APPLICATION DE TECHNIQUES DE REDUCTION DE LA DIMENSIONNALITE A LA MAINTENANCE PREDICTIVE DE TURBOMACHINES AERONAUTIQUES

Jean-Loup Loyer[1]

[1] Instituto Superior Técnico, *Avenida Rovisco Pais 1, 1049-001 Lisbon, Portugal*
jean-loup.loyer@tecnico.ulisboa.pt

Résumé. Le suivi de turbines à gaz peut permettre d'optimiser leur performance, améliorer la sécurité aérienne et diminuer les coûts opérationnels à travers une meilleure maintenance prédictive. A cette fin, jusqu'à plusieurs centaines de variables internes et externes au moteur sont mesurées en temps réel afin de mettre au point des modèles statistiques de maintenance prédictive. La plupart de ces paramètres sont des séries temporelles multivariées comportant des motifs très complexes, à partir desquels de l'information doit être extraite par un premier ensemble de techniques de réduction de la dimensionnalité dites de « Features Extraction ». Cette première phase génère des variables explicatives pour les modèles statistiques. Toutefois, de telles variables explicatives peuvent se compter par centaines et un second ensemble de techniques de réduction de la dimensionnalité, connues dans la littérature comme « Feature Subset Selection », doit être mis en œuvre afin de sélectionner les 10 variables explicatives les plus appropriées. La communication présente une revue des techniques disponibles pour chacune des deux phases de la réduction de la dimensionnalité et compare leur performance sur un jeu de données en grandes dimensions comprenant des données industrielles réelles provenant d'un fabricant de moteurs d'avion.

Mots-clés. Big data, réduction de la dimensionnalité, données en grandes dimensions, maintenance prédictive, suivi de turbines à gaz.

Abstract. Monitoring of gas turbines can lead to optimize their performance, improve safety and lower operating costs through better predictive maintenance. To do so, up to a few hundreds of internal engine and external environmental parameters are measured in real time in order to building statistical models of predictive maintenance. Most of such parameters are multivariate time series exhibiting very complex patterns, from which information has to be extracted by a first set of dimension reduction techniques known as Feature Extraction. This first phase generates predictors for subsequent statistical models. However, such predictors can number by the hundreds and a second set of dimension reduction techniques, known as Feature Subset Selection in the literature, has to be applied to select the 10 most relevant predictors. The communication will present an overview of the techniques available on each of the two phases of dimensionality reduction and compare their performance on a high-dimensional dataset comprising real industrial data from a leading manufacturer of jet engines.

Keywords. Big data, high-dimensional data, predictive maintenance, dimensionality reduction, gas turbine health monitoring

1 Introduction to predictive maintenance based on high-dimensional data

Monitoring of gas turbines allows industrial companies to optimize their performance, improve safety and lower operating costs through better predictive maintenance and lower fuel burnt. To meet this objective, dozens of sensors are typically installed at several stations in the rotating machinery in order to measure up to a few hundreds of internal engine thermodynamical and mechanical parameters: absolute and marginal temperatures, pressure, shaft rotation speeds, vibration levels, fuel flow, oil characteristics, opening of valves, alerts... They are complemented by external environmental parameters related to the atmospheric conditions the engines are evolving in: air temperature, pressure, humidity, wind speed, wind gusts…

Those hundreds of parameters are acquired continuously and recorded as multivariate time series over the operating life of the engine exhibiting complex patterns (Figure 1). For instance, one can see that some engines have been subject to only two maintenance shop visits (curves a-c) while others had up to 6 (curve e). Some engine parameters stay almost constant (curve c), evolve almost linearly (curve e) while other exhibit nonlinear trends (curve f), erratic patterns (curve b) or cluster of values (curve a). Some engine parameters exhibit more variability (curves a, c) than others (curves e, f). Finally, some curve present many discontinuities either as steps (curves b, f) or gaps in the data (curves e, f). Given the potential combinations in all such dimensions, extract meaningful features from the time series is challenging: the mean fails at capturing variability, the variance doesn't necessarily take into account trends in the data, the trend doesn't consider variability and is not particularly robust and so on... It becomes necessary to confront all those features and extract more meaningful ones from the time series for subsequent statistical modelling.

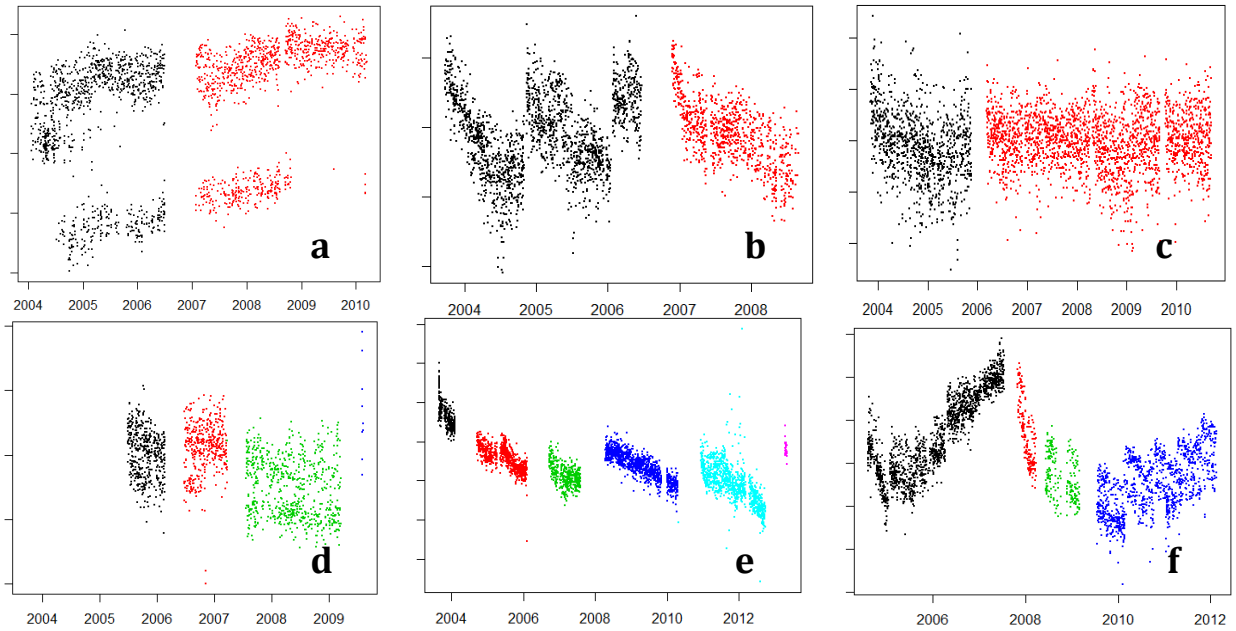

Figure 1 - Examples of time series corresponding to internal engine parameters (fictitious data[1])

The high number of predictors and their functional nature render the use of regression models and machine learning algorithms particularly challenging. To remedy these two critical issues for predictive maintenance, adequate techniques have to be selected to reduce the dimensionality of the

[1] To respect confidentiality agreement with the industrial partner, fictitious data has been simulated, with changes in dates and absence of variables' names. However, the plots are representative of the variety of situations encountered in jet engines. Colors correspond to interval of engine operations between two maintenance shop visits (MSV), corresponding to a period of the life of the engine.

dataset. The objective of this communication is to compare two sets of techniques of dimensionality reduction of high-dimensional data: 1) Feature Extraction (FE) from multivariate time series and 2) Feature Subset Selection (FSS) to generate relevant predictors for subsequent models of predictive maintenance.

The purpose of this communication is not to present a new technique but rather to present an overview and application of several methods for reducing the dimensionality of high-dimensional data and compare their performance in terms of accuracy in the prediction of the maintenance needs. One of its main interests consists in the application of the techniques to a real case study involving large volume of real industrial data from hundreds of Rolls-Royce's jet engines over the period 2002-2012.

2 Techniques for reducing high-dimensional data for predictive maintenance

Building a regression model of engine maintenance – as measured by the scrap rates of the components - as a function of engine parameters and external events requires two different types of dimensionality reduction: 1) reduction of the time dimension of the multivariate time series by extraction relevant features from them, followed by a 2) selection of the best subset of features for improving the statistical model.

Table 1 - Type of features extracted from the time series

Simple features	More complex features
Moments and "location": mean, variance, median, minimum and maximum values	Number and amplitude of gaps and jumps in the time series
Trends: linear slope of scatterplot of the curve according to time, delta between initial and final values of the time series over a period of time	Correlation and clusters of time series
	Class of the time series according to patterns

The first step in reducing the dimensionality consists in Feature Extraction from the multivariate time series, whose types and categories are summarized in Table 1. Meaningful information is captured as a few key defining features to be used novel predictors that would be as relevant as possible to explain and predict the scrap rates of the engine components. Simple engineering features from the time series are related to the basic structure of the series (minimum/maximum values, mean, variance, quantiles, linear trends, deltas measuring the evolution of a parameter over a period), which can be seen as "proxies" for the deterioration of the engine. More complex features directly extracted from time series include number and amplitude of gaps and jumps in the time series. Cross/auto-correlation/covariance analyses contribute to identify correlated components in multivariate time series and discard some engine parameters very early in the analysis, as shown by Peña and Poncela (2006) or even cluster time series, as shown by Caiada, Crato and Peña (2006). Another solution consists in classifying time series into classes according to patterns they might exhibit as proposed by Guerts (2001) (Figure 2) using techniques such as time series clustering[2], Discrete Wavelet Transform (DWT), Discrete Fourier Transform (DFT) or more recent symbolic methods such as SAX/iSAX reviewed notably by Shieh and Keogh (2009). However, projection

[2] To obtain relevant time series clustering, the measure of the distance or dissimilarity is paramount. Several measures are available: Euclidean distance, Manhattan distance, Maximum norm, Hamming distance or Dynamic Time Warping (DTW) distance...

techniques such as Principal Component Analysis and Independent Component Analysis of time series are not retained for FE because we want to keep the physical meaning of the predictors: for instance, a linear combination of a temperature, pressure and vibration would be too confusing for engineers and domain experts.

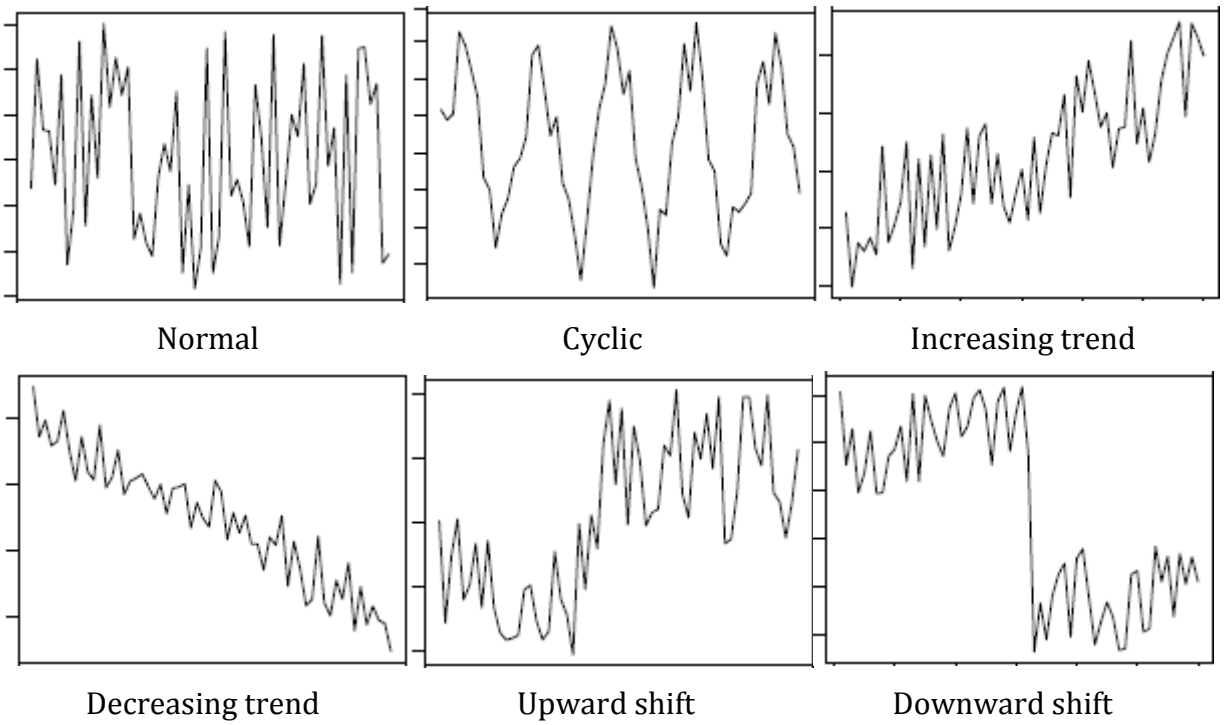

Figure 2 - Examples of 6 simple classes of time series based on pattern detection

The features extracted from the previous step can be added as new predictors to the regression model of maintenance. The next challenge consists in ranking the hundreds of those potential predictors, according to a second step in dimensionality called Feature Subset Selection (FSS). The analyst aims at obtaining a limited set - up to 10 – of understandable predictors of the components' scrap rates, out of the hundreds of potential predictors. The two main objectives for good FSS consist in improving model accuracy and lowering computing resources on high dimensional datasets. Usual techniques for FSS include stepwise regression (forward, backward, bidirectional), shrinkage methods (ridge regression, LASSO) or model selection via the usual criteria (AIC, BIC, Cp, adjusted R^2…). More recent and computational-intensive methods are leveraging the properties of decision tree-based models (random forests, gradient boosted trees…) to rank the importance of the input variables, as proposed by Genuer, Poggi and Tuleau-Marot (2010). FSS has been more extensively covered in the literature compared to FE from complex time series; Hastie et al. (2009) is a classical text on the subject while Guyon and Elisseeff (2003) provides a sound introduction to the topic.

3 Evaluation of the dimensionality reduction techniques

The main objective of the research being to improve current statistical methods for predictive maintenance, prediction accuracy appears as the most natural criterion for evaluating the aforementioned dimensionality reduction techniques. Since the model output is the scrap rate of engine parts as expressed as binary variables (i.e. "failed/not failed" part), criteria for evaluating the

performance of binary classifiers such as misclassification rates and ROC curves have been selected.

The performance of each of the two dimension reduction steps have to be assessed. First, the influence of the features generated by the FE step on the output variable is measured by importance variable ranking and visualization of scatterplots, boxplots or conditional plots. Second, after the set of original predictors has been defined, the best subset of features is selected by comparing goodness-of-fit and information criteria on a series of models obtained by stepwise regression and by assessing the importance of the variables by tree-based techniques. Results show that extracting complex features from the time series yields more accurate results than extracting simpler features such as the mean, median, standard deviation or trend. Regarding the second step of dimension reduction, the measure of the correlation between the predictors and ranking of the variable importance offer good performance in the FSS phase and contribute to a more accurate predictive model of maintenance.

Techniques for reducing the high dimension of industrial datasets are particularly relevant to improve the accuracy of predictive models. However, predictive performance should not be the only object of the extraction and selection of features. The understandability and interpretability of the features should also be an objective, since the users of the predictive models of maintenance are typically engineers who possess an extensive domain or product knowledge. In such industrial cases, the optimization of "black-box" models should be balanced by other considerations such as model interpretability, easiness of maintenance or model's ergonomics.

Bibliographie

[1] Caiadoa, J., Cratoa, N. et Peña, D. (2006), A periodogram-based metric for time series classification, *Computational Statistics & Data Analysis*, 50(10), 2668-2684

[2] Genuer, R., Poggi, J.M., et Tuleau-Malot, C. (2010), Variable selection using random forests, *Pattern Recognition Letters*, 31(14), 2225-2236.

[3] Geurts, P. (2001), Pattern Extraction for Time Series Classification, *Principles of Data Mining and Knowledge Discovery - Lecture Notes in Computer Science*, 2168, 115-127

[4] Guyon, I. et Elisseeff, A. (2003), An introduction to variable and feature selection, *The Journal of Machine Learning Research*, 3, 1157-1182.

[5] Hastie, T., et al (2009), *The elements of statistical learning*. Vol. 2. No. 1., Springer, New York.

[6] Peña, D. et Poncela, P. (2006), Dimension Reduction in Multivariate Time Series, *Advances in Distribution Theory, Order Statistics, and Inference Statistics for Industry and Technology*, 433-458

[7] Shieh, J. et Keogh, E. (2009), iSAX: disk-aware mining and indexing of massive time series datasets, *Data Mining and Knowledge Discovery*, 19, 24-57