Available online at www.sciencedirect.com

ScienceDirect

IFAC-PapersOnLine 49-23 (2016) 512–517

Marine Engine Operating Regions under Principal Component Analysis to evaluate Ship Performance and Navigation Behavior

Lokukaluge P. Perera and Brage Mo

Norwegian Marine Technology Research Institute (MARINTEK), Energy Systems and Technical Operations, Trondheim, Norway (e-mail: prasad.perera@marintek.sintef.no, brage.mo@marintek.sintef.no).

Abstract: Marine engine operating regions under principal component analysis (PCA) to evaluate ship performance and navigation behavior are presented in this study. A data set with ship performance and navigation information (i.e. a selected vessel) is considered to identify its hidden structure with respect to a selected operating region of the marine engine. Firstly, the data set is classified with respect to the engine operating points (i.e. operating modes), identifying three operating regions for the main engine. Secondly, one engine operating region (i.e. a data cluster) is analyzed to calculate the respective principal components (PCs). These PCs represent various relationships among ship performance and navigation parameters of the vessel and those relationships with respect to the marine engine operating region are used to evaluate ship performance and navigation behavior. Furthermore, such knowledge (i.e. PCs and parameter behavior) can also be used for sensor fault identification and data compression/expansion types of applications as a big data solution in shipping.

Keywords: Principal Component Analysis, Big Data, Marine Engine Operations, Ship Performance Monitoring, Structured Data.

1. INTRODUCTION

Ship performance monitoring is with sensors and data acquisition systems facilitated to collect large data sets. These large data sets, often categorized as "big data", are analyzed in real-time to extract the information of ship performance and navigation behavior. Furthermore, the respective data handling process should consist of several improvement steps that include sensor fault identification, data compression and expansion to improve the quality of the process. These large scale data sets are often as unstructured formats considered due to various nonlinearities among ship performance and navigation information parameters. Therefore, an appropriate structure should be investigated to implement the improvement steps in a real-time data handling process of these data sets. That data structure can also be used to improve the quality of the data sets under the proposed improvement steps (i.e. sensor fault identification, data compression and expansion). Hence, the results of ship performance and navigation data can further be improved.

Such unstructured data sets are associated with various conventions ship performance and navigation models to estimate their respective parameters. However, the respective parameter estimation processes may degrade due to the complexities in such empirical models and these conventions ship performance and navigation models may have difficulties in handling large data sets. Furthermore, the system-model uncertainties under ship performance and navigation parameters can further degrade the outcome of the respective estimation process. Therefore, a methodology to identify the respective structure in a selected data set of ship performance and navigation information is proposed and that may initiate the respective steps towards future ship performance and navigation models.

This study proposes to learn the respective models from large data sets of ship performance and navigation information by considering Principal component analysis (PCA) (Shlens, 2014). PCA is a non-parametric method that extracts relevant information from chaotic type data sets and reduces the initial size of the data set to improve the content visibility. In general, the respective variance values are among the parameters in a selected data set identified by PCA. These variance directions that are orthogonal represent the respective principal components (PCs) of the data set. The top and bottom PCs held the most and least important information of the data set. Therefore, the most important information of the data set accommodates to the top PCs and that is by projecting the same data set into the selected top PCs done. Therefore, the new data set is a representation and that consists of the most important information of the old data set. Therefore, the bottom PCs are often ignored during this process because that may not consist of any important information of the data set.

The descending order of the PCs represents the order of significance (i.e. the order of variance) in the data set. When the same data set is into the least PCs projected, sensor faults and other erroneous data regions are often into these least PCs separated (Perera, 2016). Therefore, the new data set that is projected into the top principal components may have less sensor faults and other erroneous data regions. The same approach can also improve the data quality, where the least PCs can be used to identify sensor faults and other

Peer review under responsibility of International Federation of Automatic Control.
10.1016/j.ifacol.2016.10.487

erroneous data regions in the same data set. Furthermore, that information can be used to improve the quality of the data set and identify the respective faulty sensors.

The PC structure is as the basis for the respective models used and those models can be both for sensor fault identification used and data compression/expansion types of applications in shipping as a big data solution. However, a proper structure of the data set in ship performance and navigation information should be used in both situations (i.e. sensor fault identification and data compression/expansion). An inadequate structure of the data set can further degrade the outcome of the respective models and sensor fault identification and data compression/expansion steps. One should note that PCA has some limitations on finding accurate parameter relationships in some situations. This parameter inaccuracy may relate to the data point distribution and that may result in unusual parameter relationships within the data set. To overcome such challenges, a data clustering approach around main engine operating points is proposed in this study.

There are several steps to implement before PCA. In general, the respective data set should have an approximate Gaussian type distribution (i.e. appropriate mean and variance) to get appropriate results from PCA. If the data set consists of various data clusters, then that can introduce erroneous conditions in PCA. Therefore, the respective data clusters should be identified, so that PCA can be implemented for each data cluster, separately to improve results. Since marine engines of vessels are operating around various operating points, operating points are appropriate mean values for such data clusters in ship performance and navigation information. The proposed data clustering approach consists of Gaussian mixture models (GMMs) with an expectation maximization (EM) algorithm and uses to identify such marine engine operating regions (Perera and Mo, 2016a). Then, the data set should equally be centered and scaled (i.e. standardized), where each parameter is subtracted and divided by the sample mean and standard deviation values. The parameter variance related erroneous conditions can be avoided by this step of PCA. If the same parameters represent unusual relationships, then the data set should be further investigated to capture additional clustering dimensions. Therefore, a methodology to identify the respective structure in a data set of ship performance and navigation information is presented in the following sections and that can handle large data sets and implement in real-time data handling process (Perera and Mo, 2016b, c & d).

2. PRINCIPAL COMPONENT ANALYSIS

An overview of PCA is in this section presented. A ship performance and navigation data set (i.e. m number of parameters), denoted as:

$$X(t) = \begin{bmatrix} x_1(t) & x_2(t) & \dots & x_m(t) \end{bmatrix} \tag{1}$$

where $x_1(t), x_2(t), \dots, x_m(t)$ with $x_i(t) \in R^n$ are the respective parameters of ship performance and navigation information.

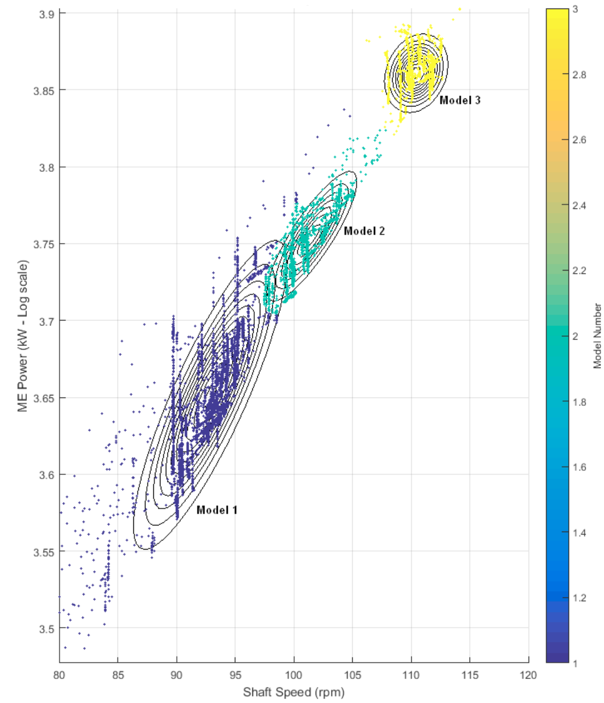

Fig. 1. Marine engine operating region classification.

The sample mean, $\bar{x}$, and variance S_x of the same data set are denoted as:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$
$$S_x = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x})^T \tag{2}$$

This data set is into a new data set transformed by considering the following transformation steps under PCA, denoted as:

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = u^T \bar{x}$$
$$S_y = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})(y_i - \bar{y})^T = u^T S_x u \tag{3}$$

where $\bar{y}$ is the mean of the new data set, and S_y is the respective variance of the transformed data set and u is a unit variance vector that uses to project the old data set into the new data set. Hence, that also satisfies:

$$u^T u = I \tag{4}$$

PCA maximizes the variance of each variance direction (i.e. principal component direction) of the new data set. Hence, the trace of S_y should be maximized and denoted as:

$$\text{Max. } trace(S_y) = \text{Max. } trace(u^T S_x u) \tag{5}$$

The Lagrange multiplier that satisfy (5) can be written as:

$$L = trace(u^T S_x u) = \sum_{i=1}^{n}\left[u_i^T S_x u_i + \lambda(1 - u_i^T u_i)\right] \tag{6}$$