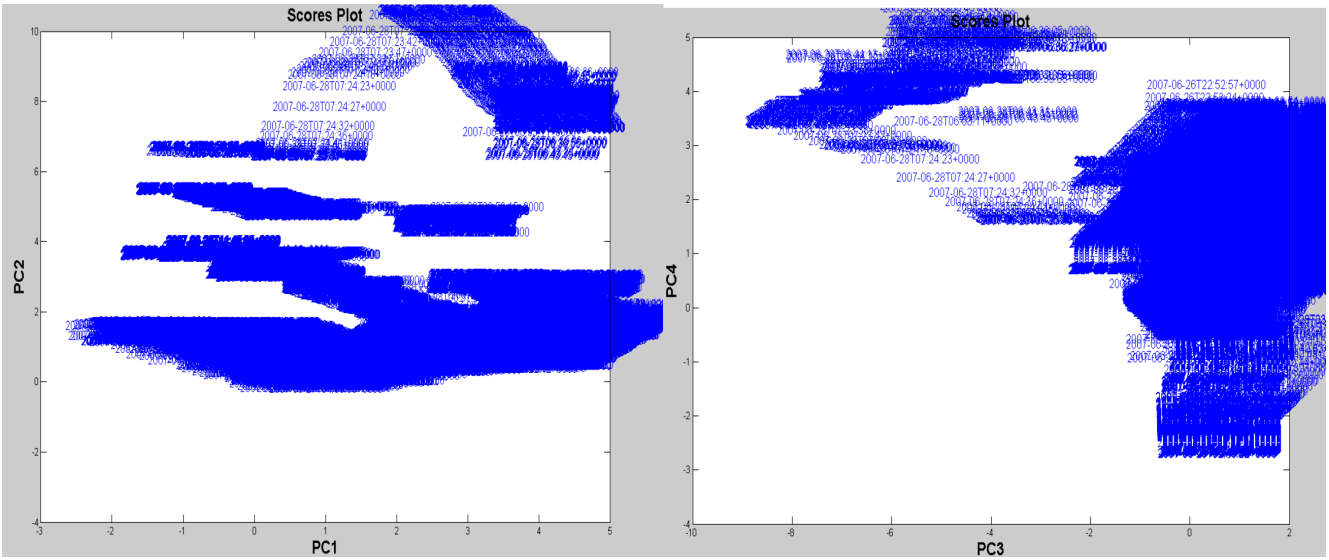

FIG. 3: (Color online) Transformed variable values corresponding to a particular data point (score plot). The left and right panels are corresponding to the left and right panels of Fig. 4.

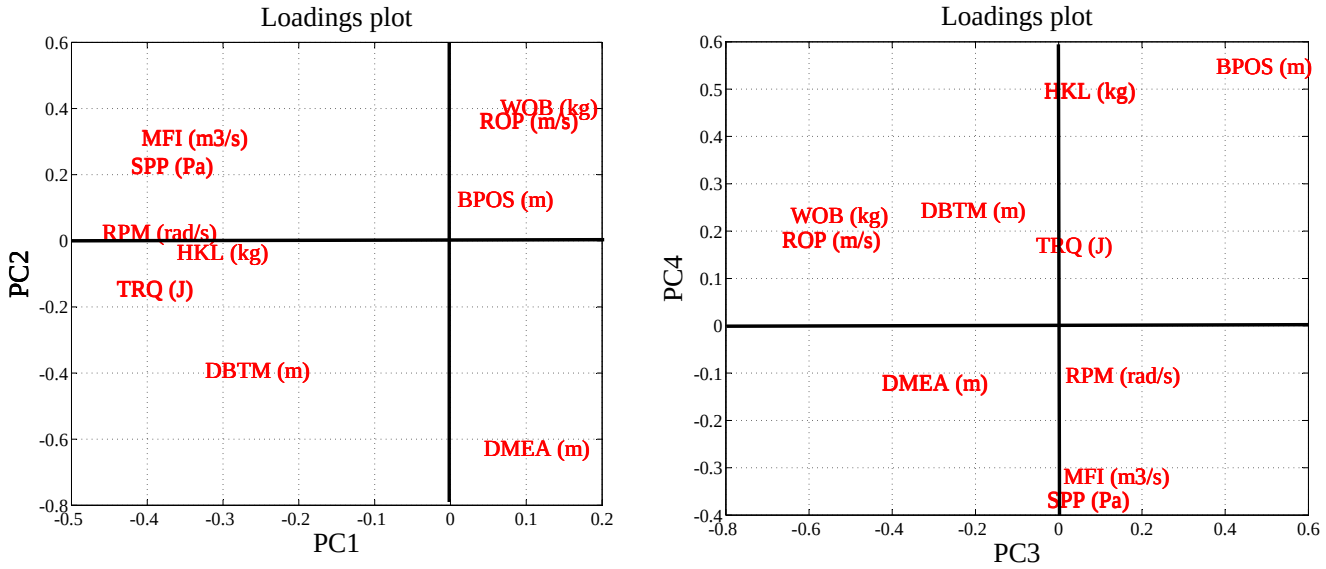

FIG. 4: (Color online) The weight that each original variable should be multiplied to get the component score which is called loading plots. Since we have considered four first principal components, we present them in two separate panels.

values of PC2 which describes 16.75% of data variations. This indicates an important information regarding the variables. The samples therefore should be distributed around some specific variables and it means those variables should have more weights in our analysis. Whereas those variables where the sample distributions are low for them should be attributed less wights in data analysis. To determine specific relations among our variables and data and then specify/predict the global picture for the drilling operation we need to have loading plots as well.

Now, the main task is that we determine the position of each variable in PC1, PC2, PC3, and PC4 space. To make interpretations as simple as possible, we have presented the projections of the variables on each PC in a 2D language in Fig. 4. From the loading plots of variables in Fig. 4 it is obvious that all the involved variables are important for describing the variations because we have no variable close to origin $(0, 0)$. Some variables are correlated together too. These variables are close together and thus they vary at least within $40.30\%(\text{PC1}) + 16.75\%(\text{PC2}) = 57.05\%$ of the variation that the two comports explain. Figure 4 illustrates the loading plots that together with scores plot (see Fig. 3) can be use to fully analyze the drilling real-time data

set. Different variables are shown by their abbreviations introduced in table I. The weight that each original variable should be multiplied to get the component score is called as 'loading'. Important information regarding the drilling operation can be extracted from this figure. Since the main variation can be captured by two PCs (57.05%), we here analyze the data by using the panel made of PC1 and PC2. Those variables which are close to each other are correlated variables. The correlated variables should have the same weights in our analysis and they basically behave the same as each other. Those variables which make exactly 90* together are uncorrelated variables. It means such variables have no specific relation together. Finally, those variables that are located in opposite directions of a PC are negatively correlated variables. Behavior of such variables are opposite. If one variable decrease the opposite ones are expected to increase while correlated variables are expected to decrease in a similar way. For the sake of simplicity, we just focus on the left panel of Fig. 4 in our following analysis. ROP and WOB (MFI and SPP) are highly correlated variables. RPM, HKL, and TRQ looks to be correlated too. However, WOP/ROP should be independent on RPM/HKL. ROP and WOP are in opposite signs of PC2 with DMEA. It means the two variables operate oppositely with DMEA. This is completely obvious from raw data plot too (Fig. 1). When DMEA decreases at a certain value of time ($\sim 3.5\times10^4$) the other two variables (ROP and WOB) increases. This also demonstrates that ROP and WOB are correlated variables. Comparing Fig. 3 and Fig. 4, DMEA has less importance since there is no sample in this region in scores plot. This is in stark contrast with WOB and ROP that many samples seem to be distributed in their region.

SUMMARY

In summary, we have used the principal component analysis (PCA) to interpret a few days of drilling operation. The number of principle components (PCs) needed to analysis the given data set is obtained. This is shown that with two PCs, 57.05% of variations can be captured. From the loading and score plot, we conclude the BPOS, RPM, HKL, ROP and WOP variables are among most important variables in the drilling operation. This is inconstant to DMEA and DBTM which have the less importance weighs. ROP and WOB are correlated variables while these two variables are negatively correlated to DMEA. Also we may conclude that RPM/HKL are approximately independent of ROP and WOP.

* Electronic address: phymalidoust@gmail.com