대한조선학회논문집 Research Paper
Journal of the Society of Naval Architects of Korea
eISSN:2287-7355, Vol. 61, No. 1, pp. 8-18, February 2024
https://doi.org/10.3744/SNAK.2024.61.1.8

주성분 분석을 통한 선박 기관 상태의 차수 축소 모델링

이승범[1] · 서정화[1,2†] · 김동환[3] · 한상민[4] · 김관우[4] · 정성욱[4] · 유병우[2]
충남대학교 선박해양공학과[1]
충남대학교 자율운항시스템공학과[2]
충남대학교 첨단수송체연구소[3]
(주)삼성중공업 조선해양연구소 자율운항연구센터[4]

Reduced Order Modeling of Marine Engine Status by Principal Component Analysis

Seungbeom Lee[1] · Jeonghwa Seo[1,2,†] · Dong-Hwan Kim[3] · Sangmin Han[4] · Kwanwoo Kim[4] · Sungwook Chung[4] · Byeongwoo Yoo[2]
Department of Naval Architecture and Ocean Engineering, Chungnam National University[1]
Department of Autonomous Vehicle System Engineering, Chungnam National University[2]
Institute of Advanced Transportation Vehicles, Chungnam National University[3]
Autonomous Ship Research Center, Samsung Heavy Industries Co., LTD.[4]

The present study concerns reduced order modeling of a marine diesel engine, which can be used for outlier detection in status monitoring and carbon intensity index calculation. Principal Component Analysis (PCA) is introduced for the reduced order modeling, focusing on the feasibility of detecting and treating nonlinear variables. By cross-correlation, it is found that there are seven non-linear data channels among 23 data channels, i.e., fuel mode, exhaust gas temperature after the turbocharger, and cylinder coolant temperatures. The dataset is handled so that the mean is located at the nominal continuous rating. Polynomial presentation of the dataset is also applied to reflect the linearity between the engine speed and other channels. The first principal mode shows strong effects of linearity of the most data channels to show the linearity of the system. The non-linear variables are effectively explained by other modes. second mode concerns the temperature of the cylinder cooling water, which shows small correlation with other variables. The third and fourth modes correlates the fuel mode and turbocharger exhaust gas temperature, which have inferior linearity to other channels. PCA is proven to be applicable to data given in binary type of fuel mode selection, as well as numerical type data.

Keywords : Singular value decomposition(특이값 분해), Marine engine(선박 기관), Principal component analysis(주성분 분석)

1. 서 론

선박 운항에서 주기관의 성능 유지는 가장 중요한 요소이며, 이의 고장 진단에 대한 연구는 꾸준히 이뤄져왔다. 그리고 선박의 무인화와 탄소집약도지수(Carbon Intensity Indicator, CII)에 기반한 운항 규정이 적용되면서 선박 기관의 모니터링에 대한 중요도는 더욱 커져가고 있다 (Korean Register, 2021).

선박 기관의 고장 진단에서 가장 기본적으로 사용된 기법은 각 변수 간의 상호상관도(correlation coefficient) 해석이다. Kim et al. (2006)의 연구에서는 선박 기관의 각 계통에 대해 상호상관도가 1에 가까워 선형성이 강하게 확인된 계통들 간의 비교를 통한 고장 진단법을 제안한 바 있다. Park et al. (2016)과 Park et al. (2017)의 연구에서는 Kim et al. (2006)의 연구와 유사하게 선형성이 확인된 데이터 간의 회귀분석 결과에서 벗어나는 데이터를 이상치로 지정하여 고장 진단을 하는 방안을 제안하였다. 이러한 선형성에 기반한 접근은 비선형적인 거동을 보이는 변수에 대해 고장진단이 어렵다는 문제와, 여러 변수의 복합적인 이상 보고 시 장비 고장과 센서 신호의 이상을 분간할 수 없다는 문제가 있다.

이러한 문제에 대응할 수 있는 기법이 차수 축소 모델(reduced order modeling)이다. 차수 축소 모델은 주어진 데이터에서 주요한 변화 양상이나 패턴을 뽑아내어 더 적은 수의 변수로도 문제를 근사하는 기법이다. 차수 축소 과정에서 국부적인 데이터 변

Received : 4 October 2023 | Revised : 30 November 2023 | Accepted : 8 December 2023
† Corresponding author : Jeonghwa Seo, j.seo@cnu.ac.kr

동은 무시되므로 데이터의 잡음 제거에 효과적이다 (Kim et al., 2020a). 앞서 설명한 결측치나 이상치 문제에 차수 축소 모델을 적용한다면, 차수 축소 모델을 통해 기존 데이터에서 인식된 통상적인 패턴과 현재의 데이터의 차이를 감지하여 센서 신호의 이상 판별 문제에 접근할 수 있다 (Park et al., 2023).

많은 채널의 데이터에 대한 차수 축소 모델을 통한 고장 진단 방법으로, Kim et al. (2020b)의 연구에서는 Gaussian mixture model을 이용한 이상치 탐지가 제안되었다. Kim et al. (2022)은 autoencoder 기법에 기반한 선박 운항 데이터 분석을 수행하였다. 여러 데이터 채널에 대해 상호상관도 해석을 통해 데이터 그룹을 생성하고, 이 그룹에 대한 autoencoder 기반의 데이터 재구성을 통해 이상치를 탐지해내는 방안을 제안하였다. 이외에도 Park et al. (2023)은 장단기메모리(long short-term memory) 알고리즘을 이용하여 2행정 저속기관의 시계열 데이터의 신뢰성을 평가하는 연구를 수행하였다.

이런 데이터 기반 모델링은 데이터 해석의 중간 과정을 사용자가 이해할 수 없는 블랙박스(black box) 접근법을 주로 취하고 있다 (Savage, 2022). 하지만 사용자의 도메인 지식(domain knowledge)이 충분히 갖춰지고 각 채널 간의 물리적 상관관계가 충분히 이해되는 선박 기관 시스템에 대해서는 해석 과정을 사용자가 추적하며 이해할 수 있는 화이트박스(white box) 접근법이나 설명가능한 인공지능(explainable artificial intelligence, XAI)의 활용이 더 유용할 것으로 판단된다 (Loyola-Gonzalez, 2019).

사용자의 이해가 가능한 데이터 기반 접근 방법 중 하나로 주성분분석(Principal Component Analysis, PCA)이 있다. 이는 데이터를 몇 개의 주요 데이터 패턴의 선형결합으로 분해하는 기법으로, 크지 않은 비선형성을 가진 시스템에 대해 높은 신뢰도로 정상 패턴을 벗어난 데이터가 발생했는지를 검출할 수 있다 (Kim et al., 2003). 주성분 분석은 분야에 따라 특이값 분해(singular value decomposition), 적합 직교 분해(proper orthogonal decomposition) 등으로 불리기도 한다 (Lee, 2017). 조선해양공학 분야에서의 PCA 적용 사례로는 난류 유동장의 주요 모드 분석 연구가 있는 정도로, 선박 외의 분야에서의 적용 사례와 비교했을 때 기관의 고장 진단이나 센서 신뢰성의 검증에 적용된 바는 없는 것으로 판단된다 (Paik et al., 2010; Lee et al., 2012; Shin et al., 2017).

선박 기관은 기본적으로 엔진 부하에 따라 다른 물리량의 변화가 수반되기 때문에 비선형성이 지배적이지는 않을 것으로 기대된다. 하지만 일부 변수들에 대해 비선형성이 보고되었으며, 이는 Kim et al. (2006) 이래의 상호상관도 기반 접근의 한계로 작용하고 있다. 따라서 주성분분석을 적용해 기관 운용 데이터의 비선형적인 변수를 다룰 수 있는지가 문제의 핵심이라 할 수 있다.

본 연구에서는 선박의 운항 데이터 분석과 관련해, 비선형적인 변수들을 식별하고 이를 주성분 분석을 통해 설명하는 과정을 보였다. 그리고 주성분 분석을 이용해 이상치를 식별해내는 방안을 제안하였다. 최신의 해석 기법을 적용하기보다는 전통적인 해석 기법중 하나인 주성분 분석을 이용해, 데이터의 선형성을 강조하기 위한 데이터 전처리 기법의 제안, 도메인 지식을 활용해 해석 결과의 물리적 의미를 설명, 주성분 분석을 통한 데이터 재구성 결과의 검증의 세 가지 측면에 초점을 두었다.

논문의 구성은 다음과 같다. 2장에서는 본 연구에서 사용된 해석 기법인 PCA를 소개했다. 3장에는 해석 대상이 된 운항데이터의 개략적인 내용과 전처리 과정의 고려사항을 다뤘다. 4장에는 특이값 분해를 통해 얻은 운항 모드에 대한 설명과, 주성분 분석을 통해 재구성한 운항 데이터와 원본 데이터의 비교를 통한 검증을 보였다. 5장에는 본 연구의 결론을 수록하였다.

2. 주성분 분석

PCA는 시간에 따른 불규칙한 데이터의 변동을 포함한 복잡한 물리 현상에서 시간에 독립적인 패턴을 추출해내는 기법으로, 카루넨-루베(Karhunen-Loeve) 변환으로도 알려져있다. 이에 대한 상세한 내용과 적용 사례는 Bishop (2006)에서 다루고 있다. 본 논문의 부록에서 본 연구에서 사용된 수식과 기호로 PCA의 수행 과정을 제공하였으므로, 2장에서는 3장과 4장 내용을 설명하는 데 필요한 사항 위주로만 정리하였다.

특정 시간 t에 얻어진 n개 채널의 데이터를 $n\times 1$의 열벡터(column vector) $\boldsymbol{x}(t)$로 나타낸다고 했을 때, 이를 시간과 무관한 $n\times 1$ 열벡터 $\boldsymbol{v}_i$와 시간에 따라 변하는 계수 $a_i(t)$의 선형결합으로 분해할 수 있다.

$$\boldsymbol{x}(t)=\begin{bmatrix}x_1(t)\\ \vdots \\ x_n(t)\end{bmatrix}=\sum_{i=1}^{n}a_i(t)\boldsymbol{v}_i \tag{1}$$

PCA에서는 n보다 작은 수의 k개의 정규직교벡터를 $\boldsymbol{v}_i$로 이용하여 생성(span)된 공간에서 $\boldsymbol{x}(t)$를 근사한다. Fig. 1에 2차원의 데이터에 대한 PCA 적용의 개념을 나타냈다. 두 변수 간의 상관관계가 강한 $\psi-\eta$ 평면 상의 데이터 $\boldsymbol{x}(t)$에 대해, 전체 변수의 변화를 가장 잘 나타낼 수 있는 새로운 직교 좌표축 ψ', η'을 찾는다. 이제 ψ' 축 1개만 기저로 이용해 더 낮은 차원의 공간을 만들어도 데이터를 어느 정도 설명할 수 있게 되므로 전체 데이터를 ψ' 축에 사영(projection)하여 1차원의 데이터로 근사하여 축소할 수 있다. 사영의 크기는 식 (1)의 $a_i(t)$에 대응되며, 데이터 벡터 $\boldsymbol{x}(t)$와 ψ' 축 방향의 단위벡터 $\hat{\psi}'$의 내적으로 구할 수 있다.

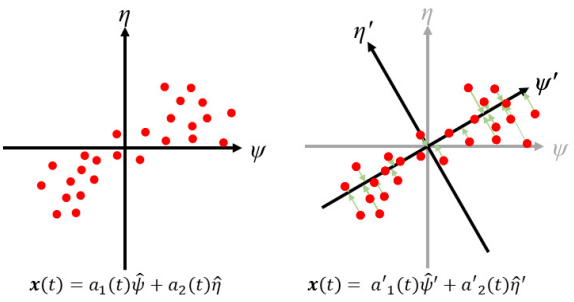

Fig. 1 Principle of principal component analysis of two-dimensional data

PCA에서 정규직교기저 $\boldsymbol{v}_i$는 데이터의 공분산 행렬 $\boldsymbol{A}_C$의 고유벡터 $\boldsymbol{p}_i = [p_1, \cdots, p_n]_i$와 이에 대응되는 고유값 λ_i를 이용해 구한다. 만약 λ_i가 크기순으로 정렬되어 있다면 n보다 작은 k번째까지의 고유값과 그에 대응되는 고유벡터만으로도 $\boldsymbol{x}$에 근사한 열벡터를 복원할 수 있게 된다. 이때 식 (2)를 이용해 k번째의 고유값까지를 통해 복원(reconstruction)한 근사 결과와 원래의 $\boldsymbol{x}$와의 유사도(similarity) s_k를 정량화할 수 있다. 본 연구에서는 k의 선택 기준을 s_k가 99%가 되는 값으로 정했다.

$$s_k = \sum_{i=1}^{k} \lambda_i / \sum_{i=1}^{n} \lambda_i \tag{2}$$

$\boldsymbol{p}_i$는 서로 직교한다는 특성을 이용해 $\boldsymbol{x}(t)$는 식 (3)과 같이 k개의 $\boldsymbol{p}_i$에 대한 투영(projection)을 통해 근사할 수 있다. 원래 n개 차원의 데이터가 k개 차원으로 축소되었음을 알 수 있다.

$$\boldsymbol{x}(t) \approx \sum_{i=1}^{k} a_i(t)\boldsymbol{p}_i = \sum_{i=1}^{k} (\boldsymbol{x}(t) \cdot \boldsymbol{p}_i)\boldsymbol{p}_i \tag{3}$$

앞에서 설명한 고유벡터 기반의 차원 축소를 적용하기 위해서는 주의해야 할 사항이 있다. 우선, PCA를 위한 데이터의 평균이 0이 되도록 조정해야 한다. 평균의 조정이 없다면 데이터의 재구성 과정에서 주요 모드의 공헌도가 올바르게 얻어지지 않는다 (Alexandris et al., 2017; Moeini et al., 2023).

Fig. 2의 좌측은 평균이 0인 2차원의 데이터세트이고, 우측은 $\overline{\boldsymbol{x}}$의 평균을 갖는 데이터세트이다. 이 둘의 공분산 행렬은 서로 동일하므로, 두 공분산 행렬의 고유값 분해를 통해 얻는 고유벡터 $\hat{\psi}'$와 $\hat{\eta}'$은 같게 된다. 새 기저 $\hat{\psi}'$와 $\hat{\eta}'$중 지배적인 $\hat{\psi}'$로 데이터를 근사하는 것이 PCA의 주요 내용이나, 해석 대상 데이터의 평균이 0이 아닌 경우에는 $\hat{\psi}'$와 $\hat{\eta}'$를 함께 사용하지 않고서는 $\boldsymbol{x}(t)$를 근사할 수 없게 된다. 따라서 실제 데이터의 주요한 변동을 표시하지 않는 $\hat{\eta}'$가 데이터의 재구성에 계속해서 동원되므로 데이터의 저차원화가 올바르게 이뤄지지 않음을 알 수 있다.

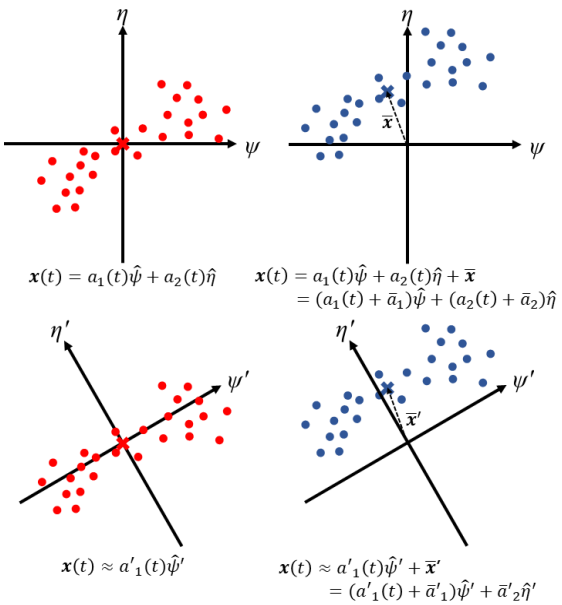

Fig. 2 PCA for two datasets: mean at the origin (left) and non-zero mean (right)

주요 모드의 공헌도와 관련하여 추가로 고려할 것은 변수의 적절한 전처리를 통해 변수 간의 선형성을 강조해야 한다는 점이다. 변수 간 선형성이 강할수록 첫 번째 주요 모드에 데이터의 전체적인 경향이 집중하여 나타난다. 그리고 나머지 주요 모드에서 비선형적인 특성이 더 잘 식별될 수 있다.

또 다른 고려사항으로, 특정 채널의 데이터의 편차가 크다면 고유값과 고유벡터의 해석 결과가 편향될 수 있다. 따라서 식 (4)와 같이 공분산을 정규화한 피어슨상관계수(Pearson correlation coefficient) $r_{i,j}$를 이용할 필요가 있다. 물론 특정 변수의 영향을 강조하기 위해서라면 정규화 과정을 조정하여 대상 변수의 편차가 다른 변수에 비해 커지도록 조정할 수 있다.

$$r_{i,j} = \frac{\mathrm{Cov}(x_i, x_j)}{\sqrt{\mathrm{Cov}(x_i, x_i)\mathrm{Cov}(x_j, x_j)}} \tag{4}$$

앞의 내용을 바탕으로, 본 연구에서의 PCA 활용은 다음 세 가지 질문의 답을 구하는데 초점을 두었다.

- 데이터의 전처리를 어떻게 할 것인가?: 주어진 운항 데이터 $\boldsymbol{X}$로부터 $\boldsymbol{p}_i$를 구하는 과정에서 선형성을 최대한 확보할 수 있는 방법의 개발
- 얻어진 주요 고유벡터의 물리적 의미는 무엇인가?: $\boldsymbol{p}_i$를 구성하는 계측 채널 데이터 간의 물리적 상관관계를 해석하고, 주 기관의 주요 운용 조건과 연계하여 설명
- 비선형적 변수의 이상치 분석이 가능할 것인가?: PCA를 통한 비선형적 변수들의 재구성 결과 평가

3. 해석 대상 데이터

3.1 데이터 목록

해석에 사용한 운항 데이터는 삼성중공업에서 제공된 선박 1척의 운항 기록으로, 1분마다 기록되어 육상으로 전송된 값이다. 선박 기관의 모니터링에 대한 Park et al. (2023)의 연구에서 최소 1Hz의 데이터 계측을 통해 시계열 변화로부터 데이터 해석을 수행한데 반해, 본 연구에서 사용된 데이터는 시계열 해석에 충분한 수준의 시간 해상도를 가지지 못했다. 다만 PCA 해석은 특성 상 시계열의 데이터 변화는 고려치 않으므로, 본 연구를 수행하는데 데이터의 시간 해상도는 문제가 되지 않는 것으로 판단하였다.

선박 기관 운항 데이터의 시계열을 검토한 결과, 기관이 완전히 정지한 상황에서 계측된 데이터의 비율이 작지 않았다. 이때의 데이터는 선박 기관의 상태를 식별하는데 별 의미는 없지만

Table 1 List of sensor channels of the marine engine

No.	Group	Channel name	Unit	Minimum	Maximum	Dynamic range	Standard deviation	Mean value at NCR
1	Engine load	Speed	rpm	−53.5	80.1	133.5	13.7	79.6
2		Power	%MCR	0.0	77.8	77.8	17.7	75.0
3	Fuel mode	LSMGO	–	0.0	1.0	1.0	0.441	0.3
4	Exhaust gas temperature	Temperature (Cyl. 1)	℃	45.3	394.0	348.7	58.0	360.6
5		Temperature (Cyl. 2)	℃	46.4	390.4	344.0	57.6	374.2
6		Temperature (Cyl. 3)	℃	45.7	376.0	330.4	54.8	366.5
7		Temperature (Cyl. 4)	℃	45.9	372.0	326.1	54.0	362.8
8		Temperature (Cyl. 5)	℃	46.5	377.7	331.2	53.8	362.5
9		Temperature (Cyl. 6)	℃	46.3	385.6	339.3	55.4	366.5
10	Turbocharger	Speed	rpm	0.0	12428.1	12428.1	2575	12223.6
11		Exhuast gas temp. before turbocharger	℃	23.6	406.3	382.7	59.9	393.9
12		Exhaust gas temp. after turbocharger	℃	22.0	251.6	229.6	25.8	219.2
13	Cylinder cooling water temperature	Water outlet temperature (No.1)	℃	70.4	92.9	22.6	2.60	87.5
14		Water outlet temperature (No.2)	℃	70.7	94.1	23.4	2.49	88.0
15		Water outlet temperature (No.3)	℃	70.5	93.8	23.3	2.51	87.7
16		Water outlet temperature (No.4)	℃	70.5	94.3	23.8	2.51	87.6
17		Water outlet temperature (No.5)	℃	70.4	93.8	23.3	2.49	87.6
18	Piston cooling water temperature	Water outlet temperature (Cyl. 1)	℃	43.4	56.6	13.2	1.80	54.9
19		Water outlet temperature (Cyl. 2)	℃	43.9	56.9	13.0	1.77	55.1
20		Water outlet temperature (Cyl. 3)	℃	43.4	56.5	13.1	1.79	54.6
21		Water outlet temperature (Cyl. 4)	℃	43.9	56.9	13.0	1.75	55.1
22		Water outlet temperature (Cyl. 5)	℃	43.8	57.3	13.5	1.85	55.3
23		Water outlet temperature (Cyl. 6)	℃	43.6	56.8	13.2	1.75	54.6

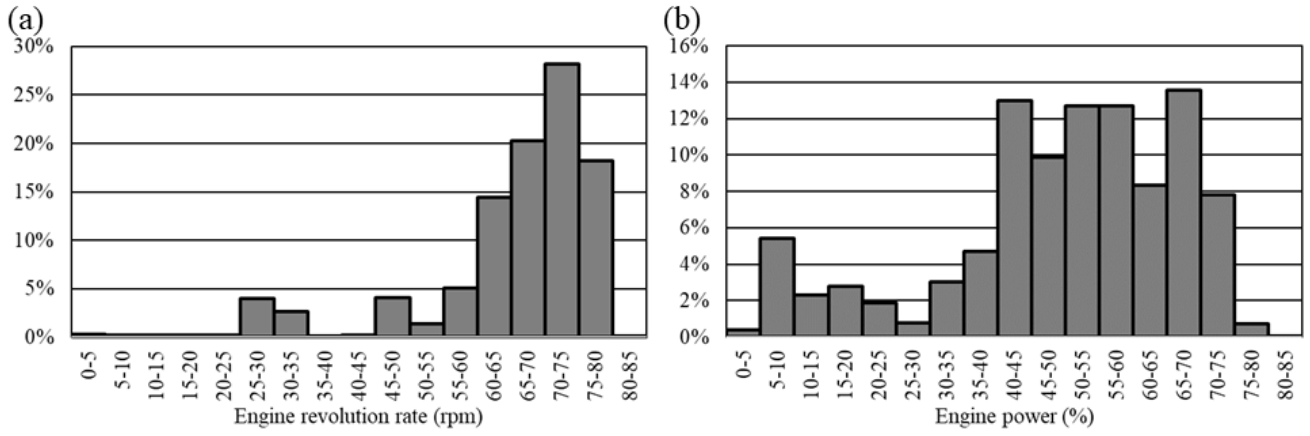

Fig. 3 Histogram of engine operation: (a) revolution rate and (b) power

데이터의 비율이 크기 때문에 주요 모드가 기관 정지 상태로 편향된다는 문제가 있어 해석 대상 데이터에서 제거하였다. 본 연구에서는 정상적인 운항 조건으로 판단되는 35,086개의 데이터를 이용할 수 있었다. 이는 약 24일의 운항데이터에 상응한다.

실제 데이터는 항로, 운항 상태, 화물탱크 데이터 등을 포함하나, 본 연구에서는 주기관과 관련한 계측값만을 사용하였다. 대부분의 데이터는 기관 각 부분의 온도로, 기관 회전수나 출력과 양의 상관관계를 강하게 가질 것으로 생각되었다. Table 1에 주기관에서 계측된 데이터의 종류를 정리하였다.

데이터는 크게 6개의 그룹으로 나눌 수 있다. 첫 번째 그룹의 엔진의 회전수, 출력은 엔진 부하 상태를 직접 알려주는 항목이다. 1번 채널이 역수가 나온 경우는 기관이 역회전하는 경우로, 이런 경우에는 회전수의 절대값을 취해 양의 회전수를 갖도록 하였다. 두 번째 그룹은 연료 종류의 선택 상태이다. 저유황 연료유(Low

Sulphur Heavy Fuel Oil, LSHFO)와 저유황 경유(Low Sulphur Marine Gas Oil, LSMGO)의 연료 중 어느 것을 사용하는지 알려주는 것으로, 불리언 형식(boolean type)으로 표현되어야 하나 0과 1의 바이너리 형식(binary type)의 숫자로 표시하였다. 세 번째 그룹은 각 연소실을 통해 나오는 배기가스의 온도이다. 과급기인 터보차저 관련 데이터는 네 번째 그룹이 속한다. 실린더와 피스톤의 냉각수 온도는 각각 다섯 번째, 여섯 번째 그룹이다.

Fig. 3은 운항 중의 엔진 회전수와 출력 분포에 대한 히스토그램이다. 회전수는 상용출력(Normal Continuous Rating, NCR) 회전수인 79.6rpm 영역 근처에 집중되어 있으나, 기관 출력은 40 – 75% 영역에 걸쳐 분포되어 있었다. 추진기의 회전수 외에도 선박의 전진 속도에 따라 전진비가 변하며 기관의 부하가 달라지므로, 본 연구에서 다루지 않은 기관 외의 선박 전반의 운항 정보에 대한 종합적인 해석으로 그 특성을 파악할 수 있을 것으로 생각된다.

Fig. 4에 엔진 출력(P)과 회전수(N)의 관계를 나타내었다. N과 P의 다항식 관계에 대해서는 선체의 유효동력(P_E)이 선속(V_S)의 세제곱에 비례해 얻어진다는 식에 착안하여 그 관계를 식 (5)와 같이 제안하였다.

$$P_E = R_{TS} V_S = (0.5\rho V_S^2 S_S C_{TS}) V_S \propto V_S^3 \quad (5)$$

여기서 ρ는 물의 밀도, S_S는 선박의 침수표면적, C_{TS}는 전저항계수이다. 선속에 따른 C_{TS}나 자항 요소, 슬립비의 변화가 크지 않다면 식 (5)로부터 $N^3 \propto P$의 관계를 적용할 수 있다. 따라서 Fig. 4의 (b)에서 보인 바와 같이, 출력과 회전수의 세제곱은 선형적 관계를 보였다. PCA를 위해 다른 계측값들이 엔진의 출력과 회전수 중 어느 쪽에 더 선형적인 관계를 보일지를 뒤에서 확인하였다.

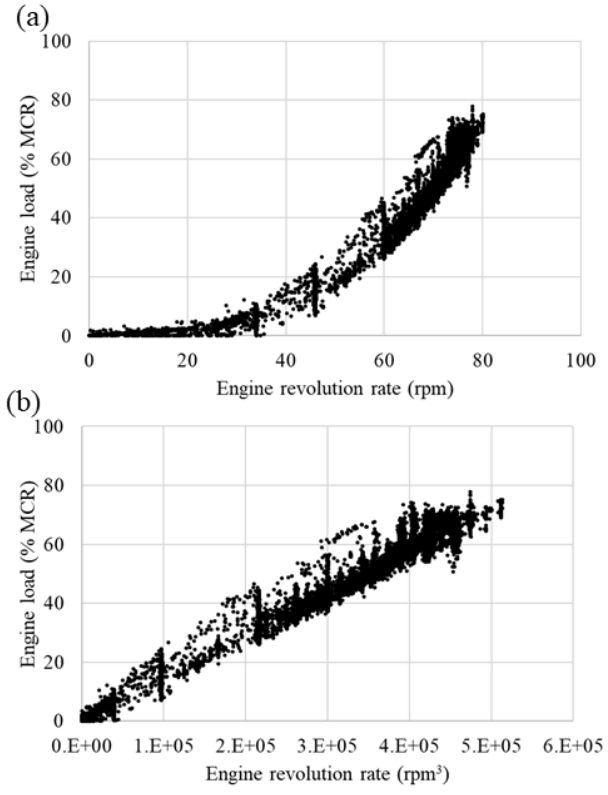

Fig. 4 Relationship between the main engine load and revolution rate: (a) N and P and (b) N^3 and P

3.2 데이터 전처리

Fig. 5에 본 연구의 PCA의 전반적인 수행 과정을 정리해 나타냈다. PCA는 주요 모드를 찾기 위한 작업과 주요 모드를 이용해 데이터를 재구성하고 데이터를 검증, 보정하는 과정으로 나뉜다. 제공된 데이터의 80%를 임의로 골라 주요 모드의 식별에 이용하였고, 나머지 20%는 4.2장에서 다룰 특이값 분해를 통해 얻은 주요 모드의 선형결합을 검증하는데 사용하였다.

2장에서 설명한 대로, 실제 데이터의 평균이 어떻게 평행 이동하든 주요 모드를 구성하는 고유벡터는 유지된다. 하지만 데이터 재구성 과정에서는 각 주요 모드의 비중에 대한 차이가 발생해 주요 모드를 인식하는데 문제가 발생할 수 있기 때문에 데이터의 평균을 0으로 조정해야한다. 여기서 한 가지 더 고려해야 할 점은, 본 연구에서 얻어낸 기관의 특성을 같은 기관을 사용하는 다른 선박에도 적용할 수 있도록 해야 한다는 것이다. 따라서 기관의 운항 기준 상태를 공칭 상태(nominal condition)인 NCR로 통일하는 것이 타당하므로, 데이터 전처리를 통해 데이터 분포의 원점에 NCR이 위치하면서 평균값이 0이 되도록 조정(mean-centering)할 필요가 있다.

P와 N을 예로 들어 데이터의 평균을 조정하는 과정을 Fig. 6에 나타내었다. Table 1에서 보인 NCR에서의 평균 데이터를 $\boldsymbol{x}_{NCR}$로 정의하고, 계측된 데이터 $\boldsymbol{x}(t)$에 식 (6)을 통해 데이터의 원점을 NCR로 평행이동한 새 데이터 $\boldsymbol{y}(t)$를 구한다.

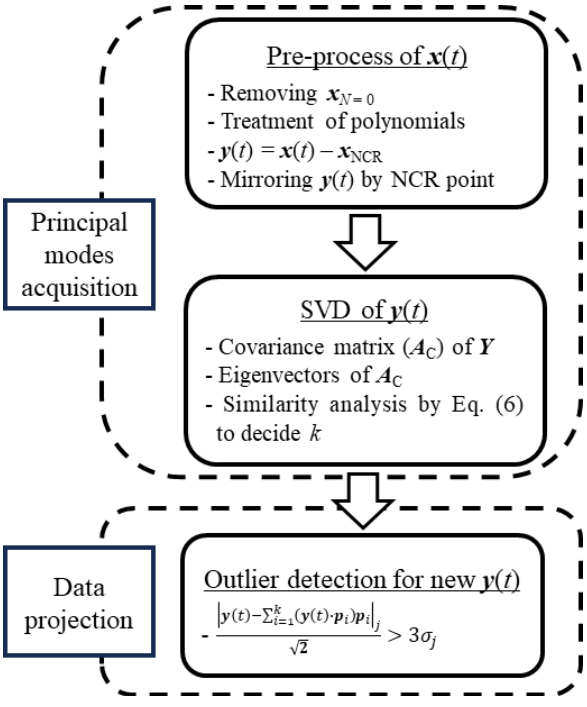

Fig. 5 Procedure of PCA and outlier detection

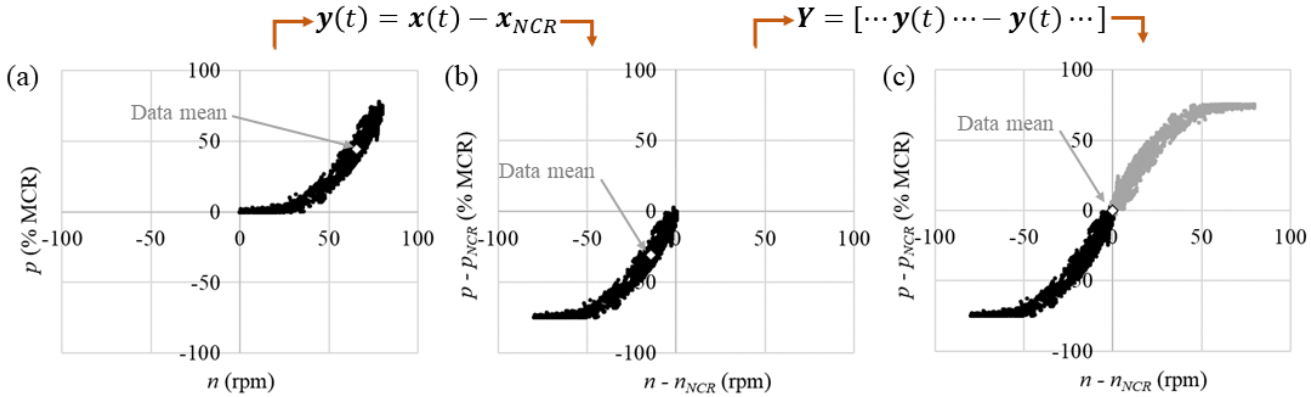

Fig. 6 Pre-processing of data mean (white symbol): (a) original data, (b) Transition of the origin to NCR by Eq. (6), and (c) mirrored data by Eq. (7) (gray symbols)

$$\boldsymbol{y}(t) = \boldsymbol{x}(t) - \boldsymbol{x}_{NCR} \tag{6}$$

이렇게 평행이동한 데이터 세트에 원점 대칭의 데이터 짝인 $-\boldsymbol{y}(t)$가 추가되어야 데이터의 평균이 0이 된다. 본 연구에서는 입력 데이터 행렬 $\boldsymbol{Y}$를 다음과 같이 전체 계측 시간 중 얻어진 $\boldsymbol{y}(t)$, $-\boldsymbol{y}(t)$를 열벡터로 갖는 행렬로 나타내었다.

$$\boldsymbol{Y} = [\cdots \boldsymbol{y}(t) \cdots - \boldsymbol{y}(t) \cdots] \tag{7}$$

변수간의 다항식 관계를 판단하기 위해 Table 2에 N, P의 다항식 해석에 따른 다른 그룹과의 상호상관도의 평균을 정리해 나타냈다. 여기서 회전수는 그대로 사용하되, 출력의 1/3제곱을 이용하는 것이 다른 변수와의 선형적 결과가 가장 잘 나타나는 것을 알 수 있다. 5번 그룹에 대해서는 다른 그룹들과 다른 경향이 나타났으나, 상호상관도 자체가 작기 때문에 변수의 다항화에서 고려하지는 않았다.

Fig. 7에 앞의 절차대로 전처리한 데이터의 공분산행렬 $\boldsymbol{A}_C$를 나타냈다. 전체적으로 같은 그룹 내부에서는 강한 상관관계가 잘 나타났고, 비선형성은 2번, 4번, 5번 그룹에서 확인되었다.

2번 그룹은 연료 모드에 대한 것으로 다른 변수간의 유의미한 상관관계가 드러나지는 않았기 때문에 운항 데이터에 대한 PCA로 연료 모드를 식별할 수 있는지가 주요한 관심의 대상이 되었다. 이에 대해서는 4.2장의 데이터 재구성에서 다뤘다.

4번 그룹의 과급기 관련 데이터에서는 과급기의 출구 온도의 상관관계만 낮게 나타났다. 과급기는 배기가스의 에너지를 빼앗아 흡기에 전달하는 기구이므로, 기관 출력과 강한 상관관계를 보이는 과급기 유입 가스의 온도와는 달리 과급기에서 에너지를 빼앗긴 결과는 출력과 직접적인 비례관계를 보이지 않게 된다.

5번 그룹인 실린더 냉각수 온도는 Table 1의 Dynamic range

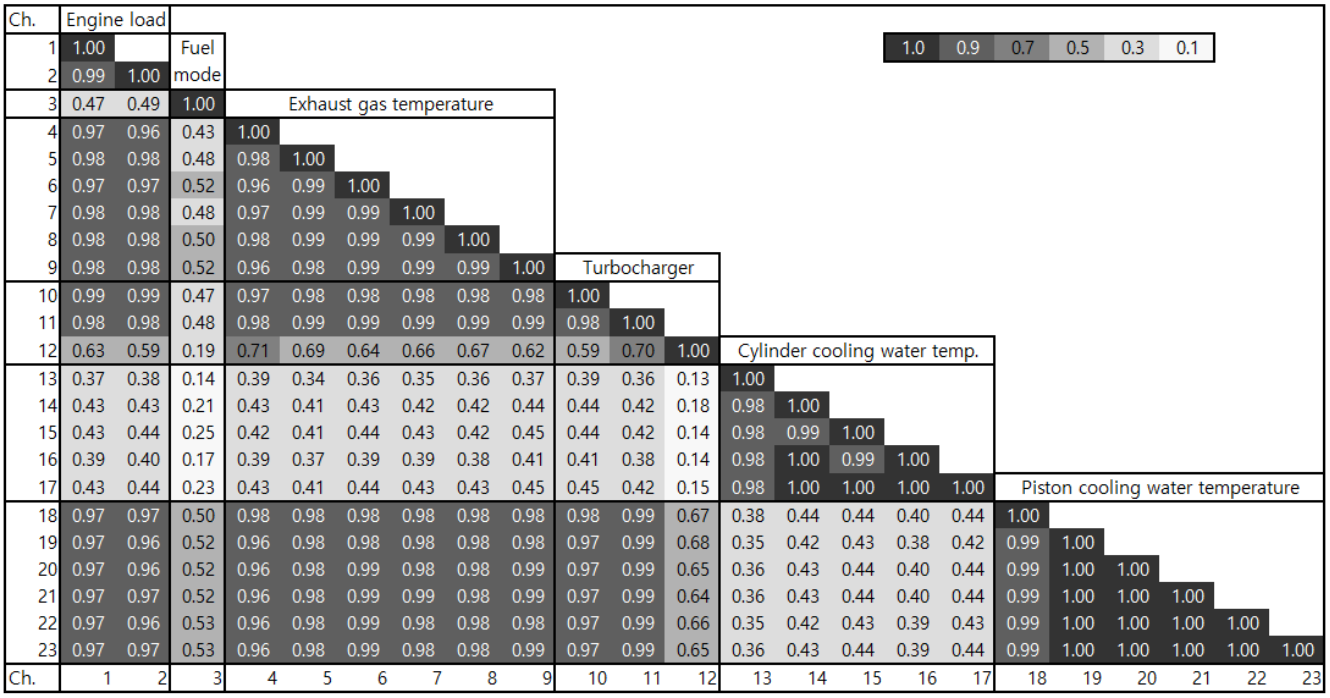

Fig. 7 Covariance matrix of main engine data

Table 2 Mean of Pierson correlation coefficient of Groups 3 to 6 with variation of order of n and P

	Group 3	Group 4	Group 5	Group 6
n	**0.978**	**0.870**	0.409	**0.968**
n^2	0.976	0.836	0.420	0.968
n^3	0.958	0.799	**0.426**	0.953
$P^{1/3}$	**0.975**	**0.853**	0.418	**0.965**
$P^{2/3}$	0.969	0.821	0.429	0.963
P	0.950	0.784	**0.434**	0.948

에서도 확인할 수 있듯, 기관의 구동 조건에 따른 변화가 작고, 최적의 실린더 온도를 유지하도록 구동되기 때문에 다른 변수와의 상관관계가 잘 드러나지 않았다. 다른 변수들은 전체적으로 1에 가까운 상호상관도를 보였으므로, 기관의 회전수에 비례하여 증가하는 것으로 볼 수 있다.

4. 데이터 해석 결과

4.1 고유벡터 해석

Fig. 8에 식 (2)에 따른 유사도(s_i)를 사용된 모드 수(k)에 대한 변화로 나타내었다. 이는 한정된 숫자의 주요 모드로 실제 데이터를 얼마나 잘 재구성할 수 있는지를 정량적으로 보인다. 첫 번째 주요 모드로는 76.2%의 유사도를 보였으며, 유사도 95%, 99%는 각각 세 번째, 다섯 번째 주요 모드에서 달성하였다. 본 연구에서는 99%의 유사도로 데이터를 재구성하기 위해 5개의 고유벡터를 주요 모드로 사용하였다.

주요 모드($\boldsymbol{p}_i$)의 각 채널 별 요소의 크기(p_i)를 Fig. 9에 나타내었다. 모든 $\boldsymbol{p}_i$는 그 길이가 1로 정규화되어있다. 한 주요 모드에서 각 점들이 0에서 균일하게 떨어져 있을수록 해당 모드에서 각 변수간의 상관관계가 크다는 의미이다. 빨간색으로 나타낸 첫 번째 모드에서 그룹 1, 3, 4, 6에 대한 정규화된 변수가 서로 비슷한 크기를 보이는데에서 강한 선형관계를 확인할 수 있다. 이는 이들 그룹의 관계에서는 첫 번째 모드만 이용해도 그 관계를 거의 설명할 수 있음을 의미하는데, Fig. 7의 공분산행렬에서 확인한 내용을 뒷받침한다.

앞서 Fig. 7에서 데이터 간의 선형성이 나타나지 않은 경우, 즉 그룹 2(연료 모드)와 그룹 4의 채널 12(과급기 출구 온도), 그룹 5(실린더 냉각수 온도)는 추가적인 모드에서 그 값을 설명할 수 있다. 그룹 5는 노란색으로 표시한 두 번째 모드에서 그 특성이 강하게 나타났으며, 다른 변수들과는 반대의 부호를 가졌다. 이는 다른 변수들의 변화에 비해 일정한 값을 유지하는 그룹 5의 특성을 잘 설명한다. 특이하게도 그룹 2와 그룹 4의 데이터 12는 녹색으로 표시한 세 번째와 네 번째 모드에서 상관관계가 나타났다. 이는 과급기의 출구 온도를 이용해 연료 모드를 식별할 수 있는 가

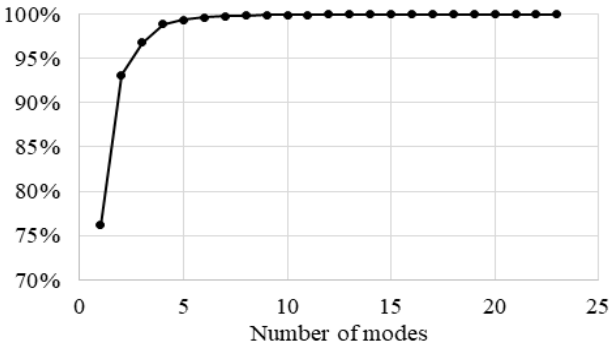

Fig. 8 Variation of similarity (s_k) by number of modes (k)

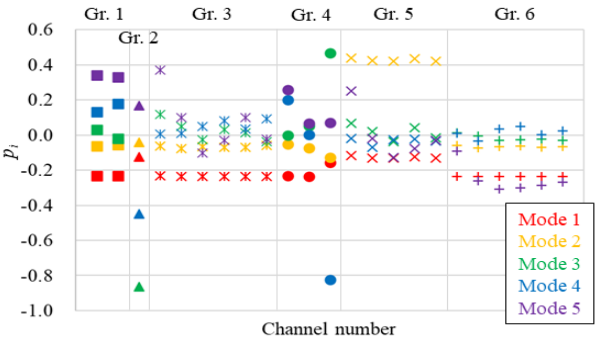

Fig. 9 Comparison of five dominant modes

능성을 의미한다. 이는 연료 종류에 따라 배기가스의 수증기와 이산화탄소 조성이 달라지므로, 배기가스의 열용량이 달라지면서 과급기를 거치면서 잃는 열에너지가 달라지는 것으로 생각된다.

다섯 번째 모드에서는 한 그룹 내에서 그룹을 구성하는 채널 간의 계측값 차이를 설명하는 성분들이 나타났으므로, 다른 변수 그룹 간의 상관관계를 식별하는데 이용하기는 어려워 보인다.

4.2 데이터 재구성

앞에서 주요 모드 식별에 사용하지 않은 20%의 데이터에 대해 식 (3)을 이용해 순간 데이터 $\boldsymbol{x}(t)$를 4.1장의 5개의 주요 모드를 이용해 재구성하고, 그 결과를 원래 값과 비교하였다. Table 3은 그 결과로, 원 데이터와 재구성한 데이터의 산포도에서 추세선의 기울기(slope)와 결정계수(R^2)를 나타냈다. 두 값이 1에 가까울수록 PCA를 통한 데이터의 재구성이 잘 이뤄졌음을 뜻한다.

다섯 개의 주요 모드만 사용하여 추정한 엔진 데이터는 실제 데이터와 잘 일치하였다. Fig. 7에서 비선형성이 크게 나타난 채널 3, 채널 12, 그룹 5 중 채널 13의 추정 결과를 Fig. 10에 나타내었는데, 이런 비선형적 변수들에 대해서도 추정 정확도가 우수함을 확인하였다.

비선형적 변수의 재구성과 관련해 특히 중요한 점은 채널 3의 연료 종류이다. 채널 3은 다른 채널과 달리 0과 1로 이산화된 형식으로 데이터가 제공되며, Fig. 7에서 보인 바와 같이 연료 종류의 선택이 기관의 물리량에 직접적으로 상관관계를 가지지는

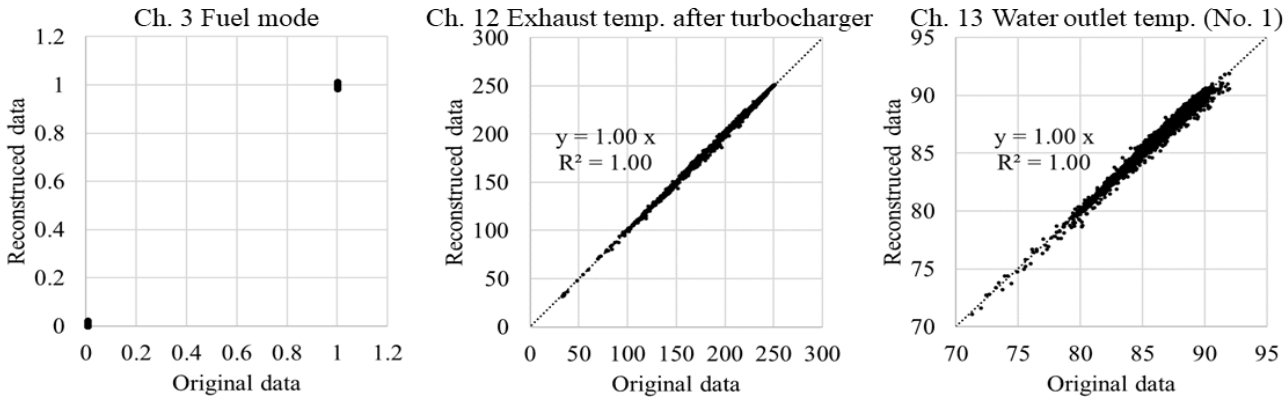

Fig. 10 Comparison of original and reconstructed data: fuel mode LSMGO (Ch. 3), Exhaust gas temperature after turbocharger (Ch. 12), and Water outlet temperature No. 1 (Ch. 13)

않는다는 점에서 PCA 해석에 어려움이 있을 것으로 예상되었다. 다만 이전의 연구들에서 logistic PCA를 통해 바이너리 형식의 데이터 해석이 가능함이 제기된 바 있었는데 (Leeuw, 2006; Landgraf and Lee, 2020), 본 연구에서는 특별한 기법 없이도 바이너리 형식의 데이터를 재구성할 수 있었다. 그리고 Fig. 9에서 보인 세 번째와 네 번째 주요 모드를 이용해 연료 모드와 과급기의 출구 온도간의 내재된 의미를 확인할 수 있었는데, Fig. 7에서 두 채널 간의 상호상관도는 0.19로 가장 낮은 축에 들었던 것을 고려하면 단순한 상호상관도의 해석으로는 이러한 관계를 찾아낼 수 없을 것이다.

Table 3 Validation of PCA: comparison of original and reconstructed data

No.	Group	Unit	Slope	R^2	σ
1	Engine load	rpm	1.00	0.98	1.36
2		%MCR	1.00	0.98	1.78
3	Fuel mode	–	1.00	1.00	0.0021
4	Exhaust gas temperature	℃	1.01	0.97	6.90
5		℃	1.00	0.98	6.19
6		℃	0.99	0.97	6.85
7		℃	0.99	0.98	6.04
8		℃	0.99	0.98	5.19
9		℃	0.99	0.97	6.62
10	Turbocharger	rpm	1.01	0.99	162
11		℃	1.00	0.99	3.30
12		℃	1.00	1.00	0.660
13	Cylinder cooling water temperature	℃	1.00	0.99	0.192
14		℃	1.00	1.00	0.114
15		℃	1.00	1.00	0.107
16		℃	1.00	1.00	0.0942
17		℃	1.00	1.00	0.0843
18	Piston cooling water temperature	℃	1.00	0.97	0.205
19		℃	1.00	0.99	0.125
20		℃	1.00	1.00	0.0914
21		℃	1.00	0.99	0.089
22		℃	1.00	0.99	0.100
23		℃	1.00	0.99	0.108

데이터와 추세선의 거리의 표준편차는 Table 3에 함께 나타내었다. 이 표준편차 값을 Table 1의 NCR 값과 비교하면 그룹 1, 3, 4에서는 NCR 값의 1-2% 정도의 크기를 보였다. 그룹 5, 6에서는 1% 미만의 표준편차 값을 보여 데이터 추정의 신뢰도가 아주 뛰어남을 알 수 있었다.

5. 결 론

선박 기관의 운용 시 센서 신뢰도를 평가하고 이상치를 식별하기 위한 차수 축소 모델링 기법으로 주성분 분석을 도입하였고, 비선형적인 상태변수에 대한 적용 가능성을 판단하였다. 상호상관도 분석을 통해 비선형적인 변수로 연료 모드, 과급기 배출가스 온도, 실린더 냉각수 온도가 식별되었다.

주성분 분석의 유효성을 증가시키기 위한 데이터의 전처리 기법을 제안하였다. 전처리를 통해 데이터의 평균값을 기관의 NCR 조건에 위치시켰다. 그리고 선형이 아닌 다항식 관계를 갖는 기관 회전속도와 동력이 선형 관계를 갖도록 하여, 변수 간의 선형적 관계가 첫 번째 주요 모드에만 한정되도록 하였다.

주성분 분석을 통해서 각 채널 데이터의 공분산 행렬의 고유벡터를 얻고, 고유값 비교를 통해 주요 모드를 식별하였다. 본 연구에서는 5개의 주요 모드를 이용해 기관 센서 데이터를 재구성하였고, 식별된 비선형적 변수간의 관계를 두 번째에서 네 번째 주요 모드를 이용해 설명할 수 있었다. 특히 바이너리 형식이면서 비선형성을 갖는 연료 종류를 주성분 분석을 통해 식별해낼 수 있는 것을 확인하였다.

본 연구에서는 기관의 주요 센서 채널에 대한 주성분 분석을 수행하였다. 이를 운항 데이터 전체로 확장한다면 선박의 기관 운용 중 특이사항을 선박의 항로, 화물 하역 조건 등과 연계하여 해석할 수 있을 것으로 기대된다.

부록: 주성분 분석의 상세 내용

특정 순간 t에서 n개 채널을 통해 계측한 데이터 $\boldsymbol{x}(t)$는 $n\times1$의 열벡터(column vector) 형식으로 주어진다. 그리고 이는 시간과 무관한 $n\times1$ 열벡터 $\boldsymbol{v}_i$와 시간에 따라 변하는 계수 $a_i(t)$의 선형결합으로 나타낼 수 있다.

$$\boldsymbol{x}(t)=\begin{bmatrix}x_1(t)\\ \vdots \\ x_n(t)\end{bmatrix}=\sum_{i=1}^{n}a_i(t)\boldsymbol{v}_i \tag{8}$$

식 (8)과 같이 총 n개의 $a_i(t)$로 순간의 데이터 $\boldsymbol{x}(t)$를 표현한다면 앞의 원 데이터와 같은 n의 자유도를 가지므로, 차원 축소 관점에서는 의미가 없게 된다. 따라서 n보다 작지만 데이터를 충분히 잘 나타낼 수 있는 정도의 k를 선택하여 $\boldsymbol{x}(t)$를 n개의 $x_i(t)$ 대신 k개의 $a_i(t)$로 나타내 데이터의 자유도를 낮춘다. PCA에서는 n개의 정규직교벡터를 찾아 그 중 의미가 있는 k개를 $\boldsymbol{v}_i$로 이용하여 생성(span)된 공간에서 $a_i(t)$를 구해 $\boldsymbol{x}(t)$를 근사한다.

고유값 분해

PCA를 통해 대표적인 기저 $\boldsymbol{v}_i$를 구하기 위해서는 고유값 분해 과정(eigen value decomposition)의 이해가 필요하다. $n\times n$ 행렬 $\boldsymbol{A}$가 주어졌을 때 식 (9)를 만족하는 $\boldsymbol{A}$의 고유값(eigen value) λ_i과 이에 대응하는 $n\times1$ 고유벡터(eigen vector) $\boldsymbol{p}_i$를 구한다.

$$\boldsymbol{A}\boldsymbol{p}_i=\lambda_i\boldsymbol{p}_i \tag{9}$$

$\boldsymbol{p}_1$에서 $\boldsymbol{p}_n$까지 n의 고유벡터들을 열벡터(column vector)로 갖는 $n\times n$ 행렬 $\boldsymbol{P}$와 고유값 λ_i만을 대각성분의 원소로 갖는 대각행렬 $\boldsymbol{\Lambda}$를 이용해 식 (9)를 다음과 같이 표현할 수 있다.

$$\begin{aligned}\boldsymbol{A}\boldsymbol{P}&=\boldsymbol{A}[\boldsymbol{p}_1\cdots\boldsymbol{p}_n]=[\lambda_1\boldsymbol{p}_1\cdots\lambda_n\boldsymbol{p}_n]\\&=[\boldsymbol{p}_1\cdots\boldsymbol{p}_n]\begin{bmatrix}\lambda_1 & & 0\\ & \ddots & \\ 0 & & \lambda_n\end{bmatrix}=\boldsymbol{P}\boldsymbol{\Lambda}\end{aligned} \tag{10}$$

이제 행렬 $\boldsymbol{A}$는 다음과 같이 표현할 수 있는데, 이를 대각화라고 한다.

$$\boldsymbol{A}=\boldsymbol{P}\boldsymbol{\Lambda}\boldsymbol{P}^{-1} \tag{11}$$

행렬 $\boldsymbol{A}$가 $\boldsymbol{A}^T=\boldsymbol{A}$를 만족하는 대칭행렬이라면, 행렬 $\boldsymbol{P}$는 전치행렬($\boldsymbol{P}^T$)이 역행렬($\boldsymbol{P}^{-1}$)과 같은 직교행렬(orthogonal matrix)이 된다. $\boldsymbol{A}$는 식 (11)로부터 다음과 같이 표현할 수 있다.

$$\begin{aligned}\boldsymbol{A}&=\boldsymbol{P}\boldsymbol{\Lambda}\boldsymbol{P}^T=[\boldsymbol{p}_1\cdots\boldsymbol{p}_n]\begin{bmatrix}\lambda_1 & & 0\\ & \ddots & \\ 0 & & \lambda_n\end{bmatrix}\begin{bmatrix}\boldsymbol{p}_1\\ \vdots \\ \boldsymbol{p}_n\end{bmatrix}\\&=\sum_{i=1}^{n}\lambda_i\boldsymbol{p}_i\boldsymbol{p}_i^T\end{aligned} \tag{12}$$

만약 λ_i가 크기순으로 정렬되어 있다면, n보다 작은 k번째까지의 고유값과 그에 대응되는 고유벡터들으로도 $\boldsymbol{A}$에 근사한 행렬을 복원할 수 있게 된다. 이때 식 (13)을 이용해 k번째의 고유값까지로 근사해 복원(reconstruction)한 행렬과 원래의 $\boldsymbol{A}$와의 유사도(similarity) s_k를 정량화할 수 있다.

$$s_k=\sum_{i=1}^{k}\lambda_i/\sum_{i=1}^{n}\lambda_i \tag{13}$$

여기까지 해서, 대칭행렬이 주어졌을 때 이에 대한 고유벡터와 고유값을 구하면 고유벡터를 직교기저로 이용한 데이터의 표현이 가능함을 알 수 있다.

이제 m번의 계측을 통해 $n\times m$행렬로 주어진 데이터 $\boldsymbol{X}$로부터 적절한 대칭행렬을 구해야 하는데, PCA에서는 공분산 행렬(covariance matrix)을 이용한다. $t=0$에서 T까지 얻어진 전체 데이터 $\boldsymbol{X}$의 각 채널의 계측결과들로부터 구한 공분산 행렬 $\boldsymbol{A}_C$는 식 (14)와 같이 나타낼 수 있다.

$$\boldsymbol{A}_C=\begin{bmatrix}\mathrm{Cov}(x_1,x_1)\cdots\mathrm{Cov}(x_1,x_n)\\ \vdots \quad\ddots\quad \vdots \\ \mathrm{Cov}(x_n,x_1)\cdots\mathrm{Cov}(x_n,x_n)\end{bmatrix} \tag{14}$$

여기서 $cov(x_i,x_j)$는 $\boldsymbol{x}(t)$를 구성하는 변수 $x_i(t)$, $x_j(t)$의 기댓값 $E(x_i)$, $E(x_j)$를 이용해 식 (15)와 같이 계산된다. 기댓값은 데이터의 평균으로 볼 수 있다.

$$\begin{aligned}\mathrm{Cov}(x_i,x_j)&=\mathrm{E}(x_ix_j)-\mathrm{E}(x_i)\mathrm{E}(x_j)\\&=\mathrm{Cov}(x_j,x_i)\end{aligned} \tag{15}$$

공분산 자체는 두 변수간의 상호상관도를 의미하여, 그 값이 커지면 강한 선형적 상관관계를 의미한다. 값이 0에 가까우면 두 변수의 상관도는 무의미해져 서로 무관한 변화를 보인다고 본다. 음수값을 가지면서 절대값이 커지면 선형적으로 음의 상관관계를 보인다는 뜻이다. $i=j$ 인 경우에는 한 데이터 변수의 시간에 따른 분산(variance)값을 얻게되므로, 행렬 $\boldsymbol{A}_C$의 대각성분은 각 변수의 분산으로 구성된다. 식 (15)를 통해 행렬 $\boldsymbol{A}_C$는 대칭행렬임을 알 수 있으므로, 공분산 행렬은 직교행렬이 되고, 따라서 식 (12)의 적용이 가능해진다.

공분산 행렬 $\boldsymbol{A}_C$의 고유벡터 $\boldsymbol{p}_i$는 길이가 1인 $n\times1$ 열행렬로, 서로 간에 직교하는 성질을 가진다. 이를 식 (8)의 $\boldsymbol{v}_i$로 이용하여 $\boldsymbol{x}(t)$를 표현할 수 있다. 그리고 고유값 λ_i는 식 (8)의 $\boldsymbol{x}_i$의 표현에서 사용된 계수 a_i의 시계열상의 분산값이 된다. 식 (12)에서 λ_i의 크기순으로 배열을 하는 과정에서 λ_i가 큰 경우는 분산이 큰 고유벡터이므로, $\boldsymbol{x}(t)$를 복원하는 과정에서 데이터의 변화를 잘 나타내는 벡터임을 알 수 있다.

식 (8)의 $a_i(t)$를 구하기 위해서는 특정 순간(t)에 계측된 데

이터로 구성된 행벡터(row vector) $\boldsymbol{x}^T(t)$와 고유벡터 $\boldsymbol{p}_i$를 열벡터로 구성된 $n\times n$ 행렬 $\boldsymbol{P}$의 행렬곱인 $\boldsymbol{x}^T(t)\boldsymbol{P}$를 구해야 한다. 행렬 $\boldsymbol{P}$가 직교행렬이므로, 이를 구성하는 모든 열벡터($\boldsymbol{p}_i$)의 크기는 1이며 서로 직교한다. 이 고유벡터와 특정 순간의 데이터 $\boldsymbol{x}^T(t)$의 행렬곱은 두 벡터의 내적과 같고, 그 내적은 $\boldsymbol{x}(t)$를 고유벡터 $\boldsymbol{p}_i$의 방향으로 사영(projection)한 크기가 되므로 이는 식 (8)의 $a_i(t)$와 같다. 다만 n개의 $\boldsymbol{p}_i$를 모두 사용하는 대신 의미 있는 k개만을 추리기로 하였으므로, k개의 $\boldsymbol{p}_i$로 구성된 $n\times k$ 행렬 $\boldsymbol{P}'$와 $\boldsymbol{x}^T(t)$의 곱을 대신 사용하여 식 (16)과 같이 근사한다. $\boldsymbol{x}^T(t)\boldsymbol{P}'$는 k개의 $a_i(t)$들을 원소로 갖는 $1\times k$ 행벡터이므로, 데이터가 축소되었음을 알 수 있다.

$$\boldsymbol{x}(t)\approx\sum_{i=1}^{k}a_i(t)\boldsymbol{p}_i=\sum_{i=1}^{k}(\boldsymbol{x}(t)\cdot\boldsymbol{p}_i)\boldsymbol{p}_i \tag{16}$$

특이값 분해

PCA과정에서 자주 사용되는 특이값 분해(singular value decomposition, SVD)는 공분산을 따로 계산하는 대신 행렬곱을 통해 공분산에 대한 정방행렬을 바로 얻는다. Fig. 2에서 설명한 바와 같이, n개 채널 데이터 $\boldsymbol{x}(t)$에서 m번의 계측을 통해 얻은 $\boldsymbol{x}$의 시간평균을 뺀 요동성분을 열벡터로 하여 나타낸 $n\times m$ 행렬 $\boldsymbol{X}$로부터 다음 두 가지 행렬 $\boldsymbol{B}$, $\boldsymbol{\Gamma}$를 구한다. 두 행렬은 각각 $m\times m$, $n\times n$ 크기의 대칭행렬이 된다.

$$\boldsymbol{B}=\frac{1}{n}\boldsymbol{X}^T\boldsymbol{X} \tag{17}$$

$$\boldsymbol{\Gamma}=\frac{1}{m}\boldsymbol{X}\boldsymbol{X}^T \tag{18}$$

행렬 $\boldsymbol{X}$의 요소 x_{ij}에서 i와 j는 각각 데이터 채널의 순서와 시간상의 순서로, 최대값은 n과 m이다. $\boldsymbol{B}$와 $\boldsymbol{\Gamma}$의 요소 β_{ij}와 γ_{ij}는 각각 식 (19)와 (20)처럼 나타낼 수 있다. 이는 식 (15)의 $E[x_i,x_j]$에 대응하여, $\boldsymbol{B}$, $\boldsymbol{\Gamma}$가 공분산 행렬임을 알 수 있다.

$$\beta_{ij}=\frac{\sum_{k=1}^{n}x_{ki}x_{kj}}{n} \tag{19}$$

$$\gamma_{ij}=\frac{\sum_{k=1}^{m}x_{ik}x_{jk}}{m} \tag{20}$$

이렇게 구한 $\boldsymbol{B}$는 특정 시각의 데이터 세트가 다른 시각의 데이터 세트와 얼마나 닮았는지를 나타낸다. $\boldsymbol{\Gamma}$는 두 개의 채널에서 얻어진 시계열 데이터가 서로 얼마나 비슷하게 변화하는지를 보인다. 앞서 본문에서 보인 바와 같이 본 연구에서는 후자의 관점에서 데이터를 다뤘다.

$\boldsymbol{B}$와 $\boldsymbol{\Gamma}$는 대칭행렬이므로 각각 고유값 분해를 적용하면 식 (21)과 (22)로 표현할 수 있다.

$$\boldsymbol{B}=\frac{1}{n}\boldsymbol{X}^T\boldsymbol{X}=\boldsymbol{V}\boldsymbol{\Lambda}_B\boldsymbol{V}^T \tag{21}$$

$$\boldsymbol{\Gamma}=\frac{1}{m}\boldsymbol{X}\boldsymbol{X}^T=\boldsymbol{U}\boldsymbol{\Lambda}_\Gamma\boldsymbol{U}^T \tag{22}$$

여기서 $\boldsymbol{U}$와 $\boldsymbol{V}$는 각각 $\boldsymbol{\Gamma}$와 $\boldsymbol{B}$ 행렬의 고유벡터인 $\boldsymbol{u}_i$, $\boldsymbol{v}_i$를 열벡터로 갖는 $n\times n$, $m\times m$ 행렬이다. 이제 식 (23)과 같이 $\boldsymbol{X}$를 $\boldsymbol{U}$, $\boldsymbol{V}$, $\boldsymbol{\Sigma}$로 나타낼 수 있다고 가정한다. 여기서 $\boldsymbol{\Sigma}$는 정방형은 아니나 대각성분만 있는 $n\times m$ 크기의 직사각 대각행렬(rectangular diagonal matrix)이다.

$$\boldsymbol{X}=\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T \tag{23}$$

$\boldsymbol{\Sigma}$는 $\boldsymbol{\Lambda}_B$, $\boldsymbol{\Lambda}_\Gamma$를 이용해 나타낼 수 있으며 그 관계는 식 (24), (25)와 같다.

$$\boldsymbol{\Sigma}^T\boldsymbol{\Sigma}=n\boldsymbol{\Lambda}_B \tag{24}$$

$$\boldsymbol{\Sigma}\boldsymbol{\Sigma}^T=m\boldsymbol{\Lambda}_\Gamma \tag{25}$$

$\boldsymbol{U}$와 $\boldsymbol{V}$는 직교행렬로, 전치행렬과 역행렬이 같다는 성질을 이용하면 식 (26)과 (27)을 유도하여 위의 가정이 타당함을 확인할 수 있다.

$$\begin{aligned}\boldsymbol{B}&=\frac{1}{n}\boldsymbol{X}^T\boldsymbol{X}=\frac{1}{n}(\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T)^T(\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T)\\&=\frac{1}{n}\boldsymbol{V}\boldsymbol{\Sigma}^T\boldsymbol{U}^T\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T=\frac{1}{n}\boldsymbol{V}\boldsymbol{\Sigma}^T\boldsymbol{\Sigma}\boldsymbol{V}^T\\&=\boldsymbol{V}\boldsymbol{\Lambda}_B\boldsymbol{V}^Tn\end{aligned} \tag{26}$$

$$\begin{aligned}\boldsymbol{Y}&=\frac{1}{m}\boldsymbol{X}\boldsymbol{X}^T=\frac{1}{m}(\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T)(\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T)^T\\&=\frac{1}{m}\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T\boldsymbol{V}\boldsymbol{\Sigma}^T\boldsymbol{U}^T=\frac{1}{m}\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^T\boldsymbol{U}^T\\&=\boldsymbol{U}\boldsymbol{\Lambda}_\Gamma\boldsymbol{U}^T\end{aligned} \tag{27}$$

$\boldsymbol{\Sigma}$의 대각성분을 특이값이라 하고, σ_{ii}로 나타내면 식 (24)와 (25) 모두 대각성분이 σ_{ii}^2로 동일한 대각행렬임을 알 수 있다. $\boldsymbol{X}\boldsymbol{X}^T$, $\boldsymbol{X}^T\boldsymbol{X}$의 고유값 λ_i은 서로 같다는 성질과 연관해, 식 (28)과 같이 $\boldsymbol{\Sigma}$의 요소를 구할 수 있다.

$$\sigma_{ii}=\sqrt{\lambda_i}\ (i\le\min(m,n)) \tag{28}$$

앞에서 공분산 행렬을 통해 보인 고유값 분해의 경우와 같이, 작은 λ_i에 대응되는 데이터를 배제하여 데이터의 차원 축소를 수행할 수 있다.

후 기

본 연구는 2023년 삼성중공업의 재원으로 충남대학교에서 수

행된 '자율운항시스템 핵심 기술 개발'과제 및 22023년도 정부(산업통상자원부)의 재원으로 한국산업기술진흥원의 지원을 받아 수행된 연구임(P0017006, 2023년 산업혁신인재성장지원사업).

References

Alexandris, N., Gupta, S. and Koutsias, N., 2017. Remote sensing of burned areas via PCA, Part 1; centering, scaling and EVD vs SVD. *Open Geospatial Data, Software and Standards,* 2, Article No. 17.

Bishop, C.M., 2006. *Pattern Recognition and Machine Learning.* Springer-Verlag.

Kim, S.-H., Lee, Y.-S. and Han, Y.-J., 2003. A study on the design of fault diagnostic system based on PCA. *Journal of Korean Institute of Intelligent Systems,* 13(5), pp.600-605.

Kim, Y.-I., Oh, H.-K. and Yu, Y.-H., 2006. The development of diesel engine room fault diagnosis system using a correlation analysis method. *Journal of the Korean Society of Marine Engineering,* 30(2), pp.253-259.

Kim, K.-S., Roh, M.-I., Jeong, Y.-J. and Na, G.-J., 2020a. Development of an application for predicting and visualizing the ocean weather. *Korean Journal of Computational Design and Engineering,* 25(4), pp.434-444.

Kim, D., Lee, S. and Lee, J., 2020b. Abnormaly detection of vessel main engine big data using Gaussian mixture model. *Journal of the Korean Data Analysis Society,* 22(4), pp.1473-1489.

Kim, D., Han, Y., Kim, H., Kang, S., Kim, K. and Bae, H., 2022. Outlier detection and labeling of ship main engine using LSTM-AutoEncoder. *The Korea Journal of Big Data,* 7(1), pp.125-137.

Korean Register, 2021. Guidebook for CII Regulation. Korean Register R&D Division.

Landgraf, A.J. and Lee, Y., 2020. Dimensionality reduction for binary data through the projection of natural parameters. *Journal of Multivariate Analysis,* 180, Article No. 104668.

Lee, S.B., Han, B.W., Park, D.W., Ahn, Y.W., Go, S.C. and Seo, H.W., 2012. Proper orthogonal decomposition of pressure fluctuations in moonpool. *Journal of the Society of Naval Architects of Korea,* 49(6), pp.484-490.

Lee, J.H., 2017. Proper orthogonal decomposition and its application: parametric reduced order models. *Computational Structural Engineering,* 30(1), pp.29-35.

Leeuw, J., 2006. Principal component analysis of binary data by iterated singular value decomposition. *Computational Statistics and Data analysis,* 50(1), pp.21-39.

Loyola-Gonzalez, O., 2019. Black-box vs white-box: understanding their advantages and weaknesses from a practical point of view. *IEEE Access,* 7, 154096-154113.

Moeini, B., Avval, T.G., Gallagher, N. and Linford, M.R., 2023. Surface analysis insight note. Principal component analysis (PCA) of an X-ray photoelectron spectroscopy image. The importance of preprocessing. *Surface and Interface Analysis,* 55(11), pp.798-807.

Paik, B.-G., Kim, K.-Y., Kim, K.-S., Lee, J.-Y. and Lee, S.-J., 2010. Analysis of the unstable propeller wake using POD method. *Journal of the Society of Naval Architects of Korea,* 47(1), pp.20-29.

Park, J.-H., Jang, M.-K., Lee, G.-H., Oh, E.-K. and Hur, S.-W., 2016. Forecasting algorithm for vessel engine failure. *Journal of Korean Institute of Information Technology,* 14(11), pp.109-117.

Park, J.-H., Oh, E.-K., Jang, M.-K., Seo, Y.-W. and Hur, S.-W., 2017. Improved forecasting algorithm for vessel engine failure. *Journal of Korean Institute of Information Technology,* 15(11), pp.175-185.

Park, J.-C., Kwon, H.-C., Kim, C.-H. and Jang, H.-S., 2023. The study of failure mode data development and feature parameter's reliability verification using LSTM algorithm for 2-stroke low speed engine for ship's propulsion. *Journal of the Society of Naval Architects of Korea,* 60(2), pp.95-109.

Savage, N., 2022. Breaking into the black box of artificial intelligence. *Nature*, DOI:10.1038/d41586-022-00858-1.

Shin, S.-Y., Jung, K.-H., Kang, Y.-D., Suh, S.-B., Kim, J. and An, N.-H., 2017. A study on the effect of large coherent structures to the skin friction by POD analysis. *Journal of the Society of Naval Architects of Korea,* 54(5), pp.406-414.


이 승 범


서 정 화


김 동 환


한 상 민


김 관 우


정 성 욱


유 병 우