ISBN: 978-3-95450-132-8 doi:10.18429/JACoW-IPAC2014-THPME188

USING PRINCIPAL COMPONENT ANALYSIS TO FIND CORRELATIONS AND PATTERNS AT DIAMOND LIGHT SOURCE

C. Bloomer, G. Rehm, Diamond Light Source, Oxfordshire, UK

Abstract

Principal component analysis is a powerful data analysis tool, capable of reducing large complex data sets containing many variables. Examination of the principal components set allows the user to spot underlying trends and patterns that might otherwise be masked in a very large volume of data, or hidden in noise. Diamond Light Source archives many gigabytes of machine data every day, far more than any one human could effectively search through for correlations. Presented in this paper are some of the results from running principal component analysis on years of archived data in order to find underlying correlations that may otherwise have gone unnoticed. The advantages and limitations of the technique are discussed.

INTRODUCTION

Principal component analysis (PCA) is a technique used to find underlying correlations that exist in a (potentially very large) set of variables. The objective of the analysis is to take a set of n variables, $Y_1, Y_2, Y_3, ..., Y_n$, and to find correlations. The most important of these correlations are called the *principal components* (PCs). The analysis will return vectors Z_1, Z_2, Z_3, ..., Z_n, each describing a different underlying variation or trend found in the initial data set. The vectors of Z are ordered by their importance; that is to say the component Z_1 is the most prevalent trend seen throughout the data, and accounts for more variation than Z_2. Z_2 is a component uncorrelated with Z_1, and will account for the second largest trend seen in the data. Z_3 describes the third largest component, and so on. The 'importance' of each Z is determined by it's variance [1].

The rationale behind performing PCA on a data set is the idea that hopefully much, or perhaps even most, of the variation seen can be attributed to just a few of the most important principal components. A highly correlated data set can often be described by just a handful of principal components. Equally, it is possible for the analysis to produce no useful results at all if the original variables are highly uncorrelated.

Consider a simple case with two variables from Diamond Light Source, Y_1 and Y_2 (beam positions from two adjacent BPMs). These are plotted against one another in Fig 1. The first principal component, Z_1, is marked in blue and indicates the predominant correlation between the variables. The second component, Z_2, is marked in red and represents the variation between the variables that is uncorrelated with the first component. The length of the blue and red vectors are the standard deviation of each component ($\sqrt{\text{variance}}$), and thus indicate that component's importance. In this case, it is clear that the relationship between Y_1 and Y_2 can be primarily described by Z_1. Z_2 represents noise in this data.

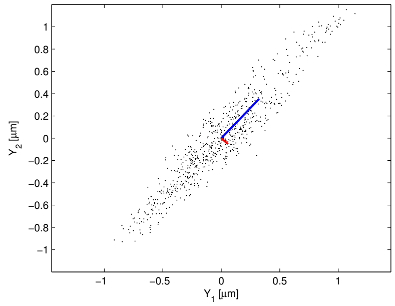

Figure 1: PCA for two highly correlated variables: two adjacent beam position monitors at Diamond Light Source. The blue and red lines indicate the first and second principal components respectively.

The chosen method to calculate the PCA presented in this paper is the singular value decomposition (SVD). This is a factorization of a matrix, based on a theorem from linear algebra. It states that a rectangular matrix, $\mathbf{A}$, of size $m \times n$, where m is the number of datapoints and n is the number of variables, can be broken down into a product of three matrices. Formally, this is usually written

$$\mathbf{A} = \mathbf{USV^T}$$

The computation of the SVD itself is beyond the scope of this paper, but is discussed in great detail elsewhere [2] [3] [4]. For the purposes of this paper, it is enough to know that the SVD is a very general method of computing the principal components of a data set, and that the SVD of a matrix can be robustly and quickly computed in many software packages (MATLAB, Python, C++, to name a few).

DATA SETS FROM DIAMOND LIGHT SOURCE

Diamond Light Source archives over 100,000 process variables (PVs), resulting in gigabytes of data being stored every day. For this study 1113 archived PVs relating to Diamond Light Source storage ring parameters were chosen for analysis (temperatures, vacuum pressures, X-ray beam positions, electron BPM quadrupolar differences $Q = (A{+}C)\text{-}(B{+}D)$, length encoders, loss monitors, and pinhole camera measurements have all been included).

Two years of data was retrieved from the archiver. A simple Gaussian low pass filter was then applied, the data is 'zeroed' by subtracting the mean value from each variable, and normalised so that the standard deviation of each variable is always 1. This normalisation process is important

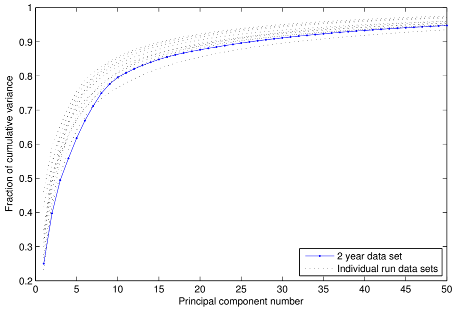

Figure 2: Cumulative variance accounted for by principal components for the whole two year data set, and for each individual 6-8 week user run.

as the PCA results can be influenced by the absolute magnitude and magnitude of variation in the input data.

In order to reduce the computational requirements, the data was then processed to remove periods with no beam (machine shutdown).

The PCA results for two timescales are presented in this paper. The full two year data set has been analysed for very long-term trends and correlation, and smaller subsets of the data have been analysed for short term variation seen over the course of a single user run (typically 6-8 weeks).

THE PRINCIPAL COMPONENTS OF TWO YEARS DATA

By performing PCA we obtain information about the contributions that each principal component makes to the total variance of the data. Over the complete two year data set it is found that a handful of principal components make up the bulk of the variance.

Figure 2 shows the *cumulative variance* accounted for by each principal component. Across 1113 variables just the first four principal components alone account for 50% the total variance, and 12 components account for an enormous 80% of the total variance. The 'total variance' is defined in this case to be the *sum of the variances of each variable in our data set*. It is truly remarkable that so few principal components account for so much of the variation seen in the original data set.

Figure 3 shows the two most important principal components identified for the two year span, along with a selection of the normalised variables themselves (shown here are the variables that exhibit the highest correlation with each shown principal component).

This is useful information as it allows effort to be focussed on investigating just the few most important trends. As the bulk of the variation seen in Diamond Light Source storage ring PVs can be accounted for by just a relatively small number of components it makes sense to concentrate efforts on investigating the causes behind just these largest contributors. Table 1 lists a few of the storage ring variables most closely associated with the first two principal components.

Table 1: A partial list of variables found to contain the largest contributions from the first and second principal components for two years of data. Steerer magnet strengths (STR), electron BPM quadrupolar differences (EBPM Q), PIN diode beam loss monitors (PIN), and physical length encoders measuring the distance between the electron BPM button blocks and the experimental floor (LENC) are all found to be correlated.

PVs correlated with:	
1st PC	**2nd PC**
VSTR 06-7	LENC 09-3
EBPM 16-5 Q	VSTR 13-3
VSTR 02-2	VSTR 13-4
VSTR 06-5	VSTR 14-4
HSTR 13-1	LENC 07-13
EBPM 04-1 Q	HSTR 24-6
PIN 12-17	EBPM 17-5 Q
LENC 18-4	EBPM 12-1 Q
LENC 15-3	LENC 19-6
...	...
25% of total variance	15% of total variance

ANALYSIS OF SHORTER TIMESCALES

Analysing two years of data will very likely result in very long term trends dominating the PCA results. It is instructive to also consider a smaller data set to determine whether or not the same lists of variables are found to be correlated over different timescales. Figure 4 shows the first and second principal components for a single user run (Run 2, 2013 - chosen to present here as it shows the longest continuous 300 mA storage ring run time).

It is interesting to note that at these (shorter) timescales the PCA has resulted in different principal components, with a different set of correlated variables. This illustrates an obvious aspect of PCA: one must carefully choose a sensible data set to use as input. If the analysis of long-term trends are of most importance, then one must input an appropriately long data set. If variation on short timescales are important, then providing a huge, year long, data set will not necessarily yield useful results.

Comparing the PCA results for each individual run within the complete two year data set shows some general trends: the same groups of PVs tend to be correlated, although the exact ordering of these groupings among the principal components is subject to change.

CONCLUSIONS

PCA is a valuable data analysis tool for investigating large volumes of data, and for identifying the principal trends and their related variables. It can quickly identify which principal components provide the largest contribution to variation in the data. Thus effort can be concentrated in trying to identify and understand these few most important components, rather than facing the daunting task of trying to guess which of the 1113 input variables might be of most importance.

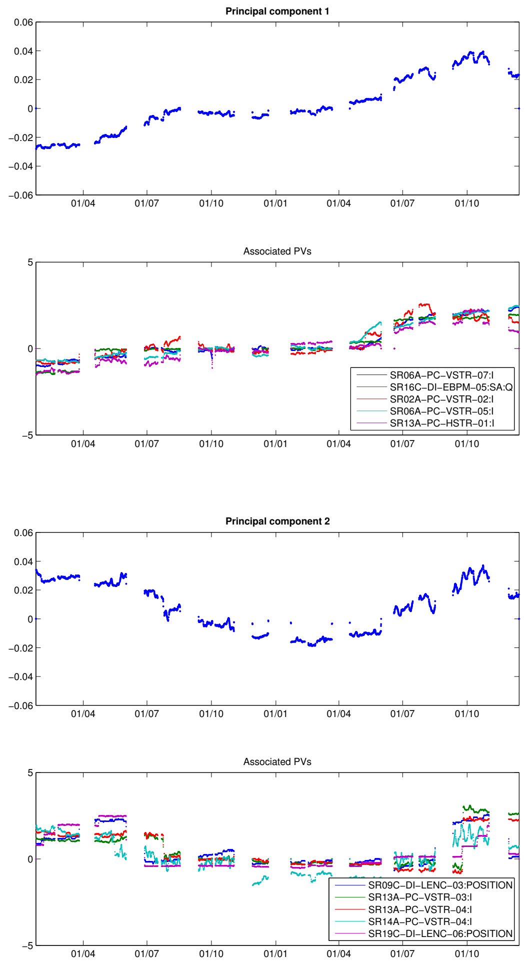

Figure 3: Two pairs of plots, showing the first and second principal components respectively, and those variables with the largest contribution from these components. The top two plots show the first PC; the bottom two plots show the second PC. The timespan shown is Jan 2012 - Jan 2014.

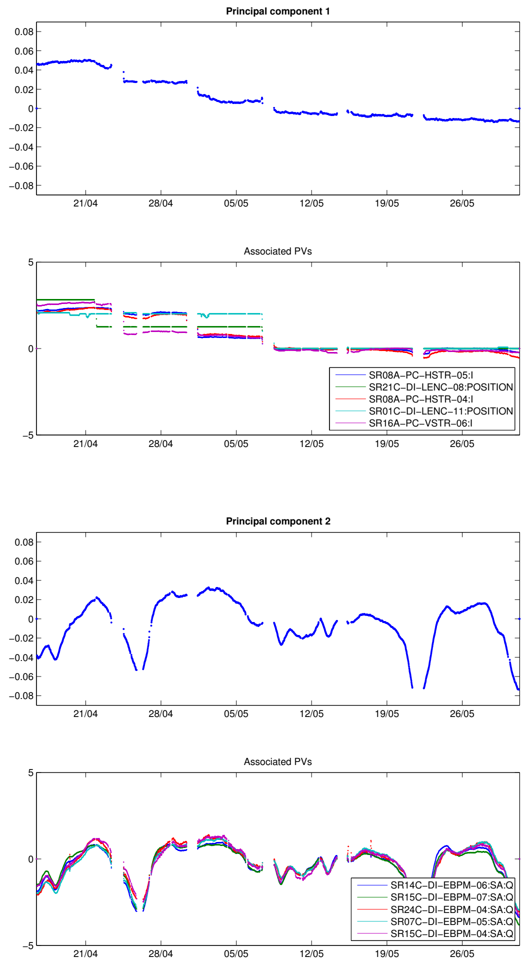

Figure 4: As per Fig 3. The top two plots show the first principal component; the bottom two show the second principal component. Their respective correlated variables are plotted underneath. The plots cover 7 weeks, Apr - Jun 2013.

However, these results also serve as an important example of the limitations of PCA: the analysis only provides information about correlations, it says nothing conclusive regarding causation. (Recall the important maxim *correlation does not imply causation*!) For example, the variables identified as being correlated with the principal components in Fig 3 and Fig 4 include corrector magnets, electron BPM Q, and storage ring length encoders, but the underlying cause of the variation is not identified. Human intelligence is still required to scrutinize the results and to draw conclusions.

Of particular interest to the authors is the correlation found between corrector magnet strengths, electron BPM Q, and in particular the length encoder measurements. Correlation between electron BPM electrical stability and corrector magnet variation has been long known, but the results of this analysis show that the physical movement of the electron BPM blocks is also correlated with the very long term trends in magnet variation. The correlation between length encoders and magnets strengths in entirely different sections of the storage ring is noteworthy. Applying our human intelligence to the PCA results, we can reason that the origin of this variation could be real, physical, machine movement over time, previously thought to be negligibly small.

REFERENCES

[1] B.F.J. Manly, *Multivariate Statistical Methods*, (Chapman & Hall/CRC, 2005).

[2] K. Baker, Singular Value Decomposition Tutorial: http://www.ling.ohio-state.edu/~kbaker/pubs/Singular_Value_Decomposition_Tutorial.pdf

[3] M. Richardson, Principal Component Analysis: http://people.maths.ox.ac.uk/richardsonm/SignalProcPCA.pdf

[4] J. Shlens, A Tutorial on Principal Component Analysis: http://www.cs.cmu.edu/~elaw/papers/pca.pdf