



UNIVERSIDADE ESTADUAL DE CAMPINAS
Faculdade de Engenharia Mecânica

Paulo Yoshio Kuga

**Identificação de anomalias em bombas centrífugas
submersas usando dados reais de operação**

Campinas
2026

Paulo Yoshio Kuga

**Identificação de anomalias em bombas centrífugas submersas usando
dados reais de operação**

Monografia de qualificação apresentada à Faculdade de Engenharia Mecânica da Universidade Estadual de Campinas como parte dos requisitos para a obtenção do título de Mestre em Engenharia Mecânica, na área de Engenharia Mecânica.

Orientador: Prof. Dr. Alberto Luiz Serpa

Este trabalho corresponde à versão final da Monografia de qualificação defendida por Paulo Yoshio Kuga e orientada pelo Prof. Dr. Alberto Luiz Serpa.

Campinas
2026

FOLHA DE APROVAÇÃO

Abaixo se apresentam os membros da comissão julgadora da sessão pública de defesa de dissertação para o Título de Mestre em Engenharia Mecânica na área de concentração Engenharia Mecânica, a que se submeteu o aluno Paulo Yoshio Kuga, em 28 de fevereiro de 2026 na Faculdade de Engenharia Mecânica – FEM/UNICAMP, em Campinas/SP.

Prof. Dr. Alberto Luiz Serpa
Presidente da Comissão Julgadora

Profa. Dra. Segunda Avaliadora
Instituição da segunda avaliadora

Dr. Terceiro Avaliador
Instituição do terceiro avaliador

Ata da defesa, assinada pelos membros da Comissão Examinadora, encontra-se no SIGA/Sistema de Fluxo de Dissertação/Tese e na Secretaria de Pós-graduação da Faculdade de Engenharia Mecânica.

Esta dissertação é dedicada a minha
mãe, Me. Elaine Aparecida Justino
Kuga (*in memoriam*).

*Não tenha uma arma favorita.
Usar demais é tão ruim quanto não saber usar.
Não imite — use o que domina.*
(Miyamoto Musashi - Go Rin no Sho)

Agradecimentos

Em desenvolvimento.

Resumo

A principal abordagem utilizada para identificação de falhas em Bombas Centrífugas Submersas (BCS) é o aprendizado supervisionado. Com dados rotulados entre falhos, ou não, determina-se um modelo que prediz a existência da falha. Portanto, existem oportunidades com relação à utilização de métodos de aprendizado não-supervisionado, eliminando a necessidade de realizar uma rotulação manual. Com isso, o objetivo desta pesquisa é investigar um método que consiga evidenciar anomalias, possivelmente relacionadas a falha, baseando-se em dados reais de operação de um sistema de bombeamento centrífugo submerso para extração de petróleo.

Estes dados possuem problemas como ruído, dados faltantes e incoerências de intervalo amostral com relação a rotulação das falhas. Para lidar com estes problemas, a principal técnica utilizada é um *Z-score* modificado, que se utiliza de uma filtragem por uma média móvel exponencial, a mediana calculada ponto a ponto, buscando robustez a *outliers*, e o desvio padrão, também ponto a ponto.

A partir da realização do *Z-score* é proposta a utilização de transformações lineares conhecidas, para evidenciar as relações numéricas entre as variáveis medidas no banco de dados. As duas transformações escolhidas foram a Análise de Componentes Principais (PCA), que busca reduzir a dimensionalidade dos dados, isto é, o número de variáveis envolvidas na identificação e a Análise de Componentes Independentes (ICA), que busca uma combinação de variáveis que maximize a independência estatística entre elas.

Em seguida, o Modelo Oculto de Markov Gaussian (GHMM) é proposto para a identificação de anomalias. Ele é um modelo probabilístico de transição de estados ocultos a partir observação dos dados medidos da bomba. O objetivo é avaliar se estes estados conseguem indicar a normalidade ou anormalidade dos sinais. Para realizar a análise, é aplicada uma normalização euclidiana dos dados, buscando representar a distância das entradas com relação à média. Para escolher o número de estados desta distribuição, é utilizado um Modelo de Misturas Gaussianas (GMM). Com esta metodologia, os padrões anômalos são ressaltados com maior nitidez, notando-se uma incidência alta de estados anômalos fornecidos pelo modelo antes e durante a marcação da falha pelos operadores.

Nesta dissertação, conclui-se que os dados das bombas possuem relações entre si e que o GHMM, combinado com o GMM para encontrar condições iniciais, é uma ferramenta útil para a detecção de anomalias em séries temporais de bombas centrífugas submersas, possuindo uma flexibilidade para lidar com dados em diferentes distribuições estatísticas, sendo, portanto, um método versátil.

Palavras-Chave: Bombas Centrífugas Submersas, Aprendizado Não-Supervisionado, Identificação de Anomalias, Análise de Componentes Princiais, Análise de Componentes Independentes, Modelos Ocultos de Markov Gaussianos.

Abstract

The main approach used to identify faults in Electrical Submersible Pumps is supervised learning. With data labeled as faulty or not, a model is trained to predict the existence of the fault. Therefore, there are opportunities for using unsupervised learning methods, eliminating the need for manual labeling. With that, the objective of this research is to investigate a method that can identify anomalies, possibly related to faults, based on real operating data from a ESP system.

These data have problems such as noise, missing data and sampling interval inconsistencies regarding fault labeling. To address these issues, the primary technique used is a modified Z-score, which uses filtering by an exponential moving average, the expanding median, seeking robustness against outliers, and the expanding standard deviation.

Based on the Z-score calculation, the proposed method uses known linear transformations to highlight the numerical relationships between the variables measured in the database. The two transformations chosen were Principal Component Analysis (PCA), which seeks to reduce data dimensionality—that is, the number of variables involved in identification—and Independent Component Analysis (ICA), which seeks a combination of variables that maximizes statistical independence.

Furthermore, to allow anomaly identification, is proposed the usage of Gaussian Hidden Markov Models (GHMM), a probabilistic model of hidden state transition based on observation of the pump's measured data. It is expected that these states may be an indicator of normality or abnormality of the data provided. In this sense, the model is applied to a euclidean normalization of the data, seeking to represent the distance of the inputs in relation to the mean. For this distribution, a Gaussian Mixture Model (GMM) is used, intending to provide initialization parameters.

Next, a Gaussian Hidden Markov Model (GHMM) is proposed for anomaly identification. It is a probabilistic model of hidden state transitions based on observations of pump measurement data. The objective is to assess whether these states can indicate normality or abnormality in the signals. To perform the analysis, Euclidean normalization is applied to the data, seeking to represent the distance of the inputs from the mean. To select the number of states in this distribution, a Gaussian Mixture Model (GMM) is used. With this methodology, anomalous patterns are highlighted more clearly, noting a high incidence of anomalous states provided by the model before and during the operators' fault marking.

In this dissertation, it is concluded that pump data are related to each other and that GHMM, combined with GMM to find initial conditions, is a useful tool for detecting anomalies in time series of submersible centrifugal pumps, having the flexibility to deal with data in different statistical distributions, therefore being a versatile method.

Keywords: Electrical Submersible Pumps, Unsupervised Learning, Anomaly Identification, Principal Component Analysis, Independent Component Analysis, Guassian Hidden Markov Models.

Lista de Figuras

2.1	Diagrama de uma instalação da BCS - adaptado de Takacs (2018).	25
2.2	Dados originais da corrida B-18 2.	26
2.3	Dados originais da corrida B-18 2 - escala logarítmica.	26
2.4	Histograma dos dados medidos (os valores limites das abscissas contemplam todos os pontos).	31
2.5	Histograma das variáveis padronizadas (de -6σ até 6σ).	32
2.6	Gráfico Q-Q das variáveis das bombas - a linha vermelha indica a distribuição normal associada.	34
3.1	Diagrama de Bode do filtro de médias móveis.	41
3.2	Diagrama de Bode do filtro EWM.	42
3.3	Sinais fictícios 1 e 2.	46
3.4	$Z\text{-score}$ modificado aplicado aos sinais fictícios.	46
3.5	Primeiro exemplo com PCA e ICA	54
3.6	Segundo exemplo com PCA e ICA	56
3.7	Ajuste do modelo de misturas gaussianas para a norma euclidiana no sinal fictício.	67
3.8	Identificação de estados com o sinal da norma euclidiana.	67
3.9	Diagrama de fluxo de dados para o pré-processamento.	71
3.10	Diagrama de fluxo de dados para a aplicação do $Z\text{-score}$ modificado.	72
3.11	Diagrama de fluxo de dados para a realização das transformações.	73
3.12	Diagrama de fluxo de dados para o treinamento do modelo GHMM.	74
4.1	O sinal Original da pressão na corrida B-18 2 (verde) e o sinal Filtrado com o filtro passa-baixa (azul).	76
4.2	Média (azul) e Mediana (laranja) comparadas para a pressão da corrida B-18 2. .	76
4.3	Comparação dos métodos de desvio: Desvio-Padrão contra MAD.	77
4.4	Sinal filtrado (azul), mediana em expansão (azul claro), desvio padrão em expansão (amarelo) e sinal normalizado (vermelho).	77
4.5	$Z\text{-score}$ modificado para a corrida 2 da bomba B-18.	78
4.6	Matriz de covariância do resultado da PCA.	78
4.7	A matriz de correlação do resultado da PCA.	79
4.8	Comparação entre os sinais originais padronizados e os obtidos pela PCA da matriz de correlação.	80
4.9	A matriz de separação da ICA - referência 19971215.	81
4.10	Comparação entre os sinais originais padronizados e os obtidos pela ICA. . . .	82
4.11	Gráfico Quantil-Quantil para as componentes do $Z\text{-score}$	83
4.12	Gráfico Quantil-Quantil para as componentes da PCA.	83
4.13	Gráfico Quantil-Quantil para as componentes da ICA.	84

4.14	Aplicação da Norma euclidiana para as transformações.	84
4.15	Gráfico Quantil-Quantil para a distribuição Chi do Z-score (ZSC), da PCA e da ICA.	85
4.16	Histogramas para a norma euclidiana e suas misturas gaussianas correspondentes.	86
4.17	Dados da corrida B-18 2 inferidos de janeiro de 2012 até abril de 2013. O gráfico é feito com as variáveis originais padronizadas.	87
4.18	Dados da corrida A-12 2 inferidos de 2014 até 2019. O gráfico é feito com as variáveis originais.	88
4.19	Dados da corrida A-29 2 inferidos de janeiro de 2016 até abril de 2018. O gráfico é feito com as variáveis originais.	88
4.20	Gráfico de barras com os resultados da persistência de estados para o valor referência 19971215.	89
4.21	Gráfico de barras com os resultados da persistência para as amostras de teste.	90
4.22	Gráfico de barras com os resultados da persistência de estados para múltiplas referências aleatórias.	90
4.23	Gráfico de barras com os resultados da persistência de estados das amostras de teste para múltiplas referências aleatórias.	91

Lista de Tabelas

2.1	Descrição das colunas dos dados	24
2.2	Propriedades dos dados originais	28
2.3	Tabela de vazios	28
2.4	Tabela de propriedades das variáveis pré-processamento	29
4.1	Correlação Explicada	80
4.2	Comparação entre o AIC e BIC	86

List of Symbols

Scalars

l	Número de saídas consideradas em uma equação de recorrência
a_l	Coeficiente da saída l na equação de recorrência
p	Número de entradas consideradas em uma equação de recorrência
b_p	Coeficiente da entrada l na equação de recorrência
k	Índice representativo de uma entrada escolhida
y_k	Saída genérica de uma equação de recorrência
x_k	Entrada genérica de uma equação de recorrência
z	Operador da Transformada \mathcal{Z}
α	Parâmetro de ajuste da EWM
S	Saída de um filtro EWM
n	Número total de entradas
m	Número total de colunas
λ	Autovalor
p	Número de colunas correspondente a 95% da variância
g	Graus de liberdade de uma distribuição

Operators

\odot	Produto de Hadamard
\leftarrow	Operador Atualização
:	Operador de fatiamento

Statistics and moments

μ	Média
M	Mediana
σ	Desvio-padrão
Z	Valor do Z -score

Funções, Densidades de Probabilidade e Operadores

- $H(z)$ Função de transferência
 $\mathcal{F}(x, k)$ Distribuição Qui com k graus de liberdade
 Γ Função Gamma
 $\|\cdot\|_2$ Operador norma euclidiana
 $\mathcal{D}_{(\cdot)}(x)$ Mistura Gaussiana sobre um conjunto de dados (\cdot)

Intervalos

- $0 : k$ Indicação de intervalo equivalente a $[0, k]$
 $\mathbb{C}_{0:k}$ Série temporal de 0 até k
 \mathbb{O} Rol

Matrizes

- Z Matriz banco de dados padronizado
 C Matriz de correlação
 Λ Matriz diagonal de autovalores
 V Matriz diagonal de autovetores

Probabilísticos Escalares

- X Variável aleatória genérica
 A Variável gaussiana
 $P(X)$ Probabilidade de uma variável aleatória
 $p(X)$ Densidade de probabilidade de uma variável aleatória
 $A | B$ Ocorrência do evento A dado o evento B .
 w_i Peso na ponderação da mistura
 γ_{ki} Probabilidade posterior do estado i , na amostra k
 Π_k Probabilidade dos estados no instante k
 \sim “segue à distribuição”

Probabilísticos Matriciais

- A Matriz de transição de estados
 B Vetor de probabilidade de observações
 α_k Vetor de probabilidade do acontecimento da sequência
 β_k Vetor de probabilidade da sequência mais provável
 γ_k Vetor de probabilidades posteriores
 δ_k Probabilidade de observação

Lista de Abreviaturas e Siglas

AIC	<i>Akaike Information Criterion</i>
AVV	Acionador de Velocidade Variável
BCS	Bombas Centrífugas Submersas
BIC	<i>Bayesian Information Criterion</i>
CSV	<i>Comma Separated Values</i>
EM	<i>Expectation-Maximization Algorithm</i>
ESP	<i>Electrical Submersible Pumps</i>
EWM	<i>Exponential Weighted Moving Average</i>
GHMM	<i>Gaussian Hidden Markov Model</i>
GMM	<i>Gaussian Mixture Model</i>
ICA	<i>Independent Component Analysis</i>
LISSA	<i>Layer of Integration with Scikit-learn and Signal Analysis</i>
MAD	<i>Mean Absolute Deviation</i>
PCA	<i>Principal Component Analysis</i>
SVM	<i>Support Vector Machine</i>

Sumário

1	Introdução	17
1.1	Motivação	19
1.2	Objetivos	22
1.3	Estrutura da Dissertação	22
2	Desafios na análise de dados operacionais das BCS	23
2.1	Coleção de dados no processo de elevação artificial	23
2.2	Descrição dos Dados	26
2.3	Características dos dados	30
3	Metodologia Proposta	35
3.1	Variável Aleatória	35
3.2	Distribuição Gaussiana	35
3.3	Momentos	36
3.4	Filtragem de Dados - Janelamento e Ponderação Exponencial	37
3.5	Cálculos por Expansão	42
3.6	<i>Z-score</i> modificado	44
3.7	Exemplo de uso do <i>Z-score</i> modificado	45
3.8	Análise de Componentes Principais	45
3.9	Análise de Componentes Independentes	50
3.10	Exemplos com a PCA e ICA	53
3.11	Norma euclidiana e distribuição Chi	56
3.12	Modelo de Misturas Gaussianas	57
3.13	Modelo Oculto de Markov Gaussiano	59
3.14	Critérios de seleção de modelos	65
3.15	Exemplo de identificação de anomalias com o sinal fictício	66
3.16	Implementação	68
3.17	Procedimento proposto	71
4	Resultados	75
4.1	Resultados da Filtragem e da Padronização	75
4.2	Resultados da PCA	78
4.3	Resultados da ICA	81
4.4	Distribuições Resultantes	82
4.5	Resultados do ajuste do modelo de Misturas Gaussianas	85
4.6	Resultados das Cadeias Gaussianas Ocultas de Markov	87
4.7	Resultado de múltiplos treinos	89
5	Conclusões e Trabalhos Futuros	92

Referências bibliográficas	96
A Demonstraçāo das equaçōes de atualizaçāo	101
A.1 Equaçāo de recorrênciā da média	101
A.2 Equaçāo de recorrênciā do desvio-padrāo	102
B Algoritmo de Maximizaçāo da Expectativa	103
C Códigos desta dissertação	107

Capítulo 1

Introdução

Na atual cadeia produtiva, o petróleo é uma matéria-prima importante. Sua utilização é diversa, abrangendo inúmeros setores da economia, e como principal, o setor energético. Diversas empresas atuam na exploração deste recurso, e todas buscam aumentar sua competitividade melhorando sua produtividade. O petróleo, por sua vez, localiza-se essencialmente em poços. Estes poços podem ser tanto *onshore* (perfuração em terra firme), como *offshore*, perfuração de um poço abaixo no leito marítimo (Devold, 2013).

Alguns poços possuem pressão própria, ou seja, o fluido acumulado nos mesmos possui uma pressão intrínseca que permite que eles sejam expelidos por conta própria. Outros, necessitam de uma fonte de trabalho para serem extraídos. Neste sentido, uma das principais ferramentas na operação de extração são as bombas de extração de petróleo. A função destas é realizar a elevação artificial do fluido, ou seja, retirar o óleo de um poço que não possui pressão intrínseca para expeli-lo por conta própria.

Alguns poços ficam em um nível abaixo do mar, e as bombas utilizadas neste contexto são denominadas como Bombas Centrífugas Submersas (BCS). As BCS atuam diretamente em contato com o reservatório, portanto, permanecendo em um ambiente de difícil acesso para uma manutenção corretiva. No presente caso desta dissertação, as bombas são elétricas, que são usadas em poços de alto fluxo de extração, sendo muito versáteis em diversas aplicações (Takacs, 2018). Este tipo de bomba é também denominado na literatura como *Electrical Submersible Pumps* (ESP), em inglês.

Desta forma, a manutenção preditiva torna-se importante neste contexto, visto que a predição de falhas é importante para a previsão da execução da manutenção ou remoção da bomba. A remoção de uma bomba submersa é um processo caro que pode impactar

diretamente na operação do poço, que se não for bem planejado, pode causar sérios prejuízos. As causas de problemas em BCS são muito variadas, uma vez que as condições de operação destes equipamentos são adversas (Al-Khalifa; Cox; Saad, 2015).

Fakher et al. (2021) cita três principais classificações de falhas nas BCS: elétricas, mecânicas e operacionais. As falhas elétricas seriam aquelas relativas aos componentes elétricos ou às cargas aplicadas à bomba, como, por exemplo, falha no cabo por degradação, falha no motor, sobrecarga da bomba e perda dos sensores. Falhas mecânicas podem ser consideradas como falhas que ocorrem nas partes físicas da bomba, como, por exemplo, corrosão da bomba, deslocamento de alguma peça, vazamento e quebra de algum componente. Por sua vez, as falhas operacionais são aquelas relativas à operação, como, por exemplo, temperatura no poço, pressão, múltiplas fases em um fluido e sólidos entrando na bomba e a danificando.

Ainda de acordo Fakher et al. (2021), a maioria das falhas é reportada de maneira manual, de modo que a identificação da mesma ocorre através de um operador classificando-a. Neste sentido, seria bom ter um método que as identificasse, ou melhor, que executasse a detecção de anomalias que podem indicar a falha. Estes métodos podem apoiar a manutenção preventiva, executando a troca das bombas de maneira planejada, sem haver parada da operação.

O desenvolvimento de técnicas de aprendizado de máquina permitiu que novas maneiras de detectar anomalias e falhas fossem descobertas, e levou a novas abordagens para o aprimoramento da manutenção preditiva em diversas máquinas. Como base de registros, a utilização dos dados dos sensores de máquinas, em geral, tem sido utilizada como fonte para detecção de anomalias. Leukel, González e Riekert (2021) fazem um levantamento dos principais métodos utilizados para a predição de falhas em máquinas no geral. No caso, para a previsão, os métodos mais utilizados são os de aprendizado supervisionado.

Os métodos de aprendizado supervisionado são aqueles que se utilizam de uma base de dados de entrada, ou seja, a informação disponível para fazer uma inferência, e a base de dados de saída, que é o objetivo a ser inferido. Deste modo, o método de aprendizado supervisionado é aquele que busca otimizar uma função (ou um modelo), que conforme uma determinada entrada, possa inferir uma saída (Gori; Betti; Melacci, 2023).

No presente contexto desta dissertação, existem dados relacionados à bomba e dados relacionados à falha. Os dados de falhas, usualmente, são registrados manualmente, ou seja, houve algum tipo de classificação manual do comportamento da bomba, registrado por

algum operador. Porém, esta rotulação de dados é usualmente manual. Não somente, a rotulação de um dado pode ser feita de forma equivocada ou atrasada (considerando uma série temporal), de tal maneira que o resultado do aprendizado supervisionado pode não ser preciso (Speight (2007), Gerón (2019)).

Há dois tipos de dados registráveis na operação de BCS: dados dos sensores das BCS e dados de produção de campo. O primeiro está relacionado às variáveis medidas nas bombas, como pressões, temperaturas e variáveis elétricas. Essas medições são fornecidas por vários sensores da BCS, permitindo o monitoramento em tempo real. Quanto aos dados de produção, as relações mais comumente medidas são razão gás-líquido, razão água-óleo, razão gás-água, sedimentos básicos e água e propriedades de pressão, viscosidade e temperatura do óleo extraído. A maior parte da literatura usa dados de sensores da bomba, em vez de dados de produção, pois dados de produção podem não estar disponíveis, visto que podem revelar informações importantes da empresa.

Nesta dissertação, os dados disponíveis são informações reais da operação de 57 corridas (intervalo entre a instalação da bomba e sua remoção) da empresa Equinor. Eles são compostos por dois conjuntos de dados diferentes. O primeiro é um banco de dados de medições dos sensores da BCS com intervalos de uma hora por entrada, que lista diversas propriedades das bombas, como pressões, temperaturas, vibrações e propriedades elétricas. O segundo é uma planilha de falhas indicando quando a falha foi relatada pelo operador.

As falhas neste conjunto de dados são decorrentes da operação real, o que muitas vezes as torna menos aparentes. Elas podem não apresentar padrões claros ou marcadores distintos, aumentando a dificuldade de sua detecção pelos especialistas de bombas. Neste sentido, o principal desafio desta dissertação é tentar identificar anomalias nestes conjuntos de dados que possam ser indicadores da ocorrência iminente da falha.

1.1 Motivação

Ao longo dos anos, diversos estudos já foram realizados no contexto da operação de BCS, utilizando novos métodos para detectar falhas e comportamentos anômalos, a maior parte com modelos de aprendizado supervisionado. Na literatura investigada, os modelos que identificam as anomalias são, em sua maioria, derivados do modelo de Máquinas Vetor de Suporte (em inglês, *Support Vector Machine* (SVM)). Por exemplo, o trabalho de Peihao Yang, Chen, Zhang

et al. (2022) utiliza de dados de produção como entrada e classificações de anomalia como saída para um modelo de SVM combinado com um autocodificador de ruído como filtro. Já o trabalho de Awsan (2023) classifica sinais de operação da bomba com base na rotulação de falha ou não. Ambos, possuem dados obtidos de laboratório, que são ambientes controlados, sendo possível ter maior controle e discernimento entre a caracterização dos estados.

Por sua vez, o trabalho de Lastra e Xiao (2022) utiliza modelos combinados, ambos baseados em SVM, florestas aleatórias e K-vizinhos mais próximos para realizar a classificação. Eles utilizam dados operacionais reais em seu trabalho, mas a rotulagem dos dados foi feita por especialistas, o que, para uma quantidade razoável de dados, exige uma quantidade significativa de trabalho. No entanto, a caracterização de falhas e anomalias por este método pode não fazer sentido quando o operador não consegue discernir o comportamento, ou o volume de dados torna esta rotulagem manual impraticável.

Por outro lado, alguns trabalhos realizam uma abordagem utilizando aprendizado não-supervisionado. Peihao Yang, Chen, Wu et al. (2022) utiliza uma técnica de classificação não-supervisionada de anomalias em seu trabalho. Através de um limiar da distribuição de dados, a entrada é classificada entre *outlier* ou não. Neste sentido, é possível analisar duas questões. A primeira é que como considerar isto em uma entrada multivariada e que, na classificação de falhas em si, as relações de séries temporais não são consideradas. A segunda abordagem é a proposta por Abdalla et al. (2022), onde é proposta uma redução de dimensionalidade através da Análise de Componentes Principais (PCA). Esta técnica é uma transformação algébrica, mas, é usualmente classificada como aprendizado não-supervisionado, visto que na redução de dimensionalidade, é possível melhorar a identificação de padrões. Para tentar simplificar o número de variáveis, a PCA é utilizada para tentar encontrar eixos ortogonais na dimensão dos dados onde haja maior variância, ou correlação, e é possível recombinar as variáveis de modo a produzir estas componentes. Os autores usam esta técnica para reduzir a dimensionalidade e fazer com que o algoritmo *XGBoost* realize a identificação com menos dados, melhorando o tempo de processamento. Como será mostrado nesta dissertação, o princípio de funcionamento desta análise é a decomposição ortogonal das matrizes de covariância, ou de correlação.

Outra técnica possível de ser utilizada para encontrar relações é a Análise de Componentes Independentes, proposta por Kuha (2004), a qual busca encontrar fontes no sinal que sejam independentes uma das outras. O autor da presente dissertação, ao investigar

a utilização deste método voltado às BCS, encontrou um artigo utilizando a técnica, Ypma, Tax e Duin (1999), combinada com uma técnica similar a SVM para a detecção de falhas em ambiente experimental, e em outro contexto, Ajami e Daneshvar (2012) utiliza o método para detectar falhas em plantas termoelétricas. Entretanto, na presente dissertação, estas técnicas serão exploradas para avaliar se existem correlações entre as variáveis medidas se é possível evidenciar anomalias.

Frente ao objetivo de detectar anomalias, é desejável um método que não exija conhecimento a priori das falhas e consiga inferir relações de séries temporais. Na literatura de Aprendizado de Máquina é difícil encontrar um método de aprendizado não supervisionado para classificação de séries temporais. Usualmente, os métodos utilizados para detecção de padrões (Hodge; Austin, 2004) não possuem relações no tempo, o que implicaria em uma desconsiderar uma informação importante.

Entretanto, uma técnica de aprendizado não-supervisionado que considera esta relação são os Modelos Ocultos de Markov, estabelecidos por Baum et al. (1970). Eles já foram usados em contextos de decodificação de sequências a partir de sinais de séries temporais, como reconhecimento de fala (Rabiner, 1989), decodificação genética (Stanke; Waack, 2003), identificação de padrões em séries temporais financeiras (Mamon; Elliott, 2010) e aplicações em Teoria de Controle (Elliott; Moore; Aggoun, 2008). Simplificadamente, este método busca classificar as entradas da série temporal com base em estados (variáveis ocultas), que seriam a "causa" da observação daquela entrada em específico.

No contexto de detecção de falhas, este método já foi utilizado para a detecção de anomalias em máquinas, como em Smyth (1994), que utiliza diversos modelos para realizar uma identificação, e em Zhou et al. (2010) que busca utilizar o modelo para tentar entender um sistema sobre mudanças externas em tempo real. Nesta dissertação, a proposição será utilizar os estados como indicativos de eventos normais ou anormais, de tal maneira a prover uma classificação dos equipamentos aos operadores sobre uma entrada em específico.

Considerando a natureza dos dados operacionais, é sabido que a medição de dados, de forma geral, apresenta ruído e dados faltantes devido a falhas nos sensores (Balbinot; Brusamarello, 2019). De forma geral, na comunidade de Ciência de Dados, existe uma série de práticas adotadas para lidar com estes problemas. De forma geral, a obra de Kotu e Bala (2019) as compilam, servindo de referência para o tratamento no contexto de exploração de dados.

1.2 Objetivos

Esta dissertação, tem, como seu objetivo principal, a identificação de anomalias em sinais de bombas elétricas centrífugas submersas usadas na elevação artificial de petróleo. Especificamente, outros objetivos são:

- Estabelecer um método de processamento dos dados das 57 corridas fornecidas pela empresa Equinor.
- Identificar relações entre os sinais dos sensores aplicando técnicas como PCA e ICA.
- Aplicar o método de Modelo Oculto de Markov Gaussiano para identificação de anomalias nos sinais.
- Analisar o resultado dos métodos propostos.

Os códigos desenvolvidos nesta dissertação utilizam bibliotecas estáveis e amplamente adotadas na linguagem Python, cuja manutenção é assegurada por comunidades consolidadas de desenvolvedores. A adoção de bibliotecas consolidadas em Python visa garantir a reproduzibilidade, a estabilidade e o suporte contínuo das ferramentas utilizadas.

1.3 Estrutura da Dissertação

Além deste capítulo, o presente trabalho possui 4 outros. No Capítulo 2, serão introduzidos os desafios na exploração e análise de dados operacionais. Para tal, será conduzida uma análise exploratória dos dados, avaliando as características dos dados, realizando considerações e propondo caracterizações. Posteriormente, no Capítulo 3, a metodologia proposta para aplicação dos modelos computacionais escolhidos para esta dissertação serão apresentados e discutidos. Aspectos da implementação dos algoritmos desta dissertação também serão apresentados. Em seguida, no Capítulo 4, seus resultados serão revelados. Por fim, no Capítulo 5, a eficácia dos modelos será discutida e avaliada, apresentando as principais conclusões do trabalho.

Capítulo 2

Desafios na análise de dados operacionais das BCS

2.1 Coleção de dados no processo de elevação artificial

O processo de elevação artificial pode utilizar bombas centrífugas submersas, que são máquinas rotativas que geram diferença de pressão através da rotação de um rotor. Ao conectá-las em um poço de petróleo, que não possua pressão suficiente para expelir seu conteúdo por conta própria, a bomba estabelece um diferencial de pressão que permite a extração do conteúdo do poço. Uma bomba pode possuir vários estágios, que possibilitam aumentar a pressão do líquido utilizando de menos velocidade de rotação. Neste processo, um sistema de alimentação provê energia para a bomba operar (Takacs, 2018).

Para o monitoramento desta operação, alguns sensores são colocados sobre a carcaça da bomba, como os sensores de vibração, outros colocados para medir as pressões e temperaturas e também há sensores utilizados para medir as propriedades elétricas, como por exemplo, a tensão e corrente no motor. Isto permite que as informações da bomba, estando submersa, seja coletada, possibilitando o monitoramento da operação (Takacs, 2018). Entretanto, como em qualquer processo de medição, existirá ruído e interferências alheias à operação da bomba. É possível que alguns dados não consigam ser registrados por falhas nos sensores, o que pode levar a trechos faltantes nas informações das medições (Balbinot; Brusamarello, 2019).

Todos os dados utilizados nesta tese são relativos ao processo de elevação artificial de 57 corridas ao longo dos anos de 2011 a 2020, fornecidos pela empresa Equinor. Uma corrida é o período entre a instalação da bomba (seja pela primeira vez, ou pós-manutenção) e sua

remoção. O outro tipo de dado é uma planilha de falhas compilada pelo time de manutenção da Equinor. Esta planilha aponta o dia em que a falha foi reportada na corrida. Na Tabela 2.1 existe uma descrição de todas as grandezas contidas nos dados dos sensores e suas respectivas unidades.

Tabela 2.1: Descrição das colunas dos dados das 57 corridas.

Propriedade	Unidade
Corrente do motor - Fase A	A
Corrente do motor - Fase B	A
Corrente do motor - Fase C	A
Frequência do AVV	Hz
Temperatura do motor da BCS	°C
Pressão diferencial da BCS	Bar
Pressão na entrada da BCS	Bar
Fração de água a 20°C	%
Temperatura na entrada da BCS	°C
Pressão de descarga da BCS	Bar
Sensor de temperatura na descarga da BCS	°C
Abertura da válvula	%
Pressão na Cabeça do Poço	Bar
Temperatura na Cabeça do Poço	°C
Poço Alinhado a A	Booleano
Poço Alinhado a B	Booleano
Tensão do Motor da BCS	V
Vibração da BCS em X	g
Vibração da BCS em Y	g
Poço desativado	Booleano

Na Figura 2.1 é mostrado o esquema de uma instalação comum de uma BCS, adaptado de Takacs (2018). Em azul, é representado o fluido. Em vermelho, é representado o fluxo do fluido no poço para dentro da coluna de produção. Esta região é considerada como sendo a entrada, onde, para o banco de dados fornecido, são medidas temperatura e pressão. Nota-se que o motor é a peça mais abaixo da instalação. Do banco de dados, há o registro das amplitudes das correntes trifásicas, da tensão do motor da BCS e, em algumas corridas, sua temperatura.

A seção final da bomba é a descarga, que é a seção após o fluido ter passado por todos os estágios da BCS. Nesta saída do processo de bombeamento, apenas a pressão é medida. Com

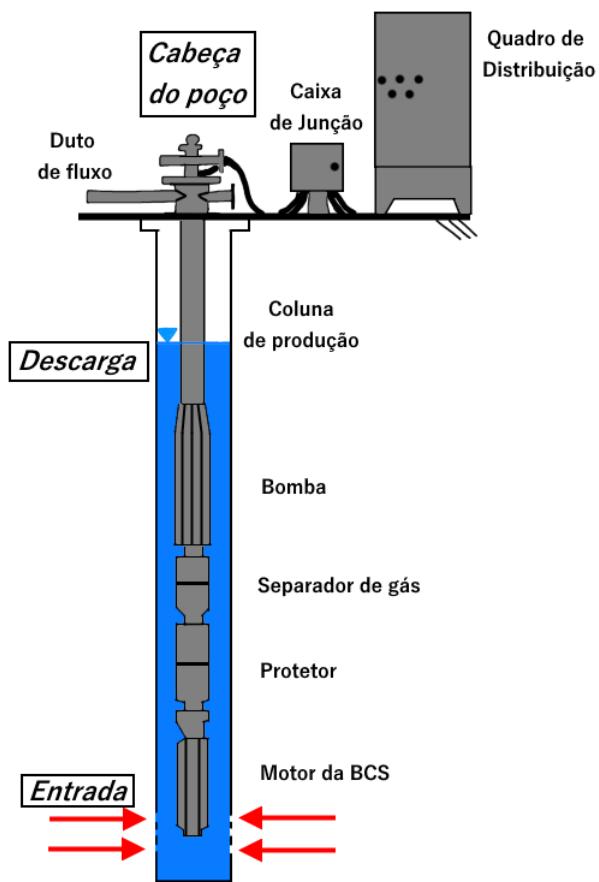


Figura 2.1: Diagrama de uma instalação da BCS - adaptado de Takacs (2018).

relação à vibração da bomba, esta é medida em duas componentes, formando um plano cujo vetor normal é paralelo ao eixo longitudinal da bomba.

Posteriormente, o fluido chega aos dutos de fluxo que o levam para o armazenamento, na cabeça do poço. Ela é composta por equipamentos que ajudam a vedar o poço, para que nenhum gás ou líquido saia para o mar (Speight, 2007). Na cabeça do poço também são medidas temperatura e pressão. Na região da cabeça do poço nota-se elementos como o quadro de distribuição e a caixa de junção, que são elementos elétricos. Com relação a eles, apenas a frequência do Acionador de Velocidade Variável (AVV) (que indica a rotação do motor) é listada no banco de dados.

Outras informações que também constam no banco de dados é a razão entre água e óleo a 20°C do fluido obtido, a abertura da válvula no sistema para recebimento do conteúdo extraído, e se o poço estava ativado ou não, isto é, se estava havendo o processo de extração nele e por consequência, se as bombas estavam ou não ligadas. Também consta a pressão diferencial que é um resultado calculado entre a diferença das pressões de descarga e de entrada.

Um exemplo de dados de sensores de uma corrida pode ser visto na Figura 2.2. Um aspecto que chama atenção são as escalas diferentes dos dados. O exemplo mais notável é a tensão do motor, evidentemente muito maior do que os outros dados. Na Figura 2.3 é possível avaliar que esta diferença é realmente bem grande. Entre os dados de vibração e tensão do motor, a diferença é da ordem de grandeza de 10^4 .

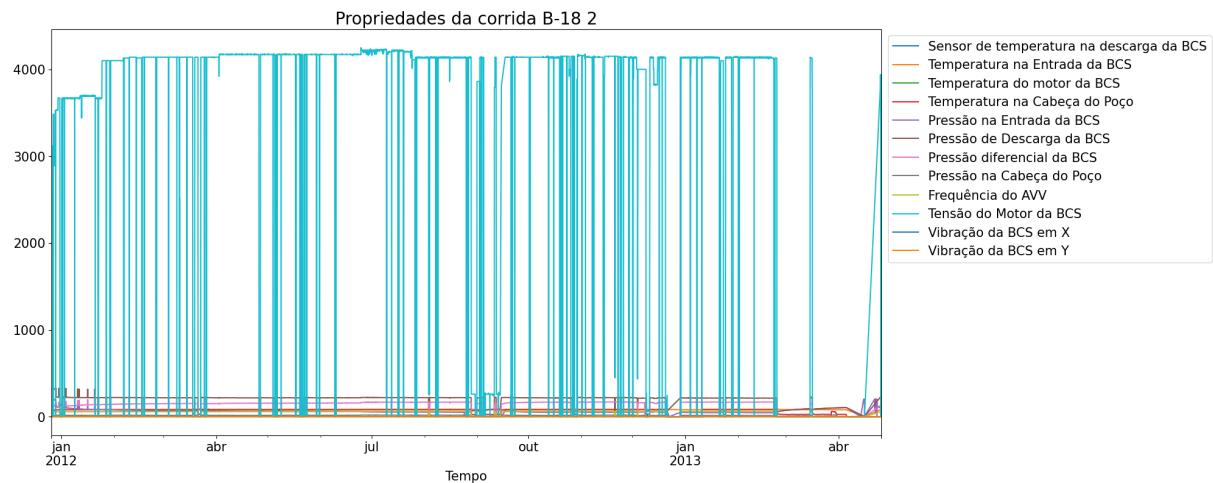


Figura 2.2: Dados originais da corrida B-18 2.

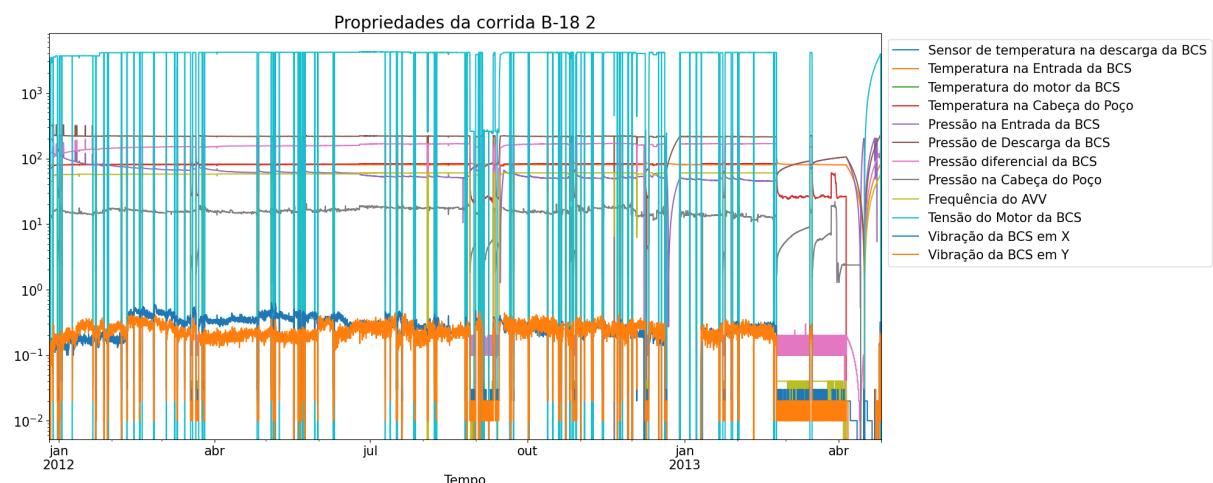


Figura 2.3: Dados originais da corrida B-18 2 - escala logarítmica.

2.2 Descrição dos Dados

Os dados dos sensores possuem algumas características que fazem a análise desafiadora, como por exemplo, dados perdidos, duplicados e ruído nas medições. Com relação aos dados duplicados, é possível os identificar e removê-los, de modo a não permitir que as duplicatas

interfiram nos resultados. No caso deste trabalho, foi utilizado um *script* na biblioteca *pandas* para executar esta tarefa.

Após remover os dados duplicados, é possível obter estatísticas descritivas, como a média e o desvio-padrão. A média é uma medida de tendência central dos dados e o desvio-padrão informa a dispersão. Desta maneira, na Tabela 2.2 são mostradas as propriedades de cada uma das variáveis listadas no banco de dados. Um fato desta tabela é notar que valores booleanos, quando obtidos sua média, retornam o percentual de elementos com valor 1 (verdadeiro). Com isto, nota-se que, considerando a propriedade de poço desativado, 38% dos dados são relativos às bombas desligadas.

Outro aspecto da Tabela 2.2 é que as três correntes possuem propriedades estatísticas muito similares entre si. Isto também ocorre para a vibração. Neste sentido, seria possível usar apenas o módulo das propriedades em conjunto, ou seja, a raiz quadrada da soma dos quadrados. No caso da vibração, isto é intuitivo, pois ela é uma grandeza espacial e pode ser somada vetorialmente. No caso, como temos as componentes no plano, podemos obter o tamanho do vetor vibração. Para a questão da corrente, é análogo a obter um valor eficaz, pois os valores medidos são as amplitudes das correntes.

Posteriormente, na Tabela 2.3 é possível avaliar o percentual de dados vazios. Nesta, com relação à Tabela 2.2, são mostradas variáveis com mais de 1% de dados vazios. Nota-se que o valor numérico com mais dados faltantes é a temperatura na descarga (saída da bomba). Desta forma, opta-se por não considerar esta variável nos cálculos posteriores, pois como pode ser explorado nos dados, em alguns modelos de bomba não há sensores para medir esta propriedade. Os demais valores numéricos percentuais faltantes são em decorrência de ausências de medição, possivelmente, falhas nos sensores.

Também são desconsideradas outras propriedades não relativas as bombas, como por exemplo o alinhamento do poço, que é uma variável booleana, considerando o poço como "alinhado"ou "não alinhado".¹ Com relação a abertura da válvula e a fração de água a 20°C, estas não são consideradas nos cálculos à frente, pois não são medições propriamente nas BCS. Outra propriedade desconsiderada é a pressão diferencial, pois é obtida diretamente entre a pressão de entrada e a pressão de descarga.

¹Nota para a qualificação: Aguarda-se uma explicação da empresa com relação a variável “alinhamento do poço”. A interpretação do nosso grupo de pesquisa foi de que era uma questão com relação ao alinhamento de produção entre dois poços distintos da Bacia de Peregrino, por isto, não foi considerada nos cálculos. Na versão final, a explicação será incluída.

Tabela 2.2: Propriedades dos dados fornecidos originalmente.

Propriedade	Unidade	Média	Desvio Padrão
Corrente do motor - Fase A	A	62,89	50,87
Corrente do motor - Fase B	A	64,18	51,77
Corrente do motor - Fase C	A	62,78	51,24
Frequência do AVV	Hz	35,67	28,13
Temperatura do motor da BCS	°C	103,56	25,11
Pressão diferencial da BCS	Bar	83,99	73,42
Pressão na entrada da BCS	Bar	119,98	55,67
Fração de água a 20°C	%	48,31	31,90
Temperatura na entrada da BCS	°C	81,40	7,92
Pressão de descarga da BCS	Bar	211,48	43,47
Sensor de temperatura na descarga da BCS	°C	5,49	20,51
Abertura da válvula	%	23,69	20,66
Pressão na Cabeça do Poço	Bar	13,54	8,73
Temperatura na Cabeça do Poço	°C	58,20	28,97
Poço Alinhado à A	Booleano	0,48	0,50
Poço Alinhado à B	Booleano	0,42	0,49
Tensão do Motor da BCS	V	2676,43	2095,46
Vibração da BCS em X	g	0,18	0,35
Vibração da BCS em Y	g	0,16	0,24
Poço desativado	Booleano	0,38	0,48

Tabela 2.3: Tabela com percentual de dados vazios por coluna.

Propriedade	Percentual
Temperatura do motor da BCS	10,2%
Pressão na Entrada da BCS	8,0%
Fração de água a 20°C	10,0%
Temperatura na Entrada da BCS	7,9%
Pressão de Descarga da BCS	7,9%
Sensor de temperatura na descarga da BCS	24,4%
Poço Alinhado à B	55,3%

Como dito anteriormente, as três correntes (A,B,C) e as duas vibrações (X,Y) podem ser unidas através do módulo de seus valores. Com isso, é possível gerar a Tabela 2.4, resultando em 10 componentes numéricas para serem analisadas.

Tabela 2.4: Tabela das variáveis após o pré-processamento de redução e aglutinação.

Propriedades	Unidade	Média	Desvio Padrão
Frequência do AVV	Hz	35,67	28,13
Temperatura do motor da BCS	°C	103,56	25,11
Pressão na entrada da BCS	Bar	119,98	55,67
Temperatura na entrada da BCS	°C	81,40	7,92
Pressão de descarga da BCS	Bar	211,48	43,47
Pressão na Cabeça do Poço	Bar	13,54	8,73
Temperatura na Cabeça do Poço	°C	58,20	28,97
Poço Alinhado à A	Booleano	0,48	0,50
Poço Alinhado à B	Booleano	0,42	0,49
Tensão do Motor da BCS	kV	2,68	2,10
Poço desativado	Booleano	0,38	0,48
Média da corrente	A	109,67	88,78
Módulo da Vibração na BCS	g	0,25	0,42

Além dos dados dos sensores, foi fornecida uma planilha informando quais as datas e razões das falhas. As falhas típicas reportadas são curto circuitos, devido a corrosão do cabo de energia ou de proteções da BCS. Também é reportada ausência de fluxo para superfície devido a quebra dos eixos das bombas. Entretanto, o detalhamento do porquê da falha só é obtido após a remoção da bomba para inspeção. Avaliando a data da falha, sua descrição e o momento registrado, analisou-se visualmente a região dos dados próxima das falhas, buscando por padrões que possam indicar as mesmas. Esta foi a análise preliminar, pois não se sabia exatamente como os operadores fazem a identificação da falha, imaginando-se que eles verificam o grau de flutuação dos valores.

A planilha de informações de falhas foi analisada e simplificada, considerando apenas a data e a causa da falha. Esta simplificação foi feita com o intuito criar uma classificação de dados “antes da falha” e “depois da falha”. Entretanto, é importante ressaltar que as falhas são reportadas apenas nos seus dias, e não o momento exato do dia em que a falha foi identificada.

Posteriormente, os dados com esta simplificação foram exportados para um arquivo de valores separados por vírgulas.

Desta maneira, é possível realizar uma combinação dos dois bancos de dados, para que a informação da falha esteja junto com o banco de dados dos sensores, o que permitirá a análise das duas informações de forma conjunta. Assim, foi possível criar a classificação entre “dado anterior ao dia da falha” e “dado posterior ao dia da falha”. A intenção inicial desta classificação era a utilização de algum método de aprendizado supervisionado, tal como Support Vector Machine (SVM) e XGBoost (XGB), que poderiam ser utilizados para classificar dados “com” ou “sem” falhas. Entretanto, isto não é adequado, em decorrência do fato de que as falhas possuem um registro único por dia, enquanto os dados dos sensores são horários. Mesmo se uma hora em específico seja escolhida para representar a falha, existe o risco de classificar dados normais como falhos e vice-versa. A data que consta como a falha, é a data do registro da mesma, e não, necessariamente, de sua observação, ou, até mesmo, de sua ocorrência. Portanto, a rotulação da série temporal baseada nos dados de falha disponíveis não seria adequada. Desta maneira, como sua referência no tempo não é precisa, um método de aprendizado supervisionado, que se utilizaria de dados temporais rotulados entre falhos e não-falhos, não seria adequado. Desta forma, opta-se pela utilização de métodos de aprendizado não supervisionado na identificação de falhas neste trabalho.

2.3 Características dos dados

Para continuar a análise exploratória de dados, é importante analisar o histograma das medições nos sensores. Os histogramas mostram o formato das distribuições, de tal maneira que seja possível caracterizar visualmente os dados. A análise visual apenas não é adequada para definir a distribuição, mas é suficiente para selecionar os possíveis tipos de distribuição que os dados podem assumir.

Na Figura 2.4, os histogramas são mostrados, de maneira que é possível analisar, que as propriedades possuem distribuições diferentes entre si. Uma vez desenhados, é possível notar que propriedades como a temperatura de entrada na BCS, pressão na cabeça do poço e média da corrente estão com valores discrepantes muito distantes da média, o que torna difícil inferir algum tipo de distribuição. De certa forma, a presença de valores muito discrepantes da própria

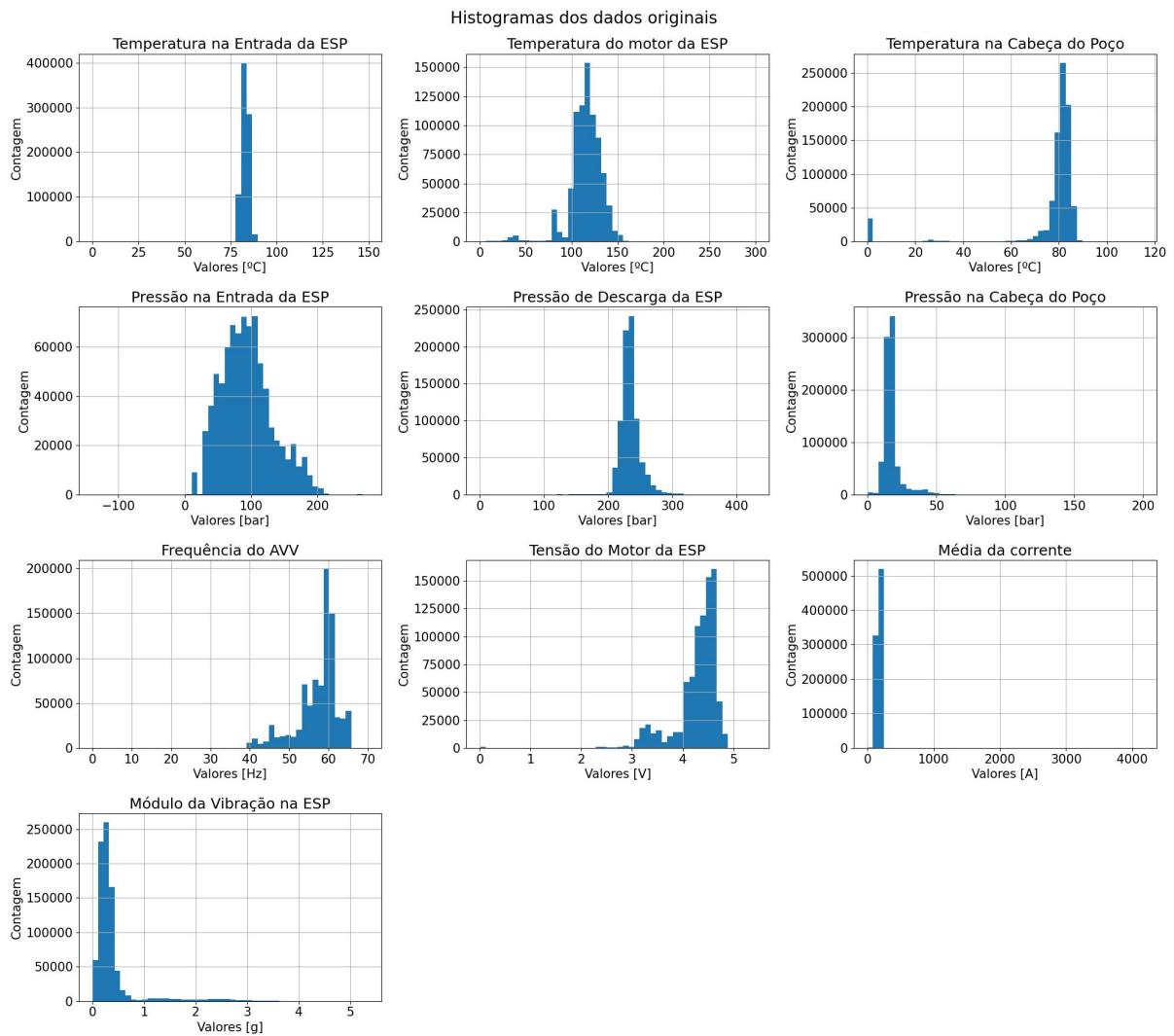


Figura 2.4: Histograma dos dados medidos (os valores limites das abscissas contemplam todos os pontos).

distribuição acabam interferindo na visualização do histograma, que foi gerado considerando todos os dados.

Neste caso, o procedimento ideal seria tomar um conjunto de valores da distribuição que correspondesse a um intervalo que contenha a maior parte destes. Desta maneira, Kay (2006) refere-se ao Teorema do Limite Central, que enuncia que para múltiplos experimentos sobre uma variável aleatória, teoricamente, a distribuição converge em uma distribuição gaussiana. Desta maneira, seria possível centralizar as variáveis em torno da média e escalá-las pelo desvio-padrão, de modo a comparar com uma Gaussiana Padrão. Este procedimento é denominado como *Z-score* e será melhor discutido no Capítulo 3. Com isto, é possível obter os histogramas da Figura 2.5.

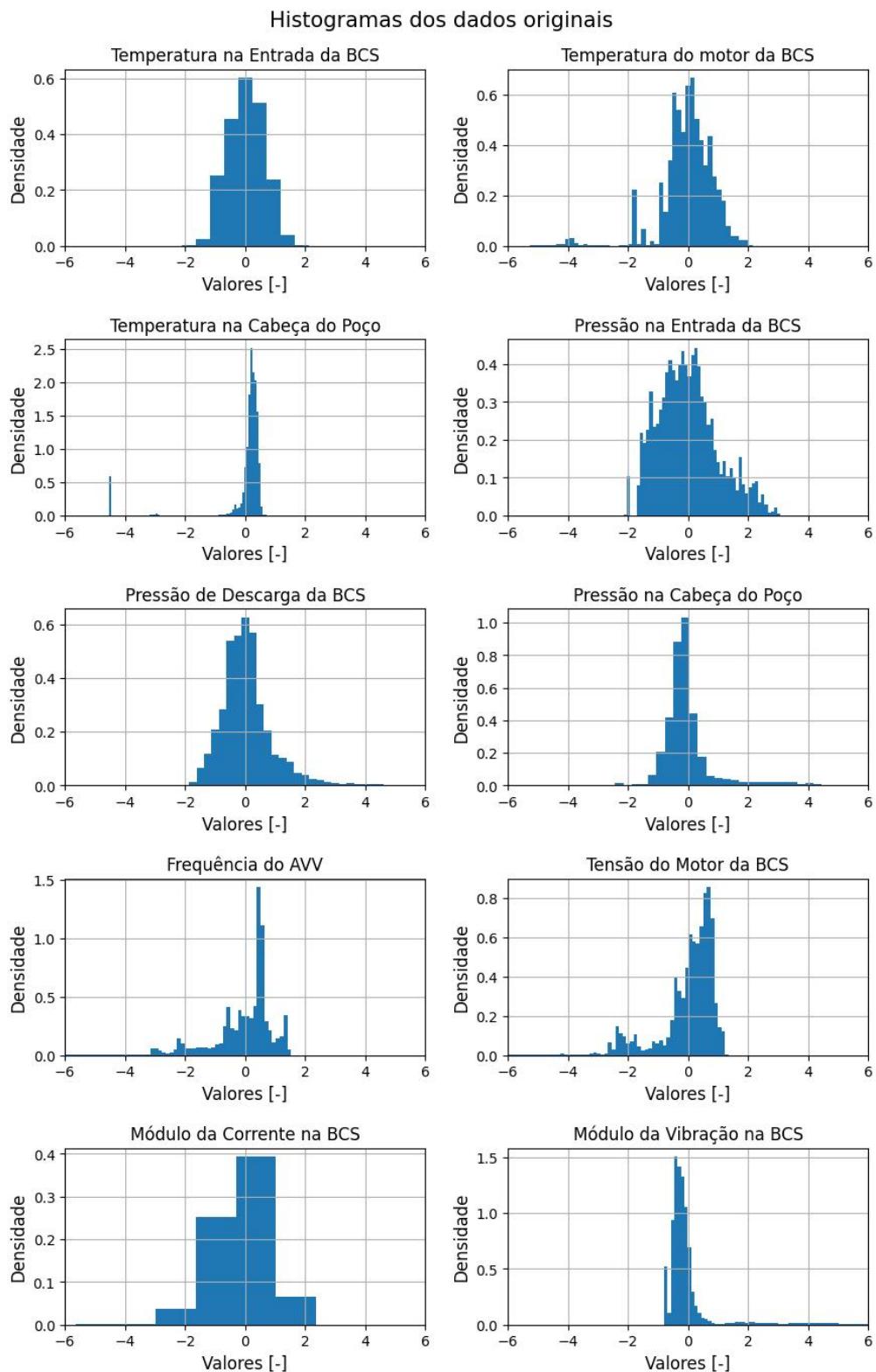


Figura 2.5: Histograma das variáveis padronizadas (de -6σ até 6σ).

Nesta abordagem, será escolhido um intervalo de dados entre -6σ e 6σ , onde σ é o desvio-padrão da variável. Caso os fenômenos possam ser descritos como gaussianos, este valor corresponderia a 99,99994266979% dos dados contidos, o que implicaria que grande parte dos dados estaria dentro deste intervalo. Nesta figura, é possível notar que, visualmente, é bem claro que a maioria das distribuições não são gaussianas, pois a maioria delas possui grande grau de assimetria em torno da média. As únicas exceções seriam a temperatura na entrada e o módulo da corrente.

Para avaliar o grau de normalidade destas distribuições, uma das sugestões propostas por Thode (2002) é realizar o Gráfico Q-Q (quantil-quantil). Este analisa os quantis de distribuições distintas, realizando uma comparação entre elas. Desta forma, é possível comparar as distribuições das variáveis da bomba com uma distribuição normal. Caso as distribuições sejam normais, os pontos da variável estarão alinhados com uma linha que representa os quantis esperados da distribuição.

Com o gráfico da Figura 2.6, é possível notar que, tal como foi observado visualmente, as variáveis não possuem um alinhamento significativo com o da distribuição normal na linha vermelha, o que é um indicativo muito forte de não normalidade. Com isto, métodos baseados em distribuições gaussianas devem ser utilizados com cautela. É possível avaliar que o problema de interesse desta dissertação necessita de uma análise mais específica. No próximo capítulo, serão propostas metodologias para tentar evidenciar as anomalias nas séries temporais.

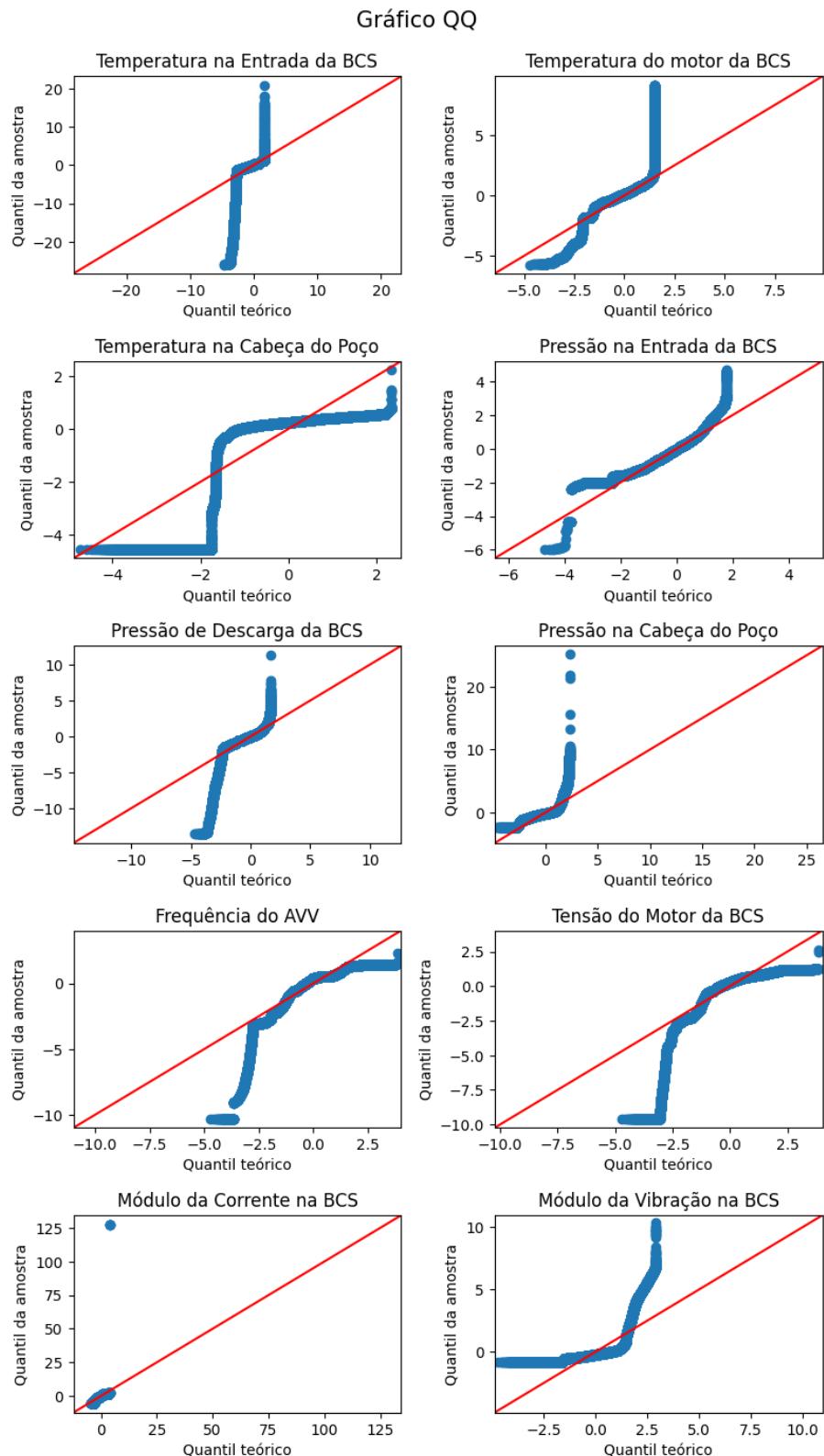


Figura 2.6: Gráfico Q-Q das variáveis das bombas - a linha vermelha indica a distribuição normal associada.

Capítulo 3

Metodologia Proposta

3.1 Variável Aleatória

A definição de variável aleatória envolve uma formulação matemática densa, que não está no escopo desta dissertação. Entretanto, é relevante entender que uma variável aleatória mapeia um conjunto de eventos para uma variável numérica. Neste sentido, para um conjunto de eventos Ω , pode existir um mapeamento para \mathbb{R} , tal que $X(\Omega) \in \mathbb{R}$, o conjunto dos reais, onde $X(\cdot)$ é chamada de variável aleatória. No presente contexto, seria possível interpretar os acontecimentos físicos na bomba sendo manifestados numericamente pelas variáveis dos sensores com o conjunto:

$$X = \{x_1, x_2, x_3, \dots, x_n\} \quad (3.1)$$

onde n é o número de elementos em X e x_k são os elementos de X .

3.2 Distribuição Gaussiana

Neste trabalho, por mais que a maioria dos fenômenos não obedeça a distribuição gaussiana, ela é uma referência para o desenvolvimento dos métodos. Ela é conhecida por representar fenômenos de diversas naturezas em diversos contextos. Pelo Teorema do Limite Central, é esperado que diversas repetições de um experimento convirjam em uma distribuição gaussiana. A distribuição Gaussiana Padrão é dada como:

$$\mathcal{N}(Z) = \frac{1}{2\pi\sigma} e^{-\frac{1}{2}Z^2} \quad (3.2)$$

onde o termo elevando o número de Euler pode ser definido como Z , sendo uma medida definida como:

$$Z(x, \mu, \sigma) = \frac{x - \mu}{\sigma} \quad (3.3)$$

em que x é a variável, μ é a média e σ é o desvio-padrão.

Esta medida Z é importante, pois permite que para diferentes conjuntos de dados, uma mesma métrica de probabilidade seja calculada. Por exemplo, suponha uma distribuição gaussiana A de $\mu = 3$ e $\sigma = 4$. Para uma variável com valor $x = 1$, a sua probabilidade acumulada é 0,30854. Para uma distribuição gaussiana B, de $\mu = 94$ e $\sigma = 105$, a mesma probabilidade acumulada ocorreria para $x = 41,5$. Isto permite estabelecer referências entre distribuições diferentes, de modo a comparar probabilidades entre valores distintos em distribuições distintas. Neste sentido, a Gaussiana Padrão é uma distribuição gaussiana que já leva em consideração os valores padronizados.

3.3 Momentos

A média e o desvio-padrão são características da distribuição de dados. Na seção anterior, estas variáveis foram interpretadas como parâmetros, entretanto, é possível calculá-las a partir dos dados. De acordo com Kay (2006), a média é definida como a esperança da variável aleatória. Com isto, a operação da esperança é definida, para dados discretos, como o somatório das probabilidades da variável ao longo de todos seus valores possíveis:

$$\mu = \mathbb{E}[X] = \sum_{x \in X} P(x_k)x_k = \sum_{k=1}^n P(x_k)x_k \quad (3.4)$$

onde X é a variável aleatória, $p(x)$ sua distribuição de probabilidades, x a variável do espaço em que as variáveis aleatórias estão contidas e o operador $\mathbb{E}[\cdot]$ denota a soma ponderada pelas probabilidades. No caso, para uma distribuição de dados obtida experimentalmente, a probabilidade de um dado elemento pode ser calculada como:

$$P(x_k) = \sum_{x_k \in X} \frac{1}{n} \quad (3.5)$$

onde x_k é um determinado valor possível dentro das variáveis listadas em X . A esperança mede o valor esperado na realização de um experimento aleatório, ou, no contexto desta dissertação,

da medição de um sensor. Neste sentido, os momentos centrais são definidos como sendo métricas para medir a dispersão em torno da média. Um momento central discreto de grau g pode ser escrito como:

$$\mathbb{E}[(x - \mu)^g] = \sum_{k=1}^n (x_k - \mu)^g P(x_k) \quad (3.6)$$

A variância é o segundo momento central. Ela é o primeiro indicador a ser utilizado neste sentido, pois o momento com $g = 1$ é zero. Deste modo, ela é definida como:

$$\mathbb{E}[(x - \mu)^2] = \sigma^2(x) = \sum_{k=1}^n (x_k - \mu)^2 P(x_k) \quad (3.7)$$

onde, desenvolvendo o somatório, é possível obter:

$$\sum_{x \in X} (x^2 - 2x\mu + \mu^2) P(x_k) = \mathbb{E}[X^2] - 2\mathbb{E}[x]\mu + \mu^2 = \mathbb{E}[X^2] - 2\mu^2 + \mu^2 = E[X^2] - \mu^2 \quad (3.8)$$

e o desvio-padrão é medido como a raiz quadrada deste valor (ou seja, apenas σ). Neste sentido, para o banco de dados:

$$\sigma = \sqrt{\sum_{k=1}^n P(x_k) (x_k^2 - \mu^2)} \quad (3.9)$$

3.4 Filtragem de Dados - Janelamento e Ponderação Exponencial

Em medições de fenômenos físicos, é comum que haja ruído e, usualmente, gaussiano (Bendat; Piersol, 2010). O ruído pode interferir nos cálculos dos algoritmos de identificação de anomalias, pois um pico ocasional pode ser detectado como uma anomalia, sendo que, em alguns casos, ele pode ser causado por fatores exógenos à bomba. Desta maneira, é preciso propor um filtro para o ruído, mas sem a perda da informação principal do sinal. Existem formas como, por exemplo, remover *outliers* e interpolá-los com a média. Entretanto, existe o risco do *outlier* poder ser a anomalia a ser detectada.

No presente contexto, os dados são discretos, e portanto, será escolhida a abordagem relativa a filtros digitais. Matematicamente, os filtros podem ser definidos como sendo uma equação de recorrência entre as saídas filtradas y e as variáveis originais x :

$$a_0 y_k + a_1 y_{k-1} + \dots + a_l y_{k-l} = b_0 x_k + b_1 x_{k-1} + \dots + b_p x_{k-p} \quad (3.10)$$

$$\sum_{i=0}^l a_i y_{k-i} = \sum_{i=0}^p b_i x_{k-i} \quad (3.11)$$

onde p e l são o número de entradas e de saídas consideradas além da entrada presente k . Os coeficientes a e b são associados às variáveis e às variáveis filtradas, respectivamente. Para facilitar a descrição destes filtros, a transformada \mathcal{Z} é utilizada.

A transformada \mathcal{Z} permite analisar a resposta em frequência para sistemas lineares invariantes no tempo discreto (Schlichthärle, 2011). Ela pode ser definida como:

$$X(z) = \sum_{i=0}^{\infty} x_i z^{-i} \quad (3.12)$$

onde ω é uma variável de frequência e para todo $i < 0$, $x_i = 0$ (sistema causal).

Uma propriedade importante é que a aplicação da transformada para um único ponto no tempo, pode gerar uma relação de deslocamento no tempo. Isto é possível de ser avaliado através da transformada para uma sequência X' atrasada em uma unidade temporal:

$$X'(z) = \sum_{i=0}^{\infty} x_{i-1} z^{-i} \quad (3.13)$$

se o índice $i - 1$ for trocado pelo índice v , temos que $i = v + 1$ e portanto:

$$X'(z) = \sum_{v=0}^{\infty} x_v z^{-(v+1)} = z^{-1} \sum_{v=0}^{\infty} x_v z^{-v} = z^{-1} X(z) \quad (3.14)$$

Com isto, avaliando cada elemento como uma série independente, é possível definir uma relação, para uma variável na série temporal, tal como:

$$x_k = x_{k-1} z \rightarrow x_{k-1} = x_k z^{-1} \quad (3.15)$$

a qual é possível analisar, recursivamente, no caso de um máximo de atrasos p :

$$\left\{ \begin{array}{l} x_{k-1} = x_k z^{-1} \\ x_{k-2} = x_{k-1} z^{-1} = x_k z^{-1} z^{-1} = x_k z^{-2} \\ \vdots \\ x_{k-(p-1)} = x_{k-(p-2)} z^{-1} = x_k z^{-1} \dots z^{-1} = x_k z^{-(p-1)} \end{array} \right. \quad (3.16)$$

Desta forma, os filtros podem ser escritos da seguinte forma:

$$\sum_{i=0}^l a_i y_{k-i} = \sum_{i=0}^p b_i x_{k-i} = \sum_{i=0}^l a_i y_k z^{-i} = \sum_{i=0}^p b_i x_k z^{-i} \quad (3.17)$$

onde, colocando x_k e y_k em evidência:

$$y_k \sum_{i=0}^l a_i z^{-i} = x_k \sum_{i=0}^p b_i z^{-i} \rightarrow H(z) = \frac{y_k}{x_k} = \frac{\sum_{i=0}^k b_i z^{-i}}{\sum_{j=0}^n a_j z^{-j}} \quad (3.18)$$

com $H(z)$ sendo a função de transferência do filtro, que é importante para estudarmos o filtro independentemente dos valores numéricos da entrada e da saída.

Para manipular os dados foi escolhida a biblioteca *pandas*. Ela é uma biblioteca de manipulação de dados, tendo utilização em mais de 2,7 milhões de projetos no GitHub (McKinney et al., 2024a). É um projeto *open source*, o que permite que a própria comunidade a mantenha atualizada.

Na biblioteca *pandas*, existem três métodos para trabalhar com séries temporais: janelamento, ponderação exponencial e expansão (McKinney et al., 2024b). No janelamento, a partir de um conjunto \mathbb{C} qualquer, um subconjunto $\mathbb{C}_k = \{c_{k-(p-1)}, c_{k-(p-2)}, \dots, c_k\}$ é escolhido com relação a uma entrada de índice k e suas p entradas anteriores. Desta forma, para cada entrada da série temporal, haverá um subconjunto relacionado, no qual as operações desejadas (como por exemplo, média e variância), são aplicadas. De acordo com Krollner, Vanstone e Finnie (2010), um dos filtros mais comuns que utiliza janelamento é a média móvel simples. Matematicamente ela é definida como:

$$M_p(k) = \frac{1}{p} (x_k + x_{k-1} + \dots + x_{k-(p-1)}) = \frac{1}{p} \sum_{i=k-(p-1)}^k x_i \quad (3.19)$$

onde p é controlado pelo usuário, considerado como a janela de análise, $l = 0$ (não possuir termos relativos a filtragens anteriores) e $b_i = \frac{1}{p}$, com relação à Equação 3.17. Considerando a Transformada \mathcal{Z} , a média móvel simples poderia ser representada como:

$$M_p(k) = \frac{1}{p} \sum_{i=k-(p-1)}^k x_i = \frac{1}{p} \sum_{i=0}^{p-1} x_k z^{-i} \quad (3.20)$$

Outro método proposto de filtragem é o de ponderação exponencial. Ao invés de tomar um subconjunto, os valores anteriores são preservados incluindo os resultados das filtragens anteriores nos resultados das filtragens atuais. Brown (1956) propõe a Média Móvel Ponderada Exponencial, ou, em inglês, *Exponential Weighted Moving Average* (EWM) como:

$$S_k = \alpha x_k + (1 - \alpha)S_{k-1} \quad (3.21)$$

onde S é o sinal filtrado e α é o peso dado a entrada atual x_k . Para uma determinada janela p , α pode ser definido como $\alpha = \frac{2}{p+1}$. Excepcionalmente, $S_0 = x_0$. Com relação à recorrência citada anteriormente, é possível avaliar que:

$$\left\{ \begin{array}{l} S_0 = x_0 \\ S_1 = \alpha x_1 + (1 - \alpha)S_0 \\ S_2 = \alpha x_2 + (1 - \alpha)S_1 = \alpha x_2 + \alpha(1 - \alpha)x_1 + (1 - \alpha)^2 x_0 \\ S_3 = \alpha x_3 + (1 - \alpha)S_2 = \alpha x_3 + \alpha(1 - \alpha)x_2 + \alpha(1 - \alpha)^2 x_1 + (1 - \alpha)^3 x_0 \\ \vdots \\ S_k = \alpha x_k + (1 - \alpha)S_{k-1} = \alpha \left(\sum_1^k (1 - \alpha)^{k-i} x_i \right) + (1 - \alpha)^k x_0 \end{array} \right. \quad (3.22)$$

sendo possível notar que, nesta média, o valor ponderado de todos os valores da série está “acumulado” no termo S_{k-1} . Em Schlichthärle (2011), é possível tornar análogo esta média como um filtro passa-baixa de primeira ordem. Neste sentido, seria assumido que o ruído está em alta frequência e o filtro conseguiria mitigar o efeito deste. Com isto, o filtro seria matematicamente descrito como:

$$S_k = \alpha x_k + (1 - \alpha)S_{k-1}z^{-1} \rightarrow S_k(1 - (1 - \alpha)z^{-1}) = \alpha x_k \rightarrow \frac{S_k}{x_k} = \frac{\alpha}{(1 - (1 - \alpha)z^{-1})} \quad (3.23)$$

Para analisar o comportamento dos filtros, é possível utilizar o diagrama de Bode, que mostra como ele se comporta ao longo de diferentes frequências dos sinais. O diagrama de

Bode mostra, em seu gráfico superior, a magnitude da resposta do sistema, e no gráfico inferior, a fase da resposta para uma determinada frequência na abscissa.

Na Figura 3.1 e na Figura 3.2 apresenta-se o gráfico para o filtro de média móvel e média móvel exponencial, respectivamente. Para ambos os filtros foi definida a utilização de um período de 7 intervalos de tempo e um intervalo de frequência entre $0 \text{ rad}/t$ a $\pi \text{ rad}/t$, onde t denota unidade de tempo. Como intervalo de amostragem do exemplo, foi escolhido 1 unidade de tempo, de tal maneira que a frequência de Nyquist seja meia unidade de tempo, e por consequência, em radianos por unidade de tempo, seja $\pi \text{ rad}/t$. A escolha de uma frequência maior que a de Nyquist exibiria um espelhamento do conteúdo em frequência do filtro.

Para a Figura 3.1, nota-se que a magnitude possui diversos lóbulos, ou seja, intermitências na magnitude ao longo das frequências, que fazem com que, cicличamente, algumas frequências sejam mais atenuadas em detrimento de outras.

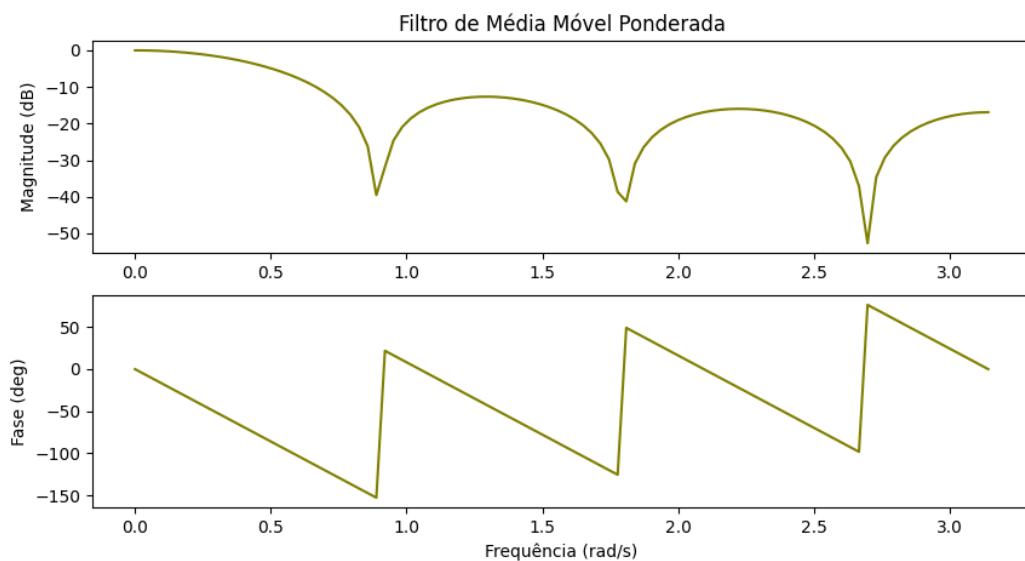


Figura 3.1: Diagrama de Bode do filtro de médias móveis.

Já na Figura 3.2, observa-se que a magnitude da resposta do filtro realmente é mais suave e atenua as frequências mais baixas. Desta forma, dos métodos disponíveis, a EWM acaba sendo mais adequada, pois, em comparação com o filtro anterior, para os mesmos parâmetros, a resposta não possui lóbulos que priorizariam frequências em detrimento de outras.

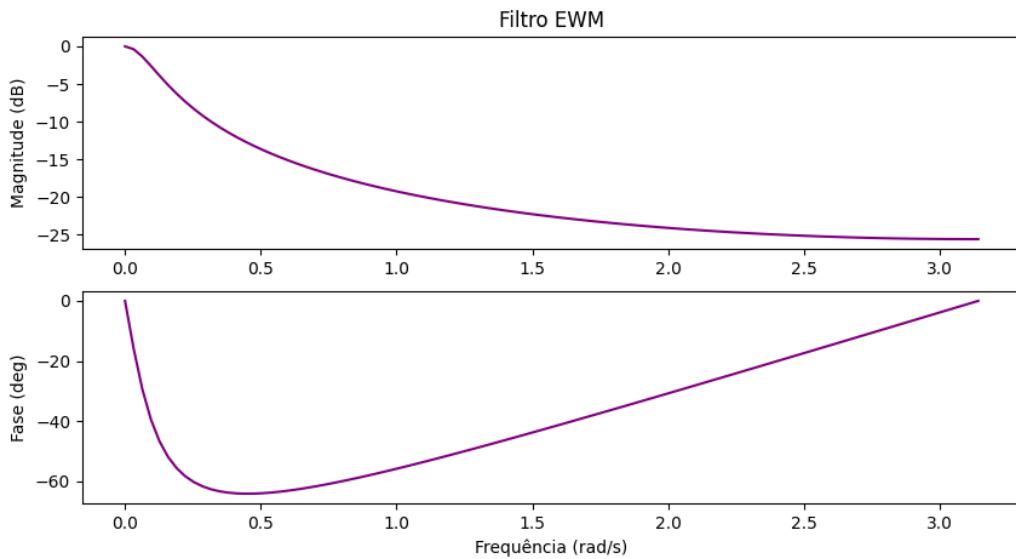


Figura 3.2: Diagrama de Bode do filtro EWM.

3.5 Cálculos por Expansão

No contexto desta dissertação, as séries temporais possuem valores que foram sendo inseridos ao longo do tempo, pois são medições de sensores. Neste sentido, mesmo que seja possível a previsão, os valores futuros das medições são desconhecidos. Portanto, é necessário desenvolver um método de escala que utilize as informações coletadas em cada inserção, de forma contínua. Nesta seção, será apresentada uma metodologia de cálculo de propriedades estatísticas considerando a inserção contínua, que será tratada como “expansão”.

A primeira análise a ser feita é avaliar um caso onde o banco de dados possua distribuições heterogêneas das variáveis e dimensões distintas. Seria conveniente trazê-las todas a uma mesma base de comparação para poder permitir uma análise. Uma das normalizações mais utilizada é a mínimo-máximo ($MM(x)$). Ela é definida como:

$$MM(x) = \frac{x - x_{min}}{x_{max} - x_{min}} \quad (3.24)$$

de tal maneira que essa razão percentual é utilizada para colocar todos os dados em uma mesma base de comparação. Mesmo que a normalização (min-max) atinja este objetivo, ela é sensível a novos dados. Uma inclusão de um dado maior que o máximo, por exemplo, poderia fazer com que toda a distribuição tivesse que ser reescalada.

Neste sentido, o *Z-score*, equação Equação 3.3, permitiria representá-los em uma medida adimensional, apenas referente ao tamanho do seu desvio, o que seria mais robusto em termos de expansão. Entretanto, é preciso adaptar o cálculo do *Z-score*, para o contexto desta dissertação e o contexto de expansão.

O *Z-score* original utiliza uma média e desvio-padrão fixos, tomados como sendo as propriedades calculadas sobre todos os valores das variáveis no banco de dados. Caso no conjunto de dados existam dois equipamentos com médias diferentes em suas operações, a média obtida do conjunto inteiro não poderá representar bem o comportamento, nem de um, nem de outro. Desta maneira, utilizar uma média geral do banco de dados não é adequado, pois pode gerar equívocos na medição da entrada, e consequentemente, se isso é uma anomalia ou não. O ideal seria utilizar a média do processo de cada corrida, pois mesmo entre modelos, devido às circunstâncias, podem haver variações. Desta forma, no contexto, as propriedades que estão sendo calculadas, são aplicadas apenas sobre as informações relativas a uma corrida singular.

Com isto, propõe-se a média por expansão de uma determinada variável a ser calculada continuamente como:

$$\mu_k = \frac{1}{k} x_k + \left(1 - \frac{1}{k}\right) \mu_{k-1} \quad (3.25)$$

Entretanto, como mostrado na Equação 3.25, um valor grande e espúrio de x_k poderia impactar e distorcer a média. Desta forma, Morettin e Bussab (2017) sugere que, ao invés da média, neste caso, seja usada a mediana. A mediana é menos sensível a *outliers* e, portanto, torna-se uma referência do centro das variáveis (tal como a média é no *Z-score*) mais robusta contra os *outliers*.

Para calcular a mediana, pode-se considerar $\mathbb{C}_{0:k}$ como o conjunto que inclui todas as entradas de uma variável da corrida, do começo da série até a k -ésima entrada. Posteriormente, cria-se um novo conjunto \mathbb{O} (o rol), que é composto pelos valores de $\mathbb{C}_{0:k}$, ordenados em ordem crescente ($c_{i-1} < c_i$). Desta forma, a mediana da série é definida como:

$$M_k = \begin{cases} o_{\frac{k+1}{2}} & k \text{ é ímpar} \\ \frac{o_{\frac{k}{2}} + o_{\frac{k}{2}+1}}{2} & k \text{ é par} \end{cases} \quad (3.26)$$

onde o é um elemento de \mathbb{O} . Neste sentido, para uma entrada no instante k , sua mediana é definida como M_k . É possível notar na Equação 3.26, que independente do valor de entrada x_k , o valor da mediana está associado a quantidade de elementos no rol e o ordenamento do conjunto \mathbb{C} .

Para calcular o desvio-padrão, considerando novas entradas, é possível desenvolver a equação do segundo momento, que torna-se:

$$\sigma_k = \left(\frac{1}{k} x_k^2 + \left(1 - \frac{1}{k} \right) (\sigma_{k-1}^2 + \mu_{k-1}^2) - \mu_k^2 \right)^{\frac{1}{2}} \quad (3.27)$$

ao qual, no Apêndice A.1 há um detalhamento maior sobre a obtenção destas equações.

Existem algumas proposições mais robustas a *outliers* para avaliar a variação de um conjunto de pontos, como o Desvio Absoluto (em inglês: *Mean Absolute Deviation* (MAD)). O MAD busca avaliar a média dos desvios com relação à mediana. O intuito é que ele seja uma medida mais robusta à anomalias, com relação à dispersão. Para uma referência em k , o MAD é proposto por Kader (1999) como:

$$MAD(x_k) = \frac{1}{k} \sum_{i=1}^k |x_k - M_k| \quad (3.28)$$

Entretanto, como será visto no Capítulo 4, o MAD apresenta muitos picos quando calculado por expansão.

3.6 Z-score modificado

Com as propriedades calculadas por expansão e as variáveis filtradas, é possível desenvolver um *Z-score* modificado que é uma escala calculada continuamente com a inserção de novas informações na série. Com isto, considerando a Equação 3.3, e combinando as Equações 3.21, 3.26 e 3.27 é possível reescrever o *Z-score* como uma equação dependente do índice de uma nova entrada k , gerando um Z_k :

$$Z_k = \frac{S_k - M_k}{\sigma_k} \quad (3.29)$$

onde S_k é a entrada x_k filtrada, M_k a mediana e σ_k o desvio-padrão calculado recursivamente. Considerando que as bombas, usualmente, possuem valor 0 para algumas propriedades,

quando desligadas, estes valores da mediana e do desvio devem ser calculados separando os dados da bomba desligada e ligada.

Para dados faltantes, é possível assumir o valor de Z_k como 0, uma vez que isto seria equivalente a considerar o valor como sendo a mediana ($Z_k = 0 \rightarrow x_k = M_k$).

3.7 Exemplo de uso do Z-score modificado

Com o objetivo de ilustrar o procedimento do *Z-score* ilustrado, um exemplo é fornecido nesta seção. Dois sinais fictícios são criados. Um que possui um decaimento exponencial e outro que é constante. Em ambos, é aplicado um degrau, durante um trecho, para simular uma anomalia. Também é simulado um ruído gaussiano sobre os sinais. Para tal, são considerados 1000 pontos em um intervalo de 0 até o tempo final $T = 10$, medidas em unidades de tempo. Desta maneira, é possível definir o sinal $x_0(k)$ como:

$$x_0(k) = 100(e^{-0,03k} + 0,03r_0(k) - \frac{1}{10}(u(k - \frac{3}{10}T) - u(k - \frac{6}{10}T))) \quad (3.30)$$

e o sinal $x_1(k)$ como:

$$x_1(k) = 20(0,03r_1(k) + \frac{1}{5}u(k - \frac{3}{4}T)) \quad (3.31)$$

onde r_i é o ruído gaussiano e u é a função degrau unitário. Em ambos, é importante notar que suas dimensões são bem distintas. No sinal $x_0(t)$ existe um fator multiplicativo de 100 e no sinal $x_1(t)$ existe um fator multiplicativo de 20.

Ao implementar estes sinais, é possível obter a Figura 3.3. Nota-se que as escalas são bem distintas, e praticamente, não é possível comparar qual anomalia é maior ou menor em cada sinal. Para contornar esta situação, o *Z-score* modificado é aplicado, de tal maneira a colocar ambos os sinais em uma mesma escala de comparação.

Na Figura 3.4, é mostrada a padronização feita nestes sinais. É possível notar, que agora, a anomalia no sinal $x_0(t)$, proporcionalmente, é muito maior do que a anomalia no sinal $x_1(t)$.

3.8 Análise de Componentes Principais

A Análise de Componentes Principais, (em inglês, *Principal Component Analysis (PCA)*) é um método de análise da variância, baseado na ideia de decomposição matricial que advém das

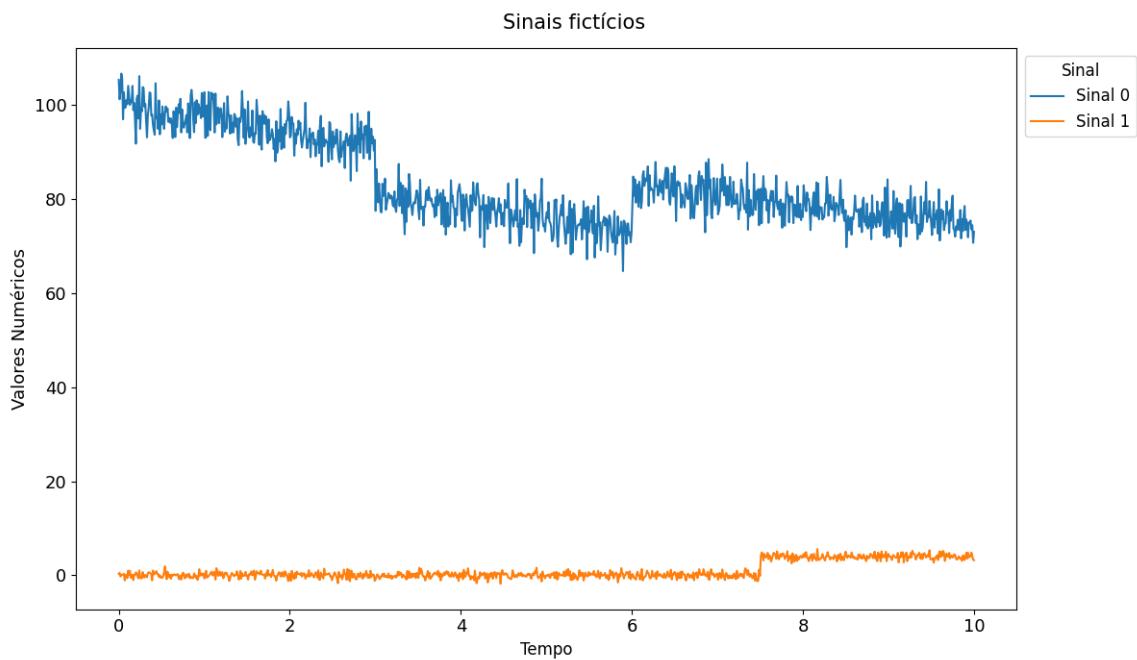
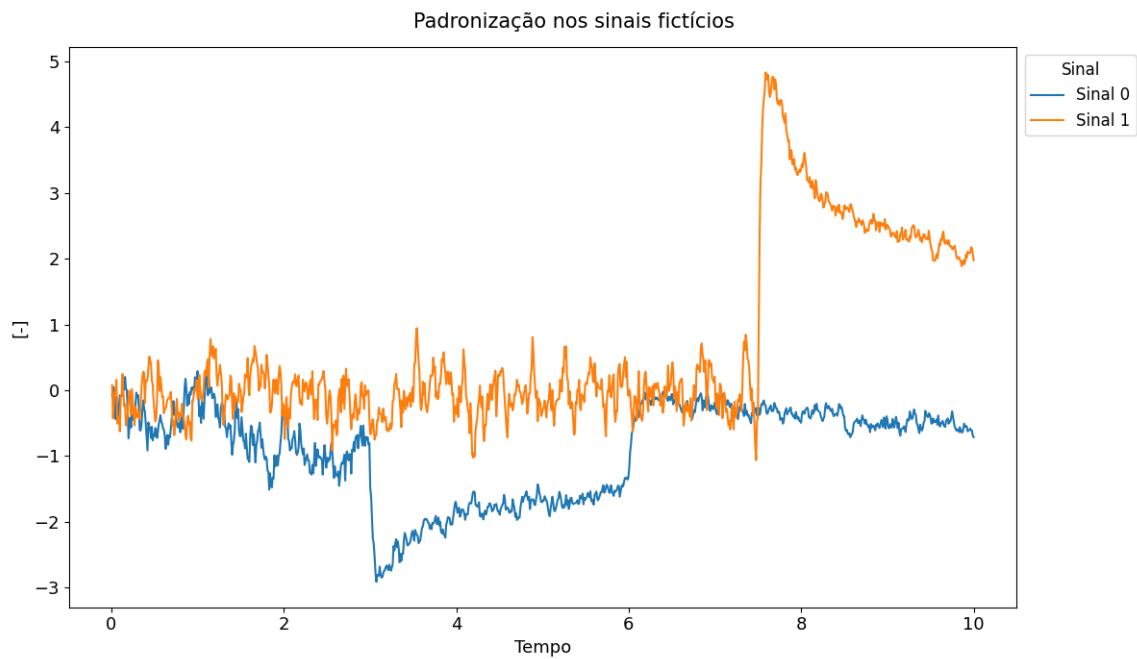


Figura 3.3: Sinais fictícios 1 e 2.

Figura 3.4: *Z-score* modificado aplicado aos sinais fictícios.

teorias de Álgebra Linear. A ideia principal da PCA é encontrar os eixos de maior variância, que podem ser combinações lineares das variáveis do banco de dados. Desta maneira, é possível descrever o conjunto de dados apenas considerando estes eixos, sem perder parte relevante da informação. No contexto desta tese, a aplicação da PCA é importante, pois os eixos principais revelam informações sobre as relações entre as variáveis numéricas do banco de dados.

A princípio, para encontrar estes eixos, é proposta a decomposição ortogonal da matriz de covariância das variáveis. Os autovalores desta matriz, somados, representam a variância inteira do banco de dados e os autovetores correspondentes representam pesos de transformações lineares que permitem que as componentes ortogonais sejam obtidas. Para tal, é possível utilizar a parte numérica do banco de dados, que é definida como \mathbf{X} :

$$\mathbf{X} = \begin{pmatrix} X_{11} & X_{12} & \dots & X_{1m} \\ X_{21} & X_{22} & \dots & X_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ X_{k1} & X_{k2} & \dots & X_{km} \\ \vdots & \vdots & \ddots & \vdots \\ X_{n1} & X_{n2} & \dots & X_{nm} \end{pmatrix} \quad (3.32)$$

onde \mathbf{X} é definido como a matriz com as séries temporais $X_{:,j}$ de cada j das m variáveis do banco de dados (de todas as corridas), notando-se que as colunas variam em torno do índice k associado ao tempo. Se esta matriz for centralizada em torno da média de cada coluna, seria possível obter uma matriz \mathbf{R} tal como:

$$\mathbf{R} = \begin{pmatrix} X_{11} - \mu_1 & X_{12} - \mu_2 & \dots & X_{1m} - \mu_m \\ X_{21} - \mu_1 & X_{22} - \mu_2 & \dots & X_{2m} - \mu_m \\ \vdots & \vdots & \ddots & \vdots \\ X_{n1} - \mu_1 & X_{n2} - \mu_2 & \dots & X_{nm} - \mu_m \end{pmatrix} \quad (3.33)$$

em que μ_i é a média de cada coluna (ou, de forma geral, o centro das variáveis). Entretanto, nesta tese, a mediana é utilizada e é calculada por expansão. Além disso, as entradas são filtradas devido ao ruído, levando a reinterpretar a matriz \mathbf{R} como sendo:

$$\mathbf{R} = \begin{pmatrix} S_{11} - M_{11} & S_{12} - M_{12} & \dots & S_{1m} - M_{1m} \\ S_{21} - M_{21} & S_{22} - M_{22} & \dots & S_{2m} - M_{1m} \\ \vdots & \vdots & \ddots & \vdots \\ S_{n1} - M_{n1} & S_{n2} - M_{n2} & \dots & S_{nm} - M_{1m} \end{pmatrix} \quad (3.34)$$

onde $M_{i,j}$ é relativo a mediana e $S_{i,j}$ a uma entrada filtrada. Com isto, a matriz de covariância (modificada) pode ser definida como:

$$\mathbf{K} = \mathbb{E} [\mathbf{R}\mathbf{R}^T] = \frac{1}{n} \mathbf{R}\mathbf{R}^T \quad (3.35)$$

Outra matriz estatística possível de ser avaliada pelo método dos autovetores e autovalores é a matriz de correlação. A matriz de correlação é obtida a partir da matriz de variáveis padronizadas \mathbf{Z} :

$$\mathbf{Z} = \begin{pmatrix} Z_{11} & Z_{12} & \dots & Z_{1m} \\ Z_{21} & Z_{22} & \dots & Z_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ Z_{k1} & Z_{k2} & \dots & Z_{km} \\ \vdots & \vdots & \ddots & \vdots \\ Z_{n1} & Z_{n2} & \dots & Z_{nm} \end{pmatrix} \quad (3.36)$$

onde Z_{kj} é uma entrada no tempo, tal que j é o índice horizontal da variável medida, n é o número de entradas e m o número de variáveis. Neste trabalho, a padronização proposta é a da Seção 3.6. Para tal, define-se a matriz de correlação como:

$$\mathbf{C} = \mathbb{E}[\mathbf{Z}^T \mathbf{Z}] = \frac{1}{n} \mathbf{Z}^T \mathbf{Z} \quad (3.37)$$

notando o leitor que esta não é a matriz de correlação convencional, uma vez que a padronização proposta foi a utilizada na Seção 3.6.

Com relação a escolha da matriz a ser utilizada no método, Jolliffe (2006) diz que se existe uma diferença muito grande de escala entre os dados, a matriz de correlação deve ser priorizada. Neste trabalho, o resultado de ambas as matrizes serão explorados, entretanto, será preferido a matriz de correlação, em virtude de que os dados possuem escalas distintas.

Como dito, o objetivo da PCA é encontrar os eixos principais dos dados através de uma decomposição ortogonal da matriz de covariância ou correlação das variáveis. É possível provar que as matrizes podem ser decompostas pelo fato de que são semi-positivas definidas. Tomando \mathbf{C} como exemplo, para um vetor $\mathbf{x} \in \mathbb{R}^m$ qualquer, procede que:

$$\mathbf{x}^T \mathbf{C} \mathbf{x} = \mathbf{x}^T (\mathbf{Z}^T \mathbf{Z}) \mathbf{x} = (\mathbf{Z} \mathbf{x})^T \mathbf{Z} \mathbf{x} = \|\mathbf{Z} \mathbf{x}\|_2 \geq 0 \quad (3.38)$$

implicando que a matriz de correlação pode ser decomposta em autovalores e autovetores, devido ao Teorema Espectral (Axler, 2024):

$$\mathbf{C} = \frac{1}{n} \mathbf{Z}^T \mathbf{Z} = \mathbf{V} \Lambda \mathbf{V}^T \quad (3.39)$$

Na Equação 3.39, Λ representa a matriz dos autovalores e \mathbf{V} representa a matriz de autovetores. Esta matriz de autovetores é importante, pois destaca como formar os eixos principais, tomando as variáveis originais como base. A partir da matriz dos autovalores, é possível avaliar como obter as componentes no espaço dos eixos principais. Para tal, reescreve-se a Equação 3.39 para isolar a matriz de autovalores:

$$\Lambda = \frac{1}{n} \mathbf{V}^T \mathbf{Z}^T \mathbf{Z} \mathbf{V} = \frac{1}{n} (\mathbf{Z} \mathbf{V})^T \mathbf{Z} \mathbf{V} = \frac{1}{n} \mathbf{Y}^T \mathbf{Y} \quad (3.40)$$

onde se nota uma relação $\mathbf{Y} = \mathbf{Z} \mathbf{V}$, que é o conjunto de dados do espaço ortogonal da variância maximizada. Esta matriz \mathbf{Y} também é um conjunto de séries temporais. Por padrão, na biblioteca *sklearn*, na matriz \mathbf{V} , a ordem das colunas é do autovetor de maior autovalor para o de menor autovalor associado.

Na PCA é possível avaliar que existem eixos com maiores autovalores (que estão associados à variância) do que outros. Esta propriedade permite representar o sistema em menos dimensões, porém, ainda sim mantendo grande parte da informação original. Isto é feito considerando quais são os autovalores que mais representam a variância, através de uma regra heurística, apresentada em Gerón (2019), que afirma que 95% do valor dos autovalores (ordenados pelo maior) pode ser utilizada como referência para redução do banco de dados, sem perda significativa de informação.

Para reduzir \mathbf{V} , multiplica-se por uma matriz identidade truncada \mathbf{I}_p de dimensão $m \times p$, onde p é o número de colunas que corresponde ao número de autovalores a serem somados para representar 95% de seu total. A percentagem P , referente à propriedade explicada, isto é, o quanto da propriedade está contido naquelas m componentes, é obtida encontrando o maior inteiro que satisfaz a relação:

$$P = \frac{\sum_{i=1}^p \lambda_i}{\sum_{i=1}^m \lambda_i} \quad (3.41)$$

onde λ_i é o autovalor da matriz de covariância, e portanto, a variância associada ao espaço dos autovetores. Desta maneira, é possível encontrar os eixos principais dos dados, que evidenciam as relações entre os dados e avaliar quais são mais importantes.

3.9 Análise de Componentes Independentes

A Análise de Componentes Independentes (em inglês, *Independent Component Analysis* (ICA)) é um método de decomposição dos sinais, tal como a PCA, entretanto, com o objetivo de separar fontes independentes. Em outras palavras, a técnica busca separar sinais de fontes individuais a partir de um conjunto de sinais misturados. Nesta seção, serão apresentados os fundamentos da formulação do algoritmo *FastICA* (Hyvärinen, 1999). Hyvärinen, Karhunen e Oja (2001) afirmam que não há um criador, em específico, da ICA, e sim, diversos autores que concomitantemente desenvolveram estas técnicas.

No presente trabalho, a ICA pode ser enxergada como um terceiro passo para a PCA. Se na PCA é encontrado uma nova base ortogonal V para o banco de dados, com relação a sua maior variância, a ICA aproveita-se desta base ortogonalizada para encontrar uma direção que isola uma fonte estatística independente. Neste sentido, o modelo que a ICA considera é que várias fontes misturadas geram os sinais ortogonalizados observados. Para tal, o modelo matemático desta mistura é descrito como:

$$S\mathbf{o} = \mathbf{y} \rightarrow \mathbf{o} = W\mathbf{y} \quad (3.42)$$

onde \mathbf{o} seriam as fontes ocultas independentes, não observáveis diretamente, e $\mathbf{y} = V^T z$ seriam as fontes misturadas (já ortogonalizadas, sendo vetores de Y), que são os dados observados. $S^{-1} = W$ é a matriz de pesos que combina as fontes misturadas para gerar as fontes originais. Deste modo, o objetivo é encontrar a matriz W que separa os sinais.

Para tal, a literatura apresenta três modos de decomposição: maximização da não-gaussianidade (Hyvärinen, 1999), maximização da probabilidade posterior (Amari; Cichocki; Yang, H., 1995) e minimização da informação mutual (Comon, 1994). Nesta dissertação, será apresentada a metodologia pela maximização da não-gaussianidade.

Para tal, é necessário apresentar alguns conceitos estatísticos, como o conceito de informação, entropia e sintropia. A informação é uma medida proposta por Shannon (1948) para quantificar a incerteza com relação a uma determinada distribuição. A incerteza é dada

como o grau de imprevisibilidade de uma variável aleatória. Neste sentido, a informação é definida por:

$$I(x_k) = -\log(P(x_k)) \quad (3.43)$$

onde $P(x_k)$ é a probabilidade da entrada x_k ocorrer com relação a uma variável aleatória X . Nota-se que quanto menor a probabilidade, maior será a informação, pois o evento será mais “inesperado”. Posteriormente, é possível determinar a entropia como sendo a esperança da informação:

$$H(X) = \mathbb{E}[I(X)] = \sum_{x \in \mathcal{X}} P(x)I(x) = -\sum_{x \in \mathcal{X}} P(x)\log P(x) \quad (3.44)$$

Deste modo, a entropia é uma medida do grau de imprevisibilidade esperado da variável aleatória. Para encontrar as componentes da matriz W através da maximização da não-gaussianidade, Hyvärinen, Karhunen e Oja (2001) propõem a sintropia. Se a entropia mensura indiretamente a incerteza esperada, a sintropia mede seu contrário, ou seja, o quanto de “certeza” existe na variável aleatória. No caso, a referência da sintropia é uma variável aleatória gaussiana A , tal que $A \sim \mathcal{N}(\mu, \sigma^2)$, pois como Cover e Thomas (2006) afirmam, a distribuição gaussiana é a distribuição estatística que maximiza a entropia. Desta maneira, ao maximizar a sintropia, maximiza-se a não-gaussianidade. Desta maneira, a entropia da variável aleatória A é:

$$H(A) = \frac{1}{2}\log(2\pi e\sigma^2) \quad (3.45)$$

e a sintropia $J(X)$ é definida como:

$$J(X) = H(A) - H(X) = \frac{1}{2}\log(2\pi e\sigma^2) - H(X) \quad (3.46)$$

onde e é o número de Euler. Portanto, se $X = A$, a sintropia é zero. Portanto, nota-se que esta medida é sempre positiva. Considerando que a distribuição do sinal pode ser desconhecida, para aproximar a sintropia, uma estimativa é proposta como:

$$J(X) \approx \frac{1}{12}\mathbb{E}[X^3]^2 + \frac{1}{48}kurt(X)^2 \quad (3.47)$$

onde a curtose $kurt(X)$ é definida como:

$$kurt(X) = \mathbb{E}[X^4] - 3(\mathbb{E}[X^2])^2 \quad (3.48)$$

Como o cálculo da curtose pode ser computacionalmente caro, uma vez que envolve um polinômio de ordem 4 (devido a presença do quarto momento), uma aproximação é proposta, em Hyvärinen, Karhunen e Oja (2001), com funções não-quadráticas G^i :

$$J(X) \approx k_1 (\mathbb{E}[G^1(X)])^2 + k_2 (\mathbb{E}[G^2(X)] - \mathbb{E}[G^1(v)])^2 \quad (3.49)$$

onde G^1 é uma função ímpar, G^2 é uma função par e v é uma variável normalizada de média 0 e desvio-padrão 1 e k_i são pesos. Se apenas uma função G for utilizada, a sintropia pode ser analisada como:

$$J(X) \propto [\mathbb{E}[G(X)] - \mathbb{E}[G(v)]]^2 \quad (3.50)$$

onde \propto significa proporcional.

Assumindo que a variável aleatória em questão seja relacionada ao conjunto de dados padronizado y , busca-se uma combinação de pesos w , que compõem a matriz W . Para restringir as soluções possíveis, restringe-se $|w|^2 = 1$, de tal maneira a encontrar apenas um vetor unitário que indique a direção, uma vez que, todos os vetores nesta direção também poderiam ser soluções possíveis. Deste modo, obtém-se uma relação de proporcionalidade para o gradiente com relação aos pesos:

$$\nabla_w J(w) \propto [(\mathbb{E}[G(w^T y)] - \mathbb{E}[G(v)]) (y \mathbb{E}[g(w^T y)])]^2 \quad (3.51)$$

onde, G , por padrão, na implementação da biblioteca *sklearn* (Pedregosa et al., 2011) é $G(x) = \log \cosh x$, cuja derivada é $g(x) = \tanh(x)$. Para resolver este problema, Hyvarinen (1999) propõe uma simplificação de ponto fixo, baseada em uma otimização lagrangiana pelo Método de Newton para encontrar os vetores coluna que compõem a matriz W :

$$w \leftarrow \mathbb{E} [z g(w^T y) - \mathbb{E} [g'(w^T y) w]] \quad (3.52)$$

e para encontrar múltiplas componentes, é proposta a ortogonalização de Gram-Schmidt. Desta maneira, é proposto o algoritmo como:

Algoritmo 1: FastICA

Entrada: Matriz de dados $X \in \mathbb{R}^{n \times m}$ com n amostras e m variáveis

Saída: Componentes independentes $W \in \mathbb{R}^{m \times m}$

1. **Aplicar a PCA com relação à correlação:** Para que os dados sejam analisados na componente de maior variância, propõe-se aplicar a PCA sobre X , de tal maneira que seja obtida Y e V ;
 2. **Para cada componente** $i = 1, \dots, m$
 - 3 Inicializar vetor w_i aleatoriamente com norma unitária;
 - 4 **Enquanto** não convergir
 - 5 $w_i \leftarrow \frac{1}{n} \sum_{j=1}^n y^{(j)} g(w_i^\top y^{(j)}) - \frac{1}{n} \sum_{j=1}^n g'(w_i^\top y^{(j)}) w_i$;
 - 6 Ortogonalizar w_i em relação a w_1, \dots, w_{i-1} via Gram-Schmidt;
 - 7 Normalizar: $w_i \leftarrow w_i / \|w_i\|$;
 - 8 **return** VW^\top ;
-

A matriz VW^\top é calculada para obter uma matriz de separação que associa diretamente as componentes independentes com as variáveis não-ortogonais dos sensores.

3.10 Exemplos com a PCA e ICA

Nesta seção serão apresentados exemplos de utilização da PCA e da ICA para identificação de relações, tentando evidenciar, a relação que cada algoritmo mostra para cada situação.

Neste primeiro exemplo, o objetivo é avaliar o comportamento de ambos os métodos em uma situação onde os sinais são dependentes, ou seja, um possui relação direta com o outro. Para tal, é proposto uma função vetorial de sinais fictícios, com duração de 0 à $T = 5$ unidades de tempo, definidos como:

$$\mathbf{y}(t) = \begin{pmatrix} \text{Original 1} \\ \text{Original 2} \\ \text{Original 3} \end{pmatrix} = \begin{pmatrix} \sin(40t)w(t) \\ \sin(5t) \\ 2w(t) \end{pmatrix} \quad (3.53)$$

onde $w(t)$ é um envelope gaussiano dos sinais definido como $w(t) = \exp\left(-\frac{10}{3}(t - \frac{T}{2})^2\right)$. Nota-se que o sinal Original 0 e Original 2 possuem o envelope como fator multiplicativo em comum. Posteriormente, define-se uma matriz de combinação Q para gerar sinais misturados. Q é definida de maneira a ser ortogonal:

$$Q = \begin{bmatrix} -0.99124783 & 0.13069384 & 0.01862429 \\ -0.10168324 & -0.66589517 & -0.73908331 \\ -0.08419181 & -0.7345085 & 0.67335652 \end{bmatrix} \quad (3.54)$$

e com isto, os sinais a serem observados após a recombinação são definidos por $\mathbf{x}(t)$, onde:

$$\mathbf{x}(t) = Q\mathbf{y}(t) + \mathbf{r}(t) \quad (3.55)$$

junto de um vetor $\mathbf{r}(t)$ de ruído gaussiano, simulando a presença de ruído na medição.

Com isto, é possível aplicar os métodos de análise no sinal \mathbf{x} e avaliar se é possível obter os sinais originais. Para atenuar o efeito do ruído e tentar melhorar a identificação, os sinais associados a $\mathbf{x}(t)$ são filtrados por um filtro passa-baixa. Desta maneira, ao aplicar os métodos, é possível obter seus resultados de avaliação. Na Figura 3.5, são ilustrados os sinais originais, a recombinação obtida, os sinais recuperados pela PCA e os recuperados pela ICA.

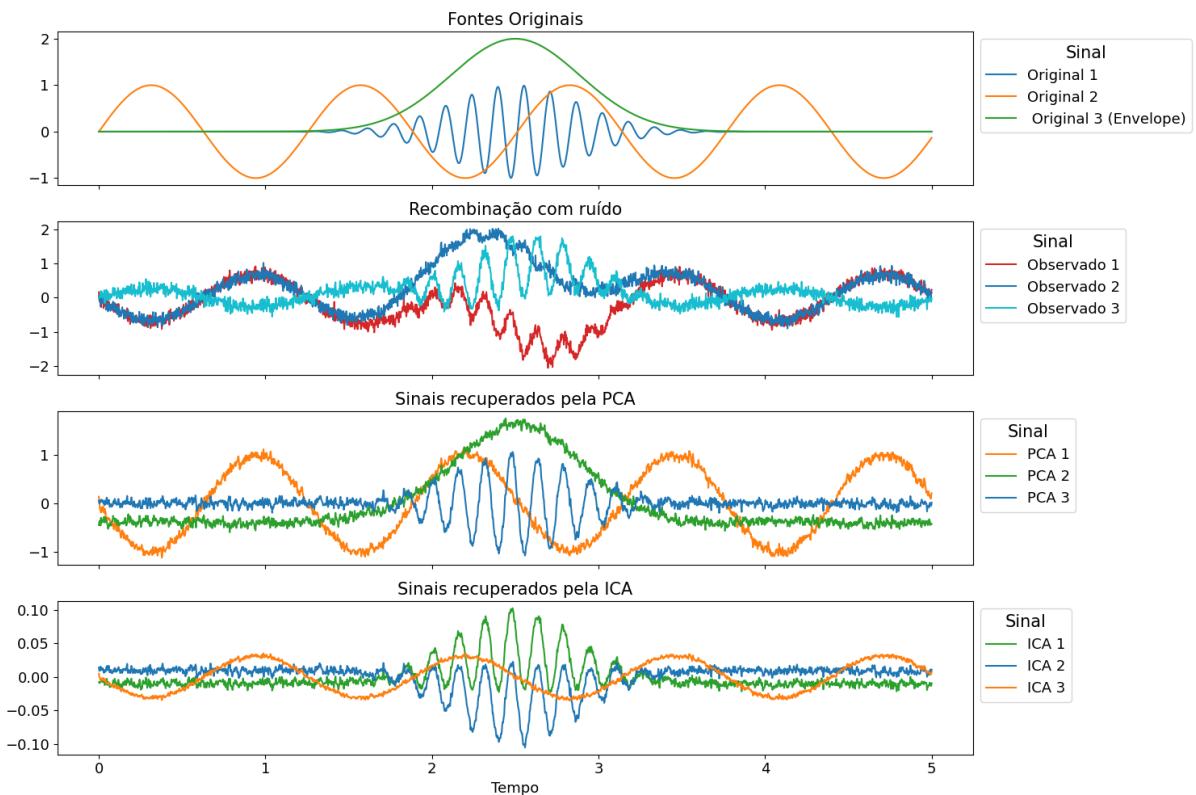


Figura 3.5: Primeiro exemplo com PCA e ICA

É possível notar que os sinais recuperados pela PCA (PCA 1, PCA 2, PCA 3) são consideravelmente fiéis ao formato original dos sinais originais. O sinal PCA 1 tem grande

similaridade com o Original 2, o PCA 2 com o Original 3 e o PCA 3 com o original 1, porém todos com inversão de fase.

Por outro lado, ao aplicar a ICA, é possível notar que, para os sinais originais interdependentes, as componentes recuperadas não correspondem ao formato original. A componente ICA 1 é a única cujo formato foi recuperado de forma mais adequada, pois não possui interdependência com os outros sinais originais. Desta maneira, os sinais ICA 2 e ICA 3 ainda são combinações entre o original 1 e o original 3.

Com isto, este exemplo mostra que se os sinais, no banco de dados, possuírem alguma dependência entre si, a ICA não conseguirá identificar bem o sinal original.

Em outro exemplo, adaptado de Pedregosa et al. (2011), é possível mostrar uma situação contrária, onde a ICA tem sucesso em identificar os sinais originais e a PCA não. Neste exemplo, um sinal senoidal, outro quadrado e outro dente de serra são recombinados. Este sinais originais podem ser definidos como:

$$\mathbf{o}(t) = \begin{pmatrix} \text{Original 0} \\ \text{Original 1} \\ \text{Original 2} \end{pmatrix} = \begin{pmatrix} \sin(2t) \\ \text{sign}(\sin(5t)) \\ \text{serra}(2\pi t) \end{pmatrix} \quad (3.56)$$

onde sign é a função sinal, que avalia se o termo é positivo ou negativo, retornando 1, 0 ou -1 . Para obter os sinais observados, é definida uma matriz S , tal como consta na Equação 3.42:

$$S = \begin{bmatrix} 1 & 1 & 1 \\ 0.5 & 2 & 1.0 \\ 1.5 & 1.0 & 2.0 \end{bmatrix} \quad (3.57)$$

Aplicando os métodos neste exemplo, é possível obter a Figura 3.6. Na identificação dos sinais originais pela PCA, nota-se que todas as componentes obtidas não correspondem com os sinais originais, de tal maneira que é até difícil colocar uma comparação entre o sinal original e o recuperado. Entretanto, para a ICA, é possível notar que os sinais foram recuperados. Nota-se que o sinal ICA 1 corresponde ao original 2, o ICA 2 ao original 3 e o ICA 3 ao original 2. Entretanto, tal como no exemplo anterior, os sinais identificados possuem uma inversão de fase.

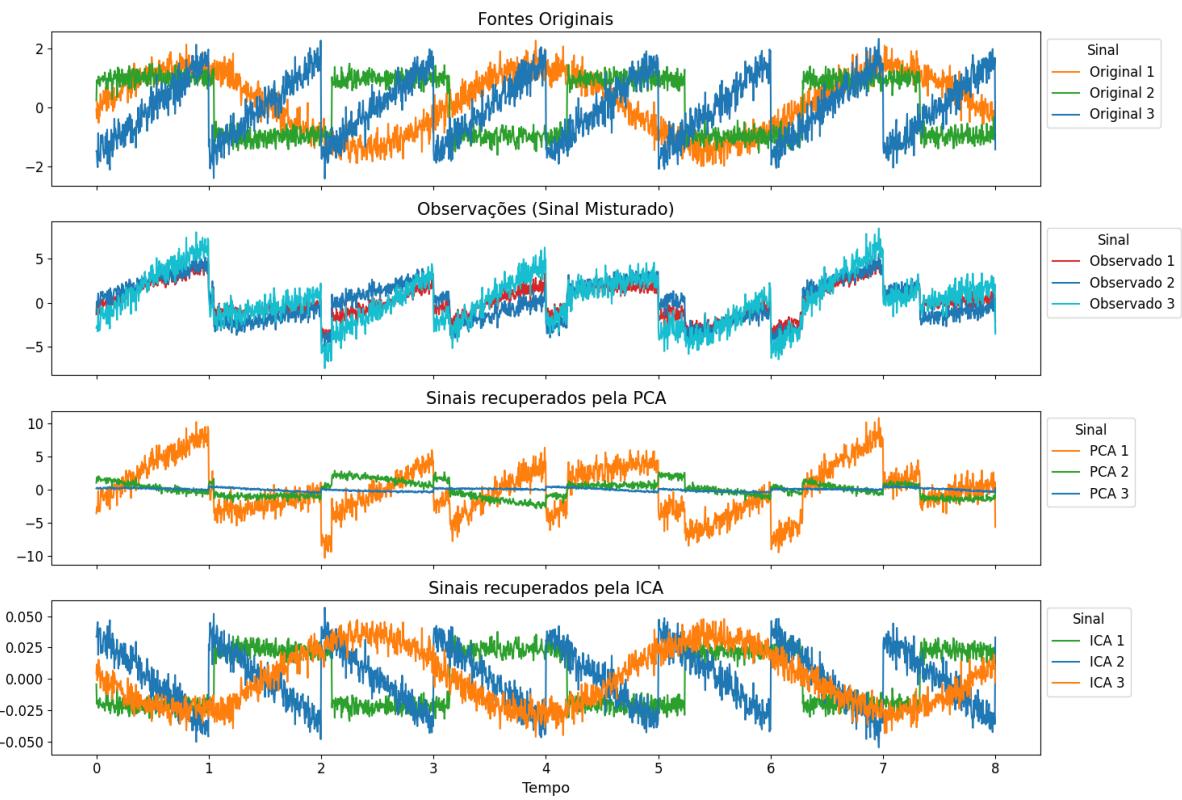


Figura 3.6: Segundo exemplo com PCA e ICA

Nota-se que, diferentemente do exemplo anterior, a PCA não conseguiu recuperar o sinal original, mas a ICA teve sucesso nesta tarefa. Isto é devido ao fato de que os sinais são bem independentes entre si e não são gaussianos, pressuposição que se faz sobre os sinais a serem identificados. Desta maneira conclui-se que, como nesta dissertação a natureza dos possíveis sinais “originais” não é conhecida, ambos os métodos podem ser aplicados para tentar obter relações e maior entendimento sobre os sinais medidos.

3.11 Norma euclidiana e distribuição Chi

De forma geral, até agora, foram propostos métodos que trabalham com transformações lineares. Entretanto, também é possível propor uma ferramenta para evidenciar os valores absolutos de todos os vetores em uma única medida. Neste trabalho, a distribuição gaussiana é usada como referência, e idealmente, as distribuições das variáveis também seriam. Desta forma, é possível avaliar os valores absolutos do *Z-score*, de modo a buscar uma evidência de quão distante uma nova entrada do vetor, de forma geral, difere da média da variável aleatória. Neste sentido, a norma euclidiana pode ser usada.

Como proposto por Axler (2024), a norma euclidiana serve para avaliar a distância de um ponto ao centro da origem do sistema de coordenadas. Em um caso tridimensional de uma esfera, a norma euclidiana de um ponto sobre a superfície da esfera é o seu raio. A norma euclidiana é definida como:

$$\|Z_{k,:}\|_2 = \left(\sum_{j=1}^m Z_{k,j}^2 \right)^{\frac{1}{2}} \quad (3.58)$$

onde $Z_{k,j}$ é a componente do vetor da entrada k do *dataset* na j variável. A norma euclidiana também pode ser entendida como sendo a raiz quadrada da energia total do sinal naquela entrada k . Desta forma, esta seria uma característica escalar do sinal na entrada.

Com isto, independentemente do número de sensores, é possível avaliar uma única grandeza representativa da entrada. Para avaliar as probabilidades relativas a esta medida, sua função densidade de probabilidade correspondente é a distribuição Chi, pois é a distribuição resultante de uma soma de quadrados de variáveis aleatórias gaussianas (Forbes et al., 2011). Ela é definida como:

$$\mathcal{F}(x, g) = \begin{cases} \frac{x^{g-1} e^{-\frac{x^2}{2}}}{2^{\frac{g}{2}-1} \Gamma(\frac{g}{2})} & x \geq 0 \\ 0 & x < 0 \end{cases} \quad (3.59)$$

onde g é o número de graus de liberdade da distribuição. No caso deste trabalho, $g = m$, o número de medições a ser considerado. Esta distribuição será utilizada como referência para a distribuição resultante da aplicação da norma euclidiana sobre os dados.

Como mostrado no Capítulo 2, as distribuições não necessariamente possuem comportamento gaussiano, portanto, não é possível esperar que a norma euclidiana delas tenha o comportamento da distribuição Chi. Entretanto, a distribuição Chi serve de referência para avaliar os trechos dos quantis das distribuições para quais valores a norma euclidiana resultante não corresponderá a uma distribuição gaussiana, mas por outro lado, avaliando a divergência não em valores absolutos, mas sim, pelo seu módulo.

3.12 Modelo de Misturas Gaussianas

O Modelo de Misturas Gaussianas, (em inglês *Gaussian Mixture Model* (GMM)) é uma maneira de interpolar qualquer distribuição utilizando a soma de gaussianas ponderadas por

pesos. Como o interesse deste trabalho é identificar anomalias nas séries temporais, a possível distribuição Chi resultante seria interpolada através deste modelo, de tal modo a poder avaliar as transições de valores entre elas. O GMM é definido por Bishop (2006) como:

$$\mathcal{D}_{\|Z\|_2}(x) = \sum_{i=1}^s w_i \mathcal{N}(\mu_i, \sigma_i)(x) \quad (3.60)$$

onde s é o número de gaussianas, $\mathcal{N}(\mu_i, \sigma_i)$ são gaussianas de média μ_i e desvio padrão σ_i . w são os pesos das gaussianas que adaptam sua dimensão à distribuição, tal como um fator de escala.

Para adaptar a distribuição resultante dos dados da norma euclidiana, uma forma do Algoritmo de Maximização de Expectativa (em inglês: *Expectation-Maximization Algorithm* (EM)) é utilizada (Dempster; Laird; Rubin, 1977). Em essência, o Algoritmo EM considera recursivamente a observação de variáveis vistas nos dados (variáveis observáveis) e variáveis latentes, que seriam variáveis não observáveis diretamente, que são intrínsecas ao modelo probabilístico e podem assumir um significado. No caso, elas são modeladas dentro dos modelos probabilísticos. Com isto, utiliza os parâmetros do modelo para atualizá-los. Uma explicação da formulação geral do algoritmo está disponível no Apêndice B.

Os parâmetros w_i , μ_i e σ_i da mistura gaussiana são inicializados aleatoriamente. A variável γ_{ki} é definida como sendo a probabilidade posterior de uma entrada k , parte da mistura, ou seja:

$$\gamma_{ki} = P(x_k | \mu, \sigma) = \frac{w_i \mathcal{N}(x_k | \mu_i, \sigma_i)}{\mathcal{D}_{\|Z\|_2}(x_k)} \quad (3.61)$$

onde i está se referindo a uma determinada gaussiana da mistura. A Equação 3.61 caracteriza o passo de Expectativa do EM para a GMM. Para o passo de maximização, o parâmetro de peso é atualizado como a média das probabilidades:

$$w_i = \frac{1}{n} \sum_{i=1}^n \gamma_{ik} \quad (3.62)$$

A média é atualizada como sendo a média ponderada das variáveis pelas suas probabilidades posteriores:

$$\mu_i = \frac{\sum_{k=1}^n \gamma_{ki} x_k}{\sum_{k=1}^n \gamma_{ki}} \quad (3.63)$$

Posteriormente, o desvio-padrão é atualizado como sendo a média do segundo momento central de cada variável, com relação à média proposta do estado, ponderada pela probabilidade posterior:

$$\sigma_i = \frac{\sum_{k=1}^n \gamma_{ki}(x_k - \mu_i)^2}{\sum_{k=1}^n \gamma_{ki}} \quad (3.64)$$

onde n é o número total de entradas relativas a variável x . Estes passos vão sendo repetidos até a convergência de um critério de seleção de modelos, que será tratado na Seção 3.14. Com isto, o Modelo de Misturas Gaussianas consegue adaptar uma distribuição qualquer utilizando uma combinação de gaussianas.

3.13 Modelo Oculto de Markov Gaussiano

O Modelo Oculto de Markov Gaussiano (em inglês, *Gaussian Hidden Markov Model* (GHMM)) é um modelo probabilístico que assume que as variáveis ocultas assumem uma dinâmica de uma Cadeia de Markov e que as variáveis observáveis (isto é, as variáveis vistas no banco de dados) assumem subpopulações gaussianas, dentro de uma única distribuição. Este modelo é capaz de classificar os trechos das séries temporais (as variáveis observadas) através da inferência de estados (as variáveis latentes). Nesta dissertação, a inferência do modelo supõe que os estados sejam uma classificação de anomalia, buscando uma predição para o estado mais anômalo.

Neste modelo, as variáveis ocultas são chamadas de estados. O estado é uma variável aleatória que mapeia eventos para um número inteiro. Entretanto, esta variável não é observável diretamente. Os estados de um sistema podem ser definidos através de um conjunto $\mathbb{S} = \{1, 2, 3, \dots, s\}$ onde s é o número de estados. Neste trabalho, o número s é adotado como o número de gaussianas que irá compor a mistura gaussiana citada na seção anterior. No presente contexto, estes estados representam as condições da bomba.

No modelo da Cadeia de Markov, existe um estado para cada entrada ao longo do tempo. Com a evolução do processo, o estado futuro possui uma chance de se alterar dependendo do estado presente. Entretanto, a hipótese a ser tomada é que o estado futuro depende apenas do estado atual (ou que o estado atual apenas depende do estado anterior). Neste sentido, a hipótese de Markov é definida como (Kay, 2006):

$$P(S_{k+1} = i | S_k, S_{k-1}, S_{k-2}, S_{k-3}, \dots, S_1) = P(S_{k+1} = i | S_k) \quad (3.65)$$

onde i é um valor possível de $\$, S_k$ é o estado efetivo no instante k , e o estado S_{k+1} é o estado futuro. Neste sentido, estabelece-se uma dinâmica apenas entre o estado atual e o futuro, sem depender dos estados passados. Com isso, é possível definir a matriz de transição A . Esta matriz avalia as probabilidades de transição de um estado para outro ou de manter o mesmo estado, considerando a probabilidade de no instante k o estado assumir um valor. Desta maneira, para todos os valores possíveis do estado atual e para todos os valores possíveis do estado futuro, é possível definir uma matriz de transição como:

$$A = \begin{bmatrix} P(S_{k+1} = 1|S_k = 1) & P(S_{k+1} = 1|S_k = 2) & \dots & P(S_{k+1} = 1|S_k = s) \\ P(S_{k+1} = 2|S_k = 1) & P(S_{k+1} = 2|S_k = 2) & \dots & P(S_{k+1} = 2|S_k = s) \\ \vdots & \vdots & \ddots & \vdots \\ P(S_{k+1} = s|S_k = 1) & P(S_{k+1} = s|S_k = 2) & \dots & P(S_{k+1} = s|S_k = s) \end{bmatrix} \quad (3.66)$$

onde, nesta matriz a relação de tempo é apresentada de k para $k + 1$, porém, também pode ser avaliada de $k - 1$ para k . Pela definição da Equação 3.65, independentemente do índice, a matriz será igual. Outro ponto importante de avaliar é a probabilidade de cada estado em cada instante de tempo. Para tal é introduzido o vetor Π_k , de dimensão $s \times 1$:

$$\Pi_k = \begin{bmatrix} P(S_k = 1) \\ P(S_k = 2) \\ \vdots \\ P(S_k = s) \end{bmatrix} \quad (3.67)$$

Com isto, é possível definir a probabilidade do estado futuro através da probabilidade do estado atual, tal como:

$$\Pi_{k+1} = A \cdot \Pi_k \quad (3.68)$$

visto que:

$$\mathbf{A} \cdot \Pi_k = \begin{bmatrix} P(S_{k+1} = 1|S_k = 1) & P(S_{k+1} = 1|S_k = 2) & \dots & P(S_{k+1} = 1|S_k = s) \\ P(S_{k+1} = 2|S_k = 1) & P(S_{k+1} = 2|S_k = 2) & \dots & P(S_{k+1} = 2|S_k = s) \\ \vdots & \vdots & \ddots & \vdots \\ P(S_{k+1} = s|S_k = 1) & P(S_{k+1} = s|S_k = 2) & \dots & P(S_{k+1} = s|S_k = s) \end{bmatrix} \begin{bmatrix} P(S_k = 1) \\ P(S_k = 2) \\ \vdots \\ P(S_k = s) \end{bmatrix} \quad (3.69)$$

onde desenvolvendo:

$$\mathbf{A} \cdot \Pi_k = \begin{bmatrix} P(S_{k+1} = 1|S_k = 1)P(S_k = 1) + P(S_{k+1} = 1|S_k = 2)P(S_k = 2) + \dots + P(S_{k+1} = 1|S_k = s)P(S_k = s) \\ P(S_{k+1} = 2|S_k = 1)P(S_k = 1) + P(S_{k+1} = 2|S_k = 2)P(S_k = 2) + \dots + P(S_{k+1} = 2|S_k = s)P(S_k = s) \\ \vdots \\ P(S_{k+1} = s|S_k = 1)P(S_k = 1) + P(S_{k+1} = s|S_k = 2)P(S_k = 2) + \dots + P(S_{k+1} = s|S_k = s)P(S_k = s) \end{bmatrix} \quad (3.70)$$

e portanto:

$$\mathbf{A} \cdot \Pi_k = \begin{bmatrix} P(S_{k+1} = 1) \\ P(S_{k+1} = 2) \\ \vdots \\ P(S_{k+1} = s) \end{bmatrix} = \Pi_{k+1} \quad (3.71)$$

Para qualquer instante de tempo $k + h$, recursivamente, avalia-se:

$$\Pi_{k+h} = \mathbf{A} \cdot \mathbf{A} \cdot \dots \cdot \mathbf{A} \cdot \Pi_k = \mathbf{A}^h \cdot \Pi_k \quad (3.72)$$

e para o instante 0, Π_0 é inicializado de forma aleatória, ou escolhida, no caso de haver informação preliminar para sua inicialização.

Para introduzir as subpopulações gaussianas dentro da distribuição dos dados, o vetor \mathbf{B} de probabilidades de observações com relação aos estados é descrito como:

$$\mathbf{B} = \begin{bmatrix} P(x_k|S_k = 1) \\ P(x_k|S_k = 2) \\ \vdots \\ P(x_k|S_k = s) \end{bmatrix} \sim \begin{bmatrix} \mathcal{N}(\mu_1, \sigma_1)(x_k) \\ \mathcal{N}(\mu_2, \sigma_2)(x_k) \\ \vdots \\ \mathcal{N}(\mu_s, \sigma_s)(x_k) \end{bmatrix} \quad (3.73)$$

onde $\mathcal{N}(\mu_i, \sigma_i)(x_k)$ é a densidade de probabilidade da observação, que assume uma distribuição normal para uma variável observável x_k . Entretanto, computacionalmente, as distribuições não são contínuas, e sim discretas. Para tal, a densidade de probabilidade pode ser multiplicada por um fator Δx que a transforma em probabilidade. Para facilitar as demonstrações seguintes, este fator será assumido, arbitrariamente, como 1.

A pergunta geral que fundamenta a necessidade do algoritmo é a avaliação de qual é a maior probabilidade por estado para cada instante de tempo. Para isto, é proposto que seja usado o algoritmo de Baum-Welch, que é um caso especial do algoritmo EM que avalia a expectativa do estado em um determinado instante de tempo, e posteriormente, determina a maximização da probabilidade para encontrar os novos parâmetros do modelo GHMM (Baum et al., 1970).

Inicialmente, supõe-se a inicialização aleatória dos parâmetros \mathbf{A} e $\boldsymbol{\Pi}$. Nesta dissertação, o GMM será utilizado como uma forma de fornecer uma estimativa inicial para os parâmetros μ_i , σ_i e $\boldsymbol{\Pi}$. De forma geral, todos os parâmetros serão representados por um vetor de parâmetros $\boldsymbol{\theta}$.

Para a etapa de Expectativa, o algoritmo *Forward-Backward* define a probabilidade da observação pertencer a cada estado, ao longo de todos os valores da sequência de dados. Neste sentido, é definido o vetor $\boldsymbol{\alpha}_k$ como sendo a probabilidade da sequência acontecer até o instante de tempo k considerando um estado em questão:

$$\boldsymbol{\alpha}_k = \begin{bmatrix} P(\{x_1, x_2, x_3, \dots, x_k\}, S_k = 1 | \boldsymbol{\theta}) \\ P(\{x_1, x_2, x_3, \dots, x_k\}, S_k = 2 | \boldsymbol{\theta}) \\ \vdots \\ P(\{x_1, x_2, x_3, \dots, x_k\}, S_k = s | \boldsymbol{\theta}) \end{bmatrix} \quad (3.74)$$

cuja expressão pode ser avaliada como:

$$\boldsymbol{\alpha}_k = (\mathbf{A} \cdot \mathbf{B}(x_k)) \odot \boldsymbol{\alpha}_{k-1} \quad (3.75)$$

onde o operador \odot é o produto de Hadamard (produto elemento a elemento), que caracteriza o aspecto *backward* (para trás). Em sequência, define-se β_k como a probabilidade da sequência mais provável para cada estado, ou seja, avaliando a série para frente de k , caracterizando o aspecto *forward* (para frente):

$$\boldsymbol{\beta}_k = \begin{bmatrix} P(\{x_k, x_{k+1}, x_{k+2}, \dots, x_n\} | S_k = 1, \theta) \\ P(\{x_k, x_{k+1}, x_{k+2}, \dots, x_n\} | S_k = 2, \theta) \\ \vdots \\ P(\{x_k, x_{k+1}, x_{k+2}, \dots, x_n\} | S_k = s, \theta) \end{bmatrix} \quad (3.76)$$

cuja expressão pode ser avaliada como:

$$\boldsymbol{\beta}_k = (\mathbf{A} \cdot \mathbf{B}(x_k)) \odot \boldsymbol{\beta}_{k+1} \quad (3.77)$$

Para as condições iniciais de α e finais de β , é definido:

$$\boldsymbol{\alpha}_0 = (\mathbf{B}(O_t) \odot \Pi) \quad (3.78)$$

$$\boldsymbol{\beta}_n = \mathbf{1}_s \quad (3.79)$$

onde $\mathbf{1}$ é um vetor de uns.

Para descobrir qual estado tem mais probabilidade de ocorrer em cada instante k , o vetor de probabilidades posteriores do estado é definido como:

$$\boldsymbol{\gamma}_k = \begin{bmatrix} P(S_k = 1 | \{x_1, x_2, x_3, \dots, x_n\}, \theta) \\ P(S_k = 2 | \{x_1, x_2, x_3, \dots, x_n\}, \theta) \\ \vdots \\ P(S_k = s | \{x_1, x_2, x_3, \dots, x_n\}, \theta) \end{bmatrix} = (\boldsymbol{\alpha}_k \odot \boldsymbol{\beta}_k) \cdot (\boldsymbol{\alpha}_k^\top \boldsymbol{\beta}_k)^{-1} \quad (3.80)$$

Desta forma, é proposto que a matriz de transição para o próximo instante de tempo, que é relativa à probabilidade de transição de estados, seja definida através da matriz Γ_k :

$$\Gamma_k = (\mathbf{A} \odot \boldsymbol{\alpha}_k)(\mathbf{B}(O_{k+1}) \odot \boldsymbol{\beta}_{k+1})^T \quad (3.81)$$

e posteriormente sua normalização, que resulta na matriz de transição para o próximo instante de tempos considerando a sequência:

$$\Xi_k = \begin{bmatrix} P(S_{k+1} = 1, S_k = 1 | \{x_1, x_2, x_3, \dots, x_n\}, \theta) & \dots & P(S_{k+1} = 1, S_k = s | \{x_1, x_2, x_3, \dots, x_n\}, \theta) \\ \vdots & \ddots & \vdots \\ P(S_{k+1} = s, S_k = 1 | \{x_1, x_2, x_3, \dots, x_n\}, \theta) & \dots & P(S_{k+1} = s, S_k = s | \{x_1, x_2, x_3, \dots, x_n\}, \theta) \end{bmatrix} = \frac{\Gamma_k}{\|\Gamma_k\|_1} \quad (3.82)$$

Desta forma, as expectativas sobre os estados foram definidas em uma iteração qualquer do algoritmo. Neste sentido, o passo de Maximização busca atualizar os parâmetros \mathbf{A} , Π_0 e os μ_i e σ_i , associados a \mathbf{B} . Para atualizar \mathbf{A} , define-se:

$$\mathbf{A} \leftarrow \sum_{i=0}^n (\Xi_i \odot \gamma_i^{-1}) \quad (3.83)$$

Para os parâmetros, a equação é a mesma fornecida na atualização das misturas gaussianas, nas equações 3.63 e 3.64. A diferença é que o parâmetro γ_{ik} é relativo a realização do *forward-backward*. Com relação a Π_0 , segue-se que:

$$\Pi_0 \leftarrow \gamma_0 \quad (3.84)$$

A cada iteração do algoritmo, um critério de seleção de modelos é utilizado para avaliar se a otimização dos parâmetros convergiu ou não. Nesta dissertação, o parâmetro de convergência foi mantido como o padrão da implementação da *hmmlearn*. Mais detalhes sobre os critérios serão apresentados na Seção 3.14.

Com os parâmetros definidos, é possível inferir os estados relativos à distribuição resultante dos $\|Z_{k,:}\|_2$ (valores de *Z-score* com norma euclidiana aplicada) usando o Algoritmo de Viterbi (Viterbi, 1967), que já está implementado na biblioteca *hmmlearn*. O objetivo do algoritmo de Viterbi é inferir a sequência mais provável de estados considerando as observações e a matriz de transição de estados.

Inicializa-se o algoritmo considerando a probabilidade inicial multiplicada pela probabilidade da observação, resultando em outro vetor de probabilidade δ :

$$\delta_0 = \Pi_0 \odot \mathbf{B}(x_0) \quad (3.85)$$

e para os termos subsequentes:

$$\delta_k = [\delta_{k-1} \odot (A \cdot B(x_k))] \quad (3.86)$$

Com isto, encontra-se a sequência de maior probabilidade considerando o estado de maior probabilidade para o instante k :

$$S_k = \arg \max_i [\delta_k] \quad (3.87)$$

3.14 Critérios de seleção de modelos

Nas funções disponíveis da biblioteca *hmmlearn* e da biblioteca *sklearn*, para os modelos anteriormente descritos, os dois critérios disponíveis para avaliação do modelo são o Critério de Informação de Akaike e o Critério de Informação Bayesiano. Ambos são critérios que visam comparar modelos, considerando o equilíbrio entre a verossimilhança e a complexidade do modelo. No caso, a propriedade da verossimilhança L é a medida de quão bem o modelo consegue explicar a série temporal, considerando os parâmetros ajustados. A ideia dos critérios é escolher o modelo que possua o critério de menor valor, comparativamente (Bishop, 2006).

Para o modelo GMM, a verossimilhança pode ser calculada como:

$$L(\boldsymbol{w}, \boldsymbol{\sigma}, \boldsymbol{\mu}) = \prod_{k=1}^n \left(\sum_{i=1}^s w_i \mathcal{N}(x_k | \mu_i, \sigma_i) \right) \quad (3.88)$$

onde os vetores \boldsymbol{w} , $\boldsymbol{\sigma}$, $\boldsymbol{\mu}$ são os vetores de parâmetros do modelo GMM. Para o modelo HMM ela pode ser expressa a partir do *Forward* no último instante de tempo:

$$L(\boldsymbol{\theta}) = \sum_{i=1}^s \boldsymbol{\alpha}(\boldsymbol{\theta})_{ni} \quad (3.89)$$

onde $\boldsymbol{\theta}$ é definido como o vetor de parâmetros do modelo (A, B, Π) e é dado pelos termos de $\boldsymbol{\alpha}$ no último instante de tempo, através da soma de suas componentes para todos os estados. Desta maneira, é possível definir os critérios implementados.

O Critério de Informação de Akaike (em inglês, *Akaike Information Criterion* (AIC)) é um critério de seleção de modelos, proposto por Akaike (1974), que visa classificar o modelo, avaliando se ele se sobreajustou aos dados ou não, isto é, se o modelo se ajusta muito bem aos dados observados, mas é muito ineficiente para prever resultados com outros dados. O critério é descrito como:

$$AIC = -2\ln(L) + 2p \quad (3.90)$$

onde p é o número de parâmetros.

Por outro lado, o Critério de Informação Bayesiana (em inglês, *Bayesian Information Criterion* (BIC)), proposto por Schwarz (1978), é um critério de seleção de modelos, avaliando o quanto bem ele pode se adequar a distribuição de dados:

$$BIC = -2\ln(L) + p\ln(n) \quad (3.91)$$

onde n é o número de entradas de dados.

Avaliando ambos os critérios é possível compará-los. É possível notar que o AIC não considera o número de dados que o modelo adapta. As únicas entradas sobre o critério são a verossimilhança e o número de parâmetros. Por outro lado, o BIC considera o número de dados, penalizando proporcionalmente o critério em através da relação multiplicativa por k .

Analizando ambos os critérios, Burnham e Anderson (2002) afirmam que não há um critério “vencedor”, mas que o AIC possui mais fundamentações axiomáticas em uma análise frequentista, do que o BIC, porém, são considerações que não precisam existir em uma análise Bayesiana. Os autores propõem que o AIC, inclusive, possa ser equiparado ao BIC através de ajustes. Por outro lado, Kuha (2004), em seu artigo, diz que os dois modelos são bem fundamentados em suas respectivas teorias e sugere o uso de ambos os critérios para seleção. Com isto, nesta dissertação, será utilizada esta abordagem, verificando os valores para ambos os modelos.

3.15 Exemplo de identificação de anomalias com o sinal fictício

Retomando os sinais utilizados na Seção 3.7, é possível tomar a norma euclidiana do conjunto de dados e obter um modelo GMM. O GMM permite um primeiro entendimento do comportamento do sinal e sua distribuição.

É possível notar que na Figura 3.7 duas gaussianas não representam acuradamente a distribuição dos dados. Entretanto, as gaussianas da mistura gaussiana obtida buscam indicar duas subpopulações da distribuição principal. Uma distribuição de valores menores, em

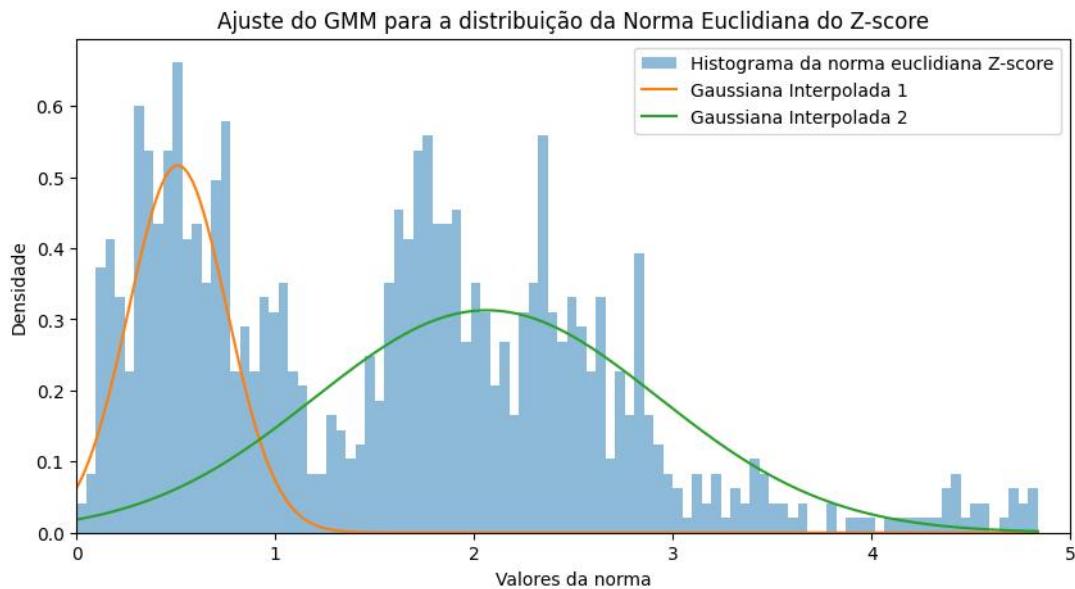


Figura 3.7: Ajuste do modelo de misturas gaussianas para a norma euclidiana no sinal fictício.

laranja, indo de 0 a 1, aproximadamente e outra que possui valores mais centrados entre 1 e 3. A primeira laranja é associada o comportamento normal do exemplo, enquanto, a segunda verde é associada o comportamento anômalo.

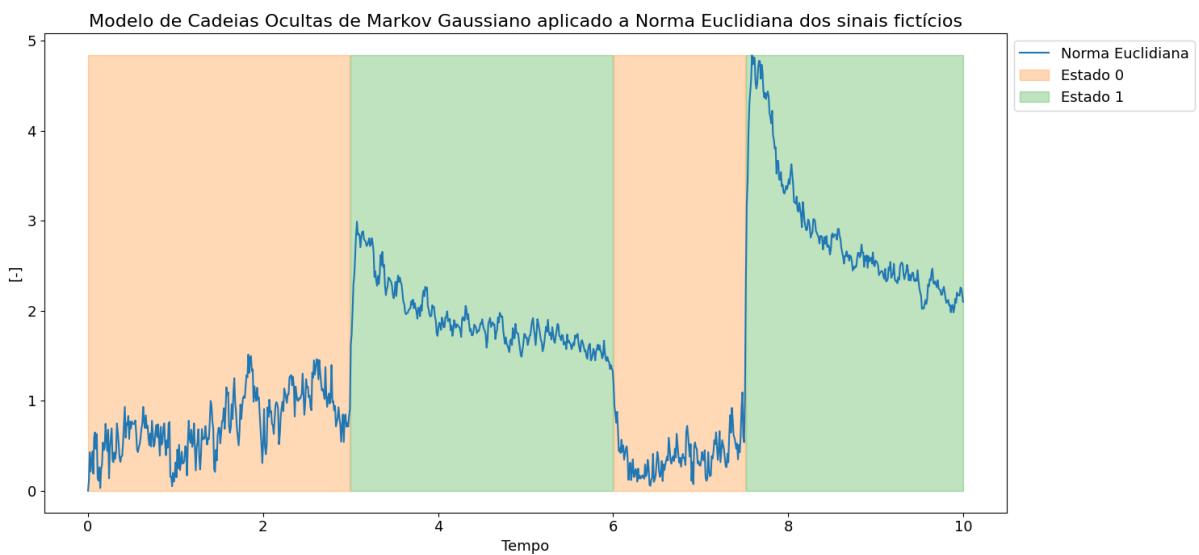


Figura 3.8: Identificação de estados com o sinal da norma euclidiana.

Fornecendo a série temporal para treinar o modelo e utilizando-o para identificar os estados, é possível obter a Figura 3.8. As anomalias na Seção 3.7 eram introduzidas a partir do instante 2,5 no sinal 1 e 7,5 no sinal 2. A norma 2 avalia a grandeza das entradas na série temporal, de tal modo que consegue capturar as duas anomalias em uma única série temporal. Nota-se, que nos instantes em que as anomalias aparecem, o estado persistente é

relativo à gaussiana verde. Desta maneira, a GMM separa a distribuição dos dados em duas gaussianas possíveis e a GHMM avalia a estrutura temporal para evidenciá-las, classificando as anomalias.

Desta maneira, com os resultados obtidos, é possível notar que o método tem sucesso em identificar as anomalias induzidas no sinal, levando em consideração a relação temporal da série, evidenciando os trechos em que elas foram detectadas.

3.16 Implementação

Nesta dissertação, é importante comentar a respeito das implementações realizadas para o processamento dos dados e da aplicação dos modelos estatísticos, de maneira geral. Já existem implementações (bibliotecas) em Python para a realização de processamento de dados (*pandas*), modelos estatísticos (*sklearn* e *hmmlearn*), modelos de transformações (*sklearn*) e geração de gráficos (*matplotlib*).

Entretanto, para utilizá-las em uma única biblioteca com as necessidades relativas aos dados fornecidos pela empresa Equinor e para os objetivos desta dissertação, foi gerada a *Layer of Integration with Scikit-learn and Signal Analysis* (LISSA) (e também *lissa* - no padrão de nomenclatura de bibliotecas desta dissertação), que, em português significa “Camada de Integração com *Scikit-Learn* e Análise de Sinais”. *Scikit-Learn* é o nome completo da biblioteca *sklearn*. A biblioteca, em si, não implementa nenhum algoritmo original, porém consolida uma série de rotinas para tratamento dos dados. Foram implementadas funções, que partindo de um arquivo CSV (em inglês, *Comma Separated Values* (CSV)) padronizado, fornecido pela empresa Equinor, realizem tarefas como:

- Transformar os dados em objeto da biblioteca Pandas, deletar dados duplicados e padronizar o formato de tempo.
- Cruzar os dados horários com os dados da planilha de falhas.
- Simplificar ou ajustar medições (por exemplo, aplicar a norma euclidiana sobre a vibração).
- Aplicar filtros e o *Z-score* proposto sobre os dados.
- Armazenar os dados em um arquivo CSV, podendo ser utilizado em análises futuras.

- Criar figuras de forma padronizada, podendo se repetir para qualquer corrida dentro dos bancos de dados fornecidos.
- Estabelecer padrões para a aplicação das transformações, reduzindo a necessidade de repetir códigos.
- Exportar o resultado das transformações para CSV.
- Determinar sequências entre a bomba ligada e desligada, utilizadas no treinamento do GHMM.
- Separar, no banco de dados, os dados das corridas entre teste e treino, separando corretamente os dados pelas suas sequências.
- Converter os estados gerados no GHMM considerando as suas proporções esperadas, uma vez que o número atribuído para representar o estado é aleatório.
- Integrar os modelos já implementados nas bibliotecas *hmmlearn* e *sklearn* com os dados processados pela *lissa*.

Um aspecto importante de ser notado é que Bishop (2006) e Gerón (2019) constatam que é importante que o conjunto de dados seja dividido entre “treino” e “teste”. O significado de ambos os termos é que o conjunto de treino é utilizado para treinar o modelo, ou seja, encontrar parâmetros para o modelo através dos dados e o conjunto de teste seria para avaliar a performance do modelo.

Entretanto, a questão do teste é uma questão complexa. Métricas, como por exemplo, acurácia (taxa de acertos sobre total de testes) e precisão (verdadeiros positivos sobre verdadeiros positivos e falsos positivos) (Gerón, 2019) são comumente utilizadas em problemas de aprendizado supervisionado. No caso, como já foi discutido anteriormente, para o presente conjunto de dados desta dissertação, não cabe a utilização de um método desta categoria. Como não se sabe se a data marcada como falha é propriamente a data da falha, não é possível comparar diretamente o resultado de uma classificação com a rotulação proposta na planilha preenchida pelos operadores. Em algumas bombas, é possível notar que houve a marcação da falha, porém, a bomba continuou operando por um longo tempo após o dado.

Desta maneira, nesta dissertação propõe-se a utilização da contagem de estados anômalos ou incomuns antes da ocorrência da remoção da bomba. Para um instante de tempo $k = T$,

onde T é a última entrada registrada, para o conjunto de dados da corrida a ser analisada, é proposto que seja avaliada a incidência de estados anômalos ou incomuns 1 dia antes da remoção. Com isto, para o conjunto de dados da corrida, quando a bomba está ligada, é proposto que seja avaliada a distribuição de estados $\{s_{T-23}, s_{T-23}, s_{T-22}, \dots, s_T\}$. Nesta dissertação, esta avaliação foi chamada de persistência, em virtude da análise de quanto um estado persists neste período.

Com relação à organização dos dados para treinamento, Gerón (2019) constata que, usualmente, existe uma proporção de 75% dos dados utilizados para treino e 25% utilizado para teste. No caso, não seria correto apenas separar os dados, desconsiderando a relação temporal entre eles e que cada linha do conjunto de dados é associada a uma corrida. Desta maneira, nesta dissertação, esta proporção é utilizada entre as 57 corridas disponíveis, ou seja, aplicada sobre os subconjuntos de corridas e não sobre o banco de dados em si. Um exemplo é, se existem 4 corridas no banco de dados $\{A, B, C, D\}$, com 200, 1100, 30 e 50 linhas, respectivamente, os subconjuntos $\{A, B, C\}$ poderiam ser utilizados para formar o conjunto de treino, totalizando em 1330 linhas. Outro ponto sobre este assunto é que cada intervalo entre a bomba desligada ou ligada forma uma sequência a ser utilizada para o treinamento. Com isto, estas linhas também são separadas entre grupos de sequências ligadas ou não.

Entretanto, é necessário salientar que a escolha das corridas para formação do conjunto de treino é pseudoaleatória. Números aleatórios no computador são gerados a partir de funções que se utilizam de uma semente (*seed*, em inglês) que é um número que é utilizado de referência para o gerador. Neste trabalho, algumas sementes serão utilizadas como forma de permitir a reprodutibilidade dos mesmos. Portanto, alguns números foram arbitrariamente escolhidos para inicializar os geradores de números aleatórios, como, por exemplo, os números 19971215, 19630926 e 20210505.

Para gerar gráficos de identificação de anomalias para todas as bombas, em um primeiro momento, o conjunto de dados inteiro tem o modelo aplicado para gerar a identificação de anomalias e realizar um primeiro gráfico de persistência, classificando os trechos das séries temporais. Em seguida, apenas o conjunto de teste é avaliado, e um gráfico de persistência é gerado para ele, podendo ser constatada a eficácia do modelo, ou não. Posteriormente, propõe-se avaliar os resultados de modelos inicializados com outras sementes de inicialização distintas, de modo a avaliar se de fato a hipótese que um ou mais estados relativos a anomalia persiste em uma região próxima a remoção da bomba.

3.17 Procedimento proposto

Desta forma, com a biblioteca implementada, é possível definir uma sequência de operações para unir os procedimentos propostos neste capítulo e consolidar os resultados desta dissertação. Nesta seção, o objetivo é explicar como é o fluxo de dados proposto. Mais detalhes sobre o código podem ser encontrados no Apêndice C.

Primeiramente, os dados horários são cruzados com os dados de falha, se tornando um único conjunto de dados unidos, tal como mostrado na Figura 3.9. Neste processo, entradas duplicadas são deletadas, de tal maneira a não considerar duas vezes a mesma informação no processo. Desta maneira, é possível avaliar para todas as bombas, em que momento da série temporal estão sendo reportadas as falhas.

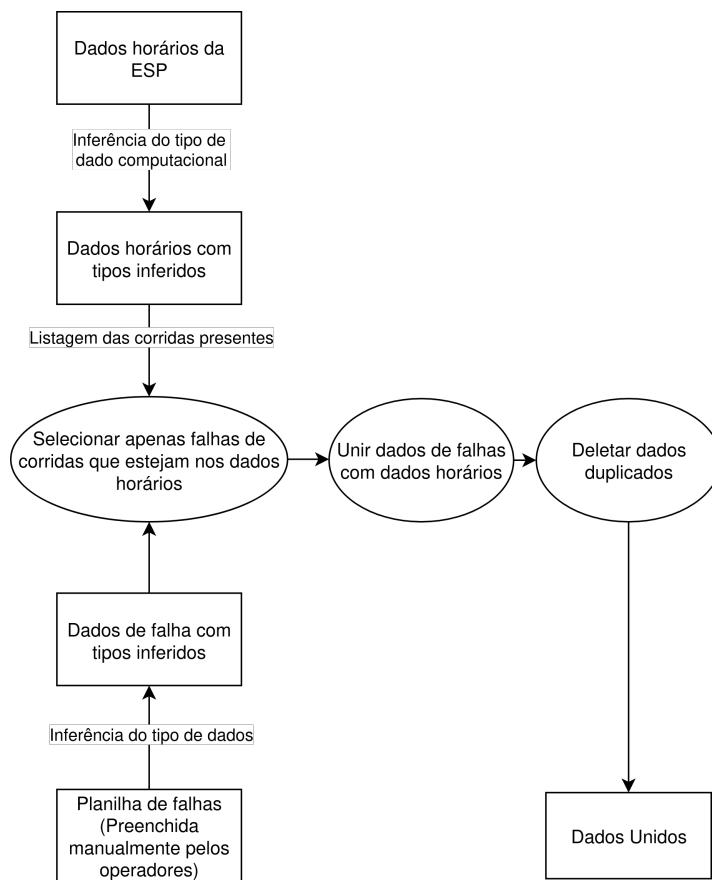


Figura 3.9: Diagrama de fluxo de dados para o pré-processamento.

Posteriormente, com os dados unidos, é possível calcular o *Z-score* modificado. Como mostrado na Figura 3.10, uma primeira corrida é selecionada, e para os dados numéricos dela o *Z-score* modificado é aplicado, e os dados faltantes são definidos como valor zero para o respectivo *Z-score*. O procedimento é feito para todas as corridas. É importante notar, que

desta forma, cada bomba tem seus parâmetros (mediana e desvio-padrão) de Z-score de forma individual.

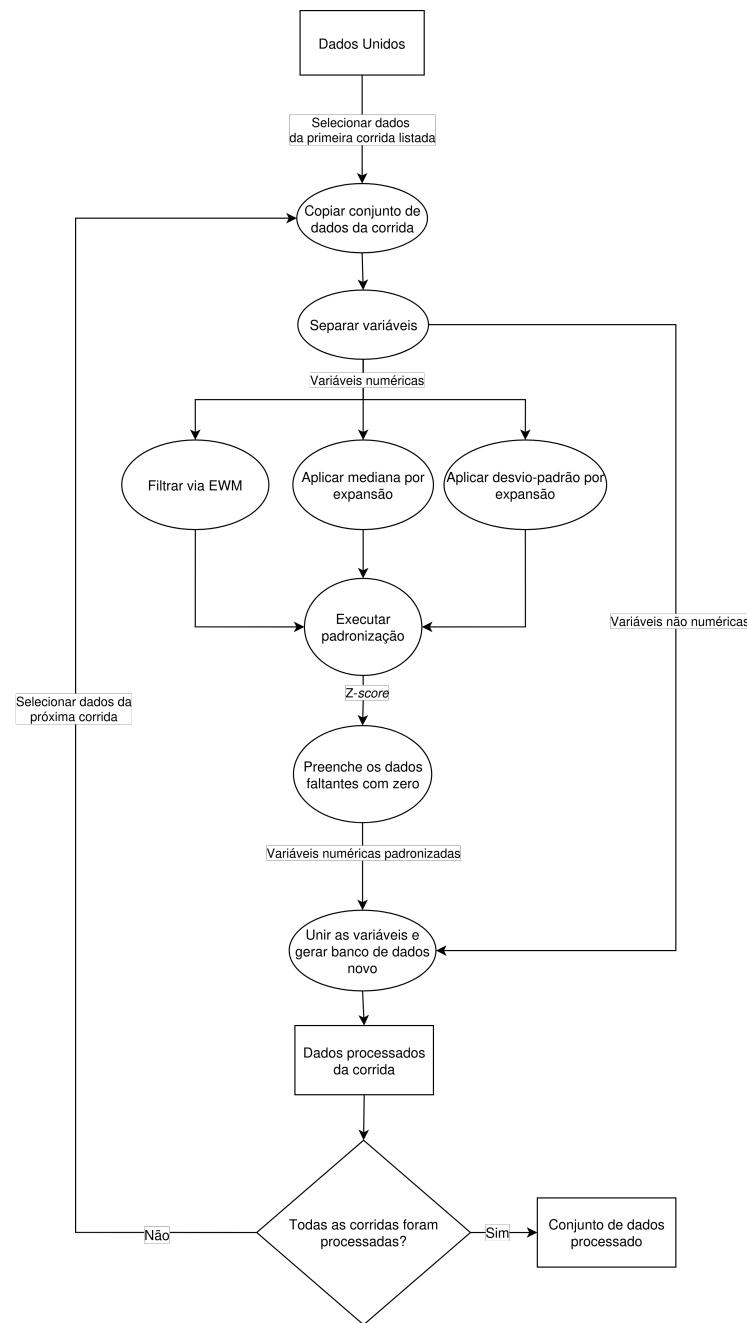


Figura 3.10: Diagrama de fluxo de dados para a aplicação do *Z-score* modificado.

Após a padronização, é possível realizar dois procedimentos distintos: treinar o modelo GHMM e encontrar as matrizes da PCA e da ICA. Para encontrar as matrizes da ICA e da PCA, o fluxo da Figura 3.11 mostra que as variáveis numéricas são colocadas sobre os modelos e os mesmos retornam as matrizes. Posteriormente, é possível realizar os gráficos Q-Q de todas as transformações e das variáveis originais, com e sem norma euclidiana aplicada.

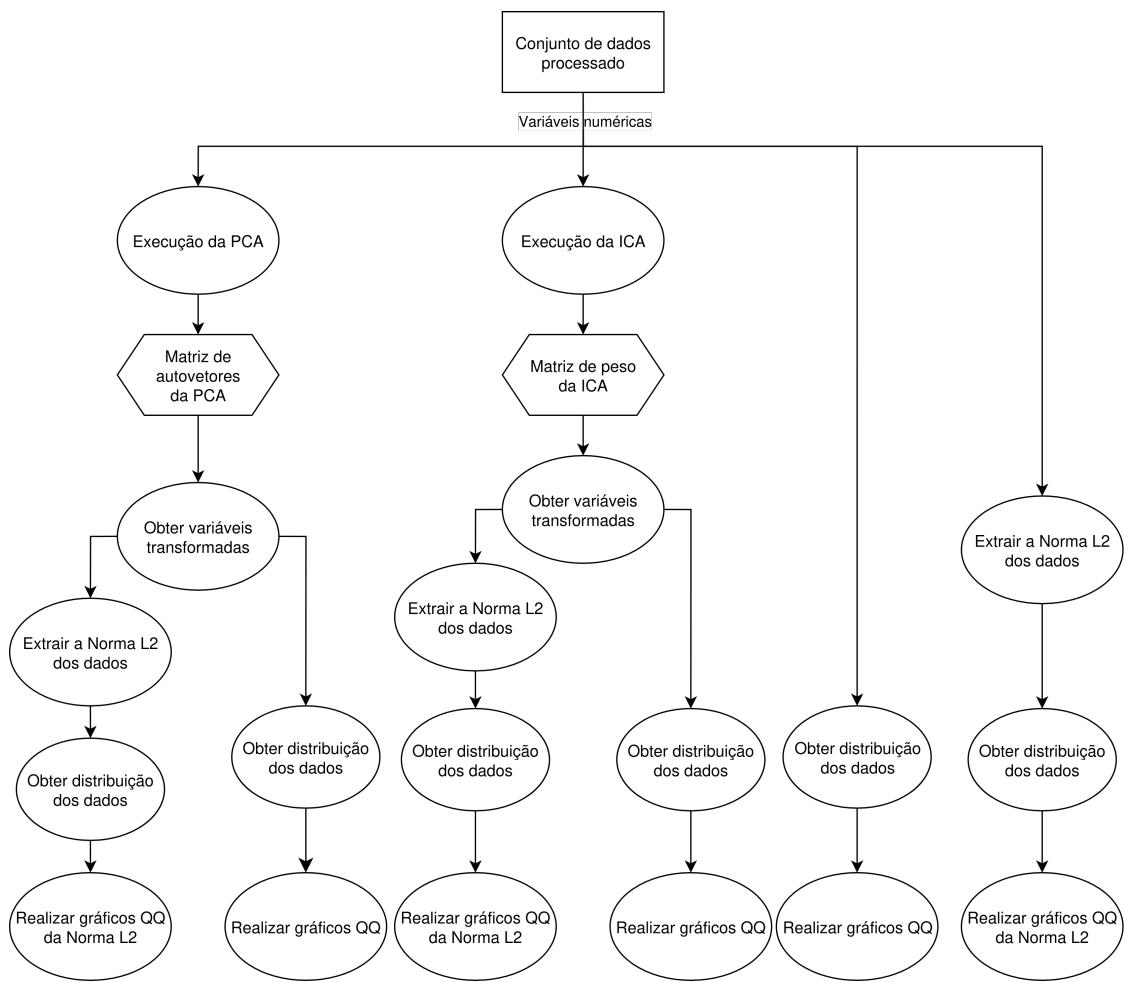


Figura 3.11: Diagrama de fluxo de dados para a realização das transformações.

Para treinar o modelo GHMM, o fluxo de dados é apresentado na Figura 3.12. O conjunto de dados processado tem a norma euclidiana extraída e a separação entre treino e teste é executada. Posteriormente, as corridas de treino são selecionadas para receber a interpolação da GMM. A GMM fornece os parâmetros iniciais e o modelo é treinado. Em sequência, todos os dados são utilizados para predizer estados em todas as corridas. Com isto, um conjunto de dados classificado é obtido e é possível gerar os gráficos de persistência para todas as corridas. Também um conjunto de testes é separado e seus gráficos de persistência são gerados.

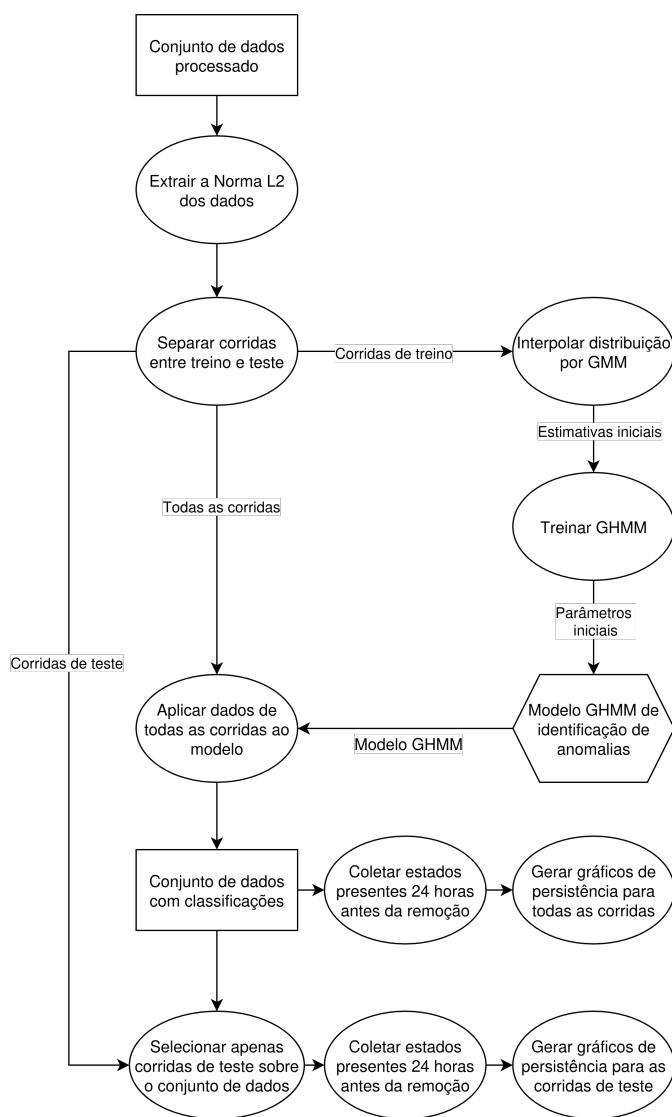


Figura 3.12: Diagrama de fluxo de dados para o treinamento do modelo GHMM.

Capítulo 4

Resultados

Após realizar a implementação das rotinas computacionais em Python, considerando a metodologia desenvolvida no Capítulo 3, é possível avaliar os resultados obtidos. Para esta seção, os dados considerados no processamento foram a temperatura do motor, módulo da corrente trifásica, tensão do motor, pressão e temperatura na cabeça do poço, pressão e temperatura na entrada, pressão de descarga, frequência do AVV, e o módulo da vibração no plano transversal da bomba.

4.1 Resultados da Filtragem e da Padronização

Como estabelecido, a filtragem é realizada por um filtro passa-baixa com período de 24 horas. No exemplo da Figura 4.1, mostra-se a pressão de entrada da corrida B-18 2 (verde) e o sinal filtrado (azul). Nota-se que picos, como os que ocorreram próximo do início de julho de 2012, são atenuados, de tal modo, que o sinal filtrado possui menos deles.

No Capítulo 3, foram propostos dois métodos de avaliação do centro dos dados e do grau de variação nos sinais. Para a pressão de entrada, é possível comparar a média e a mediana por expansão na Figura 4.2.

Nota-se que a mediana possui valores menores ao longo da corrida, o que pode ser explicado pelo fato de que ela considera apenas os elementos do meio do rol, e não todos os dados da série temporal somados de forma ponderada. Já, para o desvio-padrão por expansão e o MAD é possível compará-los na Figura 4.3. É possível notar que o MAD tem uma escala menor do que o desvio-padrão convencional. Entretanto, o MAD é mais instável temporalmente e o desvio padrão tem menos desvios ao longo do tempo. Com estes

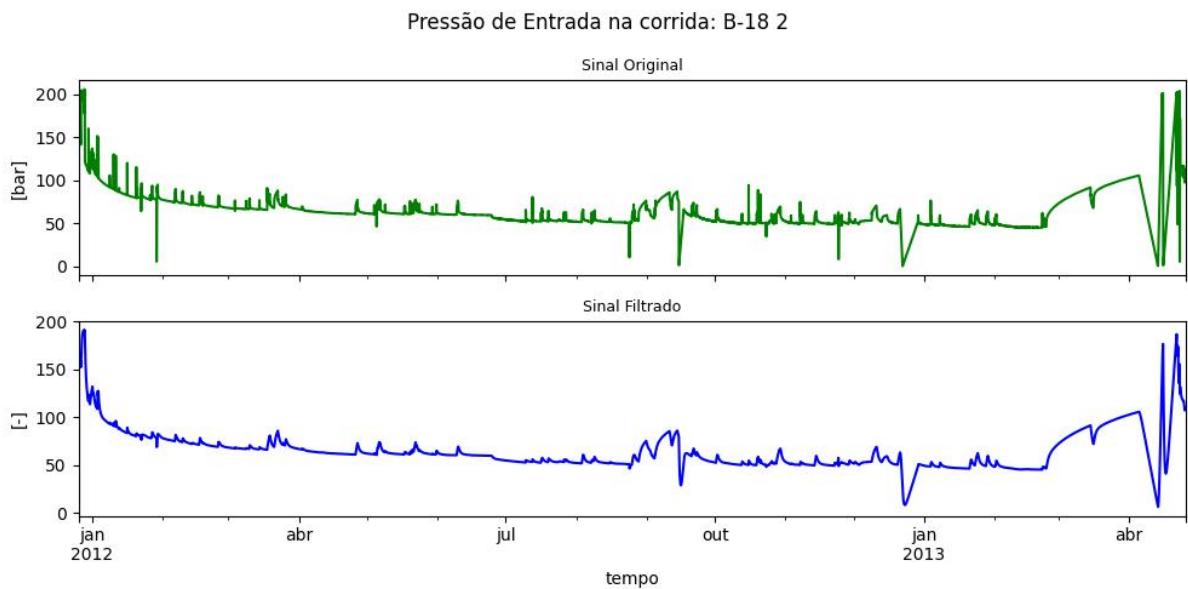


Figura 4.1: O sinal Original da pressão na corrida B-18 2 (verde) e o sinal Filtrado com o filtro passa-baixa (azul).

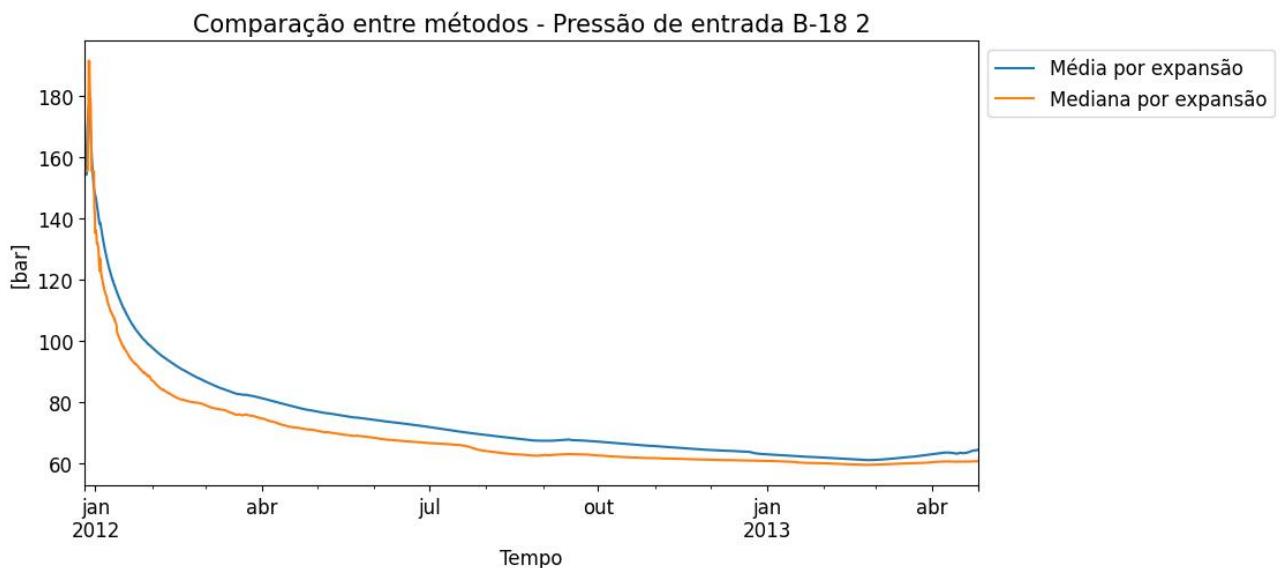


Figura 4.2: Média (azul) e Mediana (laranja) comparadas para a pressão da corrida B-18 2.

resultados, prefere-se a mediana e o desvio-padrão convencional para a consolidação do *Z-score*.

O *Z-score* modificado foi calculado para cada entrada do banco de dados. A Figura 4.4 ilustra o processo de obtenção do *Z-score* para a pressão de admissão da BCS. Primeiramente, subtrai-se do sinal filtrado a mediana em expansão (azul claro), que é o resultado do cálculo progressivo que considera os valores anteriores e o valor da entrada. Em seguida, essa diferença é dividida pelo desvio padrão em expansão (amarelo), que mensura as variações nos dados ao longo do

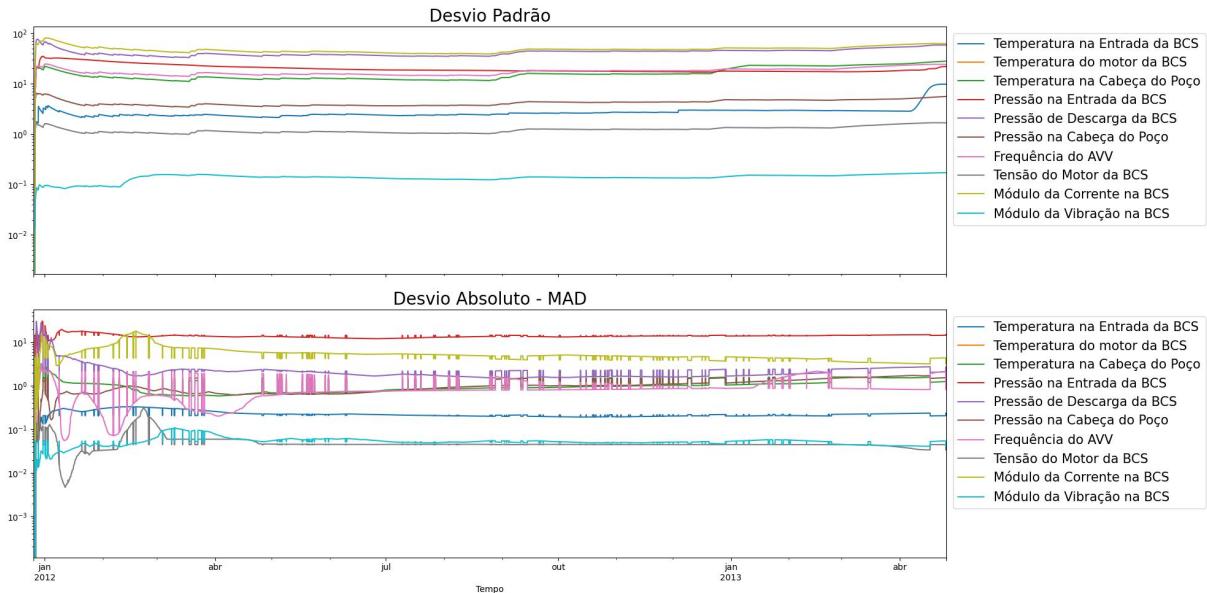


Figura 4.3: Comparação dos métodos de desvio: Desvio-Padrão contra MAD.

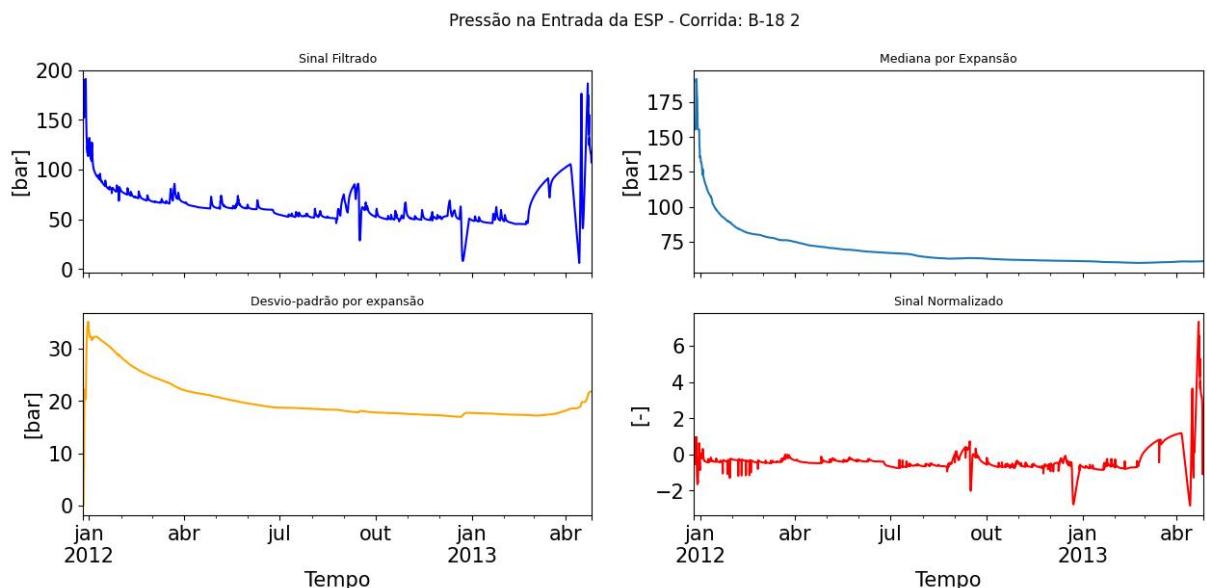


Figura 4.4: Sinal filtrado (azul), mediana em expansão (azul claro), desvio padrão em expansão (amarelo) e sinal normalizado (vermelho).

tempo. Desta forma, o *Z-score* resultante (vermelho) destaca as flutuações do sinal em relação ao seu comportamento histórico, facilitando a identificação de desvios e potenciais anomalias.

Calculando o *Z-score* para todas as variáveis, é possível obter a Figura 4.5. Notavelmente, uma anomalia aparece próximo ao final da figura. A linha tracejada vermelha marca o momento em que a falha foi registrada na planilha de falhas. Nesta análise, sugere-se que seja possível que a anomalia notada através do *Z-score* no final da série temporal, esteja relacionada com a falha, e com isso, a remoção da bomba.

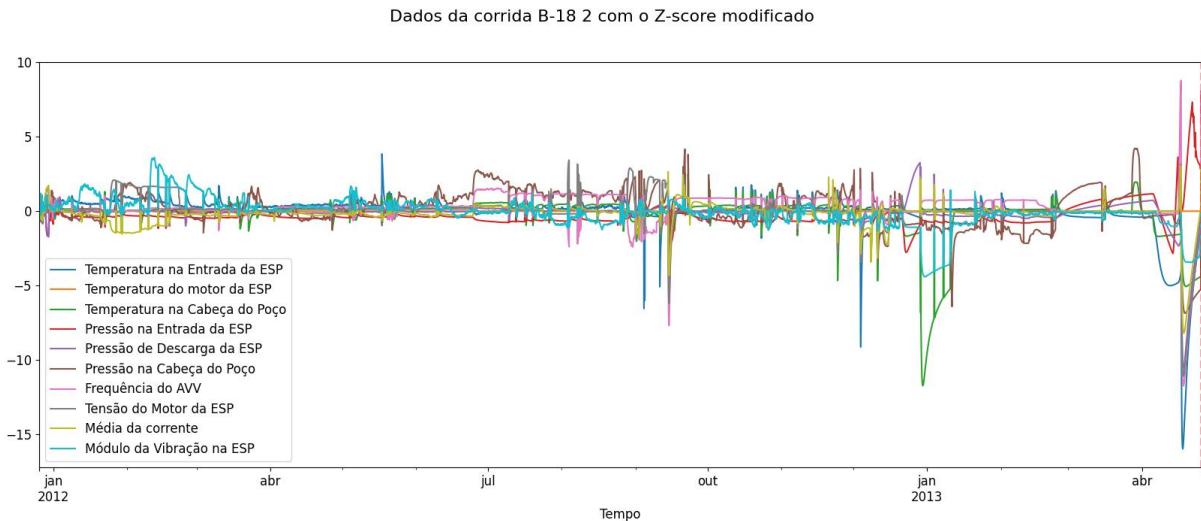


Figura 4.5: *Z-score* modificado para a corrida 2 da bomba B-18.

4.2 Resultados da PCA

Após a aplicação do *Z-score*, é possível aplicar a PCA no *dataset* e avaliar suas componentes. Neste sentido, é preciso considerar os resultados das duas abordagens. Apenas com a matriz de covariância e com a matriz de correlação.

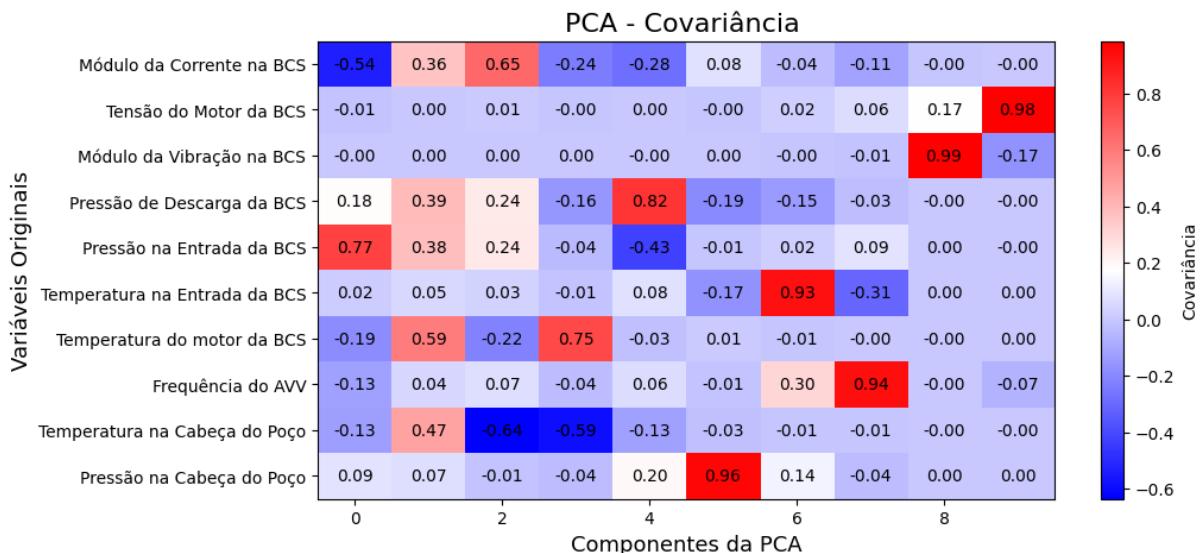


Figura 4.6: Matriz de covariância do resultado da PCA.

Primeiramente, é possível avaliar a PCA pelo método da covariância. Na Figura 4.6 é apresentada a matriz de autovetores. Notam-se várias relações entre as propriedades. Uma relação notável é na componente 8 e 9, onde a vibração é correlacionada com a tensão do motor em uma relação de 0,99 para 0,17, na componente 8, e 0,99 para -0,17 na componente

9. A outra relação notável é a componente 4, pois a PCA revela uma relação muito próxima à pressão de diferencial, que é a pressão de descarga subtraída da pressão de entrada, que não havia sido considerada nos dados processados. Nesta componente, outras variáveis interferem em seus pesos, como por exemplo, o módulo da corrente na BCS e a temperatura na cabeça do poço.

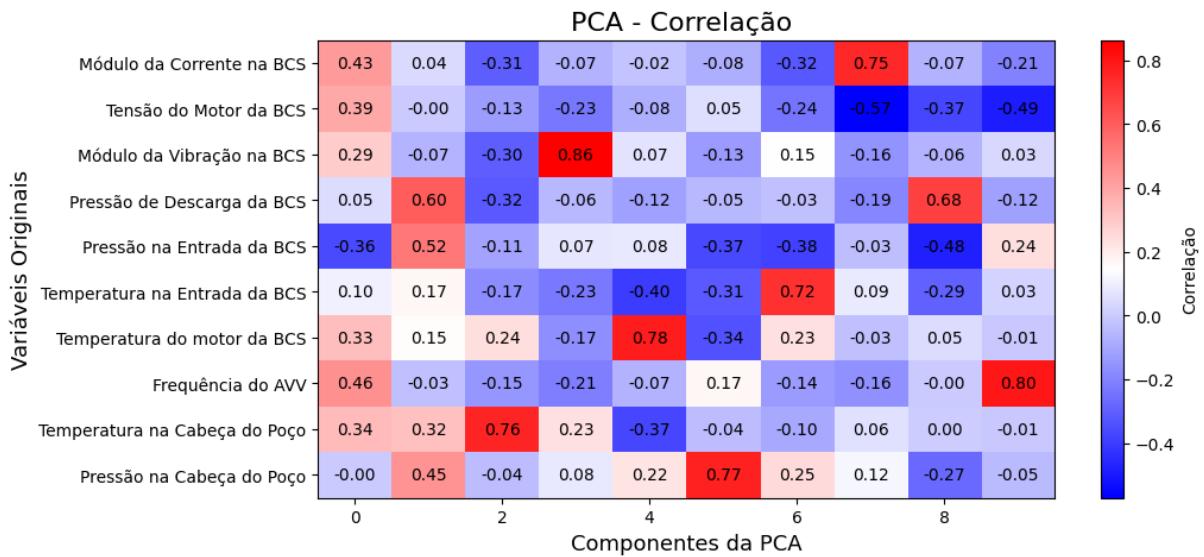


Figura 4.7: A matriz de correlação do resultado da PCA.

Posteriormente, é possível aplicar o *Z-score* conforme proposto e obter a matriz de autovetores na Figura 4.7. É possível notar que a abordagem pela correlação gera resultados distintos com relação à abordagem da covariância. Relações que não foram obtidas na abordagem da covariância são notadas. Por exemplo, na componente 10, é possível avaliar que a tensão do motor da BCS decai conforme a frequência do AVV sobe.

Além da análise das matrizes dos autovetores, é possível avaliar a Correlação e a Covariância Explícada, que são os índices gerados pela avaliação do percentual *P* dos autovalores. Como explicado anteriormente, a definição da propriedade explicada é a quantidade de componentes necessárias para expressar um percentual do total da propriedade.

Neste sentido, na Tabela 4.1 é possível notar que a partir da componente 5, a Covariância Explícada já converge acima de 95%, o limiar da heurística proposta por Gerón (2019), o que significa que com 5 componentes, é possível explicar 95% da covariância. Neste sentido, é possível notar na Figura 4.6 que existe um maior número de variáveis independentes por si só

após o limiar da heurística. No caso, como é mais adequada a abordagem pela correlação, no limiar heurístico, o *dataset* se reduziria de 10 para 8 variáveis.

Tabela 4.1: Correlação explicada através da análise dos autovalores.

Quantidade de Componentes	Covariância Explicada	Correlação Explicada
0	38,107%	31,027%
1	60,030%	47,442%
2	76,517%	59,214%
3	89,939%	69,038%
4	97,448%	77,951%
5	99,295%	85,647%
6	99,735%	91,936%
7	99,996%	95,718%
8	99,998%	98,871%
9	100,00%	100,000%

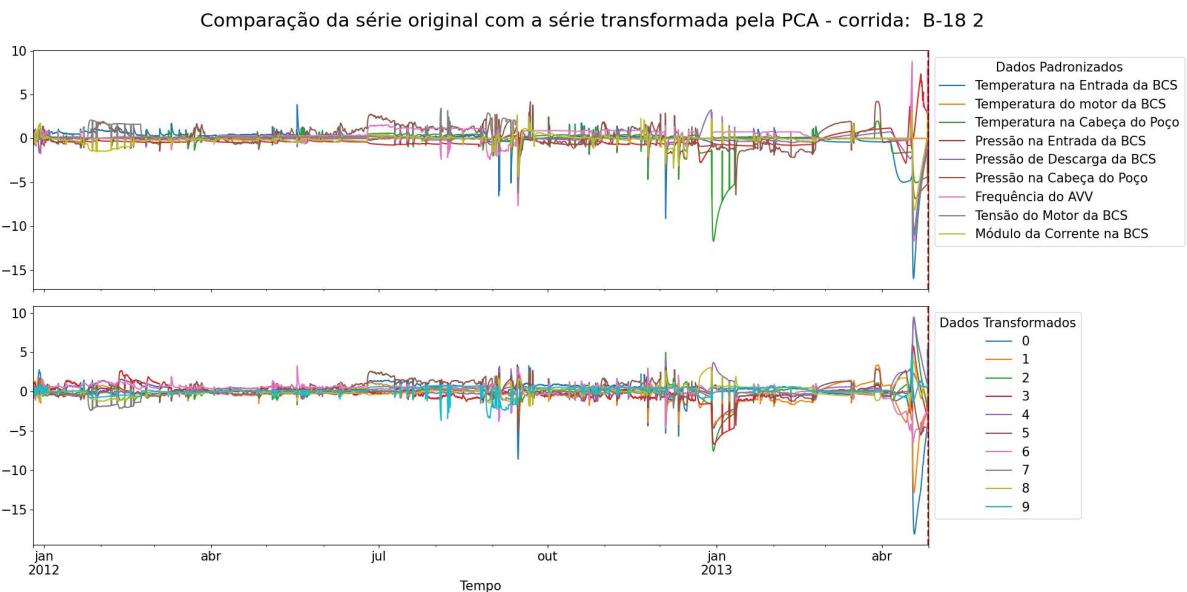


Figura 4.8: Comparação entre os sinais originais padronizados e os obtidos pela PCA da matriz de correlação.

Na Figura 4.8 é mostrada a comparação entre as séries temporais das variáveis originais e a série temporal das componentes obtidas pela PCA (todas). É possível notar que existe uma diferença nos gráficos e que, efeitos na temperatura da cabeça do poço, por exemplo, são refletidos ao longo das componentes, como é possível notar no tempo próximo a janeiro de

2013. Posteriormente, na anomalia próxima de abril do mesmo ano, as componentes mostram um formato muito parecido da curva que se formou em decorrência da anomalia, ao contrário dos dados padronizados, em que cada medição possui um comportamento diferente. Com isto, nota-se que a PCA obteve outra série temporal. Entretanto, a utilidade dela para o modelo de GHMM será avaliada na Seção 4.4.

4.3 Resultados da ICA

A ICA foi executada para as variáveis padronizadas pelo *Z-score* modificado. Tal como na PCA, ela gera uma matriz de pesos que é gerada a partir da sintropia, resultando na maximização da independência entre as componentes.

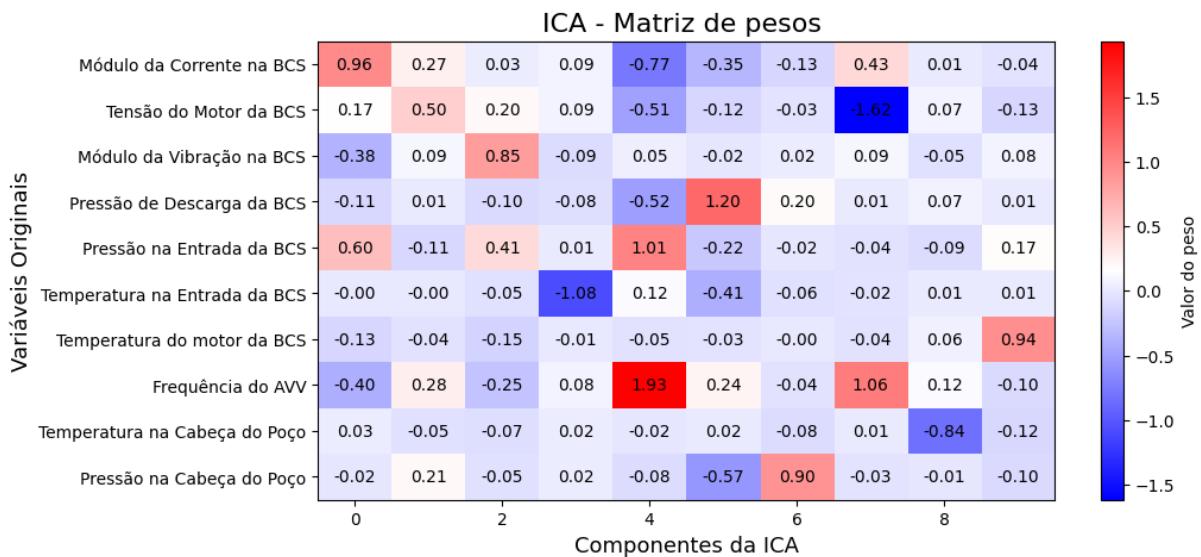


Figura 4.9: A matriz de separação da ICA - referência 19971215.

Ao calcular a ICA, um número aleatório é fornecido como referência para gerar os demais. Neste sentido, o número aleatório 19971215 foi escolhido e fixado para poder garantir a reproduzibilidade dos resultados. Com isso, é possível executar o algoritmo e verificar que na Figura 4.9 que ela encontra relações distintas à PCA. Por exemplo, na Figura 4.7, a componente 2 indica que há correlação entre a temperatura na cabeça do poço e outras componentes. Já a análise da ICA, na componente 8, encontra um peso praticamente independente para a temperatura na cabeça do poço.

Outro exemplo é a pressão na cabeça do poço. Na PCA, por correlação, a componente mais associada a esta variável mostra correlação com outras propriedades, como a pressão e

temperatura na entrada e a temperatura do motor da bomba. Já na ICA, ela aparece de forma completamente independente, na sua componente 6. Entretanto, é possível notar o contrário também, já que na componente 9 da PCA e 7 da ICA, é apresentada uma associação entre a tensão do motor e a frequência do AVV. De fato, a PCA aponta uma correlação entre a pressão na entrada e outras duas componentes, enquanto a ICA não aponta isto. Entretanto, a relação entre as duas variáveis permanece.

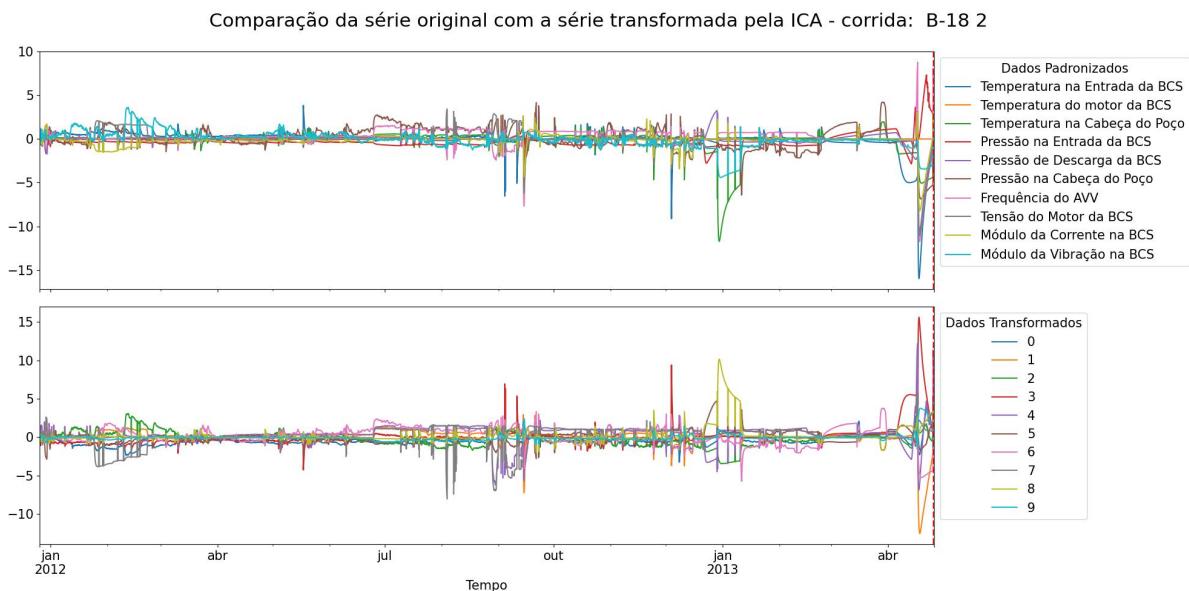


Figura 4.10: Comparação entre os sinais originais padronizados e os obtidos pela ICA.

Para a Figura 4.10, mostra-se uma comparação entre os sinais da série original e a transformação da série temporal obtida pela primeira ICA (de semente 19971215). Nota-se que a ICA gerou sinais que acompanham uma mesma tendência. Por exemplo, em julho de 2012 houve um pico para a frequência do AVV e a tensão do motor. A componente 7 da ICA acaba juntando os dois em um único fenômeno, como se o fator de pico fosse registrado em apenas um sinal. Neste sentido, a ICA agrupou eventos nas componentes, buscando com que cada uma represente a sequência de dados. Na próxima seção, será avaliada sua utilidade para o modelo de GHMM.

4.4 Distribuições Resultantes

Um aspecto importante de ser avaliado são os gráficos Quantil-Quantil (Q-Q) após as transformações. Como já tratado, os gráficos Quantil-Quantil mostram uma relação entre os

quantis esperados de uma determinada distribuição já conhecida para os quantis observados na distribuição medida.

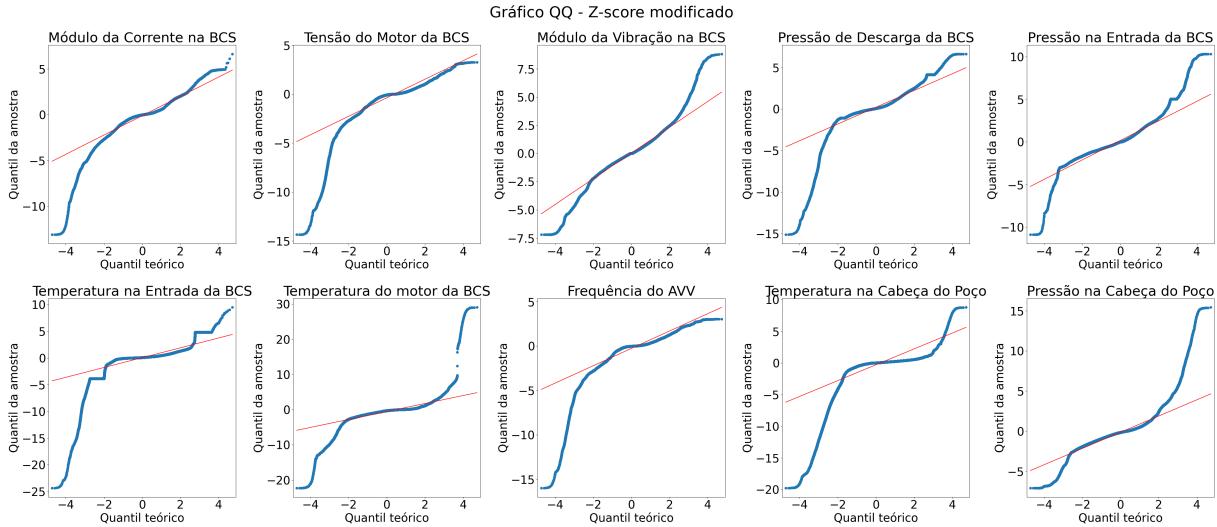


Figura 4.11: Gráfico Quantil-Quantil para as componentes do *Z-score*.

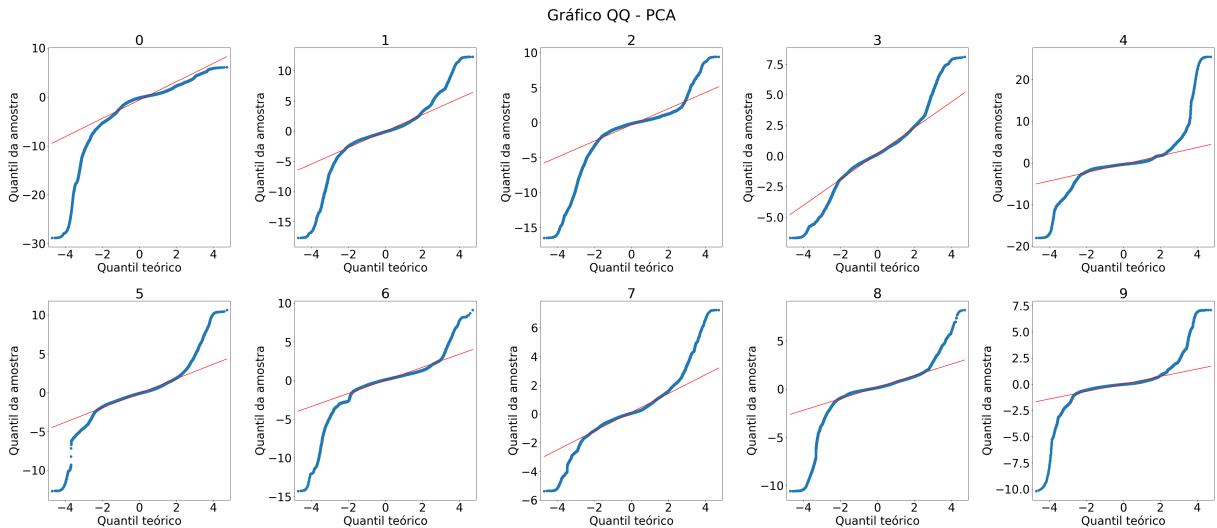


Figura 4.12: Gráfico Quantil-Quantil para as componentes da PCA.

Na Figura 4.11 é mostrado o gráfico Q-Q do *Z-score* modificado. No caso, é possível notar que as distribuições resultantes não são distribuições gaussianas, mas alguns trechos a respeitam. É possível notar que para quantis maiores, de forma geral, a cauda da curva fica mais “pesada” (em inglês, *heavy tail*), ou seja, mais incidências de menos probabilidade são notadas com relação à reta da curva gaussiana. Na Figura 4.12 e na Figura 4.13 para a PCA e a ICA, os resultados são bem similares à distribuição do *Z-score*.

Para encontrar uma distribuição conjunta entre todos os dados, a norma euclidiana é aplicada sobre todas as distribuições, tendo seu resultado apresentado na Figura 4.14.

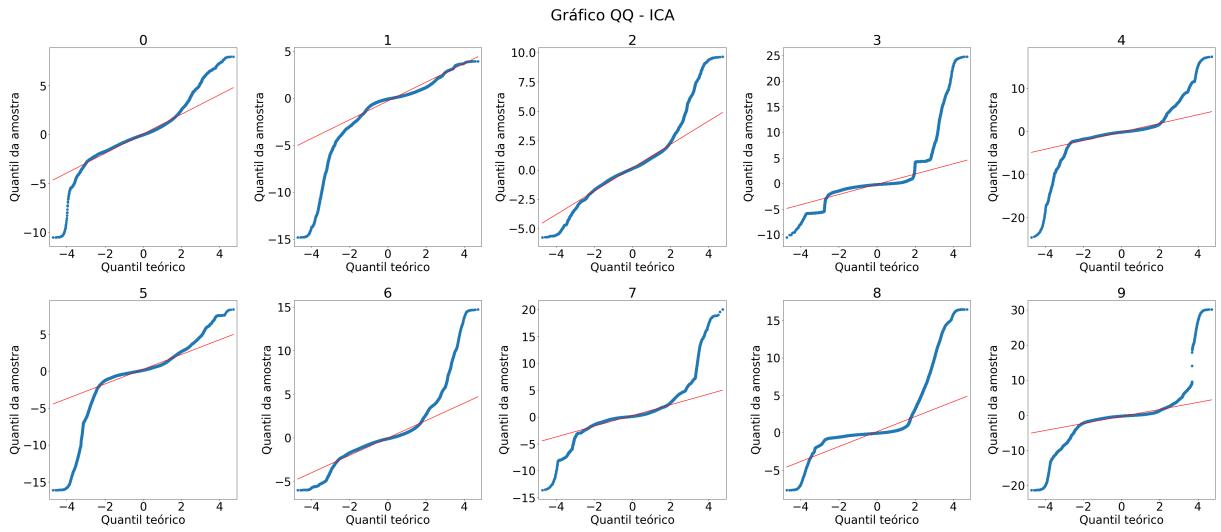


Figura 4.13: Gráfico Quantil-Quantil para as componentes da ICA.

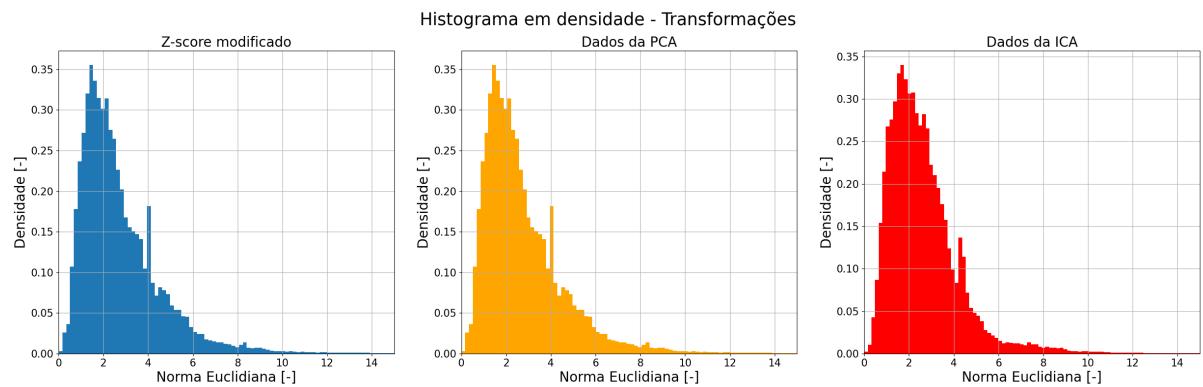


Figura 4.14: Aplicação da Norma euclidiana para as transformações.

Avalia-se que todas as distribuições são muito similares, não tendo grandes diferenças entre a distribuição do *Z-score* original e a distribuição das transformações da PCA e da ICA.

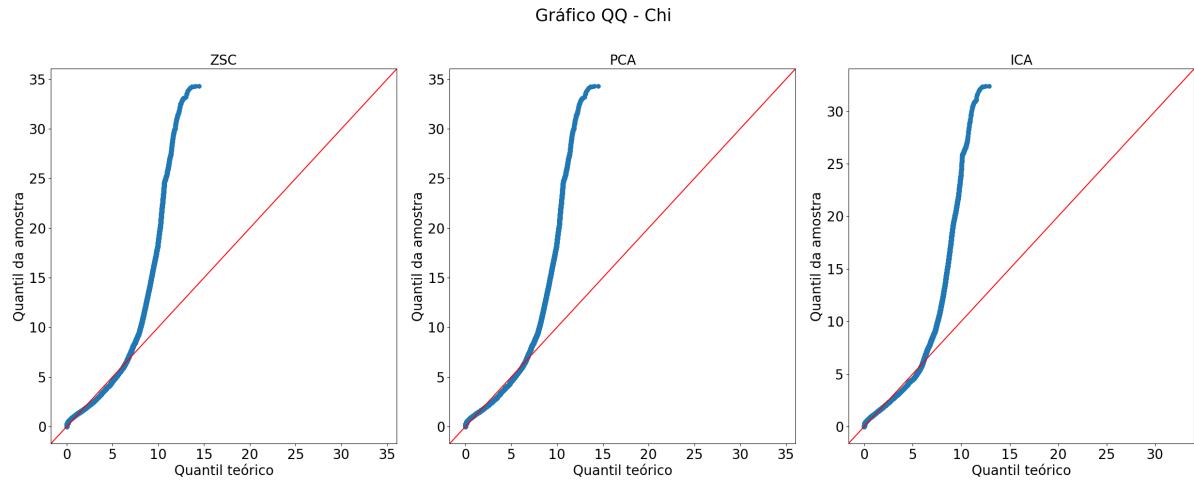


Figura 4.15: Gráfico Quantil-Quantil para a distribuição Chi do *Z-score* (ZSC), da PCA e da ICA.

Para avaliar se as distribuições resultantes possuem aderência a distribuição Chi, é possível plotar o Gráfico Q-Q. Na Figura 4.15 é possível avaliar que com o aumento do valor do quantil teórico a partir de 7, nota-se que a distribuição resultante começa a se distanciar da distribuição Chi esperada dos dados. O distanciamento ocorre, visualmente, com uma variação rápida, de forma que os quantis da amostra (curvas em azul) aumentem mais rápido que os quantis esperados (reta em vermelho). Como as distribuições são muito semelhantes entre si e o modelo de misturas gaussianas será executado sobre a norma euclidiana, será adotada a distribuição do *Z-score* modificado como sendo a distribuição utilizada nas próximas seções.

4.5 Resultados do ajuste do modelo de Misturas Gaussianas

Após selecionar a distribuição resultante da norma euclidiana do *Z-score*, é possível representá-la como uma mistura gaussiana. O número de estados s (que também é o número de gaussianas na mistura) foi escolhido conforme interpretabilidade e minimização do BIC e do AIC. A interpretabilidade é referente ao número de estados possíveis do modelo. Neste sentido, para manter um grau de significado do estado, os dois números testados foram 2 (estados “normal”, “anormal”) e 3 (estados “normal”, “incomum” e “anormal”). É possível

incluir mais estados, entretanto, esta adição implicaria em uma dificuldade de interpretar seus significados.

Tabela 4.2: Tabela de comparação entre os resultados do AIC e BIC.

Critério	Pontuação para duas gaussianas	Pontuação para três gaussianas
AIC	3207025	3089834
BIC	3207083	3089927

Uma vez selecionados 2 ou 3 estados, é possível utilizar um critério de comparação de modelos para decidir qual o melhor número. Para tal, é possível calcular os dois critérios propostos e fazer uma comparação entre eles. Esta comparação é realizada na Tabela 4.2, em que a preferência do modelo é para aquele de menor valor. Da tabela é possível notar que os valores de AIC e BIC para duas gaussianas são muito maiores que para três gaussianas. Neste sentido, a preferência seria ao modelo de três gaussianas. Além do mais, é possível notar que os valores de AIC e BIC não levaram a conclusões divergentes, pois possuem valores numéricos similares para os dois modelos propostos.

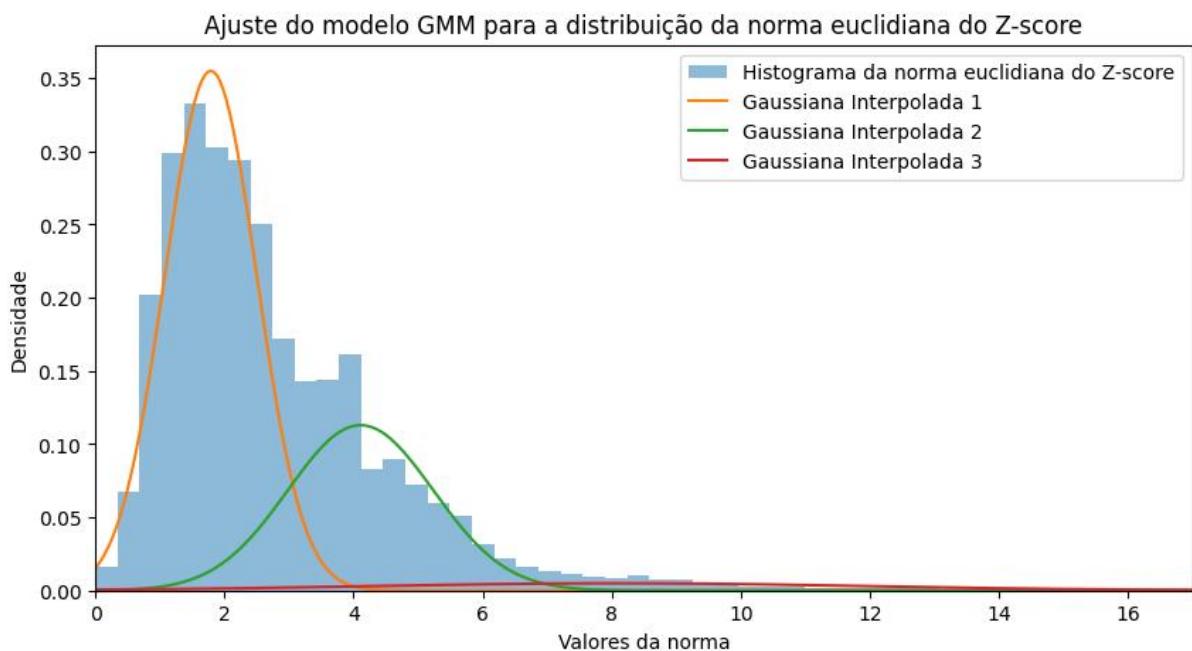


Figura 4.16: Histogramas para a norma euclidiana e suas misturas gaussianas correspondentes.

Na Figura 4.16 é apresentado o resultado do ajuste do modelo da mistura gaussiana. Nota-se que a gaussiana laranja cobre a maior parte de sua área de 0 até 4 desvios padrão,

aproximadamente, enquanto a vermelha cobre de 0 até 8 (porém com média menor) e a verde praticamente cobre todo o espaço restante. O peso da vermelha é de 64%, o da laranja, 31% e o da verde é 5%.

4.6 Resultados das Cadeias Gaussianas Ocultas de Markov

Após a adaptação do modelo GMM, é possível adaptar o modelo GHMM. Este recebe os parâmetros do modelo anterior e é treinado para os 57 conjuntos de dados de corridas recebidos. Para o modelo, um estado arbitrário 0 (Estado 0) indica se a bomba estava desligada ou não, e não é resultado do GHMM. O Estado 1 caracteriza o comportamento normal durante a operação da bomba, e pequenas anomalias da bomba são representadas no Estado 2. O Estado 3 está relacionado a anomalias graves. Esta separação entre “pequeno” e “grande” é dada pelo percentual do estado, que está associado ao peso das gaussianas.

Após explorar as identificações geradas das séries temporais, foi possível constatar que existem dois casos de incidência da falha. O primeiro é onde a remoção e a falha aparecem próximas e o outro é onde a falha é informada um considerável tempo antes da remoção. A corrida B-18 2 é um exemplo do primeiro caso e as corridas A-12 2 e A-29 2 são exemplos do segundo caso.

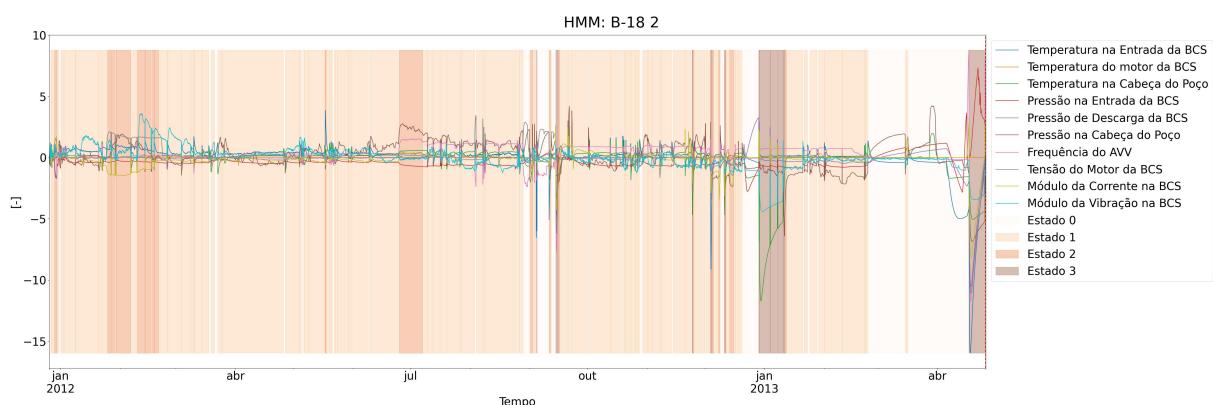


Figura 4.17: Dados da corrida B-18 2 inferidos de janeiro de 2012 até abril de 2013. O gráfico é feito com as variáveis originais padronizadas.

Um exemplo de identificação adequada é a corrida B-18 2, mostrada na Figura 4.17, onde é possível observar que o Estado 3 persiste antes da falha, representada pela linha vermelha tracejada. Neste sentido, a persistência deste estado, neste trecho, indica que um estado anormal estava acontecendo. Este resultado também pode ser avaliado pela magnitude do *Z-score* no período.

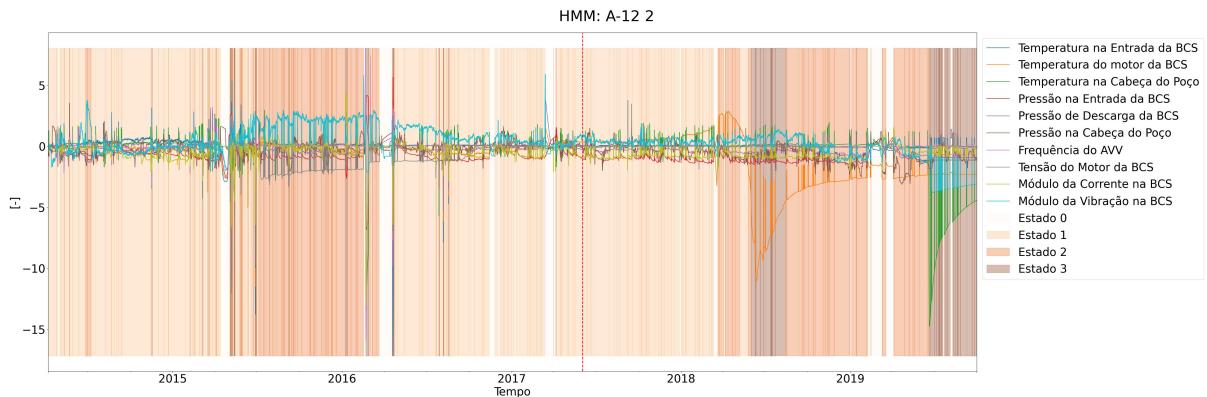


Figura 4.18: Dados da corrida A-12 2 inferidos de 2014 até 2019. O gráfico é feito com as variáveis originais.

Na corrida A-12 2, mostrada na Figura 4.18, no momento indicado pela linha tracejada vermelha, não há persistência clara do Estado 3, mas, em vez disso, uma persistência significativa do Estado 2 durante o ano anterior, 2016, e posteriormente, próximo ao final da série temporal, tanto o Estado 2 quanto o 3 permanecem ativos por longos períodos. Este é um exemplo de que, este método não necessariamente detecta a falha, mas ele detecta anomalias que podem levar à parada da bomba, visto que a não continuidade da série indica que a bomba foi removida.

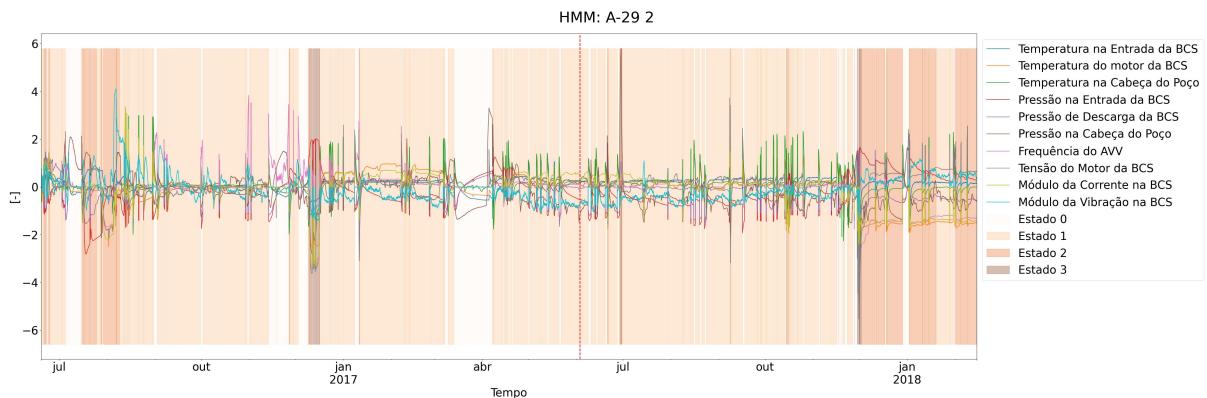


Figura 4.19: Dados da corrida A-29 2 inferidos de janeiro de 2016 até abril de 2018. O gráfico é feito com as variáveis originais.

Já por sua vez, a corrida A-29 2, ilustrada na Figura 4.19, mostra um caso onde o Estado 1 persiste no momento reportado da falha. No caso do final da série temporal, é possível notar que há uma persistência grande do Estado 2, sem grandes aparições significativas do Estado 3. Entretanto, é possível notar que o Estado 3 apareceu um pouco antes de janeiro de 2017, julho de 2017 e antes de janeiro de 2018, e que a anomalia detectada pode representar o indício da falha.

4.7 Resultado de múltiplos treinos

Nos exemplos apresentados anteriormente, as corridas foram selecionadas para ilustrar o resultado da GHMM nos dados. Após o treino do modelo e a aplicação deste para todas as corridas, é possível identificar quais são os estados mais frequentes nas 24 horas anteriores ao final do registro das séries temporais. O modelo utilizado foi baseado em uma referência de números randômicos de valor 19971215, que serve para inicializar o gerador de números randômicos da biblioteca *numpy*.



Figura 4.20: Gráfico de barras com os resultados da persistência de estados para o valor referência 19971215.

Como ilustrado na Figura 4.20, é possível avaliar que os estados mais recorrentes foram o 2 e o 3, que representavam comportamento incomum e anômalo. Entretanto, também é possível notar um número considerável de estados normais. O estado 2 possui 38% das incidências e o estado 3 possui 28%. Juntos, ambos representam 66% do total de estados.

Separando apenas o conjunto de testes, com relação a todas as corridas, é possível obter o gráfico de persistência da Figura 4.21. Nesta, é possível avaliar que os estados 2 e 3 são mais recorrentes que o estado 1.

Para levar em consideração as variações associadas à aleatoriedade das inicializações dos parâmetros, 30 execuções de treinamento foram realizadas com diferentes referências para número aleatório, indo de 19971215 a 20210505.

Com este resultado na Figura 4.22, é possível notar que os resultados permaneceram consistentes com a primeira execução, pois a proporção de estados anômalos continua maior

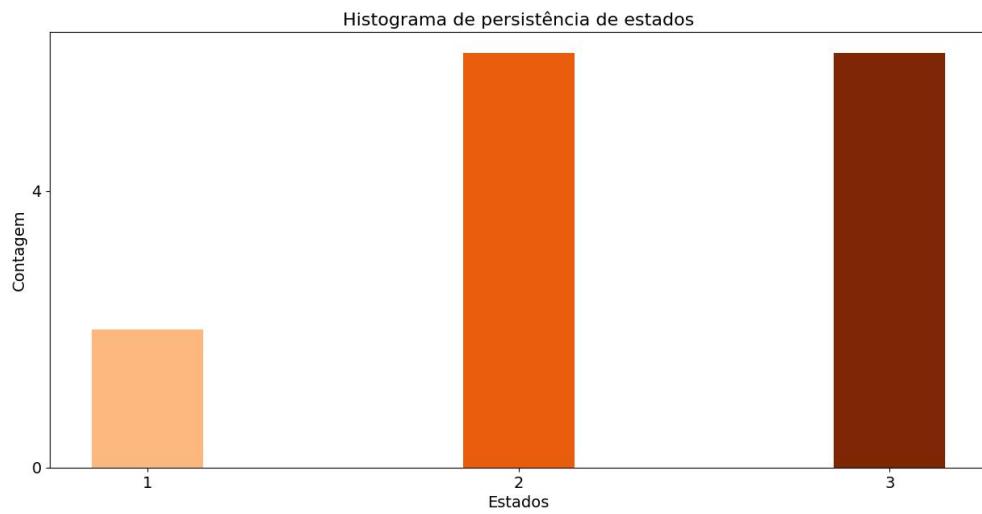


Figura 4.21: Gráfico de barras com os resultados da persistência para as amostras de teste.

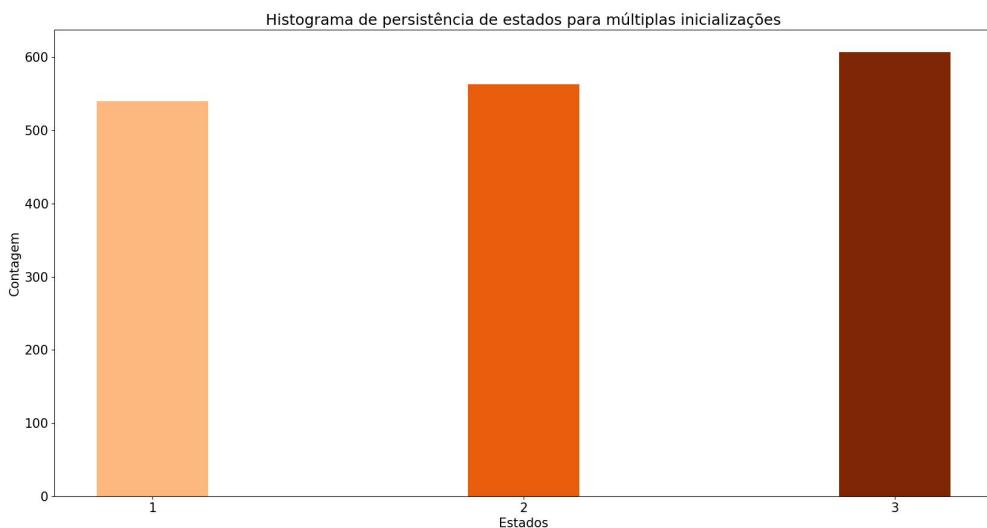


Figura 4.22: Gráfico de barras com os resultados da persistência de estados para múltiplas referências aleatórias.

que a proporção do estado normal. Não somente, nota-se que o Estado 3 é o de maior proporção entre todos. Posteriormente, para apenas as corridas de teste, nas diversas inicializações propostas, nota-se, da Figura 4.23, que a proporção de estados anômalos antes da remoção é maior do que a de estados normais.

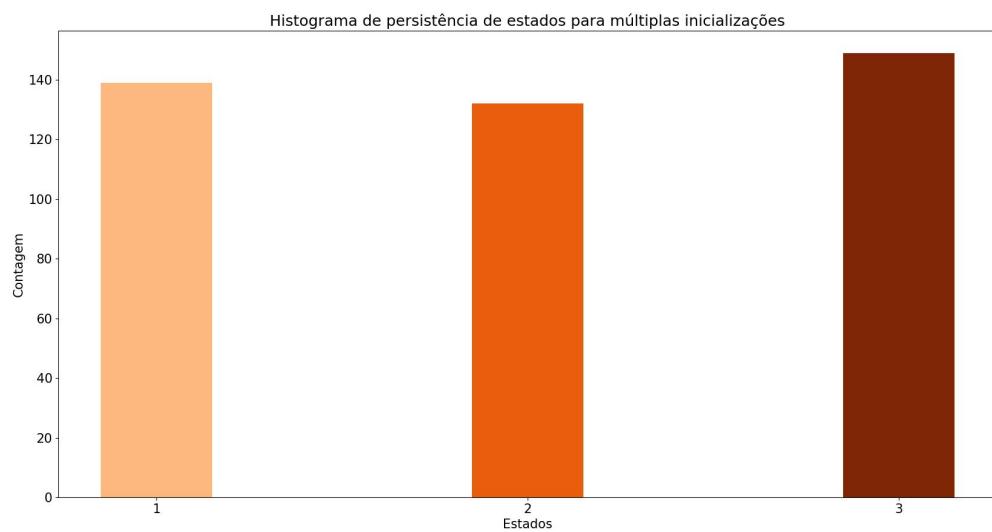


Figura 4.23: Gráfico de barras com os resultados da persistência de estados das amostras de teste para múltiplas referências aleatórias.

Capítulo 5

Conclusões e Trabalhos Futuros

Neste trabalho, foi apresentado uma metodologia de tratamento dos dados operacionais da empresa Equinor, utilizando-se de um *Z-score* modificado, aplicado a 57 informações de corridas. Após o tratamento, foram avaliadas as relações numéricas entre os dados através da Análise de Componentes Principais (PCA) e da Análise de Componentes Independentes (ICA). Posteriormente, foi aplicada a norma euclidiana nos dados, obtendo uma distribuição de dados que foi representada através de um Modelo de Misturas Gaussianas (GMM). Desta forma, com o ajuste deste modelo, o número de estados e condições iniciais foram inferidos para o Modelo de Cadeias Ocultas de Markov Gaussiano (GHMM) e foi possível se utilizar dele como método detector de anomalias.

Com relação ao tratamento de dados, foi possível concluir, que através desta metodologia, é possível colocar dados de escalas diferentes centralizando-os em torno do centro de dados. Com isto, as entradas que mais divergem desta referência são reveladas, sendo aquelas mais distantes, podendo ser consideradas como anomalias ao longo das séries temporais. No caso, na modificação do *Z-score*, foram utilizados os sinais filtrados por uma Média Móvel Exponencial (EWM) como valor a ser comparado, a mediana por expansão como centro do conjunto de dados, que é comumente utilizada para ter uma medida mais robusta do centro dos dados e o desvio-padrão como por expansão das variáveis medidas como escala, tendo sido avaliado que o desvio-padrão possui comportamento mais suave ao longo da expansão, em relação ao Desvio Absoluto Médio (MAD).

Posteriormente, com os dados transformados pelo *Z-score* modificado, foi proposta a utilização da PCA e da ICA para encontrar relações numéricas entre os dados. De forma geral, foi possível concluir que os dados coletados possuem, de fato, relações entre si. As

relações são possíveis de serem observadas visualmente, porém, estes algoritmos atestam e quantificam estas relações. No caso da PCA, avaliam-se as relações de correlação ou covariância lineares entre os sinais, isto é, como as variáveis originais constroem as componentes através de somas ponderadas.

Nesta tese, foi utilizada a análise das componentes principais pela matriz de covariância e pela matriz de correlação. Para a matriz de covariância, foi possível concluir que é possível detectar uma relação já conhecida entre a pressão de entrada e de descarga, que é a pressão diferencial. Entretanto, como dito, a maneira mais recomendada (Jolliffe, 2006) é utilizar a matriz de correlação, em que os dados são transformadas para ter a mesma escala. Nesta abordagem com a correlação, também foi possível encontrar a pressão diferencial, porém correlacionada com a temperatura da BCS , pressão na cabeça do poço, média da corrente e tensão da BCS (componente 6). Foi possível observar uma relação entre corrente e tensão (componente 7), como é esperado, pois elas estão relacionadas. Também é possível notar uma relação inversa entre a frequência do AVV e a tensão do motor da BCS (componente 9). Desta forma, conclui-se que, a partir dos resultados obtidos, é possível perceber que a matriz de correlação encontra mais relações do que a matriz de covariância.

Como continuação da avaliação das transformações, a ICA é utilizada para tentar prover componentes que evidenciem as falhas considerando a independência estatística entre os sinais. Com a ICA, é possível notar que a maior parte dos pesos que compõem as matrizes são próximos de zero, enquanto na PCA, existem mais pesos diferentes de zero. Neste sentido, a ICA encontrou outras relações a serem observadas, como por exemplo, a componente 2 que relaciona o aumento da vibração com o aumento da pressão de entrada. Outra relação possível de ser notada é na componente 5, que revela o aumento da pressão de descarga enquanto a pressão na cabeça do poço diminui. Desta maneira, com a utilização dos métodos da PCA e da ICA é possível concluir que os dados operacionais da bomba possuem relações entre si. Entretanto, diferentes métodos levam a diferentes relações observadas.

Posteriormente, no gráfico Q-Q, é possível avaliar as distribuições entre as variáveis e notar que, mesmo depois das transformações, nenhuma delas atinge uma aderência à distribuição gaussiana nas regiões de cauda. Isto mostra que, de certa forma, os fenômenos mais extremos têm uma probabilidade maior de ocorrer do que se fossem supostos por uma distribuição gaussiana.

Como o objetivo principal desta dissertação é encontrar anomalias, foi aplicada a norma euclidiana sobre cada entrada do sinal, avaliando a dimensão do conjunto medido naquele instante. Este procedimento leva em consideração que o *Z-score* indica a intensidade do desvio da amostra em torno da média da variável, e considerando a hipótese de anomalia como sendo um evento improvável, a norma euclidiana obtém um número que é relativo à intensidade do desvio geral na entrada. A distribuição resultante da norma euclidiana sobre variáveis escaladas pelo *Z-score* é a distribuição Chi.

Ao aplicar a norma euclidiana para o *Z-score* modificado e as séries temporais geradas pela PCA e a ICA, foi possível avaliar que não há muita diferença entre as distribuições. No gráfico Q-Q para a distribuição Chi, Figura 4.15, há uma grande divergência de valores a partir do quantil teórico 6, o que mostra uma baixa similaridade a distribuição Chi a partir deste quantil. Este resultado era esperado, pois a maioria dos sinais anteriores possuíam cauda pesada.

Para o cômputo dos estados, foi preferida a série apenas do *Z-score*, sem nenhuma transformação. Da forma proposta nesta dissertação, as transformações da PCA e da ICA contribuíram para a identificação de relações, mas não contribuíram de forma clara para a identificação de anomalias em suas distribuições finais. Com isto, utilizou-se apenas a distribuição resultante da norma euclidiana expressada através do Modelo de Misturas Gaussianas.

O GMM é utilizado como estimativa dos parâmetros iniciais para o Modelo de Cadeias Ocultas Gaussianas. Através dos parâmetros BIC e AIC, junto com uma análise da distribuição, foi possível concluir que a distribuição dos dados pode ser representada pela composição de três gaussianas principais. A cada uma dessas três gaussianas é associada a inicialização de três estados, que serão interpretados entre normal, incomum e anormal. Com isto, é possível executar o treinamento do modelo GHMM.

Após o treinamento, é possível aplicar o modelo para as corridas e detectar os estados. Do gráfico de persistência, nota-se que o modelo possui mais entradas incomuns e anormais do que entradas normais próximas a região da falha. Assim, conclui-se que o modelo é capaz de identificar anomalias próximas a falha. Entretanto, nem todas as remoções das bombas são antecedidas por estados anômalos, possuindo uma considerável incidência de estados normais.

Com relação à comparação destes resultados à literatura relativa às BCS, é difícil estabelecer uma comparação, visto que a mesma possui poucos tópicos sobre falhas desconhecidas, ou como encontrar anormalidades nos dados e associá-los às falhas. Com

relação a este tema, os autores citados, anteriormente, no Capítulo 1 possuem conhecimento sobre a ocorrência da falha em si e a associam à possibilidade de uma anomalia nos dados, o que gera um registro no tempo de quando a falha ocorreu. Muitos desses dados podem estar sendo gerados em ambientes controlados, possibilitando medir e controlar fatores que podem induzir a falha. Além disso, por estes motivos, as abordagens utilizadas pelos autores envolvem treinamento rotulado, implicando que a maior parte dos métodos são relacionados a aprendizado supervisionado. Outro aspecto importante é que os registros de falhas foram compilados por operadores de campo, que podem não ter registrado a falha no momento exato em que ela aconteceu. Portanto, comparar os estados detectados pelo modelo com os tempos de falha relatados pode nem sempre ser apropriado.

No entanto, apesar das limitações do GHMM, ele ainda é capaz de indicar ou fornecer uma caracterização da condição da bomba sem exigir classificação prévia. Além disso, oferece flexibilidade para lidar com diversos tipos de dados e pode ser adaptado a outros contextos. Neste trabalho, o uso de uma transformação para a norma euclidiana foi preferido à adoção da estrutura multidimensional completa do Modelo Gaussiano de Markov Oculto (GHMM).

Nesta tese, são propostas duas sugestões para trabalhos futuros. A primeira é com relação a utilização de dados de produção. Em alguns casos, hipotetiza-se que a falha pode ter sido registrada não devido a um comportamento anômalo da bomba, mas sim devido a anomalias na produção. Entretanto, no contexto desta dissertação, não temos estes dados, ficando como sugestão a utilização destes em trabalhos futuros. A segunda sugestão é com relação a tentar utilizar a estrutura multidimensional dos dados, sem a norma euclidiana, proposta nesta dissertação, para simplificação do problema. Neste caso, para Cadeias Ocultas de Markov, o modelo proposto poderia ser o mesmo (porém considerando as entradas multidimensionais), ou, dependendo do tipo de variável, um Modelo de Mistura Gaussiana associado ao Modelo de Markov Oculto (GMM-HMM), no qual múltiplas misturas gaussianas representam os estados ocultos.

Por fim, conclui-se que o método proposto demonstrou ser eficaz na identificação de padrões e na caracterização de estados operacionais em dados ESP, mesmo na ausência de rotulagem prévia. Sua natureza não supervisionada, combinada com a flexibilidade da estrutura HMM, permitiu *insights* significativos sobre o comportamento da bomba. Isso confirma a viabilidade da abordagem como uma ferramenta útil para a detecção precoce de anomalias e o monitoramento de condições em cenários reais.

Referências bibliográficas

- Abdalla, R. et al. Machine Learning Approach for Predictive Maintenance of the Electrical Submersible Pumps (ESPs). **ACS Omega**, v. 7, n. 21, p. 17641–17651, 2022. DOI: 10.1021/acsomega.1c05881. eprint: <https://doi.org/10.1021/acsomega.1c05881>. Disponível em: <<https://doi.org/10.1021/acsomega.1c05881>>.
- Ajami, A.; Daneshvar, M. Data driven approach for fault detection and diagnosis of turbine in thermal power plant using Independent Component Analysis (ICA). **International Journal of Electrical Power and Energy Systems**, v. 43, n. 1, p. 728–735, 2012. ISSN 0142-0615. DOI: <https://doi.org/10.1016/j.ijepes.2012.06.022>. Disponível em: <<https://www.sciencedirect.com/science/article/pii/S0142061512002840>>.
- Akaike, H. A new look at the statistical model identification. **IEEE Transactions on Automatic Control**, v. 19, n. 6, p. 716–723, 1974. DOI: 10.1109/TAC.1974.1100705.
- Amari, S.-i.; Cichocki, A.; Yang, H. A New Learning Algorithm for Blind Signal Separation. In: Touretzky, D.; Mozer, M.; Hasselmo, M. (Ed.). **Advances in Neural Information Processing Systems**. [S.l.]: MIT Press, 1995. v. 8. Disponível em: <https://proceedings.neurips.cc/paper_files/paper/1995/file/e19347e1c3ca0c0b97de5fb3b690855a-Paper.pdf>.
- Awsan, M. Data driven-based model for predicting pump failures in the oil and gas industry. **Engineering Failure Analysis**, v. 145, p. 107019, 2023. ISSN 1350-6307. DOI: <https://doi.org/10.1016/j.engfailanal.2022.107019>. Disponível em: <<https://www.sciencedirect.com/science/article/pii/S1350630722009864>>.
- Axler, S. Eigenvalues and Eigenvectors. In: LINEAR Algebra Done Right. Cham: Springer International Publishing, 2024. P. 132–180. ISBN 978-3-031-41026-0. DOI: 10.1007/978-3-031-41026-0_5. Disponível em: <https://doi.org/10.1007/978-3-031-41026-0_5>.
- Balbinot, A.; Brusamarello, V. **Instrumentação e Fundamentos de Medidas - Vol. 1**. [S.l.]: LTC, 2019. ISBN 9788521635833. Disponível em: <<https://books.google.com.br/books?id=zAz6zwEACAAJ>>.
- Baum, L. E.; Petrie, T.; Soules, G.; Weiss, N. A Maximization Technique Occurring in the Statistical Analysis of Probabilistic Functions of Markov Chains. **The Annals of Mathematical Statistics**, Institute of Mathematical Statistics, v. 41, n. 1, p. 164–171, 1970. DOI: 10.1214/aoms/1177697196. Disponível em: <<https://doi.org/10.1214/aoms/1177697196>>.

Bendat, J. S.; Piersol, A. G. **Random Data: Analysis and Measurement Procedures.** [S.l.]: John Wiley e Sons, Inc, 2010. (Wiley Series in Probability and Statistics). ISBN 9781118032428. Disponível em: <<https://onlinelibrary.wiley.com/doi/book/10.1002/9781118032428>>.

Bishop, C. M. **Pattern Recognition and Machine Learning.** 1. ed. [S.l.]: Springer New York, NY, 2006. ISBN 978-0-387-31073-2.

Brown, R. **Exponential Smoothing for Predicting Demand.** [S.l.]: Little, 1956. Disponível em: <https://books.google.com.br/books?id=Eo_rMgEACAAJ>.

Burnham, K. P.; Anderson, D. R. **Model Selection and Multimodel Inference.** 1. ed. [S.l.]: Springer Nature, 2002. ISBN 9780387224565. DOI: <https://doi.org/10.1007/b97636>.

Comon, P. Independent component analysis, A new concept? **Signal Processing**, v. 36, n. 3, p. 287–314, 1994. Higher Order Statistics. ISSN 0165-1684. DOI: [https://doi.org/10.1016/0165-1684\(94\)90029-9](https://doi.org/10.1016/0165-1684(94)90029-9). Disponível em: <<https://www.sciencedirect.com/science/article/pii/0165168494900299>>.

Cover, T. M.; Thomas, J. A. **Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing).** USA: Wiley-Interscience, 2006. ISBN 0471241954.

Dempster, A. P.; Laird, N. M.; Rubin, D. B. Maximum likelihood from incomplete data via the EM algorithm. **J. Roy. Statist. Soc. Ser. B**, v. 39, n. 1, p. 1–38, 1977. With discussion. ISSN 0035-9246. Disponível em: <[http://links.jstor.org/sici?&sici=0035-9246\(1977\)39:1%3C1:MLFIDV%3E2.0.CO;2-Z&origin=MSN](http://links.jstor.org/sici?&sici=0035-9246(1977)39:1%3C1:MLFIDV%3E2.0.CO;2-Z&origin=MSN)>.

Devold, H. **Oil and Gas Production Handbook: An Introduction to Oil and Gas Production.** [S.l.]: Lulu.com, 2013. ISBN 9781105538643. Disponível em: <<https://books.google.com.br/books?id=nJ2XAwAAQBAJ>>.

Elliott, R. J.; Moore, J. B.; Aggoun, L. **Hidden Markov Models - Estimation and Control.** 1. ed. [S.l.]: Springer New York, 2008. ISBN 978-0-387-84854-9.

Fakher, S.; Khlaifat, A.; Hossain, M. E.; Nameer, H. Rigorous review of electrical submersible pump failure mechanisms and their mitigation measures. **Journal of Petroleum Exploration and Production Technology**, v. 11, n. 10, p. 3799–3814, out. 2021. ISSN 2190-0566. DOI: [10.1007/s13202-021-01271-6](https://doi.org/10.1007/s13202-021-01271-6). Disponível em: <<https://doi.org/10.1007/s13202-021-01271-6>>.

Forbes, C.; Evans, M.; Hastings, N.; Peacock, B. **Statistical Distributions.** [S.l.]: Wiley, 2011. ISBN 9781118097823. Disponível em: <<https://books.google.com.br/books?id=YhF1osrQ4psC>>.

Gerón, A. **Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow.** 2. ed. [S.l.]: O'Reilly Media, Inc., 2019. ISBN 978-1-492-03264-9.

Gori, M.; Betti, A.; Melacci, S. **Machine Learning - A Constraint-Based Approach.** 2. ed. [S.l.]: Elsevier, 2023. ISBN 978-0-323-89859-1.

- Hodge, V. J.; Austin, J. A Survey of Outlier Detection Methodologies. **Artificial Intelligence Review**, v. 22, n. 2, p. 85–126, out. 2004. ISSN 1573-7462. DOI: 10 . 1007 / s10462 - 004 - 4304 - y. Disponível em: <<https://doi.org/10.1007/s10462-004-4304-y>>.
- Hyvarinen, A. Fast and robust fixed-point algorithms for independent component analysis. **IEEE Transactions on Neural Networks**, v. 10, n. 3, p. 626–634, 1999. DOI: 10 . 1109 / 72 . 761722.
- Hyvärinen, A.; Karhunen, J.; Oja, E. **Independent Component Analysis**. 1. ed. [S.l.]: John Wiley e Sons, Ltd, 2001. ISBN 9780471221319. DOI: <https://doi.org/10.1002/0471221317>. eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1002/0471221317>. Disponível em: <<https://onlinelibrary.wiley.com/doi/abs/10.1002/0471221317>>.
- Jolliffe, I. T. **Principal Component Analysis**. 2. ed. [S.l.]: Springer New York, NY, 2006. ISBN 978-0-387-22440-4.
- Kader, G. D. Means and MADs. **Mathematics Teaching in the Middle School**, National Council of Teachers of Mathematics, Reston VA, USA, v. 4, n. 6, p. 398–403, 1999. DOI: 10 . 5951 / MTMS . 4 . 6 . 0398. Disponível em: <<https://pubs.nctm.org/view/journals/mtms/4/6/article-p398.xml>>.
- Kay, S. M. **Intuitive Probability and Random Processes using MATLAB®**. 1. ed. [S.l.]: Springer New York, NY, 2006. ISBN 978-0-387-24158-6.
- [S.l.]. **Electric Submersible Pump Installation and Commissioning - Challenges and Lesson Learned from Field Development**. [S.l.: s.n.], abr. 2015. SPE Saudi Arabia Section Annual Technical Symposium and Exhibition. spe-177990-ms. DOI: 10 . 2118 / 177990 - MS. eprint: <https://onepetro.org/SPESATS/proceedings-pdf/15SATS/15SATS/SPE-177990-MS/1460468/spe-177990-ms.pdf>. Disponível em: <<https://doi.org/10.2118/177990-MS>>.
- Kotu, V.; Bala, D. **Data Science - Concepts and Practice**. 1. ed. [S.l.]: Morgan Kaufmann, 2019. ISBN 978-0-128-14761-0.
- Krollner, B.; Vanstone, B.; Finnie, G. Financial time series forecasting with machine learning techniques: A survey. English. In: PROCEEDINGS of the 18th European Symposium on Artificial Neural Networks (ESANN 2010). [S.l.: s.n.], 2010. P. 25–30. European Symposium on Artificial Neural Networks : Computational Intelligence and Machine Learning, ESANN 2010 ; Conference date: 28-04-2010 Through 30-04-2010. ISBN 2930307102.
- Kuha, J. AIC and BIC: Comparisons of Assumptions and Performance. **Sociological Methods & Research**, v. 33, n. 2, p. 188–229, 2004. DOI: 10 . 1177 / 0049124103262065. eprint: <https://doi.org/10.1177/0049124103262065>. Disponível em: <<https://doi.org/10.1177/0049124103262065>>.
- Kullback, S.; Leibler, R. A. On Information and Sufficiency. **The Annals of Mathematical Statistics**, Institute of Mathematical Statistics, v. 22, n. 1, p. 79–86, 1951. DOI: 10 . 1214 / aoms / 1177729694. Disponível em: <<https://doi.org/10.1214/aoms/1177729694>>.

Lastra, R. A.; Xiao, J. Machine Learning Engine for Real-Time ESP Failure Detection and Diagnostics. Day 2 Wed, October 26, 2022, d021s007r001, out. 2022. DOI: 10.2118/206935-MS. eprint: <https://onepetro.org/SPEMEAL/proceedings-pdf/20MEAL/2-20MEAL/D021S007R001/3021607/spe-206935-ms.pdf>. Disponível em: <<https://doi.org/10.2118/206935-MS>>.

Leukel, J.; González, J.; Riekert, M. Adoption of machine learning technology for failure prediction in industrial maintenance: A systematic review. **Journal of Manufacturing Systems**, v. 61, p. 87–96, 2021. ISSN 0278-6125. DOI: <https://doi.org/10.1016/j.jmsy.2021.08.012>. Disponível em: <<https://www.sciencedirect.com/science/article/pii/S0278612521001849>>.

Mamon, R. S.; Elliott, R. J. (Ed.). **Hidden Markov Models in Finance**. 1. ed. [S.l.]: Springer New York, 2010. ISBN 978-0-387-71163-8.

McKinney, W.; She, C.; Roeschke, M.; Van den Bossche, J. **pandas-dev/pandas: Pandas**. [S.l.: s.n.], 2024. <https://github.com/pandas-dev/pandas>. Version 2.2.2 or latest. Accessed: 2025-06-17.

McKinney, W.; She, C.; Roeschke, M.; Van den Bossche, J. **pandas-dev/pandas: Pandas**. [S.l.]: Zenodo, set. 2024. DOI: 10.5281/zenodo.13819579. Disponível em: <<https://doi.org/10.5281/zenodo.13819579>>.

Morettin, P.; Bussab, W. **ESTATÍSTICA BÁSICA**. [S.l.]: Editora Saraiva, 2017. ISBN 9788502207172. Disponível em: <<https://books.google.com.br/books?id=vDhnDwAAQBAJ>>.

Pedregosa, F. et al. Scikit-learn: Machine Learning in Python. **Journal of Machine Learning Research**, v. 12, p. 2825–2830, 2011.

Rabiner, L. A tutorial on hidden Markov models and selected applications in speech recognition. **Proceedings of the IEEE**, v. 77, n. 2, p. 257–286, 1989. DOI: 10.1109/5.18626.

Schlichthärle, D. **Digital Filters - Basics and Design**. 2. ed. [S.l.]: Springer Berlin, Heidelberg, 2011. ISBN 978-3-642-14325-0.

Schwarz, G. Estimating the Dimension of a Model. **The Annals of Statistics**, Institute of Mathematical Statistics, v. 6, n. 2, p. 461–464, 1978. DOI: 10.1214/aos/1176344136. Disponível em: <<https://doi.org/10.1214/aos/1176344136>>.

Shannon, C. E. A mathematical theory of communication. **The Bell System Technical Journal**, v. 27, n. 3, p. 379–423, 1948. DOI: 10.1002/j.1538-7305.1948.tb01338.x.

Smyth, P. Hidden Markov models for fault detection in dynamic systems. **Pattern Recognition**, v. 27, n. 1, p. 149–164, 1994. ISSN 0031-3203. DOI: [https://doi.org/10.1016/0031-3203\(94\)90024-8](https://doi.org/10.1016/0031-3203(94)90024-8). Disponível em: <<https://www.sciencedirect.com/science/article/pii/0031320394900248>>.

Speight, J. G. CHAPTER 2 - Origin and Production. In: Speight, J. G. (Ed.). **Natural Gas**. [S.l.]: Gulf Publishing Company, 2007. P. 35–59. ISBN 978-1-933762-14-2. DOI: <https://doi.org/10.1016/j.jmsy.2021.08.012>.

org/10.1016/B978-1-933762-14-2.50007-8. Disponível em: <<https://www.sciencedirect.com/science/article/pii/B9781933762142500078>>.

Stanke, M.; Waack, S. Gene prediction with a hidden Markov model and a new intron submodel. **Bioinformatics**, v. 19, p. 215–225, set. 2003. ISSN 1367-4803. DOI: 10.1093/bioinformatics/btg1080.

Takacs, G. **Electrical Submersible Pumps Manual**. 2. ed. [S.l.]: Gulf Professional Publishing, 2018. ISBN 978-0-128-14570-8.

Thode, H. **Testing For Normality**. [S.l.]: CRC Press, 2002. (Statistics, textbooks and monographs). ISBN 9780203910894. Disponível em: <<https://books.google.com.br/books?id=gbegXB4SdosC>>.

Viterbi, A. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. **IEEE Transactions on Information Theory**, v. 13, n. 2, p. 260–269, 1967. DOI: 10.1109/TIT.1967.1054010.

Yang, P.; Chen, J.; Wu, L.; Li, S. Fault Identification of Electric Submersible Pumps Based on Unsupervised and Multi-Source Transfer Learning Integration. **Sustainability**, v. 14, n. 16, 2022. ISSN 2071-1050. DOI: 10.3390/su14169870. Disponível em: <<https://www.mdpi.com/2071-1050/14/16/9870>>.

Yang, P.; Chen, J.; Zhang, H.; Li, S. A Fault Identification Method for Electric Submersible Pumps Based on DAE-SVM. **Shock and Vibration**, v. 2022, n. 1, p. 5868630, 2022. DOI: <https://doi.org/10.1155/2022/5868630>. eprint: <<https://onlinelibrary.wiley.com/doi/pdf/10.1155/2022/5868630>>. Disponível em: <<https://onlinelibrary.wiley.com/doi/abs/10.1155/2022/5868630>>.

Ypma, A.; Tax, D.; Duin, R. Robust machine fault detection with independent component analysis and support vector data description. In: NEURAL Networks for Signal Processing IX: Proceedings of the 1999 IEEE Signal Processing Society Workshop (Cat. No.98TH8468). [S.l.: s.n.], 1999. P. 67–76. DOI: 10.1109/NNSP.1999.788124.

Zhou, Z.-J. et al. A model for real-time failure prognosis based on hidden Markov model and belief rule base. **European Journal of Operational Research**, v. 207, n. 1, p. 269–283, 2010. ISSN 0377-2217. DOI: <https://doi.org/10.1016/j.ejor.2010.03.032>. Disponível em: <<https://www.sciencedirect.com/science/article/pii/S0377221710002596>>.

Apêndice A

Demonstração das equações de atualização

A.1 Equação de recorrência da média

Como estabelecido anteriormente, a média (Equação 3.4) pode ser definida como um somatório dos dados disponíveis, ponderados por $\frac{1}{k}$, onde k é o total de dados. Caso um novo dado seja incluído, o novo número será de $k + 1$ entradas. Neste sentido, a média se torna:

$$\mu_{k+1} = \sum_{i=0}^{k+1} \frac{1}{k+1} x_i \quad (\text{A.1})$$

Entretanto, é verdade que:

$$\mu_{k+1} = \frac{1}{k+1} \sum_{i=0}^k x_i + \frac{1}{k+1} x_{k+1} \quad (\text{A.2})$$

e considerando que $\sum_{i=0}^k x_i = k\mu_k$, é possível concluir que:

$$\mu_{k+1} = \frac{k}{k+1} \mu_k + \frac{1}{k+1} x_{k+1} \quad (\text{A.3})$$

e reescrevendo os termos:

$$\mu_{k+1} = \left(1 - \frac{1}{k+1}\right) \mu_k + \frac{1}{k+1} x_{k+1} \quad (\text{A.4})$$

ou, para um termo anterior:

$$\mu_k = \left(1 - \frac{1}{k}\right)\mu_{k-1} + \frac{1}{k}x_k \quad (\text{A.5})$$

sendo esta a equação de recorrência para a média, conforme apresentado em Equação 3.25.

A.2 Equação de recorrência do desvio-padrão

Analizando novamente a Equação 3.7, a esperança do quadrado das variáveis em um instante k é igual a $\sigma_k^2 + \mu_k^2$, podendo ser escrito:

$$\sum_{i=0}^k \frac{1}{k} x_i^2 = \sigma_k^2 + \mu_k^2 \quad (\text{A.6})$$

e para um incremento em k :

$$\sum_{i=0}^{k+1} \frac{1}{k+1} x_i^2 = \sigma_{k+1}^2 + \mu_{k+1}^2 = \frac{1}{k+1} \sum_{i=0}^k x_i^2 + \frac{1}{k+1} x_{k+1}^2 \quad (\text{A.7})$$

de maneira que seja possível desenvolver o somatório relativo ao termo anterior como $\sum_{i=0}^k x_i^2 = k(\sigma_k^2 + \mu_k^2)$. Com isto, obtém-se:

$$\sigma_{k+1}^2 + \mu_{k+1}^2 = \frac{1}{k+1} x_{k+1}^2 + \frac{k}{k+1}(\sigma_k^2 + \mu_k^2) \quad (\text{A.8})$$

No caso, é possível notar que o desvio-padrão no instante $k+1$ não está isolado. Para isolá-lo na equação é preciso seguir com:

$$\sigma_{k+1}^2 = \frac{1}{k+1} x_{k+1}^2 + \frac{k}{k+1}(\sigma_k^2 + \mu_k^2) - \mu_{k+1}^2 \quad (\text{A.9})$$

e depois tirar a raíz:

$$\sigma_{k+1} = \left(\frac{1}{k+1} x_{k+1}^2 + \frac{k}{k+1} \sigma_k^2 + \mu_k^2 - \mu_{k+1}^2 \right)^{\frac{1}{2}} \quad (\text{A.10})$$

e em seguida, analisando a equação para o instante k , obtém-se a Equação 3.27:

$$\sigma_k = \left(\frac{1}{k} x_k^2 + \left(1 - \frac{1}{k}\right) (\sigma_{k-1}^2 + \mu_{k-1}^2) - \mu_k^2 \right)^{\frac{1}{2}} \quad (\text{A.11})$$

Apêndice B

Algoritmo de Maximização da Expectativa

O Algoritmo de Maximização da Expectativa (EM) tem por objetivo aumentar a estimativa de máxima verossimilhança entre as observações e os parâmetros do modelo. Neste sentido, a fundamentação matemática do algoritmo é proposta neste apêndice.

A premissa inicial é que, considerando os parâmetros do modelo, a distribuição da observação pode ser encontrada integrando a distribuição conjunta da variável aleatória atrelada à observação e dos estados:

$$P(X | \theta) = \int p(X, S | \theta) dS \quad (\text{B.1})$$

onde S é o estado, X é a variável aleatória e θ são os parâmetros do modelo. Entretanto, obter diretamente $p(X, S | \theta)$ pode ser complicado. Considerando a regra de Bayes (Morettin; Bussab, 2017), é possível escrever:

$$P(S | X, \theta)P(X | \theta) = P(X | S, \theta)P(S | \theta) \quad (\text{B.2})$$

onde S é o estado, X é a variável aleatória e θ são os parâmetros do modelo. Neste sentido, reescrevendo a equação anterior, é possível obter:

$$P(X | \theta) = P(S | \theta) \frac{P(X | S, \theta)}{P(S | X, \theta)} \quad (\text{B.3})$$

de tal forma que se isola o termo desejado. Posteriormente, multiplica-se o numerador e denominador da fração por $q(S)$, a distribuição oculta do estado, isto é, a distribuição intrínseca do estado oculto, ou seja:

$$P(X | \theta) = \frac{P(X | S, \theta)P(S | \theta)}{P(S | X, \theta)} \frac{q(S)}{q(S)} \quad (\text{B.4})$$

Em seguida, é possível remanejar como:

$$P(X | \theta) = \frac{P(X | S, \theta)P(S | \theta)}{q(S)} \left(\frac{P(S | X, \theta)}{q(S)} \right)^{-1} \quad (\text{B.5})$$

Na literatura, torna-se mais conveniente a utilização da função logarítmica. Isto acontece pela propriedade de termos multiplicativos de um produtório se tornarem um somatório ao ter a função logarítmica aplicada:

$$\log(P(X | \theta)) = \log\left(\frac{P(X | S, \theta)P(S | \theta)}{q(S)}\right) - \log\frac{P(S | X, \theta)}{q(S)} \quad (\text{B.6})$$

Com isto, é possível integrar ao longo de $q(S)$, buscando obter implicitamente a integral conjunta sobre os estados:

$$\int \log(P(X | \theta))q(s)ds = \int \log\left(\frac{P(X | S, \theta)P(S | \theta)}{q(S)}\right)q(s)ds - \int \log\frac{P(S | X, \theta)}{q(S)}q(s)ds \quad (\text{B.7})$$

e como $\log(P(X | \theta))$ é constante perante s , ele pode ser retirado da integral no lado esquerdo da equação. Desta forma, o resultado da integral é 1, pois integra a distribuição $q(s)$. Portanto, é possível escrever:

$$\log(P(X | \theta)) = \int \log\left(\frac{P(X | S, \theta)P(S | \theta)}{q(S)}\right)q(s)ds - \int \log\frac{P(S | X, \theta)}{q(S)}q(s)ds \quad (\text{B.8})$$

Posteriormente, é possível notar que o termo que subtrai na equação, corresponde à chamada Divergência de Kullback-Leibler (Kullback; Leibler, 1951). Esta medida compara duas distribuições de probabilidade distintas e provê uma medida do desvio entre ambas. Ela é definida matematicamente como:

$$KL(q(s) \parallel P(S | X, \theta)) = - \int \log \frac{P(S | X, \theta)}{q(s)} q(s) ds \quad (\text{B.9})$$

de tal maneira, que substituindo na equação da integral, é possível obter:

$$\log(P(X | \theta)) = \int \log\left(\frac{P(X | S, \theta)P(S | \theta)}{q(s)}\right) q(s) ds + KL(q(s) \parallel P(S | X, \theta)) \quad (\text{B.10})$$

Portanto, a ideia é maximizar $\log(P(X | \theta))$ através do controle de θ . Desta forma, a propostação do algoritmo é inicializar randomicamente o parâmetro, de tal maneira a possuir um ponto de partida. É possível descrever o problema como também a maximização da seguinte função:

$$F(q^t(s), \theta_{t-1}) = \log(P(X | \theta)) - KL(q(s) \parallel P(S | X, \theta_{t-1})) \quad (\text{B.11})$$

onde t é um parâmetro de iteração do algoritmo. Desta maneira, se:

$$\max_q F \rightarrow KL(q(s) \parallel P(S | X, \theta)) = 0 \quad (\text{B.12})$$

então, a distribuição dos estados ocultos deve ser a distribuição dos estados com base no parâmetro anterior do modelo:

$$q^t(s) \leftarrow P(S | X, \theta_{t-1}) \quad (\text{B.13})$$

Este é o passo da expectativa (passo E) do algoritmo. Note que a expectativa com relação a função F foi justamente obtida assumindo um valor para $q^t(s)$.

Posteriormente, é necessário atualizar o valor do parâmetro. Desta maneira, é possível avaliar F novamente:

$$F(q^t(s), \theta_{t-1}) = \int \log\left(\frac{P(X | S, \theta)P(S | \theta)}{q(s)}\right) q(s) ds \quad (\text{B.14})$$

de tal maneira que:

$$F(q^t(s), \theta_{t-1}) = \int \log P(X | S, \theta)P(S | \theta)q(s) ds - \int \log(q(s)) q(s) ds \quad (\text{B.15})$$

O termo negativo a direita é justamente a entropia $H(q)$ da distribuição $q(s)$. Com isso, reescreve-se F como:

$$F(q^t(s), \boldsymbol{\theta}_{t-1}) = \int \log P(X | S, \boldsymbol{\theta})P(S | \boldsymbol{\theta})q(s)ds + H(q) \quad (\text{B.16})$$

Observa-se que $H(q)$ é um termo independente de $\boldsymbol{\theta}$, podendo a maximização se referir apenas a integral. Desta maneira, busca-se obter os parâmetros $\boldsymbol{\theta}$ que maximizem a integral. Com isto, é possível obter a atualização de $\boldsymbol{\theta}$:

$$\boldsymbol{\theta}^t \leftarrow \arg \max_{\boldsymbol{\theta}} \int q^t(s) [\log P(X | S, \boldsymbol{\theta})P(S | \boldsymbol{\theta})] ds \quad (\text{B.17})$$

com este sendo o passo de maximização (passo M) do algoritmo. A realização destes passos é feita até a convergência.

Apêndice C

Códigos desta dissertação

Para a elaboração desta dissertação, foi necessário mais de 1000 linhas de código. Desta forma, fica inviável colocar todos os códigos de forma organizada neste texto.

Neste sentido, fica a disposição, o link do GitHub (github.com/RL2-SP3/LISSA) para todos os códigos da dissertação.¹

¹ Atualmente, esta biblioteca ainda não está disponível, uma vez que o nosso repositório conjunto está privado e é preciso refatorar o código em alguns pontos.