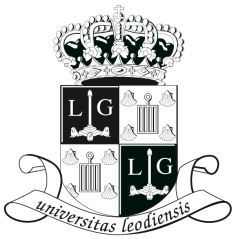
Université de Liège

Variations on the Kalman filter for enhanced performance monitoring of gas turbine engines

Sébastien J. Borguet

Department of Aerospace and Mechanics
Turbomachinery Group



Université de Liège
Faculté de Sciences Appliquées

Variations on the Kalman filter for enhanced performance monitoring of gas turbine engines

par

Sébastien J. BORGUET
Docteur en Sciences de l'Ingénieur
Ingénieur électricien-mécanicien

Thèse de doctorat

2012

Abstract

Since their advent in the 1940's, gas turbines have been used in a wide range of land, sea and air applications due to their high power density and reliability. In today's competitive market, gas turbine operators need to optimise the dispatch availability (*i.e.,* minimise operational issues such as aborted take-offs or in-flight shutdowns) as well as the direct operating costs of their assets. Besides improvements in the design and manufacture processes, proactive maintenance practices, based on the actual condition of the turbine, enable the achievement of these objectives.

Generating dependable information about the health condition of the gas turbine is a requisite for a successful implementation of condition-based maintenance. In this thesis, we focus on the assessment of the performance of the thermodynamic cycle, also known as *Module Performance Analysis*. The purpose of module performance analysis is to detect, isolate and quantify changes in engine module performance, described by so-called health parameters, on the basis of measurements collected along the gas-path of the engine. Generally, the health parameters are correcting factors on the efficiency and the flow capacity of the modules while the measurements are inter-component temperatures, pressures, shaft speeds and fuel flow.

Module performance analysis can be cast as an estimation problem that is characterised by a number of difficulties such as non-linearity of the system and noise and bias in the measurements. Moreover the number of health parameters usually exceeds the number of gas-path measurements, making the estimation problem underdetermined.

This thesis starts with a survey of the state-of-the-art in module performance analysis. We then propose enhancements to a monitoring tool for steady-state data developed by Dr. P. Dewallef during his thesis at the Turbomachinery Group. Specifically, the improvements concern the fault detection and isolation tasks, respectively handled by a hypothesis testing and a sparse estimator. As a complement, we define metrics for the selection and analysis of sensor–health parameter suites based on the Information Theory.

In a second step, we investigate the feasibility and the benefit that could be expected from the processing of data collected during transient operation of a gas turbine. We also discuss the impact of modelling errors on the estimation procedure and propose a solution that makes the health assessment robust with respect to modelling errors.

The theoretical developments are evaluated on the basis of simulated test-cases through a series of metrics that gauge the estimation accuracy and the performance of the fault detection and isolation modules.

Keywords: module performance analysis, gas turbines, health monitoring, Kalman filter, fault detection and isolation

Remerciements

Une des caractéristiques de cette thèse aura probablement été sa longueur. Mesurée à l'aune d'une crise politique belge, elle en devient cependant bien plus acceptable. Avant d'entrer dans le vif du sujet, je voudrais adresser des remerciements chaleureux

à mon promoteur, le Professeur Léonard, pour m'avoir embarqué dans cette aventure enrichissante qu'est la réalisation d'une thèse. Merci Olivier pour ta confiance et ton soutien,

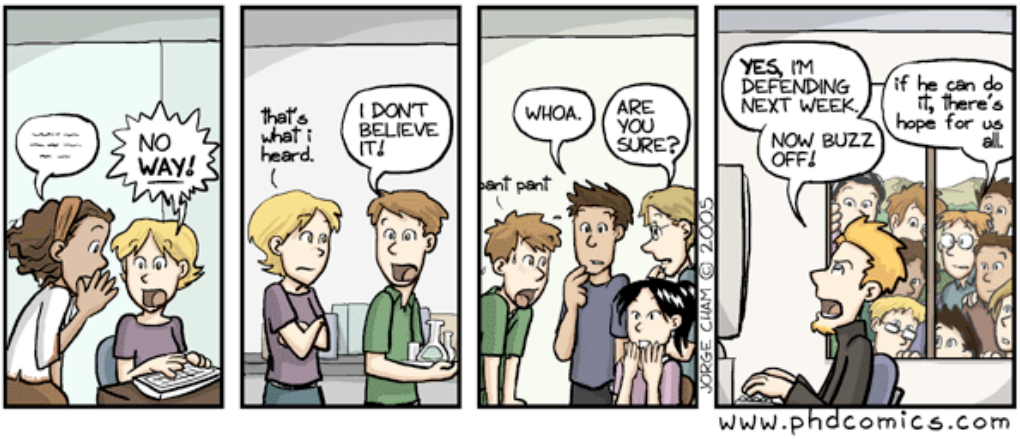
à ma famille, et plus particulièrement à mes parents pour l'éducation qu'ils m'ont donnée et pour leur soutien au fil de ces (indécemment) longues études. J'ai également une pensée pour mes frères que je rejoins – finalement – dans la grande famille des Docteurs,

à mes collègues des services Turbomachines et Aérodynamique, trop nombreux pour être cités ici, pour l'atmosphère de travail détendue et bon enfant. Je tiens à remercier plus particulièrement le Professeur Pierre Dewallef pour son assistance en début de thèse ainsi que Jean-Philippe Thomas, mon coloc' de bureau pour les bons moments de rigolade partagés au bureau et en mission. Merci également à mes co-auteurs suédois Mattias Henriksson et Tomas Grönstedt, Tack så mycket för ditt hem och din vänlighet,

à mes amis, pour m'avoir rappelé inlassablement qu'il existe bel et bien une vie en dehors la recherche, merci co cint fèyes mes djins,

aux membres du jury, pour avoir accepté de consacrer du temps à la lecture de cette thèse.

Sébastien Borguet



Tout finit par arriver, en effet !

Contents

Nomenclature

Notations

$\bar{\cdot}$	average value
$\cdot^{bl}$	nominal health condition
$\widehat{\cdot}$	estimated value
$\dot{\cdot}$	time derivative
$\cdot^{\star}$	actual value
$\delta\,\cdot$	deviation with respect to reference value
$\tilde{\cdot}$	scaled value

Vectors and matrices

$\mathbf{b}$	bias
$\mathbf{h}$	health parameters
$\mathbf{r}$	residuals (difference between actual and predicted values)
$\mathbf{u}$	control parameters
$\mathbf{x}$	state variables
$\mathbf{y}$	gas-path measurements

$\mathbf{FIM}$	Fisher information matrix
$\mathbf{G}$	Influence coefficient matrix
$\mathbf{P}$	covariance matrix of estimated quantities
$\mathbf{Q}$	covariance matrix of prior information
$\mathbf{R}$	covariance matrix of noise terms

Engine parameters

ALT	flight altitude
CDP	(high pressure) compressor delivery pressure
$\mathsf{DT}_{\mathsf{ISA}}$	offset with respect to the standard day temperature
EGT	exhaust gas temperature
NF	fan speed
NC	core speed
Pamb	ambient pressure
PCNF	percent corrected fan speed
Pi	total pressure at station i
Tamb	ambient temperature
Ti	total temperature at station i
SEi	correcting factor on the efficiency of the component whose inlet is located at station i
SWi	correcting factor on the flow capacity of the component whose inlet is located at station i
V0	true airspeed
WF	fuel flow rate

Abbreviations

BCM	bias compensation module
DKF	dual Kalman filter
EKF	extended Kalman filter
ISA	international standard atmosphere
KF	Kalman filter
map	maximum a posteriori
ml	maximum likelihood
mmse	minimum-mean-square error
pdf	probability density function
SLS	sea level static
UKF	unscented Kalman Filter

Chapter 1

Introduction

Contents

1.1 From time between overhaul to power by the hour

Since their advent in the 1940's, gas turbines have been increasingly used in a wide range of land, sea and air applications. This document focuses on turbine engines for aircraft propulsion, but the concepts developed can be applied to land-based and marine gas turbines.

In the early days of the turbojet, the level of performance was modest and the operating cost was not an important issue. Maintenance was scheduled at fixed intervals, specified by the time between overhaul (TBO). An overhaul consisted in a removal of the engine from the aircraft and a complete disassembly for individual inspection and repair of the components. After an acceptance test, the engine was returned to service.

The growth in lift and range capability of aircrafts required more powerful and efficient engines. This led to the development of more complex architectures [see Rolls-Royce, 1999] both at the engine level, such as two– or three– shafts turbofans (figure 1.1) and at the module level, such as variable stator vanes (figure 1.2) and stability bleed valves on the compressors. The manufacturers dealt with this higher complexity through a modular design, which allowed the replacement of the TBO method by on-wing maintenance.

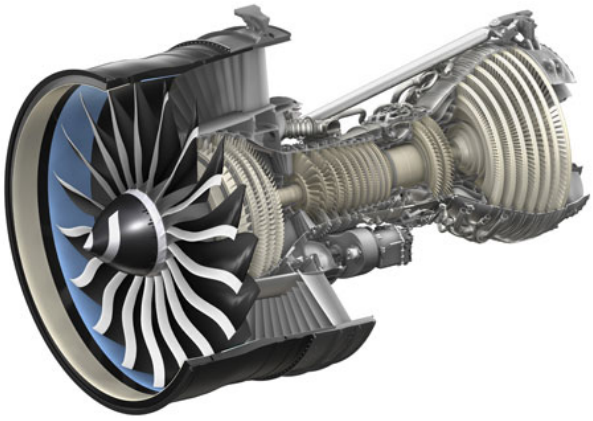


Figure 1.1: Cutaway of a General Electric GENX turbofan engine

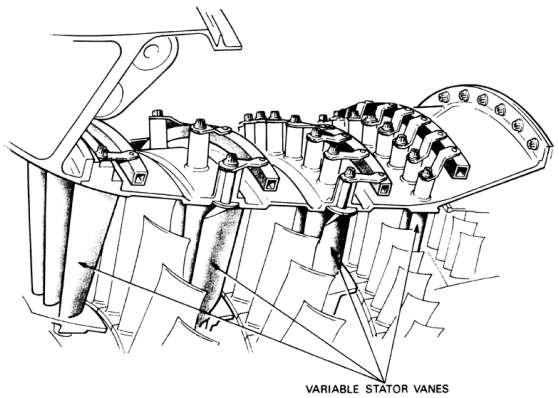


Figure 1.2: Sketch of variable stator vanes

A drawback of the TBO method is that the maintenance interval for the whole engine is set by the fastest deteriorating component – usually the high pressure turbine due to the harsh environment in which it operates – while the other modules can remain operable for longer periods of time. The combination of modular design and on-wing maintenance allows a lifetime to be specified for each major component. A module is replaced by a similar one when it reaches its time limit (figure 1.3) and the engine returns to service. The worn module is shipped to the manufacturer for inspection and maintenance. The

on-wing maintenance practice is a first progress towards improved availability and reduced costs as it decreases the engine downtime. This scheme still relies on fixed intervals which may not be optimal. Indeed, the deterioration of an engine depends largely on its usage (*e.g.*, short vs. long haul, flight over water/continental/desert areas). Besides, a number of unpredictable events such as a bird strike or a heavy landing may happen that call for an unscheduled maintenance.

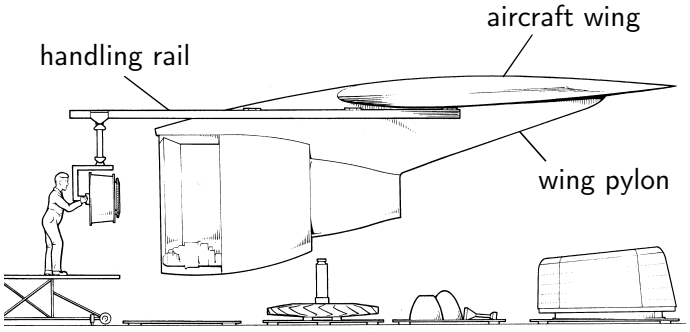


Figure 1.3: On-wing maintenance on a modular turbofan

In nowadays highly competitive market of commercial air transport, another leap in engine maintenance practice is therefore warranted on the operator as well as on the manufacturer side. For airlines, the goal is essentially twofold. On one hand, they seek to improve the dispatch reliability and safety of their fleet by reducing, among others, the in-flight shutdown and the delayed and cancelled rates. This can be achieved by moving unscheduled events into planned engine removal. Reduction in life cycle costs is another driver for enhancing the maintenance practice. According to Marinai et al. [2004], the part associated to the engines amounts to about 30% of the direct operating costs of an aircraft. Of this contribution, roughly one third is for maintenance. As paradoxical as it can be, manufacturers are also – if not more – interested in extending the "life on wing" of their engines through a wiser maintenance planning. This is due to a relatively recent evolution in the customer–supplier relationships named performance based logistics, aka. power-by-the-hour at Rolls–Royce [Kim et al., 2007]. In this framework, the customer does not buy an engine anymore, but only pays for it when it works *i.e.*, when it generates revenue.

Implementation of condition-based maintenance (CBM) [see Rajamani et al., 2004] is an essential step towards the achievement of these goals. The maintenance actions are planned on the basis of the actual health condition of the engine rather than simply based on the number of cycles. Generating a reliable information about this health condition becomes therefore a requisite for an effective application of CBM. This explains the intensive research carried out since several years in engine health monitoring.

Jaw [2005] proposes a detailed survey of the current status in the field. Engine health monitoring basically relies on the recording and processing of engine data to detect and isolate changes in the condition of the gas turbine. Given the numerous disciplines involved in gas turbine design and operation, engine health monitoring is generally subdivided into four functional areas:

- **oil and debris monitoring** detects abnormal size and number of particles representing excessive wear or fatigue failure of engine components such as gearboxes and bearings [see Green, 2005, for a detailed presentation]. Somehow related, gas-path debris monitoring is an emerging technology that scrutinises the inlet and exhaust sections for the passage of debris,
- **vibration monitoring** is interested in identifying possible unbalance at all operating speeds resulting from failure of rotating parts or loss of a blade, [see de Smet, 2005, for a detailed presentation],
- **life usage monitoring** is concerned with the computation of the level of damage and prediction of remaining life of critical components like disks and blades [see Koehl, 2001, for more information],
- **gas path performance monitoring**, also known as gas path analysis or module performance analysis, relies on measured air/gas flow properties to infer the problems in engine modules. Since the pioneering work of Urban [1972] almost forty years ago, a tremendous amount of research on gas path analysis has been published.

It becomes widely acknowledged that an efficient implementation of a complete solution for engine health monitoring shall combine the results of these functional analyses. Examples of fusion of vibration and gas path monitoring are reported in [Turso and Litt, 2004, Kyriasis et al., 2009]. Dempsey and Afjeh [2002] have combined vibration and oil analyses for detection of faults in helicopter gearboxes. NASA initiated some time ago an ambitious program on information fusion [see Volponi, 2005]. The scope was to fuse diagnosis information available from the disparate data sources listed above. For some reasons, the test campaign encountered several problems and preliminary work showed that the informational overlap between the different available data streams was quite limited. As a consequence, a fusion of two gas path analysis techniques is the only application reported.

The present document concentrates on the area of gas path performance monitoring as it is the closest discipline to the domain of expertise of the Turbomachinery Group at the University of Liège. Before introducing in more details the underlying principles of gas path analysis, the following section gives an overview of the degradation mechanisms that hurt the performance of engine modules.

1.2 Deterioration of engine components

Deterioration can be loosely defined as an increased level of losses in the engine components. From an operational point of view, deterioration is reflected in overall system performance through a loss of thrust and an increase in specific fuel consumption. Studies, both at the experimental and theoretical levels, on the sources of degradation in engine components have produced a very rich literature of which [AGARD, 1994] can be considered as a reference book. The brief description of the main deterioration mechanisms

provided below is based on information compiled from [Zaita et al., 1998, Diakunchak, 1992, Kurz et al., 2009]:

- **fouling** is the accumulation of particles on the airfoils and annulus walls. It increases surface roughness and changes the shape of the blades,
- **erosion** (fig. 1.4) is the removal of material from blades resulting from a collision with solid particles suspended in the air/gas stream,

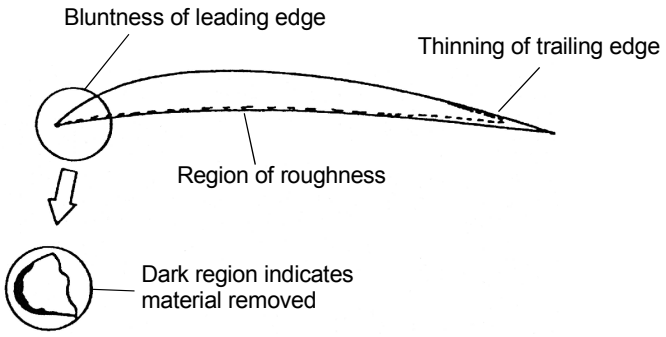
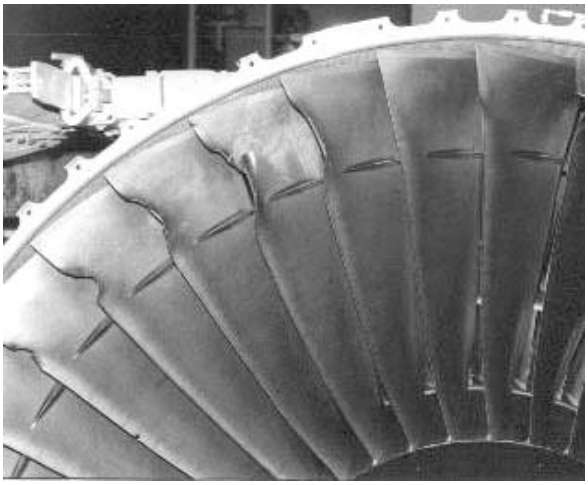


Figure 1.4: Alteration of blade profile caused by erosion

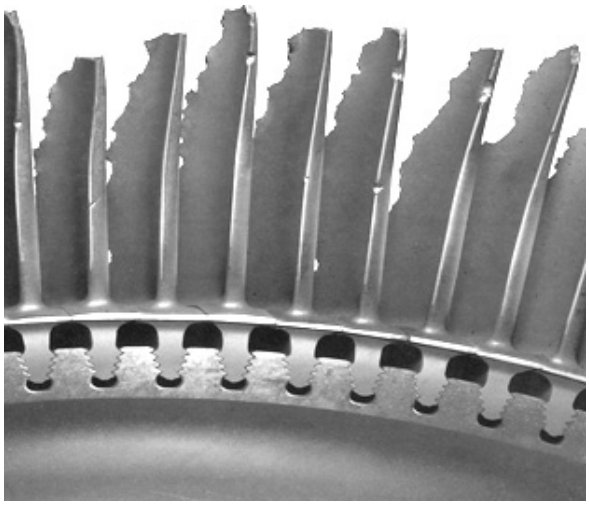
- **rubbing-wear** or **abrasion** is the removal of material from rotor blade tips and seals caused by the contact between static and rotating parts. Rubbing-wear occurs naturally during the running-in period, to establish proper clearances. Another part is age related due to a loss of stiffness of the bearings,
- **corrosion** is the loss of material from flow-path components caused by chemical reactions between the component materials and contaminants such as salts or reactive gases. It is mainly promoted by high temperatures,
- **damage** is a generic term for a severe alteration of the component. Causes are as various as a foreign object damage, resulting for instance from a bird strike (fig. 1.5(a)), a domestic object damage, caused for instance by a piece of seal hitting a downstream component (fig. 1.5(c)), a turbine overheating following a hot start (fig. 1.5(b)), or a tip rub due to improper clearance control during fast transient manoeuvres (fig. 1.5(d)).

Some of these deterioration processes are partly recoverable with on-site cleaning, such as the fouling of compressors. Others are said non-recoverable as they imply the replacement of hardware parts during an overhaul.

Deterioration is very often split up into progressive wear due to normal operation and damages due to accidental events as exemplified in figure 1.6. While the former varies on a long time-scale and affects all the components at the same time, the latter are short-time-scale processes involving a brutal drift in performance of one or two component(s) at a time.

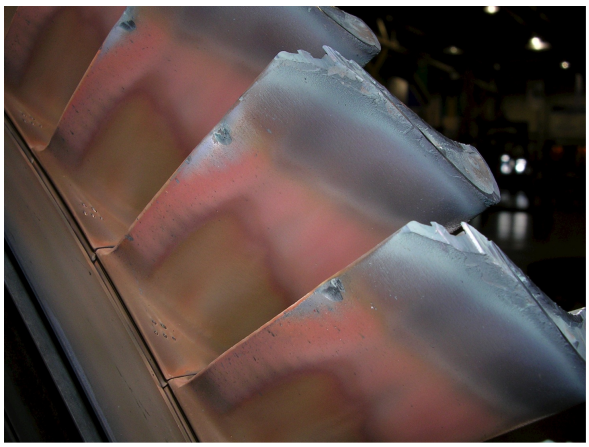


(a) Foreign object damage



(b) Overheating

(c) Domestic object damage



(d) Tip rub

Figure 1.5: Examples of damaged blades

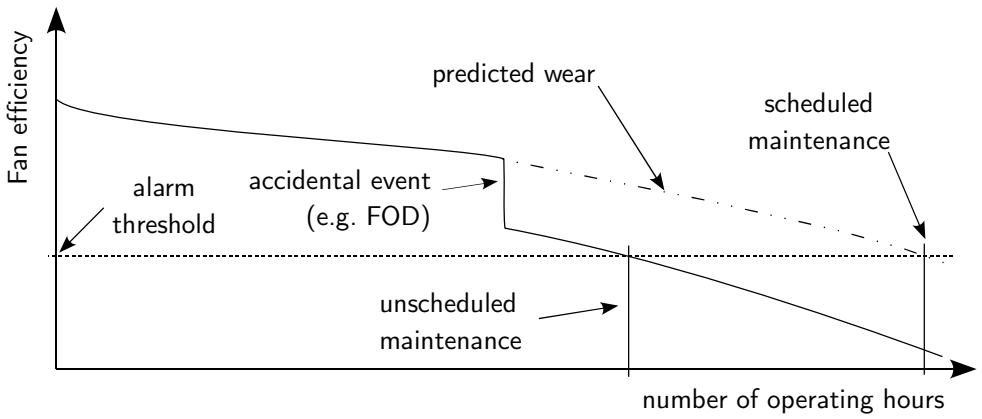


Figure 1.6: Sample deterioration of fan efficiency over time

Gradual deterioration of an engine during normal operation arises from a few sources of which the main ones are fouling, erosion and corrosion. Ogaji et al. [2002] describe the effects of these mechanisms on the major modules of an engine. For compressors, it translates into increased tip clearances and changes in airfoil geometry and surface quality. This causes a reduction in efficiency, flow capacity and work capability of the component. The combustion system is not likely to be directly hit by deterioration. The combustion efficiency usually does not decrease, except for extreme cases like plugged fuel injector. Combustor deterioration can however lead to variation in the module exit temperature profile and existence of temperature peaks that may subsequently damage the turbine section. Turbines experience mostly the same issues as compressors do, causing here a reduction in efficiency and an increase in flow capacity. These observations are supported by the analysis of engine operational data [see Sallee, 1978, Wulf, 1980].

In spite of their continuous character, the mechanisms involved in gradual deterioration do not evolve in a purely deterministic way, but depend to some extent on factors like usage, environment and time. Any engine will hence experience a unique degradation profile.

Accidental events may be of many different types as previously noticed, but above all they are unpredictable, which has a harmful effect on the maintenance planning. This is illustrated in figure 1.6 where the evolution of the efficiency of the fan is plotted versus the number of operating hours. A maintenance is scheduled based on the wear of a "mean" engine (dash-dot line). However, if an accidental event (*e.g.*, FOD) occurs, the efficiency drops (the actual engine performance is drawn in plain line) and the actual performance curve crosses an alarm threshold level well before the planned service. When a performance test is carried out on the ground, this results in an unscheduled maintenance, generating additional costs and downtime of the engine.

1.3 Principles of module performance analysis

The modifications in component health described in the previous section are neither directly measurable (there is currently no sensor of fouling or erosion), nor detectable without a deep inspection of the engine (*e.g.*, borescope check). Quoting Volponi [2003], the purpose of module performance analysis is to detect, isolate and quantify these changes in engine module performance on the basis of measurements collected along the gas path of the engine. Over time, the analysis has been extended to malfunctions of engine systems (such as the variable bleed valves and the variable stator vanes) and to instrumentation problems.

This approach to performance diagnostics can be described as relative, in the sense that it assesses changes in the gas turbine performance rather than its absolute level. Gas path analysis allows to track the evolution of a particular engine with respect to some baseline performance which may be engine specific (*e.g.*, at the time of installation) or

fleet averaged. The gas path analysis framework is illustrated in figure 1.7, and is governed by the following principles:

- the condition of the rotating components (fan, lpc, hpc, hpt, lpt) can be represented by a set of independent performance indicators called health parameters which are typically correcting factors on their efficiency and flow capacity,
- physical problems (progressive or accidental) encountered by the engine components induce a modification of their performance, which in turn produces discernible shifts of some measurements collected along the gas path (temperatures, pressures, mass flows, shaft speeds),
- a physical or mathematical model of the engine is available. It translates the aero-thermodynamic relationships between the health parameters and the quantities measured along the gas path.

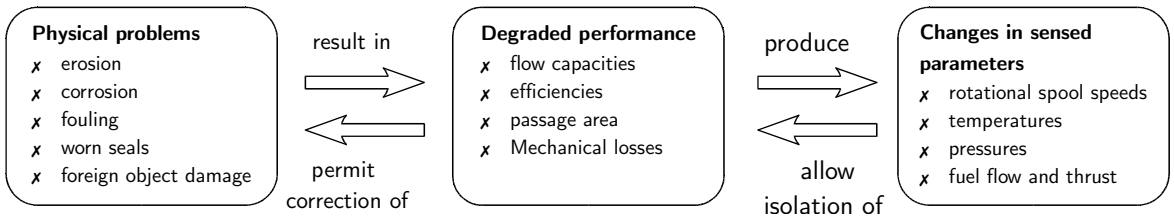


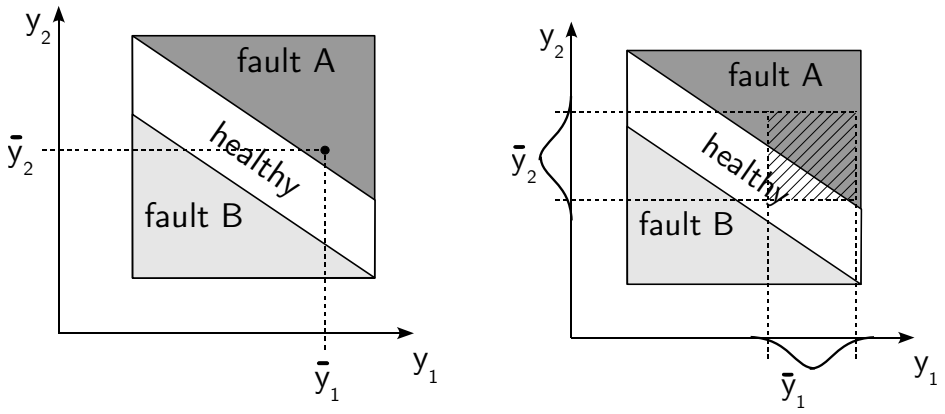
Figure 1.7: Founding principle of gas path analysis [Urban, 1972]

Gas path analysis relies on the physical meaning of the health parameters to locate the underlying physical problem. The diagnosis can then be used to infer a corrective action. This last step results from a complex trade-off between economic, technical, safety-related and legal aspects which is beyond the scope of this document.

As underlined by Marinai et al. [2004], three major difficulties complicate module performance analysis. First, the behaviour of a gas turbine is typically non-linear, with respect to both the health parameters and the operating conditions. Second, the measurements are corrupted by errors caused by noise, bias and uncertainties in the model. Third, the number of sensors is usually smaller than the number of health parameters. Such a lack of information is sometimes termed negative redundancy in the literature. The influence of the last two points is discussed qualitatively below.

Consider the sample diagnosis problem illustrated in figure 1.8(a). The hypothetical diagnosis rule associates a discrete engine health condition of the kind "healthy-fault A-fault B" to each couple of measurements $(\mathbf{y}_1,\mathbf{y}_2)$ (*e.g.*, exhaust gas temperature and core speed) to determine whether a maintenance must be carried out and which corrective action should be done. In an ideal, deterministic situation, reporting the measurement recordings $(\overline{\mathbf{y}}_1,\overline{\mathbf{y}}_2)$ in the diagram leads to diagnose fault A. In practice however, measurement inaccuracies hide the relevant information about engine faults. The uncertainties make

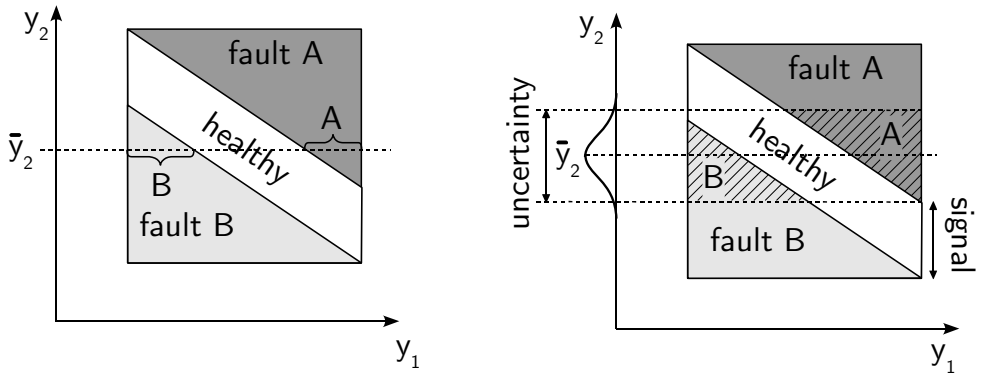
the decision rule fuzzy which is represented in figure 1.8(b) by a hatched area embedding all the possible solutions associated to the measurement readings ($\overline{\mathbf{y}}_1$,$\overline{\mathbf{y}}_2$).



(a) Without measurement noise (b) With measurement noise

Figure 1.8: Influence of the uncertainties on the diagnosis problem

The effect of negative redundancy is sketched in figure 1.9(a) where only one measurement is now available (say $\mathbf{y}_2$) instead of the two previous ones. It is clear that even if no uncertainty contaminates the measurement readings, the diagnosis rule has no unique solution. From the point of view of an external observer having only access to $\overline{\mathbf{y}}_2$, all the situations located on the dashed line are equally probable. If, based on this information, it is decided to take the corrective action corresponding to the fault A, there exist a number of situations, located in the region 'B', where the engine needs the action B. Conversely, if the fault B is considered, there exists a number of situations, located in the region 'A', where the engine needs the action A. The situation is even worse in the presence of measurement uncertainties (fig. 1.9(b)) as the number of possible wrong decisions increases (hatched areas 'A' and 'B').



(a) Without measurement noise (b) With measurement noise

Figure 1.9: Influence of the negative redundancy on the diagnosis problem

It is worth noting that the areas 'A' and 'B' in figure 1.9(b) decrease when the magnitude of the faults increases. This explains why it is generally easier to detect large faults (*e.g.*, efficiency drops of 5%) than small faults (*e.g.*, efficiency drops of 0.5%).

The rather high level of wrong decisions achieved by most of the existing methods is recognised as the primary reason for a limited application of condition-based maintenance so far [see Tumer and Bajwa, 1999].

1.4 State of the art in module performance analysis

As previously mentioned, the literature related to gas path analysis is abundant and it seems illusory to provide an exhaustive survey of the discipline. The purpose of the present section is to give a broad glimpse of the main trends that the research in engine performance monitoring has followed. Some points, stated in the next section, will be addressed more specifically later in the thesis; for the others the reader is directed to the latest available reviews [Li, 2002, Marinai et al., 2004, Jaw, 2005].

Many criteria could be used to classify the methodologies that have been applied to solve the gas path analysis problem. The division adopted here is based on the knowledge of the system. The model-based family makes use of a model of the gas turbine and leads to estimation/optimisation methods, the data-driven family extracts information solely from field data and leads to machine learning/artificial intelligence techniques.

Model-based techniques

The early GPA techniques [Urban, 1972, Danielsson, 1977] consisted basically in a parameter sensitivity analysis, implying the solution of a set of linear equations. They allowed isolation and quantification of the degradation of an engine and resulted from an idealised formulation. Indeed, they assumed an accurate model, noise free data and a well-posed problem *i.e.*, the number of health parameters was smaller than or equal to the number of sensed quantities. The basic GPA algorithm was extended to the non-linear case (through a recursive application of the linear method) in order to cope with larger deterioration magnitudes.

The issues of measurement uncertainty and negative redundancy were addressed by switching to so-called optimal estimation techniques such as weighted least-squares [Doel, 1994b] and Kalman filters [Provost, 1994, Volponi et al., 2003] and their non-linear counterparts [Mathioudakis et al., 2002, Cordoba, 2005]. A comparative study of the effectiveness of linear and nonlinear methods is proposed by Kamboukos and Mathioudakis [2005]. These algorithms recognise the stochastic character of the measurements and allow the introduction of a priori information in the problem. Due to their formulation, the optimal estimation techniques are more suited to the assessment of progressive, distributed degra-

dation but are known to spread localised faults over several components, which is termed the "smearing" effect in the literature [see Provost, 1994].

Genetic algorithms have also been applied to solve the gas path analysis problem [Gulati et al., 2001, Wallin and Grönstedt, 2004]. They have a number of distinctive features with respect to the more traditional gradient-based techniques such as global search capability, derivative-free computation and efficient handling of constraints. Their main drawbacks are a horrific computational demand and the need for highly-skilled users for tuning.

The underdetermined nature of the estimation problem resulting from the limited instrumentation fitted on production engines led the community to investigate three roads of remedy. The first approach, already mentioned above, is to complement the definition of the estimation problem by introducing prior knowledge on the parameters [see Doel, 1994a, Dewallef et al., 2004]. Basically, this drives the search for a solution in a particular area of the solution space. The second approach consists in extending the analysis to a multiple operating point formulation as in [Stamatis et al., 1991, Grönstedt, 2002]. The rationale behind this idea is to take advantage of the non-linear behaviour of the engine across its operating envelope which translates into different signatures of the degradation with the operating conditions. The third approach is concerned with the optimal selection of the parameter–measurement configuration in order to maximise the effectiveness of the diagnostics [Stamatis et al., 1992, Ogaji et al., 2002].

Data-driven techniques

Application of artificial intelligence techniques to module performance analysis is a rather young discipline with respect to the legacy GPA method. Artificial neural networks form a first subfamily of which several flavours have been evaluated for fault detection and isolation purpose. The multi-layer perceptron can be seen as a massively distributed processor made up of simple processing units that maps input and output patterns in a non-linear fashion, mimicking the brain's structure. This type seems to be the most popular in the community, see for instance [Ogaji and Singh, 2002, Barbosa et al., 2007, Fast et al., 2008]. Probabilistic neural network [Romessis and Mathioudakis, 2003] is another type that implements Bayesian statistics in a network form. Unlike the multi–layer perceptron, it does not require a learning phase, as the whole learning database is contained in the network itself. Its computational demand is on the other hand much higher at runtime. Bayesian belief networks [Romessis and Mathioudakis, 2006] consist in a graphical representation of a probability distribution describing cause–to–effect relationships between symptoms – the measurement deviations here – and causes – the faults.

Expert systems [Doel, 1990, DePold and Gass, 1999] are a second category of artificial intelligence methods. They transpose input patterns into fault classes on the basis of a knowledge base. The interpretation of the input is achieved through a set of rules defining a so–called inference engine. Expert systems actually automatise a pattern recognition task that was earlier performed by maintenance staffs.

Other feature extraction methods such as (nonlinear) principal component analysis have also been applied with a reasonable degree of success to the problem of engine fault detection and localisation [Menon et al., 2003, Mukkherjee et al., 2005, Borguet and Léonard, 2009]. The underlying principle is to extract from a learning database reduced–order features corresponding to gas turbine normal and faulty operations. Fault detection and isolation then occur in the feature space where new operational data are projected.

From the results reported in the literature, it can be stated that artificial-intelligence-based methods are immune to the smearing effect, essentially because of their classification nature. They hence deliver better performance for the isolation of accidental events. Indeed, the space of solutions is here limited to a finite number of cases. The main drawback of the data-driven techniques is that they require a massive amount of data, covering all expected fault conditions, for a proper learning/set-up phase. This development step can be quite long, but is traded-off with a very small execution time

The potential of information fusion

Marinai et al. [2004] summarise in a convenient tabular format the main features of each method like linear/nonlinear approach, ability to deal with noise/uncertainties, ability to cope with underdetermined problems, suitability for gradual or abrupt degradation. It can be concluded from this table that each technique has its own advantages and that none of them provides a satisfactory and comprehensive answer to the issues of MPA.

Similarly to the idea of coupling the diagnostics results from different functional areas (*e.g.*, oil, vibration, gas path), some research has been carried out on the fusion of gas path analysis techniques. It is currently a developing field and reported applications in the literature include in particular: the combination of neural networks and genetic algorithms, used by Kobayashi and Simon [2001] to diagnose component and sensor faults; the coupling of a Kalman filter with either a Bayesian belief network in Dewallef et al. [2006], or a classifier based on principal component analysis in Borguet and Léonard [2009], in order to improve the isolation of component faults; the application of the Dempster-Schafer theory for fusing a probabilistic neural network and a bayesian belief network in Romessis et al. [2007], leading to enhanced isolation of module faults too.

Towards a diagnosis in transient operation

The techniques described above were designed to process data recorded during a steady-state operation of the engine. In some cases, it is however difficult to get stead-state data of good quality, especially for military aircraft that may spend up to 70% of their mission with their engine in transient operation. In the late 1980's, Merrington [1989] showed that alterations in the health condition are also reflected in measurements collected during transient operation of the engine. His work triggered an endeavour to develop diagnosis techniques based on transient data. Chapter 5 will provide a detailed survey of the topic.

1.5 Main contributions of the thesis

The research activities of the Turbomachinery Group on module performance analysis dates back to the late nineties. A first thesis by Dewallef [2005] reports the development of a performance monitoring tool based on a Kalman filter that processes steady-state data from the on-board instrumentation of a gas turbine. Our research work builds on this initial effort and offers enhancements in a number of areas. More specifically, the main contributions achieved in the course of this thesis cover the four topics briefly described below.

Improved fault detection

The Kalman filter embeds a transition model that describes a slow and continuous variation of the health parameters. As a result, the Kalman filter is very capable at tracking a gradual and distributed process like engine wear. On the other hand, any short-time-scale (either rapid or abrupt) change in the engine condition is interpreted by the Kalman filter as a disturbance since it does not fit the health parameter dynamics. The transition model prevents a correct capture of this fast variation which translates into a sluggish response of the filter.

In order to reduce the detection latency of faults, we propose to use a modified transition model that accounts for possible "jumps" in the health parameters. The resulting adaptive algorithm combines a Kalman filter, which relies on the assumption of a smooth variation of the health parameters, and a secondary system that monitors the residuals. This auxiliary component implements a hypothesis test based on the generalised likelihood ratio in order to detect short-time-scale events. In that way, the performance of the Kalman filter is maintained under normal operation (long-time-scale deterioration) whereas the event detector provides a timely detection of sudden variations in the engine health.

Improved fault isolation

Least-squares-based methods have relatively poor isolation capability due to the very nature of the estimation problem. In fact, the number of health parameters exceeds oftentimes the number of measurements, making the estimation problem underdetermined. The regularisation technique addresses this issue by adding a penalty term, traditionally a quadratic form, on the deviations of the health parameters. Such a function favours solutions involving small variations of numerous health parameters and penalises heavily solutions characterised by large deviations in a few health parameters. This causes the adverse effect known as *smearing i.e.,* the estimate of a fault is spread over several components instead of being clearly concentrated.

In order to improve the isolation capability of the least-squares technique, we propose to modify the regularisation term so that it better matches the expected pattern of a fault *i.e.,* a solution vector with many nought components as accidental events impact one (maybe two) component(s) at a time. This translates mathematically into the search for a *sparse solution* that is achieved by penalising the absolute value of the health parameter deviations instead of their square. Such a regularisation term is much more large-deviation-friendly and reduces drastically the smearing effect. The sparse estimation problem is advantageously formulated as a quadratic programming problem for which efficient solvers are available.

Tools for sensor set selection and analysis

Current sensors remain intrusive and their integration on an engine is limited by constraints such as cost, weight and a harsh working environment. On production engines, the number of sensors is therefore kept to a minimum mainly driven by the requirements of the control system. On the other hand, with the growing interest in condition monitoring, some manufacturers offer optional kits to complement the sensor suite. This raises the question of the selection of the sensor suite (or part of it) that maximises the performance of the diagnosis algorithm, generally defined in terms of estimation accuracy, true positive rate and false alarm rate.

We propose to revisit the problem of sensor selection for turbine engine performance monitoring within the scope of information theory. From the *Fisher information matrix*, we define a set of metrics that relate to the aforementioned qualities. Loosely speaking, this matrix measures the amount of information conveyed by a given instrumentation about the health condition of the engine. To further increase the understanding about the sensor–parameter layout, we propose metrics to quantify the observability of a health parameter as well as the utility of a given sensor in the estimation process.

Diagnosis from transient data and robustness with respect to modelling errors

Most of the existing techniques for module performance analysis evaluate the health condition of the gas turbine from data collected at a single steady-state operating point, which leads generally to an underdetermined estimation problem. The concept of multi-point estimation was proposed as a palliative. Data are collected at several steady-state operating points to take advantage of the non-linear response of the engine. The main drawback of this technique is its practical feasibility in terms of time required to stabilise the engine at several operating points. Part of our research consisted in extending the multi-point framework to transient operation of the engine to remove the need for steady-state condition to be achieved prior to collecting the data. In transient operation, we obviously have to estimate the health parameters, but also the state variables associated to the dynamic

behaviour of the engine (such as spool speeds and metal temperatures). We solve this *double estimation problem* by means of two Kalman filters running concurrently.

In a second step, we investigate the impact of model-plant mismatch on the accuracy of the transient estimation procedure. When an imperfect, biased model of the gas turbine is used in the Kalman filter, the health parameters lose their physical meaning and become mere tuners that capture both the actual degradation and the modelling errors. It is therefore essential to take into account the model-plant mismatch in the estimation procedure. To this end, we propose the concept of *hybrid model* to improve the fidelity of the predicted measurements. The hybrid model combines the biased physics-based model and a black-box component that specifically handles the modelling errors. The empirical component is identified from data collected on a new and clean engine.

1.6 Thesis outline

Chapter 2 provides the root material for the subsequent developments. It introduces the concept of model-based diagnosis and its components. Then it formulates module performance analysis as an estimation problem. The chapter ends with the development of metrics to perform the selection and analysis of sensor/health parameter set-ups.

Chapter 3 reports the development of a performance monitoring tool for steady-state data. The main components of the tool consist in a Kalman filter that tracks gradual deterioration and specific modules for the detection and isolation of accidental events.

Chapter 4 evaluates the monitoring tool designed in chapter 3 on the basis of scenarios simulated with a computer model of a commercial turbofan. A set of metrics gauge the performance in terms of estimation accuracy and fault detection and isolation capability.

Chapter 5 extends the monitoring tool to the processing of transient data. In a second step, the robustness of the methodology with respect to model-plant mismatch is addressed. The proposed solution involves a so-called hybrid model. The potential benefit of diagnosis based on transient data are evaluated from a series of simulated test-cases.

Chapter 6 wraps up the main findings and proposes directions for future research.

Chapter 2

Model-based gas path analysis: the basics

Contents

2.1 Model-based diagnostics

2.1.1 Principle

THE purpose of model-based diagnostics consists in detecting, localising and possibly quantifying the faults that may occur on the monitored process. This can be achieved through analytical redundancy, that is by using the functional dependencies existing among the different measured signals [see Isermann, 1997]. These relationships are expressed by a model of the process, hence the model-based term.

Figure 2.1 illustrates the generic structure of a model-based diagnostics which can be divided into two major tasks [Venkatasubramanian et al., 2003]. First, generate discrepancies between the actual and the expected behaviour of the system. These residuals $\widehat{\mathbf{r}}_k^{\mathrm{hl}}$ are built as the difference between the real outputs observed on the process $\mathbf{y}_k$ and the outputs simulated with a healthy model of the engine $\widehat{\mathbf{y}}_k^{\mathrm{hl}}$, k being a discrete time index. In a second step, process the residuals, that reflect the potential deterioration of the system, with a fault detection and isolation (FDI) logic to create the diagnosis report.

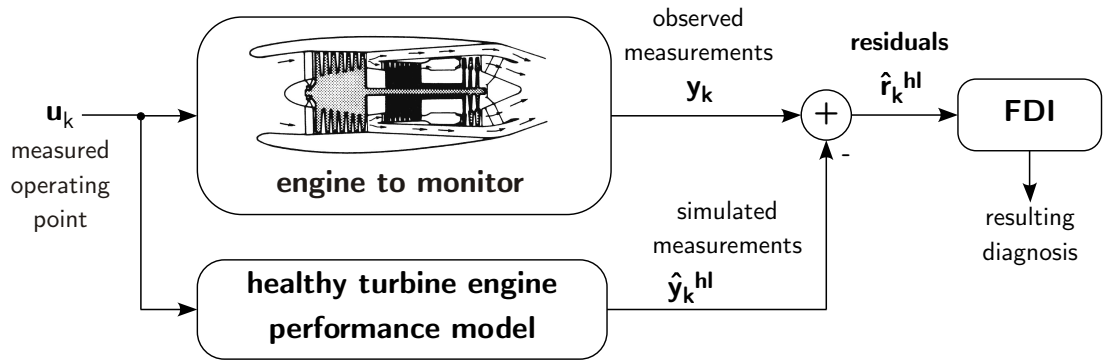


Figure 2.1: Generic scheme of model-based diagnostics including residual generation and fault detection and isolation (FDI) logic

A key component of the model-based approach is clearly the model of the process, that should be able to reproduce the healthy behaviour of the system in terms of the measured quantities. A large majority of processes, including turbine engines, respond not only to changes in their health condition, but also to input signals $\mathbf{u}_k$ that drive the system to a given performance and/or come from an interaction of the system with its surroundings.

In the case of aero-engines, the input signals gather the power setting parameter and the flight conditions as shown in figure 2.2. The power setting parameter is a measurable quantity such as fan speed or engine pressure ratio that serves as an indicator of the amount of thrust generated by the engine. It is adjusted directly by the pilot or the aircraft flight computer. The flight conditions depend on a number of factors such as aircraft

motion, flight velocity and meteorological conditions that define the thermodynamical state of the air.

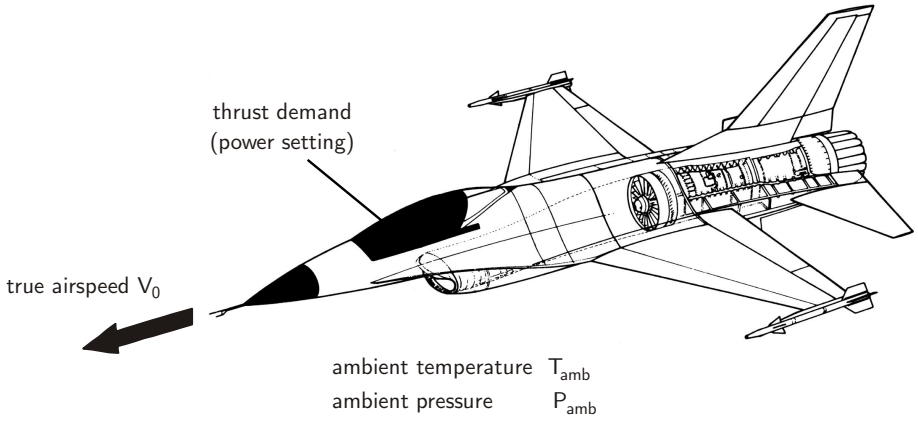


Figure 2.2: Turbine engine in its working environment

Disregarding the effect of humidity, which is negligible at least at high altitude, the flight conditions are completely defined from the knowledge of three measurable parameters: the engine inlet total pressure $\mathsf{P1}$ and temperature $\mathsf{T1}$ and the ambient pressure $\mathsf{P_{amb}}$. The triplet presented in figure 2.2, formed by the ambient pressure $\mathsf{P_{amb}}$ and temperature $\mathsf{T_{amb}}$ and the true airspeed $\mathsf{V_0}$ can be derived from the previous three.

2.1.2 Modelling a turbine engine

Physical considerations

Turbine engines can be modelled at various levels of detail, ranging from simple algebraic relations to a full 3-D description of the gas path. In this spectrum, aero-thermodynamic models (aka. 0-D models) are massively used by the manufacturers throughout an engine program: for preliminary design and performance prediction, for the synthesis of the control laws, for the engine-airframe integration as well as for condition monitoring. Their popularity is due to a moderate computational load and to some degree of flexibility in the modelling that allows numerous effects to be taken into account. These models are able to predict gas path measurements (*e.g.*, temperatures, pressures, air mass flow rates, rotational speeds) as well as global engine performance (*e.g.*, thrust, specific fuel consumption), but do not include measurements related to the lubrication system or vibrational data. For this reason, they are usually named cycle-match models. A number of reference textbooks exist on this topic, see for instance [Walsh and Fletcher, 1998, RTO, 2002].

In a 0-D approach, the engine is subdivided into its major components shown in figure 2.3, such as intake, compressors, combustor, turbines and nozzles. The behaviour of each

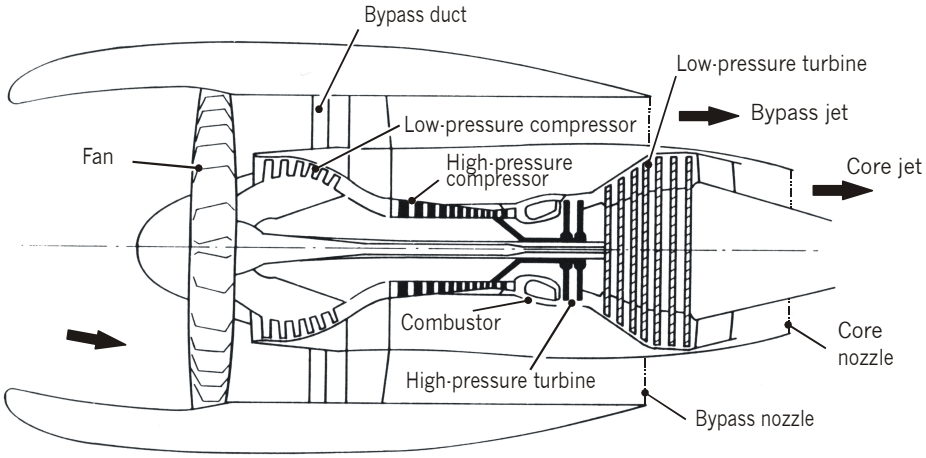


Figure 2.3: Main components of a turbofan engine

component is described with empirical information derived from rig tests or advanced CFD calculations (*e.g.*, compressor and turbine maps) and/or with equations for mass, momentum and energy balances. The components are then linked together through the same conservation equations to lead to the model of the whole engine. The time-dependent behaviour of an engine is governed by three main groups of dynamics [Horobin, 1999]:

- heat transfer between the working fluid and the walls (disks, blades and casings),
- mechanical inertia of rotating parts,
- gas dynamics.

The associated time constants range, for a current commercial turbofan, from several minutes for the heat soakage down to a couple hundredths of a second for the gas dynamics. The early models were aimed at general studies of engine operability and took into account only the spool inertia because it is the dominant effect in the acceleration characteristic. Shaft dynamics is indeed responsible for the excursion of the operating point of the compressor outside of the equilibrium working line during acceleration/deceleration.

With the growing computational power and the better understanding of their physics, heat transfers were gradually included in the performance models. The heat released during the combustion process contributes to the useful power of the turbine and to a warming of the engine parts. It is appreciated that as much as 30% of the fuel flow during an idle-max acceleration is lost for this warming. Heat soakage impact directly the tip clearances, modifying the component efficiency and the stall line.

Gas dynamics models the storage of mass, momentum and energy in the gas flow within a component. It is generally implemented as a trim on the quasi-steady state modelling of that component. Although essential for a representative simulation of high-frequency

events such as compressor stall and fuel spiking, it is oftentimes neglected in cycle-match models as it implies a small integration step (on the order of a millisecond).

Mathematical considerations

The structure of an aero-thermodynamic model can be expressed from a mathematical point of view as a system of non-linear, differential-algebraic equations (DAE):

$$\dot{\mathbf{x}}(t) = f_d\left(\mathbf{x}(t), \mathbf{x}_a(t), \mathbf{u}(t), t\right) \tag{2.1}$$
$$\mathbf{0} = f_a\left(\mathbf{x}(t), \mathbf{x}_a(t), \mathbf{u}(t), t\right) \tag{2.2}$$
$$\mathbf{y}(t) = f_y\left(\mathbf{x}(t), \mathbf{x}_a(t), \mathbf{u}(t), t\right) \tag{2.3}$$

where $\mathbf{x}$ are the dynamic state variables associated to the transient phenomena (*e.g.*, spool speeds for the shaft inertia, metal temperatures for the heat soakage), $\dot{\mathbf{x}}$ stands for the time derivatives of the dynamic state variables, $\mathbf{x}_a$ are the algebraic state variables, or iteration variables, $\mathbf{u}$ are the input signals and $\mathbf{y}$ are the gas path measurements.

The mapping f_d in equation (2.1) describes the dynamic behaviour of the engine, the mapping f_a in equation (2.2) represents the algebraic loops inside the model. Some of these algebraic constraints may result from a truncated dynamic modelling. For instance, if the gas dynamics is disregarded, conservation of the mass flow throughout the engine generates such algebraic relations. It should be noted that the number of elements in vector zero (denoted by $\mathbf{0}$) in equation (2.2) is equal to the number of algebraic states. Finally, the time-varying static mapping f_y in equation (2.3) relates the gas-path measurements $\mathbf{y}$ to the states and the inputs. The time varying nature of the three mappings accounts implicitly for possible internal changes occurring in the engine.

In practice, the solution of the algebraic constraints, by means of a root finding method like the Newton-Raphson one, is handled internally so that from a user's perspective, the model can be summarised with the following system of two vector equations:

$$\dot{\mathbf{x}}(t) = f_c\left(\mathbf{x}(t), \mathbf{u}(t), t\right) \tag{2.4}$$
$$\mathbf{y}(t) = g\left(\mathbf{x}(t), \mathbf{u}(t), t\right) \tag{2.5}$$

where subscript c of the mapping f_c in equation (2.4) underlines the continuous-time character of the model.

Introduction of health parameters

The model defined by equations (2.4-2.5) reproduces the behaviour of the engine in some reference health conditions. As a consequence, discrepancies could be observed between the simulated and the actual measurements, due for instance to manufacturing tolerances

or deterioration. The physical nature of the model offers the opportunity to incorporate adaptive factors intended to take the actual condition of the engine into account [Stamatis et al., 1990]. In the particular scope of condition monitoring, these adaptive factors are called the health parameters $\mathbf{h}$. Their definition is to some extent engine-type and manufacturer dependent. As was explained in section 1.2, the deterioration affects mainly the performance of the turbomachinery components *i.e.*, the compressors and turbines. Consequently, the health parameters are usually defined as corrections, either of multiplicative or additive type, on the flow capacity and on the efficiency of these modules. With the inclusion of the health parameters, the model becomes:

$$\dot{\mathbf{x}}(t) = f_c\left(\mathbf{x}(t), \mathbf{u}(t), \mathbf{h}(t)\right) \tag{2.6}$$

$$\mathbf{y}(t) = g\left(\mathbf{x}(t), \mathbf{u}(t), \mathbf{h}(t)\right) \tag{2.7}$$

For performance prediction studies, the temporal evolution of the health parameters $\mathbf{h}(t)$ should be specified similarly to that of the control inputs $\mathbf{u}(t)$ and the initial state $\mathbf{x}(t=0)$. The possible internal changes in the engine are now parameterised through vector $\mathbf{h}(t)$, therefore the functions f_c and g lose their explicit dependence on the time variable.

Steady-state model

The steady-state performance of the engine is obtained from the stationary solutions of the system of equations (2.6-2.7) *i.e.*, the set of solutions satisfying the condition $\dot{\mathbf{x}}(t) = 0$. From a computational standpoint, the dynamic state variables degenerate into additional algebraic state variables which are added to the root finding method. From a user's perspective, the steady-state model writes:

$$\mathbf{y}_k = \mathcal{G}(\mathbf{u}_k, \mathbf{h}_k) \tag{2.8}$$

where the time variable can be of continuous or discrete nature due to the static character of the mapping $\mathcal{G}$.

Complementary remarks

A number of remarks should be made about the transient and steady-state models. First of all, the models as described above are limited to the gas turbine itself and do not include any sensor or actuator dynamics. While irrelevant for steady-state simulations, these subsystems should be accounted for in the case of transient computations.

The complexity of the model depends to some degree on the type of engine and obviously on the required level accuracy. Typical dimensions for the control and state vectors are respectively between one and five and from two to thirty for a two-spool mixed flow military engine [Härefors, 1999]. The number of algebraic states is about ten. Real-time

simulation can be achieved on computers representative of on-board hardware (*e.g.,* Intel Pentium 100 MHz).

The steady-state and transient models are also supposed to be perfectly faithful *i.e.*, they provide an accurate description of the behaviour of the gas turbine. This implicitly implies that all physics phenomena are perfectly modelled and that the value of all parameters of the gas turbine (*e.g.*, cooling flows, fuel heating values) are known without any uncertainty.

The models provide a plane-average value of the circumferential and radial profiles of temperatures and pressures. However, in most production engines, pressures and temperatures are sensed from single-element probes [Doel, 2003]. We assume that a correction for probe position is built in the model.

For simplicity reasons, the rest of the chapter focuses on steady-state operation in which a commercial engine spends up to 90% of its time. Engine health monitoring during transient operation will be addressed in chapter 5.

2.1.3 Taking uncertainty into account: the statistical model

The user of a gas turbine interacts with the system through the input-output data sequences $\{\mathbf{u}\}_1^N$ and $\{\mathbf{y}\}_1^N$, where N is the number of samples. These signals are measured on the turbine engine and even if the data collection were repeated a number of times in exactly the same operating and health conditions, the sequences $\{\mathbf{u}\}_1^N$ and $\{\mathbf{y}\}_1^N$ would differ due to the presence of various disturbances. In such cases, it is meaningful to consider these sequences as a collection of random vectors of which different realisations are observed. A complete model of the system will then include the description of the statistical properties of the data streams [Ljung, 1999].

Systematic versus random errors

With the assumption previously stated concerning the model, the only source of disturbances is the measurement chain through which the signals are observed. Even if it is stimulated by the actual value $\mathsf{y}_k^\star$ of a thermodynamic property, the acquisition hardware provides the user with a raw measurement y_k.

The measurement error, defined as the difference between the raw measurement and its actual value, comes from imperfections and external disturbances in the measurement chain that corrupt the signal during the measurement process. A typical measurement chain is sketched in figure 2.4 where the quantity to measure (*e.g.*, a pressure) is led by a probe (*e.g.*, a Pitot tube) into a sensing element (*e.g.*, a membrane) responding to the stimulus by producing a secondary signal. A transducer (*e.g.*, a strain gage) with its associated electronics transforms this signal into a conditioned signal compatible with a

digital processing unit. The output of the digital processing unit is usually considered as the measurement reading y_k. A distinction is made between the systematic errors b and the random errors η_k:

$$\mathsf{y}_k = \mathsf{y}_k^\star + \mathsf{b} + \eta_k \tag{2.9}$$

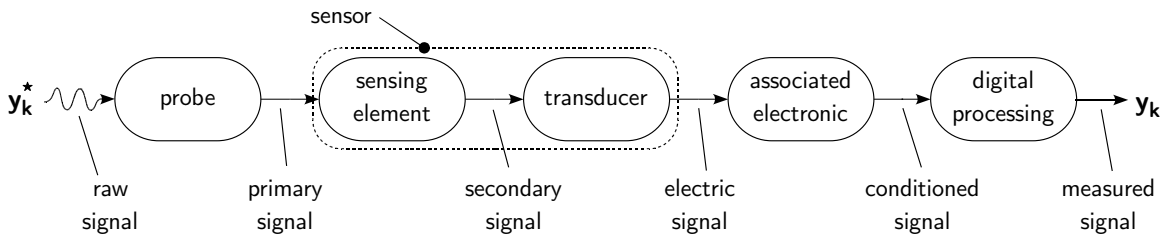


Figure 2.4: Generic structure of a measurement chain

A systematic error is either constant or exhibits a slow drift with respect to the duration of the experiment. It introduces a bias between the measurement reading y_k and the actual value $\mathsf{y}_k^\star$. A significant amount of documentation describes measurement techniques intended to maintain systematic errors as low as possible, see [Saravanamuttoo, 1993, Dudgeon, 1994].

By nature, neither the magnitude nor the sign of random errors η_k are predictable. They are therefore conveniently described by means of a probability density function (pdf) $p(\eta_k)$ that gives the probability of the random error to fall in a certain interval. Random errors are generally modelled as a Gaussian, white noise *i.e.*, a sequence of real numbers generated by independent, identically distributed Gaussian random variables. This choice is justified by both its simplicity, a scalar Gaussian variable being thoroughly defined by only two parameters, and its practical effectiveness[1]. The mean value of the random error is equal to zero (otherwise it should be accounted for in the systematic error term) and the standard deviation σ_y is an image of the scatter of the values around the mean.

The effects of the systematic and random errors on data generated by a measurement chain are illustrated with the following example. Let us try and determine the actual, but unknown value of the quantity $\mathsf{y}^\star$ from a series of recordings $\mathsf{y}_k,\ k = 1, \ldots, N$. Taking into account both systematic and random errors, the model relating the observations to the unknown writes:

$$\mathsf{y}_k = \mathsf{y}^\star + \mathsf{b} + \eta_k \quad \text{with} \quad \eta_k \in \mathcal{N}(0, \sigma_y) \tag{2.10}$$

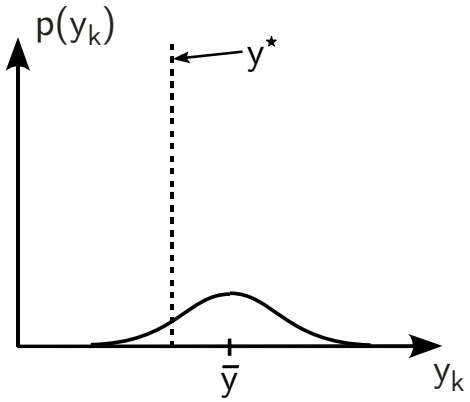
The sample mean is given by:

$$\overline{\mathsf{y}} = \frac{1}{N} \sum_{k=1}^{N} \mathsf{y}_k \tag{2.11}$$

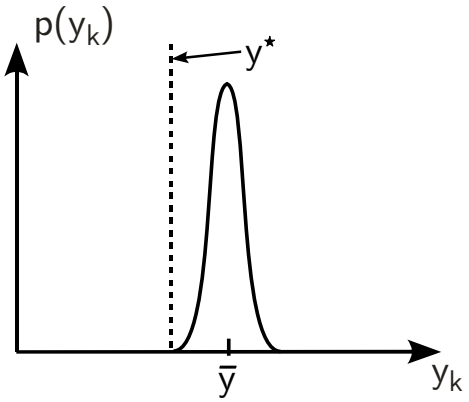
[1]This is also a consequence of the "central limit theorem" mentioned in appendix A.1

and converges towards the true value plus the systematic error as the number of samples N tends to infinity:

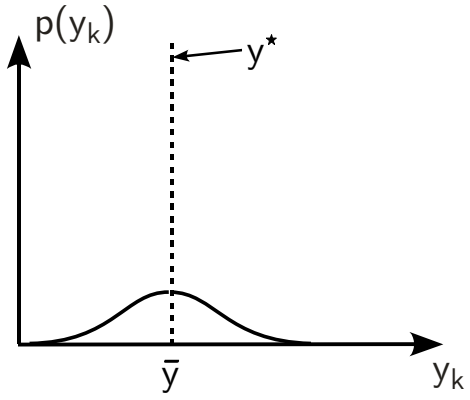
$$\bar{\mathsf{y}} \rightarrow \mathsf{y}^\star + \mathsf{b} \quad \text{as} \quad N \rightarrow \infty \tag{2.12}$$



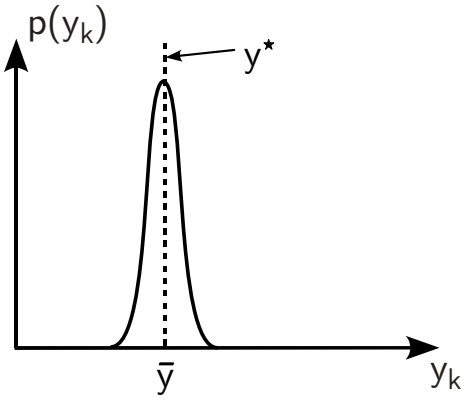
(a) Both systematic and random errors are large: the sensor is not exact and unreliable



(b) Random errors are low but systematic errors are important: the sensor is reliable but not exact



(c) Low systematic errors but important random errors: the sensor is exact but unreliable



(d) Low systematic and random errors: the sensor is accurate (exact and reliable)

Figure 2.5: Different repartitions of measurement results

A measurement chain is commonly characterised by its levels of systematic and random errors as sketched in figure 2.5:

- **reliability** is the property associated with a low level of random errors (σ_y <<). The readings are distributed closely to the mean value $\bar{\mathsf{y}}$ (figures 2.5(b) and 2.5(d)),
- **exactness** is the property associated with a low level of systematic errors ($\mathsf{b} \simeq 0$). The mean value $\bar{\mathsf{y}}$ is close to the actual value $\mathsf{y}^\star$ (figures 2.5(c) and 2.5(d)),

- **accuracy** combines exactness and reliability. Each reading is close to the actual value $\mathsf{y}^\star$ (fig. 2.5(d)).

Installations for which systematic errors are not negligible (fig. 2.5(a) and 2.5(b)) lead to biased results: irrespective of the number of data samples, the sample mean does not give a faithful estimation of the actual value.

The measurement model (2.9) can be extended to a vector form and applied to the quantities sensed in the engine gas path:

$$\mathbf{y}_k = \mathbf{y}_k^\star + \boldsymbol{\eta}_k \tag{2.13}$$

where $\mathbf{y}_k^\star$ are the true values generated with the deterministic steady-state model (2.8) and $\boldsymbol{\eta}_k$ is a vector of Gaussian random variables with mean zero and covariance matrix $\mathbf{R}_{\mathbf{y},k}$. Given the assumption of a perfectly faithful model and of an adequate installation of the sensors, the bias term vanishes ($\mathbf{b} = \mathbf{0}$).

The probability density function of the observations then writes:

$$p(\mathbf{y}_k) = \frac{1}{\sqrt{(2\pi)^{n_y}\,|\mathbf{R}_{\mathbf{y},k}|}}\,\exp\left(-\frac{1}{2}\left(\mathbf{y}_k - \mathbf{y}_k^\star\right)^T \mathbf{R}_{\mathbf{y},k}^{-1}\,\left(\mathbf{y}_k - \mathbf{y}_k^\star\right)\right) \tag{2.14}$$

with

$$E(\mathbf{y}_k) = \mathbf{y}_k^\star \tag{2.15}$$

$$E\left((\mathbf{y}_k - \mathbf{y}_k^\star)(\mathbf{y}_j - \mathbf{y}_j^\star)^T\right) = E\left(\boldsymbol{\eta}_k\,\boldsymbol{\eta}_j^T\right) = \mathbf{R}_{\mathbf{y},\mathbf{k}}\,\delta_{k,j} \tag{2.16}$$

The covariance matrix of the measurement noise $\mathbf{R}_{\mathbf{y},k}$ is a diagonal matrix built up by the collection of the variances $\boldsymbol{\sigma}_y^2$ of all the sensors as the measurement noises are assumed to be uncorrelated. The variance of each sensor is given by the manufacturer or results from an in-house calibration.

Accounting for the uncertainties in the operating conditions

The noise affecting the quantities defining the operating point of the engine should not be overlooked as pointed out by Pinelli and Spina [2002]. The measurement model for the input variables writes:

$$\mathbf{u}_k = \mathbf{u}_k^\star + \boldsymbol{\zeta}_k \tag{2.17}$$

where $\mathbf{u}_k^\star$ are the true input variables and $\boldsymbol{\zeta}_k$ is a vector of Gaussian random variables with mean zero and a diagonal covariance matrix $\mathbf{R}_{\mathbf{u},k}$ made of the variances $\boldsymbol{\sigma}_u^2$ of the sensors. The measurements of the inputs are also assumed to be unbiased ($\mathbf{b} = \mathbf{0}$).

Combining equations (2.8) and (2.13), the gas path measurements are generated according to:

$$\mathbf{y}_k = \mathcal{G}(\mathbf{u}_k^\star, \mathbf{h}_k) + \boldsymbol{\eta}_k \tag{2.18}$$

Introducing the measurement model of the inputs and considering a Taylor series expansion of the mapping $\mathcal{G}$ limited to the first-order, we have:

$$\begin{aligned} \mathbf{y}_k &= \mathcal{G}(\mathbf{u}_k - \boldsymbol{\zeta}_k, \mathbf{h}_k) + \boldsymbol{\eta}_k \qquad (2.19)\\ &\simeq \mathcal{G}(\mathbf{u}_k, \mathbf{h}_k) - \mathbf{G}_{\mathbf{u},k}\, \boldsymbol{\zeta}_k + \boldsymbol{\eta}_k \\ &\simeq \widehat{\mathbf{y}}_{k|h} - \mathbf{G}_{\mathbf{u},k}\, \boldsymbol{\zeta}_k + \boldsymbol{\eta}_k \\ \Rightarrow \boldsymbol{\epsilon}_k = \mathbf{y}_k - \widehat{\mathbf{y}}_{k|h} &\simeq -\mathbf{G}_{\mathbf{u},k}\, \boldsymbol{\zeta}_k + \boldsymbol{\eta}_k \qquad (2.20) \end{aligned}$$

where $\mathbf{G}_{\mathbf{u},k}$ and $\widehat{\mathbf{y}}_{k|h}$ are respectively the Jacobian matrix defined below and the values predicted by the model based on the sensed inputs $\mathbf{u}_k$ and arbitrary values of the health parameters $\mathbf{h}_k$.

$$\mathbf{G}_{\mathbf{u},k} = \left.\frac{\partial \mathcal{G}(\mathbf{u}_k, \mathbf{h}_k)}{\partial \mathbf{u}_k}\right|_{\mathbf{u}_k\,;\,\mathbf{h}_k} \tag{2.21}$$

The statistical properties of the random error $\boldsymbol{\epsilon}_k$ are given by:

$$\begin{aligned} E(\boldsymbol{\epsilon}_k) &= E(\boldsymbol{\eta}_k) - \mathbf{G}_{\mathbf{u},k}\, E(\boldsymbol{\zeta}_k) \\ &= \mathbf{0} \qquad (2.22)\\ E(\boldsymbol{\epsilon}_k\, \boldsymbol{\epsilon}_j^T) &= E\left[(\boldsymbol{\eta}_k - \mathbf{G}_{\mathbf{u},k}\, \boldsymbol{\zeta}_k)\, (\boldsymbol{\eta}_k - \mathbf{G}_{\mathbf{u},j}\, \boldsymbol{\zeta}_j)^T\right] \\ &= E(\boldsymbol{\eta}_k\, \boldsymbol{\eta}_j^T) - E(\boldsymbol{\eta}_k\, \boldsymbol{\zeta}_j^T)\, \mathbf{G}_{\mathbf{u},j}^T \\ &\quad - \mathbf{G}_{\mathbf{u},k}\, E(\boldsymbol{\zeta}_k\, \boldsymbol{\epsilon}_j^T) + \mathbf{G}_{\mathbf{u},k}\, E\left(\boldsymbol{\zeta}_k\, \boldsymbol{\zeta}_j^T\right) \mathbf{G}_{\mathbf{u},j}^T \\ &= \left(\mathbf{R}_{\mathbf{y},k} + \mathbf{G}_{\mathbf{u},k}\, \mathbf{R}_{\mathbf{u},k}\, \mathbf{G}_{\mathbf{u},k}^T\right)\, \delta_{k,j} \triangleq \mathbf{R}_{\mathbf{r},k}\, \delta_{k,j} \qquad (2.23) \end{aligned}$$

as the noises on the inputs and on the gas-path measurements are both zero mean and independent.

Basically, the covariance matrix $\mathbf{R}_{\mathbf{r},k}$ lumps the uncertainties in the operating point and in the gas-path measurements. In general, this matrix is not diagonal. The off-diagonal terms translate the fact that the residuals between the observed and simulated gas-path measurements are correlated. These correlations originates from the use of noisy input measurements in the simulation model. Focusing on the diagonal terms of the matrix $\mathbf{R}_{\mathbf{r},k}$, the uncertainties in the input signals contributes to a reduction in the accuracy of each gas-path sensors with respect to its nominal value:

$$\boldsymbol{\sigma}_{\mathbf{r},k}(i) = \sqrt{\mathbf{R}_{\mathbf{r},k}(i,i)} > \boldsymbol{\sigma}_{\mathbf{y},k}(i) \tag{2.24}$$

The statistical model

The statistical model combines the deterministic model (2.8) that describes the response of the gas turbine and the statistical properties of the data streams:

$$\mathbf{y}_k = \mathcal{G}(\mathbf{u}_k, \mathbf{h}_k) + \boldsymbol{\epsilon}_k \quad \text{with} \quad \boldsymbol{\epsilon}_k \in \mathcal{N}(\mathbf{0}, \mathbf{R}_{\mathbf{r},k}) \tag{2.25}$$

where the lumped noise term $\boldsymbol{\epsilon}_k$ is assumed additive and generated by a Gaussian random vector with zero mean and covariance matrix $\mathbf{R}_{\mathbf{r},k}$.

An equivalent formulation of the statistical model is given directly as a probability density function for the gas-path measurements:

$$p(\mathbf{y}_k; \mathbf{h}_k; \mathbf{u}_k) = \frac{1}{\sqrt{(2\pi)^{n_y}\,|\mathbf{R}_{\mathbf{r},k}|}} \exp\left(-\frac{1}{2}(\mathbf{y}_k - \widehat{\mathbf{y}}_{k|h})^T\, \mathbf{R}_{\mathbf{r},k}^{-1}\, (\mathbf{y}_k - \widehat{\mathbf{y}}_{k|h})\right) \tag{2.26}$$

where the notation $p(\mathbf{y}_k; \mathbf{h}_k; \mathbf{u}_k)$ emphasises the fact that the probability density is parametrised with the health parameters $\mathbf{h}_k$ and depends on the control inputs $\mathbf{u}_k$. In relation (2.26), $\widehat{\mathbf{y}}_{k|h}$ are values generated with the deterministic model (2.8) fed with the *measured* operating conditions $\mathbf{u}_k$ and arbitrary values of the health parameters:

$$\widehat{\mathbf{y}}_{k|h} \triangleq \mathcal{G}(\mathbf{u}_k, \mathbf{h}_k) \tag{2.27}$$

A transformation to whiten the noise

We have established above that the residuals for the different gas-path sensors are correlated due to the noise on the measured input signals, which mathematically takes the form of a full covariance matrix $\mathbf{R}_{\mathbf{r},k}$. We report here the so-called whitening transformation [Kay, 1993, chap. 4, p. 94] that converts the covariance matrix to the identity one. After this transformation, the residuals are white, independent and identically distributed gaussian random variables, which might be a convenient form in some occasions.

The whitening process is based on an eigenvalue decomposition [see Golub and van Loan, 1996] of the covariance matrix. The matrix being positive definite, all eigenvalues are strictly positive:

$$\mathbf{R}_{\mathbf{r},k} = \mathbf{E}_k\, \boldsymbol{\Lambda}_k\, \mathbf{E}_k^T \tag{2.28}$$

where $\boldsymbol{\Lambda}_k$ is a diagonal matrix made of the eigenvalues and $\mathbf{E}_k$ is an orthonormal matrix whose columns are the eigenvectors.

The random vector then undergoes a linear transformation through the matrix $\mathbf{Sc}_k$:

$$\begin{aligned} \boldsymbol{\epsilon}_k \in \mathcal{N}(\mathbf{0}, \mathbf{R}_{\mathbf{r},k}) \quad &\rightarrow \quad \tilde{\boldsymbol{\epsilon}}_k = \mathbf{Sc}_k\, \boldsymbol{\epsilon}_k \in \mathcal{N}(\mathbf{0}, \mathbf{I}) \\ \text{with} \quad \mathbf{Sc}_k \quad &= \quad \boldsymbol{\Lambda}_k^{-1/2}\, \mathbf{E}_k^T \end{aligned} \tag{2.29}$$

where $\mathbf{\Lambda}_k^{-1/2}$ is a diagonal matrix made of the inverse of the square root of the eigenvalues.

Geometrically, the whitening transformation can be interpreted as a rotation+norming operation to express the random vectors in the principal axis of the original covariance matrix $\mathbf{R}_{\mathbf{r},k}$, see appendix A.1 for further details.

2.2 Formulating GPA as an estimation problem

2.2.1 Motivation

At the beginning of this chapter, the residual $\widehat{\mathbf{r}}_k^{\text{hl}}$ was defined as the difference between the measurement readings $\mathbf{y}_k$ and an estimation of these measurements $\widehat{\mathbf{y}}_k^{\text{hl}}$ for a healthy engine. The inclusion of health parameters in the model allows to broaden this notion of residual. Indeed, as underlined by the subscript $k|h$ in equation (2.27), the measurements predicted with the model are now a function of the health parameters. As a consequence, the residuals built as the difference between the actual measurements $\mathbf{y}_k$ and the predicted values $\widehat{\mathbf{y}}_{k|h}$ are also a function of the health parameters $\mathbf{h}_k$:

$$\widehat{\mathbf{r}}_{k|h} \triangleq \mathbf{y}_k - \widehat{\mathbf{y}}_{k|h} \equiv f(\mathbf{h}_k) \tag{2.30}$$

This functional dependency of the residuals with respect to the health parameters naturally leads to a formulation of the gas-path analysis as an estimation problem. Basically this estimation problem can be thought of as the inverse problem to performance prediction, as sketched in figure 2.6.

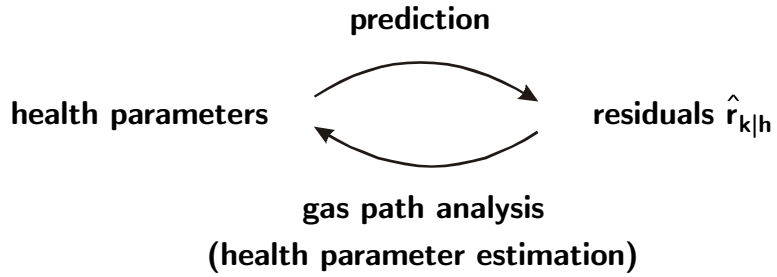


Figure 2.6: Gas path analysis as the inverse problem of performance prediction

The tenet behind this estimation approach is that the health parameters are assessed by finding the value $\widehat{\mathbf{h}}_k$ that minimises the residuals *i.e.*, the distance between the actual measurements $\mathbf{y}_k$ and the model predictions $\widehat{\mathbf{y}}_{k|h}$. Mathematically, it comes down to the solution of a minimisation problem of the type:

$$\widehat{\mathbf{h}}_k = \arg\min_{\mathbf{h}_k} \left\{\rho_k(\mathbf{y}_k - \widehat{\mathbf{y}}_{k|h})\right\} = \arg\min_{\mathbf{h}_k} \left\{\rho_k(\widehat{\mathbf{r}}_{k|h})\right\} \tag{2.31}$$

where the form of the loss function $\rho_k(\cdot)$ should be specified so as to accomplish these two objectives: first, measure the user's satisfaction adequately and second, result in a computationally tractable problem.

2.2.2 The ingredients of an estimation problem

From a general outlook, an estimation problem is composed of the following four components illustrated in figure 2.7 [van Trees, 1968]:

1. a **parameter space** of dimension n_h that defines the range of values that the health parameters may take. Any health condition $\mathbf{h}$ is therefore represented by a vector in this space,
2. a **probabilistic mapping** $p(\mathbf{y};\mathbf{h};\mathbf{u})$ from parameter space to observation space. This mapping is in fact a probability law that governs the effects of the parameters on the observations such as equation (2.26),
3. an **observation space** of dimension n_y that defines the range of values that the observations may take. Any observation $\mathbf{y}$ is represented by a vector in this space,
4. an **estimation rule** $\widehat{\mathbf{h}}(\mathbf{y})$ that assigns a value to the unknown parameters for each realisation of the observation. The estimator is a function of the data, which are random variables, so it is a random variable too. As a result, an estimator is completely described by means of its probability density function or equivalent information.

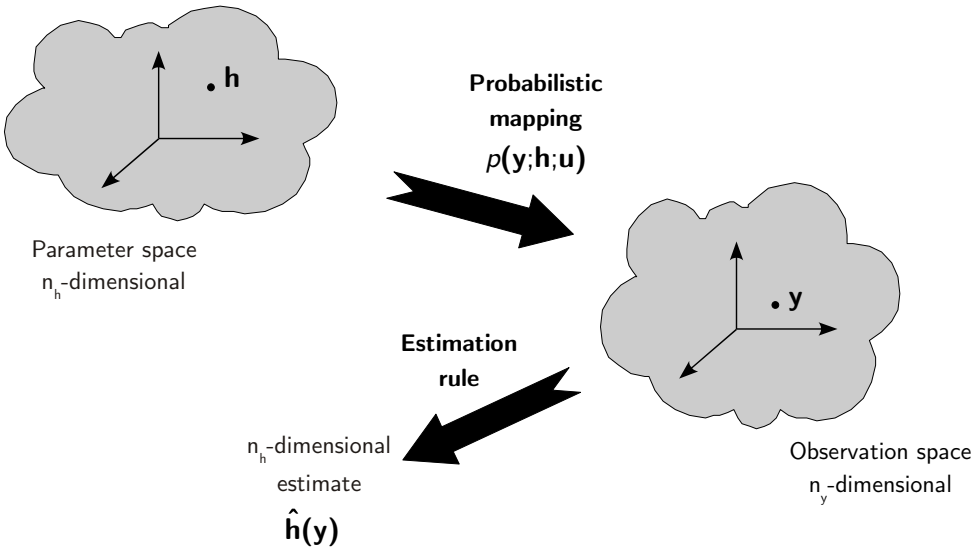


Figure 2.7: The components of an estimation problem

In the following, two celebrated estimation rules are presented. Next to their derivation and implementation, issues such as measure of errors and bounds on the estimator performance are addressed. For a complete discussion, the reader is directed to reference textbooks such as [Kay, 1993].

2.2.3 Working hypotheses

A number of working assumptions, beside the steady-state operation, are made before proceeding to the derivation of the estimators. These assumptions will progressively be waived in the course of the document.

Batch data processing: It means that the operating point and gas-path measurements are collected and stored into databases respectively denoted $\{\,\mathbf{u}\,\}_1^N$ and $\{\,\mathbf{y}\,\}_1^N$ during a given period of time represented here by the number of samples N:

$$\begin{aligned} \{\,\mathbf{u}\,\}_1^N &= [\,\mathbf{u}_1\ \mathbf{u}_2\ \cdots\ \mathbf{u}_N\,] \\ \{\,\mathbf{y}\,\}_1^N &= [\,\mathbf{y}_1\ \mathbf{y}_2\ \cdots\ \mathbf{y}_N\,] \end{aligned}$$

The health parameters are estimated afterwards from the whole databases. This implies a certain delay in the assessment of the engine condition as the databases must be available before carrying out the diagnosis.

Constant health parameters: It is supposed that the health parameters do not show significant variations during the time required to generate the databases. The health parameters associated to the whole batch of data are simply denoted by $\mathbf{h}$ as they do not depend on time.

Linear model: The steady-state behaviour of the engine with respect to the health parameters is approximated with a first-order Taylor expansion around the healthy conditions $\mathbf{h}^{\mathrm{hl}}$:

$$\begin{aligned} \mathbf{y}_k &= \widehat{\mathbf{y}}_k^{\mathrm{hl}} + \mathbf{G}_{\mathbf{h},k}\,(\mathbf{h}-\mathbf{h}^{\mathrm{hl}}) + \boldsymbol{\epsilon}_k \\ \Leftrightarrow \delta\mathbf{y}_k &= \mathbf{G}_{\mathbf{h},k}\,\delta\mathbf{h} + \boldsymbol{\epsilon}_k \end{aligned} \tag{2.32}$$

where

$$\widehat{\mathbf{y}}_k^{\mathrm{hl}} = \mathcal{G}(\mathbf{u}_k, \mathbf{h}^{\mathrm{hl}}) \quad \text{and} \quad \mathbf{G}_{\mathbf{h},k} = \left.\frac{\partial \mathcal{G}(\mathbf{u}_k, \mathbf{h})}{\partial \mathbf{h}}\right|_{\mathbf{u}_k\,;\,\mathbf{h}=\mathbf{h}^{\mathrm{hl}}} \tag{2.33}$$

are respectively the model prediction at the current operating point, for a healthy engine and the so-called influence coefficient matrix that relates the measurement drifts to any

change in the health parameters for the current operating point $\mathbf{u}_k$. In equation (2.32), $\delta\mathbf{y}_k$ is nothing else but the residual with respect to the healthy conditions.

With this linear approximation, the probability density function for the gas-path measurements writes:

$$p(\mathbf{y}_k;\mathbf{h}_k) = \frac{1}{\sqrt{(2\pi)^{n_y}\,|\mathbf{R_{r,k}}|}}\,\exp\left(-\frac{1}{2}(\delta\mathbf{y}_k - \mathbf{G}_{\mathbf{h},k}\,\delta\mathbf{h})^T\,\mathbf{R}_{\mathbf{r},\mathbf{k}}^{-1}\,(\delta\mathbf{y}_k - \mathbf{G}_{\mathbf{h},k}\,\delta\mathbf{h})\right) \tag{2.34}$$

In certain upcoming sections of the document, it will be more convenient to consider the whitened version of the linear model which is obtained by left-multiplying the statistical model (2.32) with the whitening operator $\mathbf{Sc}_k$ defined by relation (2.29):

$$\begin{aligned}\mathbf{Sc}_k\,\{\delta\mathbf{y}_k &= \mathbf{G}_{\mathbf{h},k}\,\delta\mathbf{h} + \boldsymbol{\epsilon}_k\}\\ \delta\tilde{\mathbf{y}}_k &= \tilde{\mathbf{G}}_{\mathbf{h},k}\,\delta\mathbf{h} + \tilde{\boldsymbol{\epsilon}}_k\end{aligned} \tag{2.35}$$

where the so-called scaled model is given by $\tilde{\mathbf{G}}_{\mathbf{h},k} = \mathbf{Sc}_k\ \mathbf{G}_{\mathbf{h},k}$ and the noise vector is obviously distributed as $\tilde{\boldsymbol{\epsilon}}_k \in \mathcal{N}(\mathbf{0},\mathbf{I})$.

2.3 Maximum likelihood estimation

2.3.1 Principle

As previously stated, gas path analysis is an inverse problem that consists in determining the health parameters $\mathbf{h}$ (*i.e.*, the causes) from the observations $\mathbf{y}_k$ (*i.e.*, the effects). The maximum likelihood estimate $\widehat{\mathbf{h}}_{\mathrm{ml}}$ is the value of the health parameters that maximises the so-called likelihood function, here noted $\mathcal{L}(\mathbf{h}|\mathbf{y}_k)$. As its name suggests, this function reflects the likelihood that the observations should indeed take place, it basically is the probability density of the measurements to be observed as a function of the health parameters.

With this definition, the likelihood function coincides mathematically with the probability density of the measurements (2.34): $\mathcal{L}(\mathbf{h}|\mathbf{y}_k) \equiv p(\mathbf{y}_k;\mathbf{h}_k)$. However, a major difference between the likelihood and the probability density of the measurements lies in what is fixed and what may vary. In the likelihood perspective, the observations are fixed and the health parameters are the unknowns. The maximum likelihood estimator therefore inverts the role of the observations (the effects) and the parameters (the causes).

2.3.2 Derivation of the estimator

For a batch of data $\{\,\mathbf{y}\,\}_1^N$, the maximum likelihood estimate is the solution of the following optimisation problem:

$$\widehat{\mathbf{h}}_{\mathrm{ml}} = \arg\max_{\mathbf{h}} \left\{\mathcal{L}(\mathbf{h}|\{\,\mathbf{y}\,\}_1^N)\right\} \tag{2.36}$$

The likelihood function for one of the samples in the database is, under our working hypotheses, given by:

$$\begin{aligned}\mathcal{L}(\mathbf{h}|\mathbf{y}_k) &\equiv p(\mathbf{y}_k;\mathbf{h}_k)\\ &= \frac{1}{\sqrt{(2\pi)^{n_y}\,|\mathbf{R}_{\mathbf{r},\mathbf{k}}|}}\,\exp\left(-\frac{1}{2}(\delta\mathbf{y}_k-\mathbf{G}_{\mathbf{h},k}\,\delta\mathbf{h})^T\,\mathbf{R}_{\mathbf{r},\mathbf{k}}^{-1}\,(\delta\mathbf{y}_k-\mathbf{G}_{\mathbf{h},k}\,\delta\mathbf{h})\right)\end{aligned} \tag{2.37}$$

Since the measurement noise is generated by a white and Gaussian random variable, each sample in the batch is *statistically* independent of the others. The likelihood for the N samples in the database becomes:

$$\mathcal{L}(\mathbf{h}|\{\,\mathbf{y}\,\}_1^N) = \prod_{k=1}^{N}\mathcal{L}(\mathbf{h}|\mathbf{y}_k) \tag{2.38}$$

The maximisation problem defined in (2.36) is converted into a minimisation one:

$$\begin{aligned}\mathcal{J}_{\mathrm{ml}}(\mathbf{h}) &= -\log\prod_{k=1}^{N}\mathcal{L}(\mathbf{h}|\mathbf{y}_k)\\ &= -\sum_{k=1}^{N}\log\mathcal{L}(\mathbf{h}|\mathbf{y}_k)\end{aligned} \tag{2.39}$$

as the opposite of the natural logarithm is a monotonically decreasing function.

The maximum likelihood estimate is the value that cancels the derivatives of the objective function with respect to the parameters, which gives after substituting (2.37) into (2.39):

$$\frac{\partial\mathcal{J}_{\mathrm{ml}}(\mathbf{h})}{\partial\mathbf{h}} = \sum_{k=1}^{N} -\mathbf{G}_{\mathbf{h},k}^T\,\mathbf{R}_{\mathbf{r},\mathbf{k}}^{-1}\,(\delta\mathbf{y}_k-\mathbf{G}_{\mathbf{h},k}\,\delta\mathbf{h}) = \mathbf{0} \quad @ \quad \mathbf{h}=\widehat{\mathbf{h}}_{\mathrm{ml}} \tag{2.40}$$

Aggregating the data samples column-wise into global vectors and matrices defined as:

$$\delta\mathbf{y} = \begin{bmatrix}\delta\mathbf{y}_1\\ \vdots\\ \delta\mathbf{y}_N\end{bmatrix}\;;\;\mathbf{G}_{\mathbf{h}} = \begin{bmatrix}\mathbf{G}_{\mathbf{h},1}\\ \vdots\\ \mathbf{G}_{\mathbf{h},N}\end{bmatrix}\;\text{and}\;\mathbf{R}_{\mathbf{r}} = \begin{bmatrix}\mathbf{R}_{\mathbf{r},\mathbf{1}} & & \mathbf{0}\\ & \ddots & \\ \mathbf{0} & & \mathbf{R}_{\mathbf{r},\mathbf{N}}\end{bmatrix} \tag{2.41}$$

equation (2.40) can be written as:

$$\mathbf{G}_\mathbf{h}^T \, \mathbf{R}_\mathbf{r}^{-1} \, \delta\mathbf{y} = \mathbf{G}_\mathbf{h}^T \, \mathbf{R}_\mathbf{r}^{-1} \, \mathbf{G}_\mathbf{h} \, \delta\mathbf{h} \tag{2.42}$$

which leads to the following expression of the maximum likelihood estimate:

$$\widehat{\mathbf{h}}_{\mathrm{ml}} = \mathbf{h}^{\mathrm{hl}} + \left(\mathbf{G}_\mathbf{h}^T \, \mathbf{R}_\mathbf{r}^{-1} \, \mathbf{G}_\mathbf{h}\right)^{-1} \mathbf{G}_\mathbf{h}^T \, \mathbf{R}_\mathbf{r}^{-1} \, \delta\mathbf{y} \tag{2.43}$$

The likelihood function achieves a minimum value at the maximum likelihood estimate. Indeed, the Hessian of the objective function is given by:

$$\frac{\partial^2 \mathcal{J}_{\mathrm{ml}}(\mathbf{h})}{\partial \mathbf{h}^2} = \mathbf{G}_\mathbf{h}^T \, \mathbf{R}_\mathbf{r}^{-1} \, \mathbf{G}_\mathbf{h} \tag{2.44}$$

which is a positive semidefinite matrix as $\mathbf{R}_\mathbf{r}$ is built from the covariance matrices $\mathbf{R}_{\mathbf{r},\mathbf{k}}$.

2.3.3 Complementary remarks

Equation (2.43) requires that the Hessian of the objective function be invertible. This means that the rank of matrix $\mathbf{G}_\mathbf{h}$ must be equal to its number of columns *i.e.*, to the number of health parameters. If the samples in the database are related to a single operating point $\mathbf{u}_k$, this condition can only be fulfilled if the number of gas-path sensors n_y is at least equal to the number of health parameters n_h.

Otherwise, the estimation problem is underdetermined which leads to a continuum of solutions that spans the null-space of the influence coefficient matrix $\mathbf{G}_\mathbf{h}$. Indeed, linear algebra tells us that the projection of any solution vector onto the null-space is not reflected in the measurements. This point will be addressed in more details in section 2.5.

Another possibility to guarantee a full-rank influence coefficient matrix is to collect data at different operating conditions, thereby taking advantage of the non-linearity of the engine response, this is the multi-point estimation mentioned in the review of methods. However, this multi-point framework turns out to be a rather long and therefore costly procedure on a test bed as it requires the stabilisation of the engine at several regimes. On the other hand, it is very seldom to achieve steady-state conditions outside the cruise part of a commercial flight.

The batch treatment of data acts as a basic filter of the measurement noise. For a single-point estimation, it can be shown that the maximum likelihood estimator (2.43) applied to the batch or to its sample average leads to the same results. In the next section, some performance metrics of the maximum likelihood estimator are presented.

2.3.4 Measure of performance

The quality of an estimator can be assessed through a number of metrics, the two most popular being the bias and the covariance matrix respectively defined as:

$$\mathbf{b} \triangleq E\left\{\widehat{\mathbf{h}}_{\mathrm{ml}} - \mathbf{h}\right\} \tag{2.45}$$

$$\mathbf{P} \triangleq E\left\{\left(\widehat{\mathbf{h}}_{\mathrm{ml}} - E\left\{\widehat{\mathbf{h}}_{\mathrm{ml}}\right\}\right)\left(\widehat{\mathbf{h}}_{\mathrm{ml}} - E\left\{\widehat{\mathbf{h}}_{\mathrm{ml}}\right\}\right)^{T}\right\} \tag{2.46}$$

where the expectation is taken over the measurements, the only random variables in the problem. Simply put, the bias is the mean value of the estimation error and the covariance matrix is a measure of the spread of the estimation error.

It can be shown that, for a well-posed problem, the maximum likelihood estimator is unbiased ($\mathbf{b} = \mathbf{0}$) *i.e.*, on average, the estimator yields the true value of the parameters. It also minimises the variance in the estimates, which means that the estimation error is as much as possible concentrated around zero.

However, the variance of the estimates cannot be decreased to any arbitrary level, as stated by the Cramer-Rao inequality [Rao, 1973]:

$$\sigma_h^2(i) \triangleq \mathrm{var}\left\{\widehat{\mathbf{h}}_{\mathrm{mle}}(i)\right\} \geq \mathbf{FIM}_{i,i}^{-1} \tag{2.47}$$

where $\mathbf{FIM}$ is the Fisher Information Matrix (FIM). The inequality expresses the physical impossibility to find an unbiased estimate whose variance is lower than the Cramer-Rao bound.

The FIM quantifies the extent to which the observations reduce the uncertainty regarding the unknown parameters. Loosely speaking, it is intimately linked to the sensitivity of the measurements with respect to the parameters. Mathematically, the FIM is defined as the variance of the gradient of the natural logarithm of the probabilistic model relating the observations to the parameters [Kay, 1993], leading to the following expression for the element (i, j):

$$\mathbf{FIM}_{i,j} \triangleq E\left\{\frac{\partial}{\partial \mathbf{h}_i}\log p(\mathbf{y}_k;\mathbf{h}) \cdot \frac{\partial}{\partial \mathbf{h}_j}\log p(\mathbf{y}_k;\mathbf{h})\right\} \tag{2.48}$$

In the case of the linear Gaussian model (2.34), the FIM writes:

$$\mathbf{FIM} = \mathbf{G}_{\mathbf{h}}^{T}\,\mathbf{R}_{\mathbf{r}}^{-1}\,\mathbf{G}_{\mathbf{h}} \tag{2.49}$$

The maximum likelihood estimator is hence asymptotically unbiased and of minimum variance. The combination of these two properties makes it an asymptotically efficient estimator distributed according to:

$$\widehat{\mathbf{h}}_{\text{ml}} \rightarrow \mathcal{N}(\mathbf{h}, \mathbf{FIM}^{-1}) \quad \text{as} \quad N \rightarrow \infty \tag{2.50}$$

The required number of data samples N for the asymptotic results to hold depends on the signal to noise ratio.

2.3.5 Connection with the least-squares method

The famous least-squares method introduced by Gauss at the end of the eighteenth century has been widely applied to data fitting problems in a myriad of scientific fields. The criterion for goodness of fit consists in minimising the sum of squares between the observations and the prediction of a model, without any probabilistic assumption about the data. Using the aggregated system description, the objective function of the least-squares approach writes:

$$\mathcal{J}_{\text{ls}} = (\delta\mathbf{y} - \mathbf{G_h}\ \delta\mathbf{h})^T\ (\delta\mathbf{y} - \mathbf{G_h}\ \delta\mathbf{h}) \tag{2.51}$$

The least-squares estimate is the value that minimises the above objective function:

$$\begin{aligned} \widehat{\mathbf{h}}_{\text{ls}} &= \arg\min_{\mathbf{h}}\ \{\mathcal{J}_{\text{ls}}\} \\ &= \mathbf{h}^{\text{hl}} + \left(\mathbf{G}_\mathbf{h}^T\ \mathbf{G_h}\right)^{-1} \mathbf{G}_\mathbf{h}^T\ \delta\mathbf{y} \end{aligned} \tag{2.52}$$

The least-squares estimate can be interpreted in a geometric way. Let $\mathbf{e}_y = \delta\mathbf{y} - \mathbf{G_h}\ \delta\hat{\mathbf{h}}$ be the prediction error. Then the least-squares estimate is solution of:

$$\mathbf{e}_y^T\ \mathbf{G_h} = \mathbf{0}^T \tag{2.53}$$

This equation means that the prediction error must be perpendicular to the columns of the influence coefficient matrix. The error therefore represents the part of the observations that cannot be explained by the signal model. This is known as the orthogonality principle.

A modified version of the least-squares method applies a weighting factor to each observations leading to the so-called weighted least-squares solution:

$$\widehat{\mathbf{h}}_{\text{wls}} = \mathbf{h}^{\text{hl}} + \left(\mathbf{G}_\mathbf{h}^T\ \mathbf{W}\ \mathbf{G_h}\right)^{-1} \mathbf{G}_\mathbf{h}^T\ \mathbf{W}\ \delta\mathbf{y} \tag{2.54}$$

where $\mathbf{W}$ is the matrix of weighting factors.

It can be seen that if the weighting matrix $\mathbf{W}$ is chosen equal to $\mathbf{R}_\mathbf{r}^{-1}$ *i.e.*, the inverse of the covariance matrix of the residuals, the weighted least-squares estimator has exactly the same expression as the maximum likelihood estimator. This is a consequence of the combination of a linear model for the observation–parameter relationships and a Gaussian distribution of the measurement noise.

2.4 Maximum a posteriori estimation

2.4.1 A Bayesian view to estimation

In the maximum likelihood approach, the health parameters are considered as deterministic whereas the observations are random variables, they are hence not of the same nature. Moreover, only asymptotic results are available concerning the uncertainty in the estimated parameters. The Bayesian philosophy tries to address these issues by recognising the parameters a random character too. As a result, the health parameters are described with a probability density $p(\mathbf{h})$ and the estimation problem is to assess a particular realisation of these random parameters.

The main motivation behind Bayesian estimation is that if some prior knowledge about the parameters is available (from previous experience or engineering judgement for instance), it should be incorporated into the estimation process. As explained by Rao [1973], the use of prior knowledge leads to a more accurate estimator in certain regions of the parameter space around the a priori value, provided the prior knowledge is relevant of course. By more accurate, it is meant an estimator achieving a lower mean square error than the unbiased, minimum variance estimator.

The observations and the health parameters being both random variables, they are described through the joint probability density $p(\mathbf{y}_k, \mathbf{h})$, which can be decomposed as:

$$\begin{aligned} p(\mathbf{y}_k, \mathbf{h}) &= p(\mathbf{h}|\mathbf{y}_k) \cdot p(\mathbf{y}_k) \\ &= p(\mathbf{y}_k|\mathbf{h}) \cdot p(\mathbf{h}) \end{aligned} \tag{2.55}$$

where

- $p(\mathbf{y}_k|\mathbf{h})$ is the conditional probability density of the measurements when the health parameters take the specific value $\mathbf{h}$. It expresses the dependance of the measurements with respect to the health parameters, as given by the statistical model (2.34),
- $p(\mathbf{h})$ is the so-called a priori knowledge about the health parameters namely, the information available before the measurements $\mathbf{y}_k$ are observed,
- $p(\mathbf{h}|\mathbf{y}_k)$ is the so-called posterior distribution, that summarises the knowledge about the parameters after using the information contained in the observations,
- $p(\mathbf{y}_k)$ is here a simple normalising factor since the measurements $\mathbf{y}_k$ are observed.

In the Bayesian framework, the posterior distribution appears to be the relevant quantity to determine. Rearranging the right-hand sides of (2.55) gives the so-called Bayes' rule expressing the posterior in terms of the others densities:

$$p(\mathbf{h}|\mathbf{y}_k) = \frac{p(\mathbf{y}_k|\mathbf{h}) \cdot p(\mathbf{h})}{p(\mathbf{y}_k)} \tag{2.56}$$

An optimal estimator can be defined as the one that minimises the Bayesian mean square error (bmse), leading to the minimum mean square estimator:

$$\begin{aligned} \widehat{\mathbf{h}}_{\text{mmse}} &\triangleq \arg\min \left\{ \text{bmse}(\widehat{\mathbf{h}}_{\text{mmse}}) \right\} \\ &= \arg\min \left\{ E\left\{ \left(\widehat{\mathbf{h}}_{\text{mmse}} - \mathbf{h}\right)^T \left(\widehat{\mathbf{h}}_{\text{mmse}} - \mathbf{h}\right) \right\} \right\} \end{aligned} \tag{2.57}$$

note that in the computation of the bmse, the expectation is taken with respect to the observations and the parameters as they are all random variables.

It can be shown that $\widehat{\mathbf{h}}_{\text{mmse}}$ corresponds to the mean of the posterior distribution [see *e.g.,* van Trees, 1968]. The computation of the mmse estimator involves many integrations, that's why the maximum a posteriori estimate is sometimes preferred as it comes down to a maximisation problem, generally easier to solve. The maximum a posteriori estimate is defined as the mode of the posterior *i.e.*, the location of the maximum:

$$\widehat{\mathbf{h}}_{\text{map}} \triangleq \arg\max_{\mathbf{h}} \{p(\mathbf{h}|\,\mathbf{y}_k)\} \tag{2.58}$$

As already mentioned, the choice of the prior distribution for the parameters happens to be a key element of the Bayesian approach. A wrong choice would result in a poor estimation quality. The prior $p(\mathbf{h})$ should reflect the prior knowledge in a faithful manner. However, in the specific case of performance monitoring, the prior knowledge is often rather coarse and it is difficult to find a density that renders this information. Therefore, it might be more valuable to choose a prior density that models only approximately the a priori knowledge, but that leads to a posterior density for which a closed-form solution exists.

In the present case, it is convenient to select a Gaussian prior. Indeed, combined with the conditional density $p(\mathbf{y}_k|\,\mathbf{h})$, that is also Gaussian, it leads to a Gaussian posterior $p(\mathbf{h}|\,\mathbf{y}_k)$. In this particular case, the mode coincides with the mean: $\widehat{\mathbf{h}}_{\text{map}} \equiv \widehat{\mathbf{h}}_{\text{mmse}}$.

2.4.2 Derivation of the estimator

For a batch of data $\{\,\mathbf{y}\,\}_1^N$, Bayes' rule (2.56) becomes:

$$p(\mathbf{h}|\,\{\,\mathbf{y}\,\}_1^N) = \frac{p(\{\,\mathbf{y}\,\}_1^N|\,\mathbf{h}) \cdot p(\mathbf{h})}{p(\{\,\mathbf{y}\,\}_1^N)} \tag{2.59}$$

and the map estimate is the solution of the following maximisation problem:

$$\begin{aligned} \widehat{\mathbf{h}}_{\text{map}} &= \arg\max_{\mathbf{h}} \left\{ p(\mathbf{h}|\,\{\,\mathbf{y}\,\}_1^N) \right\} \\ &= \arg\max_{\mathbf{h}} \left\{ p(\{\,\mathbf{y}\,\}_1^N|\,\mathbf{h}) \cdot p(\mathbf{h}) \right\} \end{aligned} \tag{2.60}$$

as $p(\{\,\mathbf{y}\,\}_1^N)$ does not depend on the health parameters.

Since the measurement noise is assumed to be independent and identically-distributed, the probability density of the whole batch of observations conditioned on the parameters can be decomposed as:

$$p(\{\mathbf{y}\}_1^N|\mathbf{h}) = \prod_{k=1}^{N} p(\mathbf{y}_k|\mathbf{h}) \tag{2.61}$$

As stated previously, the probability density of a data sample conditioned on the parameters $p(\mathbf{y}_k|\mathbf{h})$ has the same functional expression as the probability density (2.34) of the statistical model. Note that both distributions have the same form, but have a different statistical interpretation. In the Bayesian framework, it is a conditional distribution, computed for a particular realisation of the random parameters, whereas the statistical model is simply the probability density for the measurements, parameterised by a deterministic vector $\mathbf{h}$ of real numbers.

As indicated above, the prior knowledge is chosen as a Gaussian distribution of mean $\widehat{\mathbf{h}}_0$ and covariance matrix $\mathbf{Q}_0$:

$$p(\mathbf{h}) = \frac{1}{\sqrt{(2\pi)^{n_h}\,|\mathbf{Q}_0|}} \exp\left[-\frac{1}{2}(\mathbf{h}-\widehat{\mathbf{h}}_0)^T\,\mathbf{Q}_0^{-1}\,(\mathbf{h}-\widehat{\mathbf{h}}_0)\right] \tag{2.62}$$

The maximisation problem defined in (2.60) is converted into a minimisation one by considering the following objective function:

$$\begin{aligned} \mathcal{J}_{\text{map}}(\mathbf{h}) &= p(\mathbf{h}) \cdot \prod_{k=1}^{N} p(\mathbf{y}_k|\mathbf{h}) \\ &= -\log p(\mathbf{h}) - \sum_{k=1}^{N} \log p(\mathbf{y}_k|\mathbf{h}) \end{aligned} \tag{2.63}$$

as the opposite of the natural logarithm is a monotonically decreasing function.

Substituting the expressions (2.34) and (2.62) for the conditional and prior densities and using the stacked database quantities defined in (2.41), the objective function becomes:

$$\mathcal{J}_{\text{map}}(\mathbf{h}) = \text{cst} + \frac{1}{2}\,(\delta\mathbf{y} - \mathbf{G_h}\,\delta\mathbf{h})^T\,\mathbf{R_r}^{-1}\,(\delta\mathbf{y} - \mathbf{G_h}\,\delta\mathbf{h}) + \frac{1}{2}\,(\mathbf{h}-\widehat{\mathbf{h}}_0)^T\mathbf{Q}_0^{-1}(\mathbf{h}-\widehat{\mathbf{h}}_0) \tag{2.64}$$

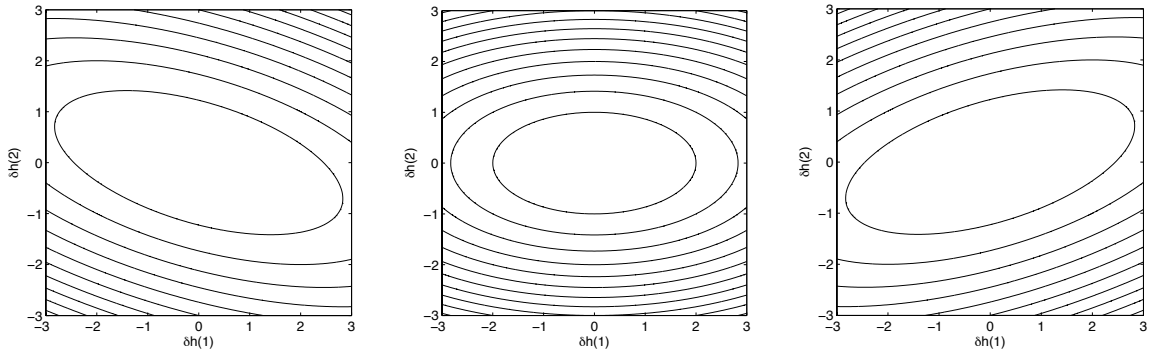
The second term in the cost function (2.64) is the same as in the maximum likelihood approach. It provides a quadratic penalisation of the residuals similar to the weighted least-squares criterion. The last term can be interpreted as a "constraint" that penalises (also in a quadratic manner) solutions far from the reference value $\widehat{\mathbf{h}}_0$. The maximum a posteriori estimation is equivalent from an optimisation standpoint to ridge regression/regularisation.

The mean $\widehat{\mathbf{h}}_0$ and covariance matrix $\mathbf{Q}_0$ of the prior can be regarded as tuning parameters of the estimator. The mean is usually taken equal to the nominal conditions (*i.e.*, $\widehat{\mathbf{h}}_0 = \mathbf{h}^{\text{hl}}$) as sensible information is rarely available. On the contrary, guidelines with a rather physical interpretation may help in building the covariance matrix $\mathbf{Q}_0$. The diagonal terms are set as $(\delta\mathbf{h}_{max}/3)^2$ where $\delta\mathbf{h}_{max}$ reflects the expected maximum deviation for each health parameter. Limiting $\mathbf{Q}_0$ to a pure diagonal matrix supposes that the health parameters vary independently. However, it is highly likely that the degradation of a given component will involve a modification of both its flow capacity and its efficiency which means that these parameters are, in some way, correlated. As explained by Doel [2003], such correlations can be introduced into the covariance matrix by setting the adequate off-diagonal terms of the matrix $\mathbf{Q}_0$ according to:

$$\mathbf{Q}_0(i,j) = \mathbf{Q}_0(j,i) = \rho_{i\leftrightarrow j}\ \sqrt{\mathbf{Q}_0(i,i)\cdot\mathbf{Q}_0(j,j)} \tag{2.65}$$

where $\rho_{i\leftrightarrow j}$ is the correlation factor.

Figure 2.8 shows the effect of the correlation factor on the contours of the prior for a 2-D example. A negative correlation factor is used in the case where the parameters evolve in opposite directions (*i.e.*, one increases, the other one decreases). A nought correlation factor leads to independent parameters. In figure 2.8(b), it is assumed that parameter $\delta\mathbf{h}(1)$ might vary twice as much as parameter $\delta\mathbf{h}(2)$. Finally, a positive correlation factor reflects the fact that both parameters vary in the same direction (*i.e.*, both increase or decrease). The correlation factor satisfies $|\rho_{i\leftrightarrow j}| \leq 1$, the equality meaning a total correlation. In any case, the resulting covariance matrix $\mathbf{Q}_0$ should be positive definite which translates geometrically by closed-contours as shown in figure 2.8. Coupling between different modules could be introduced with the same technique if appropriate.



(a) $\rho_{i\leftrightarrow j} = -0.5$ (b) $\rho_{i\leftrightarrow j} = 0$ (c) $\rho_{i\leftrightarrow j} = 0.5$

Figure 2.8: Effect of the correlation factor on the contours of the prior

The maximum a posteriori estimate is the value that cancels the derivatives of the objective function (2.64) with respect to the parameters:

$$\frac{\partial\mathcal{J}_{\text{map}}(\mathbf{h})}{\partial\mathbf{h}} = -\mathbf{G}_\mathbf{h}^T\ \mathbf{R}_\mathbf{r}^{-1}\ (\delta\mathbf{y} - \mathbf{G}_\mathbf{h}\ \delta\mathbf{h}) + \mathbf{Q}_0^{-1}\ \delta\mathbf{h} = \mathbf{0} \quad @ \quad \mathbf{h} = \widehat{\mathbf{h}}_{\text{map}} \tag{2.66}$$

which leads to the following expression of the maximum a posteriori estimate:

$$\widehat{\mathbf{h}}_{\text{map}} = \mathbf{h}^{\text{hl}} + \left(\mathbf{Q}_0^{-1} + \mathbf{G}_\mathbf{h}^T\ \mathbf{R}_\mathbf{r}^{-1}\ \mathbf{G}_\mathbf{h}\right)^{-1} \mathbf{G}_\mathbf{h}^T\ \mathbf{R}_\mathbf{r}^{-1}\ \delta\mathbf{y} \tag{2.67}$$

The objective function achieves a minimum value at the maximum a posteriori estimate. Indeed, its Hessian is given by:

$$\frac{\partial^2 \mathcal{J}_{\text{map}}(\mathbf{h})}{\partial \mathbf{h}^2} = \mathbf{G}_\mathbf{h}^T\ \mathbf{R}_\mathbf{r}^{-1}\ \mathbf{G}_\mathbf{h} + \mathbf{Q}_0^{-1} \tag{2.68}$$

which is positive definite as $\mathbf{R}_\mathbf{r}$ and $\mathbf{Q}_0$ are too.

2.4.3 Complementary remarks

From a mathematical standpoint, the major difference between relations (2.43) and (2.67) lies in the presence of the matrix $\mathbf{Q}_0^{-1}$ in the Hessian. As mentioned in section 2.3, a reliable diagnosis is obtained provided the inner product $\mathbf{G}_\mathbf{h}^T\mathbf{R}_\mathbf{r}^{-1}\mathbf{G}_\mathbf{h}$ is invertible (*i.e.*, full ranked) which can be obtained if the lines of $\mathbf{G}_\mathbf{h}$ are linearly independent. The main effect of the additive term $\mathbf{Q}_0^{-1}$ is to artificially increase the rank of the matrix $\mathbf{Q}_0^{-1} + \mathbf{G}_\mathbf{h}^T\mathbf{R}_\mathbf{r}^{-1}\mathbf{G}_\mathbf{h}$ to improve its conditionning. In the case of underdetermined problems, ridge regression can therefore be seen as a mathematical trick that drives the estimator towards one of the many solutions.

The implementation of the maximum a posteriori estimator (2.67) involves three matrix inversions. A more efficient formulation is obtained by using the following matrix inversion lemma:

$$\left(\mathbf{A} - \mathbf{B}\ \mathbf{D}^{-1}\mathbf{C}\right)^{-1} \mathbf{B}\ \mathbf{D}^{-1} = \mathbf{A}^{-1}\ \mathbf{B}\left(\mathbf{D} - \mathbf{C}\ \mathbf{A}^{-1}\mathbf{B}\right)^{-1} \tag{2.69}$$

and identifying $\mathbf{A} \equiv \mathbf{Q}_0^{-1}$, $\mathbf{B} \equiv \mathbf{G}_\mathbf{h}^T$, $\mathbf{C} \equiv -\mathbf{G}_\mathbf{h}^T$ and $\mathbf{D} \equiv \mathbf{R}_\mathbf{r}$, the maximum a posteriori estimator becomes:

$$\widehat{\mathbf{h}}_{\text{map}} = \mathbf{h}^{\text{hl}} + \mathbf{Q}_0\ \mathbf{G}_\mathbf{h}^T\ \left(\mathbf{R}_\mathbf{r} + \mathbf{G}_\mathbf{h}\ \mathbf{Q}_0\ \mathbf{G}_\mathbf{h}^T\right)^{-1} \delta\mathbf{y} \tag{2.70}$$

which involves only one matrix inversion.

Introducing the covariance matrix of the data $\mathbf{P}_{yy}$ and the cross-covariance matrix between the health parameters and the data $\mathbf{P}_{hy}$:

$$\mathbf{P}_{yy} = E\{\delta\mathbf{y}\ \delta\mathbf{y}^T\} = E\{(\mathbf{G}_\mathbf{h}\ \delta\mathbf{h} + \boldsymbol{\epsilon})\ (\mathbf{G}_\mathbf{h}\ \delta\mathbf{h} + \boldsymbol{\epsilon})^T\} = \mathbf{R}_\mathbf{r} + \mathbf{G}_\mathbf{h}\ \mathbf{Q}_0\ \mathbf{G}_\mathbf{h}^T \tag{2.71}$$

$$\mathbf{P}_{hy} = E\{\delta\mathbf{h}\ \delta\mathbf{y}^T\} = E\{\delta\mathbf{h}\ (\mathbf{G}_\mathbf{h}\ \delta\mathbf{h} + \boldsymbol{\epsilon})^T\} = \mathbf{Q}_0\ \mathbf{G}_\mathbf{h}^T \tag{2.72}$$

the maximum a posteriori estimator can be compactly rewritten as:

$$\widehat{\mathbf{h}}_{\text{map}} = \mathbf{h}^{\text{hl}} + \mathbf{P}_{hy}\ \mathbf{P}_{yy}^{-1}\ \delta\mathbf{y} \tag{2.73}$$

The estimator (2.73) is linear in the data. This structure results from two assumptions: first, the measurement noise and the prior knowledge are jointly Gaussian random variables. Second, the measurements are related linearly to the parameters. This leads to an objective function (2.64) that is quadratic in the unknown health parameters.

As previously noticed, the hypothesis of jointly Gaussian data and parameters makes the map estimator coincide with the mmse estimator. We now review three properties of the mmse estimator that will be useful for the derivation of the Kalman filter in chapter 3.

Property 1. The mmse estimator commutes over linear transformation. Let $\mathbf{h}^\star = \mathbf{A}\,\mathbf{h} + \mathbf{b}$, with $\mathbf{A}$ and $\mathbf{b}$ a known matrix and vector, then the mmse estimate of $\mathbf{h}^\star$ is:

$$\begin{aligned} \hat{\mathbf{h}}^\star &= E\left\{\mathbf{h}^\star|\delta\mathbf{y}\right\} \\ &= E\left\{\mathbf{A}\,\mathbf{h} + \mathbf{b}|\delta\mathbf{y}\right\} \\ &= E\left\{\mathbf{A}\,\mathbf{h}|\delta\mathbf{y}\right\} + E\left\{\mathbf{b}|\delta\mathbf{y}\right\} \\ &= \mathbf{A}\,\hat{\mathbf{h}} + \mathbf{b} \end{aligned} \tag{2.74}$$

because of the linearity of the expectation operator.

Property 2. The mmse estimator has an additivity property with respect to independent data sets. Assume that $\delta\mathbf{h}$, $\delta\mathbf{y}_1$ and $\delta\mathbf{y}_2$ are zero mean, jointly Gaussian and that the data vectors are independent. The mmse estimator is:

$$\begin{aligned} \delta\hat{\mathbf{h}} &= E\left\{\delta\mathbf{h}|\delta\mathbf{y}_1, \delta\mathbf{y}_2\right\} \\ &= E\left\{\delta\mathbf{h}|\delta\mathbf{y}_1\right\} + E\left\{\delta\mathbf{h}|\delta\mathbf{y}_2\right\} \end{aligned} \tag{2.75}$$

Let $\delta\mathbf{y} = [\delta\mathbf{y}_1^T \;\; \delta\mathbf{y}_2^T]^T$, the estimate is given by relation (2.73). Since $\delta\mathbf{y}_1$ and $\delta\mathbf{y}_2$ are independent, the covariance matrix of the data (and its inverse) are block-diagonal:

$$\begin{aligned} \mathbf{P}_{yy}^{-1} &= \begin{bmatrix} \mathbf{P}_{y_1y_1} & \mathbf{0} \\ \mathbf{0} & \mathbf{P}_{y_2y_2} \end{bmatrix}^{-1} \\ &= \begin{bmatrix} \mathbf{P}_{y_1y_1}^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{P}_{y_2y_2}^{-1} \end{bmatrix} \end{aligned}$$

So that the estimator finally writes:

$$\delta\hat{\mathbf{h}} = \mathbf{P}_{hy_1}\,\mathbf{P}_{y_1y_1}^{-1}\delta\mathbf{y}_1 + \mathbf{P}_{hy_2}\,\mathbf{P}_{y_2y_2}^{-1}\delta\mathbf{y}_2 \tag{2.76}$$

Property 3. The mmse estimator admits a similar geometrical interpretation as the least-squares estimator, the major difference being that the vectors are now random ones. We can think of the random vectors $\delta\mathbf{h}$, $\delta\mathbf{y}_1$, ..., $\delta\mathbf{y}_N$ lying in a vector space as illustrated on the left side of figure 2.9. Due to measurement inaccuracies (noise), $\delta\mathbf{h}$ cannot be perfectly expressed as a linear combination of the data samples which translates geometrically by the fact that vector $\delta\mathbf{h}$ has some component outside the subspace spanned by the data samples.

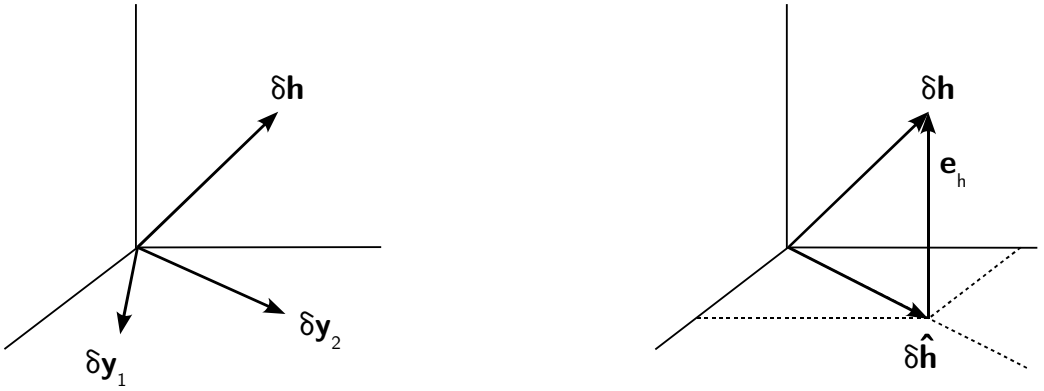


Figure 2.9: Illustration of the orthogonality principle

In such a space made of random vectors, it can be shown that an appropriate definition for the inner product between two vectors $\mathbf{a}$ and $\mathbf{b}$ is given by:

$$\langle \mathbf{a}, \mathbf{b} \rangle = E\left\{\mathbf{a}^T\, \mathbf{b}\right\} = \mathsf{Tr}(E\left\{\mathbf{a}\, \mathbf{b}^T\right\}) \tag{2.77}$$

With that definition, the squared length of a zero-mean, random vector is hence equal to the trace of its covariance matrix:

$$\|\mathbf{a}\|^2 = \langle \mathbf{a}, \mathbf{a} \rangle = \mathsf{Tr}(E\left\{\mathbf{a}\, \mathbf{a}^T\right\}) \tag{2.78}$$

and two vectors are said orthogonal if $\langle \mathbf{a}, \mathbf{b} \rangle = 0$.

The minimum mean square criterion (2.57) can be interpreted from a geometrical standpoint as minimising the squared length of the estimation error $\mathbf{e}_h = \delta\mathbf{h} - \delta\widehat{\mathbf{h}}$. From the right side of figure 2.9, the minimum length is achieved when the estimation error is orthogonal to every data sample of the batch.

Using relations (2.73) and (2.41), the estimation error $\mathbf{e}_h$ for a given realisation of the parameters is given by:

$$\mathbf{e}_h = \delta\mathbf{h} - \delta\widehat{\mathbf{h}} = \left(\mathbf{I} - \mathbf{P}_{hy}\, \mathbf{P}_{yy}^{-1}\, \mathbf{G}_\mathbf{h}\right)\, \delta\mathbf{h} - \mathbf{P}_{hy}\, \mathbf{P}_{yy}^{-1}\, \boldsymbol{\epsilon} \tag{2.79}$$

We now evaluate the covariance between the estimation error and the data samples:

$$\begin{aligned} E\{\mathbf{e}_h\ \delta\mathbf{y}^T\} &= E\left\{\left[\left(\mathbf{I}-\mathbf{P}_{hy}\ \mathbf{P}_{yy}^{-1}\ \mathbf{G_h}\right)\ \delta\mathbf{h}-\mathbf{P}_{hy}\ \mathbf{P}_{yy}^{-1}\ \boldsymbol{\epsilon}\right]\left(\mathbf{G_h}\ \delta\mathbf{h}+\boldsymbol{\epsilon}\right)^T\right\} \\ &= E\left\{\left(\mathbf{I}-\mathbf{P}_{hy}\ \mathbf{P}_{yy}^{-1}\ \mathbf{G_h}\right)\ \delta\mathbf{h}\ \delta\mathbf{h}^T\mathbf{G}_\mathbf{h}^T\right\}+ \\ &\quad E\left\{\left(\mathbf{I}-\mathbf{P}_{hy}\ \mathbf{P}_{yy}^{-1}\ \mathbf{G_h}\right)\ \delta\mathbf{h}\ \boldsymbol{\epsilon}^T\right\}- \\ &\quad E\left\{\mathbf{P}_{hy}\ \mathbf{P}_{yy}^{-1}\ \boldsymbol{\epsilon}\ \delta\mathbf{h}^T\mathbf{G}_\mathbf{h}^T\right\}- \\ &\quad E\left\{\mathbf{P}_{hy}\ \mathbf{P}_{yy}^{-1}\ \boldsymbol{\epsilon}\ \boldsymbol{\epsilon}^T\right\} \end{aligned}$$

In the previous sum, the second and third terms are nought because of the independence between the parameters and the measurement noise. As a result, the projection of the estimation error onto the subspace spanned by the observation vectors is:

$$\begin{aligned} E\{\mathbf{e}_h\ \delta\mathbf{y}^T\} &= \left(\mathbf{I}-\mathbf{P}_{hy}\ \mathbf{P}_{yy}^{-1}\ \mathbf{G_h}\right)\ \mathbf{Q}_0\ \mathbf{G}_\mathbf{h}^T-\mathbf{P}_{hy}\ \mathbf{P}_{yy}^{-1}\ \mathbf{R_r} \\ &= \mathbf{P}_{hy}-\mathbf{P}_{hy}\ \mathbf{P}_{yy}^{-1}\ \left(\mathbf{G_h}\ \mathbf{Q}_0\ \mathbf{G}_\mathbf{h}^T+\mathbf{R_r}\right) \\ &= \mathbf{P}_{hy}-\mathbf{P}_{hy}\ \mathbf{P}_{yy}^{-1}\ \mathbf{P}_{yy} \\ &= \mathbf{0} \end{aligned} \tag{2.80}$$

This last result demonstrates that the estimation error is indeed orthogonal to the data samples. This is the orthogonality principle, aka. the projection theorem.

2.4.4 Measure of performance

In the Bayesian framework, the randomness of the parameters results in a different probability density for each realisation of the parameters. A well-performing estimator should therefore have a small estimation error $\mathbf{e}_h$ for any realisation of the parameters. As a consequence, the performance is assessed by determining the probability density of the error which should ideally be concentrated about zero.

The mean of the error is zero on the average:

$$\begin{aligned} E\left\{\mathbf{e}_h\right\} &= E\left\{\left(\mathbf{I}-\mathbf{P}_{hy}\ \mathbf{P}_{yy}^{-1}\ \mathbf{G_h}\right)\ \delta\mathbf{h}-\mathbf{P}_{hy}\ \mathbf{P}_{yy}^{-1}\ \boldsymbol{\epsilon}\right\} \\ &= \mathbf{0} \end{aligned} \tag{2.81}$$

as $\delta\mathbf{h}$ and $\boldsymbol{\epsilon}$ are assumed zero-mean.

The variance of the error is given by:

$$\begin{aligned} \mathrm{var}\left\{\mathbf{e}_h\right\} &= E\left\{\mathbf{e}_h\ \mathbf{e}_h^T\right\} \\ &= \left(\mathbf{G}_\mathbf{h}^T\ \mathbf{R}_\mathbf{r}^{-1}\ \mathbf{G_h}+\mathbf{Q}_0^{-1}\right)^{-1} \end{aligned} \tag{2.82}$$

For a finite set of data samples, the Cramer-Rao bound sets a lower bound on the minimum mean square error:

$$\sigma^2_{\mathbf{e}_h}(i) \triangleq \text{var}\,\{\mathbf{e}_h(i)\} \geq \mathbf{FIM}^{-1}_{i,i} \tag{2.83}$$

where the FIM (2.48) is now computed using the joint probability density (2.55) and the expectation is taken over the measurements and the parameters:

$$\mathbf{FIM} = \mathbf{G}_\mathbf{h}^T\, \mathbf{R}_\mathbf{r}^{-1}\, \mathbf{G}_\mathbf{h} + \mathbf{Q}_0^{-1} \tag{2.84}$$

2.4.5 Connection with the maximum likelihood estimation

The 1-D example sketched in figure 2.10 compares the maximum likelihood approach to the maximum a posteriori one. The prior $p(\mathbf{h})$ is drawn in light grey together with its most likely value $\widehat{\mathbf{h}}_0$ (*i.e.*, the value of $\mathbf{h}$ for which $p(\mathbf{h})$ is maximum). When the measurements $\mathbf{y}_k$ are observed, the likelihood function $p(\mathbf{y}_k|\,\mathbf{h})$ can be drawn together with the most likely value $\widehat{\mathbf{h}}_{\text{mle}}$ (black line). In a Bayesian perspective, the posterior $p(\mathbf{h}|\,\mathbf{y}_k)$ can be found through relation (2.56) which results in the orange line. The map estimate $\widehat{\mathbf{h}}_{\text{map}}$ is the value that maximises this function.

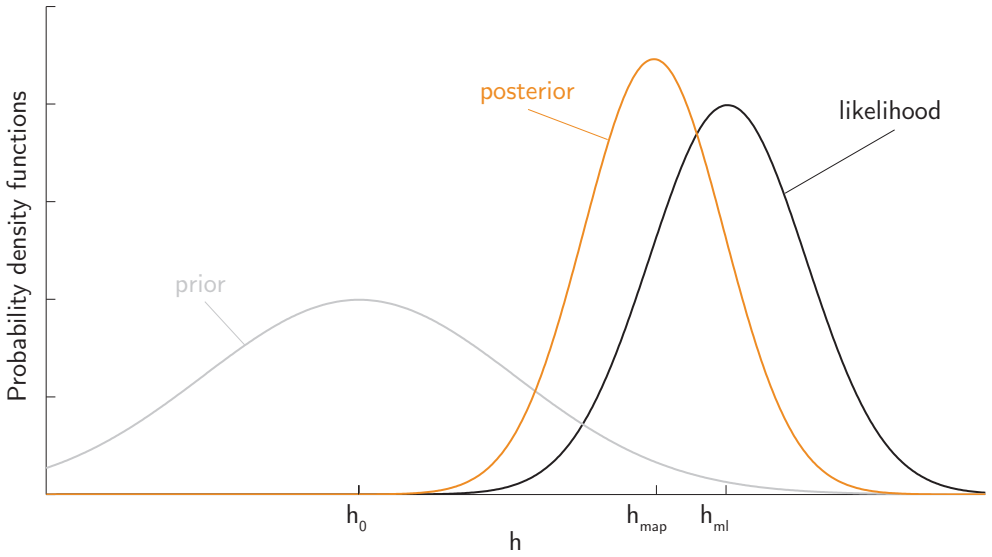


Figure 2.10: Effect of the prior knowledge on the posterior distribution

The two estimates $\widehat{\mathbf{h}}_{\text{mle}}$ and $\widehat{\mathbf{h}}_{\text{map}}$ do not coincide since the map estimate is "attracted" by the prior knowledge showing that, if chosen improperly, the prior can completely spoil the estimation. It is worth mentioning the limiting case of a flat (or non-informing) prior $p(\mathbf{h}) = \text{constant}$. In that case, the posterior confounds with the likelihood function:

$$\begin{aligned} \arg\max_{\mathbf{h}}\ \{p(\mathbf{y}_k|\,\mathbf{h})\} \quad &= \quad \arg\max_{\mathbf{h}}\ \{p(\mathbf{h}|\,\mathbf{y}_k)\} \\ \Rightarrow \quad \widehat{\mathbf{h}}_{\text{map}} \quad &= \quad \widehat{\mathbf{h}}_{\text{ml}} \end{aligned} \tag{2.85}$$

2.5 Tools for the selection and analysis of sensor-parameter suites

2.5.1 Context

Next to an efficient formulation and solution method, the selection of a relevant configuration of the sensors is an important factor towards an accurate estimation of the health parameters.

Current sensors remain intrusive and their integration on the engine is limited by several constraints like cost, weight or harsh environment. As a consequence, the number of sensors is usually kept to a minimum dictated by the control system and hence outweighed by the number of health parameters. This makes the estimation problem underdetermined. Moreover, the measurements are corrupted with noise. The efficiency of the diagnosis algorithm depends on the sensor configuration which should therefore be selected so as to optimise the amount of information that is carried about the engine condition.

Optimal sensor selection has been investigated only sparsely in the engine health monitoring community, few contributions indeed address this issue. Most of the work so far is based on linear approaches. In Kamboukos et al. [2001], a metric based on the sensitivity of the sensors with respect to the parameters is defined. The use of the condition number of the influence coefficient matrix that relates measurements and parameters is also investigated in order to perform the sensor selection in the case of a square problem (*i.e.*, as many sensors as parameters). In Mushini and Simon [2005], a figure of merit based on the steady-state error covariance matrix of a Kalman filter is used to perform the optimisation of the sensor set, again in the case of a square problem. This study also takes into account the cost of the various sensors. In Bechini et al. [2005] the strong link between sensor selection and observability is underlined. A sensor configuration is considered as informative if it avoids both redundancy between measurements and correlation between fault directions. The solution of the sensor selection problem relies on a method for observability analysis developed by Provost [1994]. In Sowers et al. [2008, 2009], sensor selection is based on a metric defined to maximise the detectability and discriminability of user-specified fault types. Sensor cost and fault criticality are complimentary metrics considered in this approach.

A closely related problem, investigated in [Litt, 2008, Simon and Garg, 2009] consists in selecting the best subset of health parameters (or linear combinations thereof) in order to minimise the estimation error of non-measurable, yet key engine performance factors such as the net thrust, the turbine inlet temperature or the stall margins for use as virtual sensors in advanced control laws [Kreiner and Lietzau, 2003].

The topic of sensor selection has received much more attention in other fields such as chemical or structural engineering. As a general trend [see for instance Kirkegaard and Brincker, 1994, van den Berg et al., 2000, Vande Wouwer et al., 2000, Papadimitriou,

2004], the sensor selection problem is addressed in the framework of information theory. Metrics based on the Observability Grammian or on the Fisher Information Matrix are defined in order to optimise the configuration of the experiment.

In the light of these considerations, we have revisited the problem of sensor selection for turbine engine performance monitoring within that scope of information theory [see Borguet and Léonard, 2008]. A metric that handles the aforementioned particularities is proposed. As complementary tools to increase the understanding about the sensor–parameter layout, we propose metrics to assess the observability of a parameter as well as the utility of a given sensor in the estimation process.

2.5.2 Theoretical foundation

The rationale behind sensor selection is to optimise the amount of information conveyed by the measurements about the parameters to be estimated. Optimal information can be defined in various ways, such as the maximum response of the measurements to a change in the health condition, the minimum uncertainty in the estimated parameters or the maximum orthogonality between the measurements to name a few. It is therefore desirable to base the optimisation on a quantity that captures these properties and that is convenient to compare different configurations.

The Fisher information matrix (FIM) introduced earlier in this chapter is a relevant mathematical entity to perform sensor selection. Indeed, the FIM quantifies the amount of information that an observation vector carries about an unknown parameter vector, which underlines the strong coupling between information, observability and estimation. We gave the definition of a generic element of the FIM in equation (2.48), but an equivalent definition for this element is the expectation of the Hessian of the natural logarithm of the probabilistic model:

$$\mathbf{FIM}_{i,j} \triangleq -E\left\{\frac{\partial^2}{\partial \mathbf{h}_i\, \partial \mathbf{h}_j} \log p(\mathbf{y}_k; \mathbf{h})\right\} \tag{2.86}$$

Geometrically, the Hessian is associated to the curvature of the vector field. As such, a high curvature of the probabilistic model with respect to one parameter means that the estimate lies in a narrow valley, that is the variance of the estimate is low, which can be interpreted as having an important amount of information about this parameter.

For sake of completeness, let us recall the expression of the FIM for the scaled probabilistic model (2.35), in the case where the parameters are seen as deterministic variables ($\mathbf{FIM}_{\text{det}}$) and Gaussian random variables ($\mathbf{FIM}_{\text{grv}}$) respectively:

$$\begin{aligned} \mathbf{FIM}_{\text{det}} &= \tilde{\mathbf{G}}_{\mathbf{h}}^T\, \tilde{\mathbf{G}}_{\mathbf{h}} \\ \mathbf{FIM}_{\text{grv}} &= \tilde{\mathbf{G}}_{\mathbf{h}}^T\, \tilde{\mathbf{G}}_{\mathbf{h}} + \mathbf{Q}_0^{-1} \end{aligned}$$

Engine performance monitoring is often characterised in practice by negative redundancy, which means that the number of parameters exceeds the number of sensors ($n_h > n_y$). In this case, the maximum likelihood formulation leads to an underdetermined estimation problem that has no unique solution. Bayesian estimation, by allowing the introduction of prior knowledge on the parameters, can be applied as a palliative. Comparing $\mathbf{FIM}_{\text{det}}$ and $\mathbf{FIM}_{\text{grv}}$, we see that both include a common part, that effectively depends on the sensor-parameter configuration. As a conclusion, the sensor selection problem will be solved by using this common part, that is by using $\mathbf{FIM}_{\text{det}}$.

2.5.3 Metrics for sensor selection

Practically, two main objectives govern the design of a diagnosis tool. On one hand, high sensitivity is desirable in order to provide an early detection of a fault. On the other hand, the user's confidence in the diagnosis system is conditioned to a minimum false alarm rate. These objectives can be achieved by determining the configuration that maximises both the orthogonality between the sensors and the orthogonality between the parameters. The former means that two sensors should not react in the same way to any engine fault. The latter means that two parameters should have a distinct signature on the observations. Various scalar figures of merit based on the FIM can describe these objectives:

1. the **condition number** CN of the FIM is defined as the ratio of the largest to the smallest singular value. The condition number is linked to the rank of a matrix and to the difficulty in performing its inversion. In the case $n_y < n_h$, the rank of the FIM is at best equal to n_y which can be ensured by minimising the condition number. In that way, no sensor is redundant with another,

2. the **trace** Tr of the FIM is defined as the sum of the singular values. In the case $n_y < n_h$, the sum is limited to the first n_y singular values. The trace is a measure of the global sensitivity of the sensors with respect to the parameters and hence has to be maximised. It is linked to some extent to the sensitivity measure defined by Kamboukos et al. [2001],

3. the **determinant** Det of the FIM is defined as the product of the singular values. Again, the product is restricted to the first n_y singular values. Indeed, as stated in point 1, at most n_y singular values are non-zero. This quantity has to be maximised too as the inverse of the determinant is a measure of the overall uncertainty on the estimated parameters. Furthermore, this criterion tends to standardise the singular values and hence the contribution of each sensor. This criterion is similar to the one used by Bechini et al. [2005].

The figures of merit are expressed mathematically through relations (2.87), where $\sigma_i, i = 1, \ldots, n_y$ are the singular values of the FIM traditionally sorted in descending order.

$$\mathsf{CN} = \frac{\sigma_1}{\sigma_{n_y}} \qquad \mathsf{Tr} = \sum_{i=1}^{n_y} \sigma_i \qquad \mathsf{Det} = \prod_{i=1}^{n_y} \sigma_i \tag{2.87}$$

Each of these metrics puts emphasis on a particular aspect of the problem. To perform the sensor selection, it is proposed to combine them in an aggregated figure of merit:

$$\mathsf{FOM} = -W_1 \log(\mathsf{CN}) + W_2 \log(\mathsf{Tr}) + W_3 \log(\mathsf{Det}) \tag{2.88}$$

where W_i are factors that allow both normalisation of the magnitude of the various components and variable weighting between the objectives. The figure of merit FOM has to be maximised with respect to the sensor configuration under the constraint of a fixed number of sensors.

2.5.4 Metrics for observability analysis

The purpose of an observability analysis as defined here is twofold: firstly assess the contribution of each sensor to the estimation problem and secondly quantify the observability level of each health parameter. In simple words, the observability level measures the possibility to estimate correctly a given parameter.

As pointed out by Brown [1966], the classical test for observability of static systems *i.e.*, checking the invertibility (or full-rankedness) of the scaled system matrix $\tilde{\mathbf{G}}_\mathbf{h}$ gives merely a boolean answer. In the case of underdetermined problems, the system is always classified as non-observable. Practically, it is however *partially* observable and it is worth determining the observability level of the different parameters as well as the unobservable fault directions. Inspired by the work of Provost [1994], a method for generating this information is proposed in the following.

Reinterpretation of the singular value decomposition

Our intent is to derive metrics for both the sensors and the parameters, therefore we base the computation on the scaled influence coefficient matrix rather than on the FIM. Singular value decomposition (SVD) [see Golub and van Loan, 1996] of $\tilde{\mathbf{G}}_\mathbf{h}$ serves as the central mathematical tool to carry out the observability analysis:

$$\tilde{\mathbf{G}}_\mathbf{h} = \mathbf{U}\,\mathbf{\Sigma}\,\mathbf{V}^T = \mathbf{U} \begin{bmatrix} \mathbf{\Sigma}_1 \mid \mathbf{\Sigma}_2 \end{bmatrix} \begin{bmatrix} \mathbf{V}_1 \mid \mathbf{V}_2 \end{bmatrix}^T \tag{2.89}$$

where

- $\mathbf{U}$ is an orthogonal $n_y \times n_y$ matrix whose columns define an orthonormal basis for the output (measurement) space,

- $\mathbf{V}$ is an orthogonal $n_h \times n_h$ matrix whose columns define an orthonormal basis for the input (parameter) space,
- $\mathbf{\Sigma}$ is an $n_y \times n_h$ matrix whose main diagonal terms are the singular values σ, which can be seen as stretching factors along the directions defined in $\mathbf{U}$ and $\mathbf{V}$.

We propose a geometrical interpretation of the SVD hereafter. The influence coefficient matrix $\tilde{\mathbf{G}}_\mathbf{h}$ can be seen as an operator that maps the space of the health parameters (input space) to the space of the measurements (output space). In this representation, each health parameter (resp. measurement) is one orthogonal base vector of the input (resp. output) space. Application of the SVD to the matrix $\tilde{\mathbf{G}}_\mathbf{h}$ consists essentially in a change of reference frame in both the input and output spaces as sketched in figure 2.11. The position of any health condition of the engine is now measured according to the columns of matrix $\mathbf{V}$ and the corresponding measurement coordinates are expressed in terms of the columns of matrix $\mathbf{U}$. The major difference between the two representations (original and SVD) lies in the mapping from the input space to the output space. In the SVD representation, the mapping is given by the diagonal matrix $\mathbf{\Sigma}$ which simplifies the description of the input-output relation.

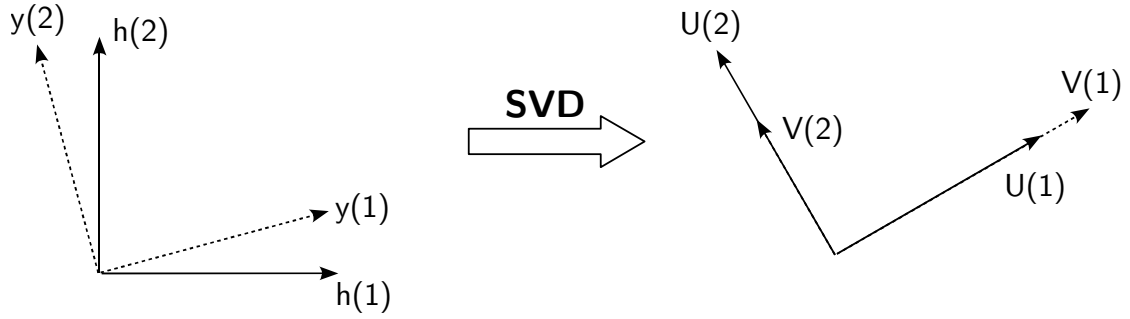


Figure 2.11: Graphical interpretation of the singular value decomposition

A sensitivity index for the sensors

Given the meaning of the columns of $\mathbf{U}$ and the diagonal terms of $\mathbf{\Sigma}$, a sensitivity index SI is defined for each sensor as a weighted norm of each row of $\mathbf{U}$:

$$\mathsf{SI}_i = \frac{\mathbf{U}_i \, \mathbf{\Sigma} \, \mathbf{\Sigma}^T \mathbf{U}_i^T}{\operatorname{trace}\{\mathbf{U} \, \mathbf{\Sigma} \, \mathbf{\Sigma}^T \mathbf{U}^T\}} \quad i = 1, \ldots, n_y \tag{2.90}$$

The sensitivity index is an image of the relative change in a given sensor reading due to a modification in the health condition. The higher the SI, the more important the influence of the associated sensor in the estimation of the health parameters. The denominator in the right-hand-side of equation (2.90) serves as a normalisation factor such that the SI's sum to one.

An observability index for the parameters

An observability index for the parameters, and to a larger extent for any degradation (*i.e.*, combination of several parameters) could be defined in a similar way as the sensitivity index is. Nonetheless, in the case of negative redundancy, a loss of information is introduced due to the fact that the number of parameters outweighs the number of sensors. It can be stated from linear algebra theory that the rank of the Jacobian matrix $\tilde{\mathbf{G}}_{\mathbf{h}}$ is at best equal to n_y. Hence, $n_h - n_y$ singular values are equal to zero. As highlighted in the second part of equation (2.89), matrix $\mathbf{\Sigma}$ is partioned in two blocks where $\mathbf{\Sigma}_1$ gathers the n_y non-zero singular values and $\mathbf{\Sigma}_2$ is a sub-matrix containing the $n_h - n_y$ nought singular values. Matrix $\mathbf{V}$ is accordingly split column-wise into the observable subspace $\mathbf{V}_1$ – aka. the range – and its unobservable counterpart $\mathbf{V}_2$ – aka. the null space – which does not carry any information about the state of the system [Castillo et al., 2005]. The observability index proposed here quantifies this loss of information. For some arbitrary deterioration $\delta\mathbf{h}$, the observability index OI is computed according to:

$$\mathsf{OI}_{\delta\mathbf{h}} = \frac{\left\|\mathbf{V}_1^T\ \delta\mathbf{h}\right\|_2^2}{\left\|\mathbf{V}^T\ \delta\mathbf{h}\right\|_2^2} \tag{2.91}$$

where the numerator is a measure of the magnitude of fault $\delta\mathbf{h}$ in the observable subspace and the denominator represents the total magnitude of fault $\delta\mathbf{h}$.

Following this definition, the observability index is bounded by zero and one. An observability index equal to unity means that the considered degradation has no component located in the null space and can therefore be estimated with a great accuracy. On the contrary, a small observabillity index characterises a large loss of information. The estimation of the associated degradation is hence less accurate.

The fault directions defined by the columns of matrix $\mathbf{V}_2$ are unobservable in the sense that they produce no shift in the measurements. Any degradation that is a linear combination of these vectors has obviously the same detrimental property. The examination of the unobservable fault directions gives clues about possible confusions between weakly observable faults.

Chapter 3

A tool for engine health management

Contents

3.1 Introduction

In this chapter, we report the development of a tool for engine health monitoring. As depicted in figure 3.1, such a tool is a key component of a fleet-wide engine health management process. Snapshot data are gathered and archived on-board at each flight for each engine in the fleet. The data are then transferred to a ground-based station that monitors the health condition of each engine. This enables operators to make informed decisions about maintenance and overhaul of their fleet [Simon, 2010].

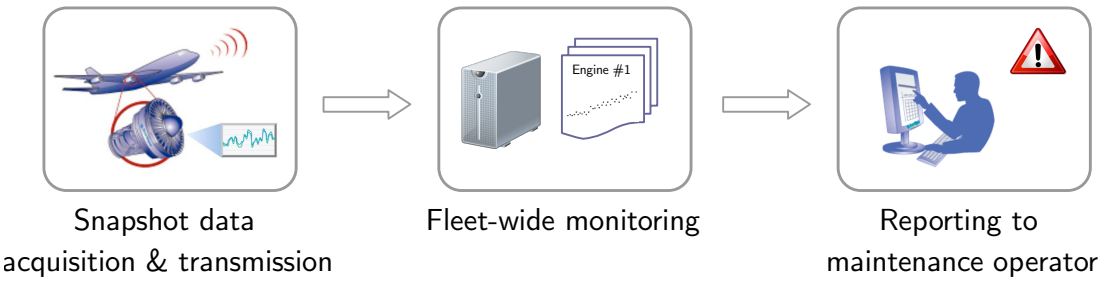


Figure 3.1: Fleet-wide engine health management

As previously mentioned, the monitoring tool shall process noisy data collected from the parsimonious on-board instrumentation. Alterations in the health condition of the flow-path components combine gradual deterioration as well as accidental events as sketched in figure 1.6. Even if running on ground-based facilities, the monitoring tool shall have reasonable computational and memory requirements, for it shall handle a whole fleet of engines that may amount to several dozens (or even hundreds) of units.

Our work relies on the experience gained during previous research on the same topic at the University of Liège [Dewallef, 2005]. The root material was presented in chapter 2 and the various limiting assumptions that were made are progressively removed in the course of this chapter in order to match the specifications of the monitoring tool.

In section 3.2, we move from batch to sequential data processing. We end up with the recursive minimum-mean-square-error estimator that can be efficiently implemented as a Kalman filter. This recursive formulation does not require to store the whole history of measurements and allows to track smoothly evolving health conditions. In section 3.3, we present extensions of the Kalman filter able to deal with nonlinear models. We then compliment our nonlinear Kalman filter design with two additional modules that improve the handling of accidental events. Section 3.4 focuses on the issue of fault detection. The implemented solution relies on a modified transition model for the health parameters that leads to a generalised likelihood ratio test to recognise accidental events. The subsequent fault isolation and quantification tasks are addressed in section 3.5. We propose an original approach to fault isolation based on the concept of sparse estimation which is implemented in the form of a Quadratic Programming problem for which efficient off-the-shelf solvers are avaible.

3.2 Tracking of gradual deterioration

3.2.1 Modelling the evolution of the health parameters

In order to formulate the estimation problem in a sequential fashion, we have to specify the temporal evolution of the health parameters, the time being represented by the discrete index k. The case of constant health parameters (*i.e.*, no variation in the engine condition) can be modelled by means of this simple equation:

$$\delta\mathbf{h}_k = \delta\mathbf{h}_{k-1} \tag{3.1}$$

Note that throughout section 3.2, we still make the assumption of a linear response of the engine to a change in its health condition (this assumption will be released in section 3.3). Therefore, we work with the variations of the health parameters (and the gas-path measurements) with respect to their values for a healthy engine: $\delta(\cdot)_k = (\cdot)_k - (\cdot)^{\text{hl}}$.

The trending of engine performance during a large number of cycles cannot be based on the assumption of a constant health condition since the engine deteriorates with usage. The deterioration trajectory is a gradual and smooth process that depends nonetheless on many factors such as operating environment or type of missions as seen in the introductory chapter. It is therefore very difficult to develop a generic ageing model. The solution is to consider an extremely simple transition model that yet captures the continuous nature of the wear profile. Such a model is the random walk:

$$\delta\mathbf{h}_k = \delta\mathbf{h}_{k-1} + \boldsymbol{\omega}_k \tag{3.2}$$

Sticking to the Bayesian approach, the health parameters are still seen as a vector of random variables, following a Gaussian distribution in the present case. The vector $\boldsymbol{\omega}_k$ is called the process noise and is composed of independent, identically distributed Gaussian random variables. Moreover, the initial health parameter vector $\delta\mathbf{h}_0$ is assumed to be zero-mean and independent of the process noise sequence:

$$\begin{cases} \delta\mathbf{h}_0 \in \mathcal{N}(\mathbf{0}, \mathbf{P}_{\mathbf{h},0}) \\ \boldsymbol{\omega}_k \in \mathcal{N}(\mathbf{0}, \mathbf{Q}_k) \end{cases} \quad \text{and} \quad \begin{cases} E\left\{\delta\mathbf{h}_0\, \boldsymbol{\omega}_k^T\right\} = \mathbf{0} & \forall k \\ E\left\{\boldsymbol{\omega}_k\, \boldsymbol{\omega}_\ell^T\right\} = \mathbf{0} & \text{if } k \neq \ell \end{cases} \tag{3.3}$$

The evolution of the health parameters is hence interpreted as the realisation of the random process (3.2) that introduces some correlation between the health condition at different times. This constraint prevents the estimate from fluctuating too widely in time.

As will appear more clearly in section 3.2.2, $\mathbf{P}_{\mathbf{h},0}$ corresponds to the prior covariance $\mathbf{Q}_0$ used in section 2.4 and represents our knowledge on the possible interval of variation of the health parameters around the baseline value before any data has been collected. On

the other hand, the covariance matrix of the process noise, $\mathbf{Q}_k$, provides some capability to track a time-changing health condition by controlling the stochastic character of the time series formed by the health parameters: low values mean slow variations while large values suppose rapid variations. The magnitude of the diagonal elements of matrices $\mathbf{P}_{\mathbf{h},0}$ and $\mathbf{Q}_k$ as well as possible cross-correlation terms can be set following a similar rule-of-thumb as given for matrix $\mathbf{Q}_0$ in section 2.4.

The health parameters being modelled as Gaussian random variables, their distribution is completely specified through their mean and covariance matrix. Applied to the random walk equation (3.2), the propagation of the mean and covariance matrix writes:

$$\begin{aligned} E\left\{\delta\mathbf{h}_k\right\} &= E\left\{\delta\mathbf{h}_{k-1} + \boldsymbol{\omega}_k\right\} \\ &= E\left\{\delta\mathbf{h}_{k-1}\right\} + E\left\{\boldsymbol{\omega}_k\right\} \\ &= E\left\{\delta\mathbf{h}_{k-1}\right\} \end{aligned} \tag{3.4}$$

since the process noise $\boldsymbol{\omega}_k$ is zero-mean.

$$\begin{aligned} \mathbf{P}_{\mathbf{h},k} &= E\left\{\delta\mathbf{h}_k\ \delta\mathbf{h}_k^T\right\} \\ &= E\left\{\left(\delta\mathbf{h}_{k-1} + \boldsymbol{\omega}_k\right)\ \left(\delta\mathbf{h}_{k-1} + \boldsymbol{\omega}_k\right)^T\right\} \\ &= E\left\{\delta\mathbf{h}_{k-1}\ \delta\mathbf{h}_{k-1}^T\right\} + E\left\{\boldsymbol{\omega}_k\ \delta\mathbf{h}_{k-1}^T\right\} + E\left\{\delta\mathbf{h}_{k-1}\ \boldsymbol{\omega}_k^T\right\} + E\left\{\boldsymbol{\omega}_k\ \boldsymbol{\omega}_k^T\right\} \\ &= \mathbf{P}_{\mathbf{h},k-1} + \mathbf{Q}_k \end{aligned} \tag{3.5}$$

where the second and third terms on the third line are nought since the health parameters $\delta\mathbf{h}_{k-1}$ are a function of the initial state $\delta\mathbf{h}_0$ and of the noises $\boldsymbol{\omega}_1, \ldots, \boldsymbol{\omega}_{k-1}$ which are all independent of the noise $\boldsymbol{\omega}_k$ by assumption.

3.2.2 Sequential minimum-mean-squared-error estimation

The goal is to derive a sequential minimum-mean-squared-error (mmse) estimator that allows the estimation of $\delta\mathbf{h}_k$ based on the sequence of observations $\{\,\delta\mathbf{y}\,\}_1^k$ as the time index k increases. This operation is known as filtering and should ideally be of recursive nature *i.e.*, the computation of $\delta\widehat{\mathbf{h}}_k$ should be based on $\delta\widehat{\mathbf{h}}_{k-1}$.

Before proceeding to the derivation of the recursive estimation rule, it is worth restating the modelling assumptions for the problem of engine performance trending that we are addressing. We consider the following state-space model composed of the state transition equation (3.2), that describes the evolution of the health parameters over time, and of the measurement equation (2.32) that relates variations in the gas-path measurements with changes in the health parameters.

$$\delta\mathbf{h}_k = \delta\mathbf{h}_{k-1} + \boldsymbol{\omega}_k \tag{3.6}$$

$$\delta\mathbf{y}_k = \mathbf{G}_{\mathbf{h},k}\ \delta\mathbf{h}_k + \boldsymbol{\epsilon}_k \tag{3.7}$$

The following assumptions are made about the different variables of the state-space model:

- the health parameters $\delta\mathbf{h}_k$ are considered as Gaussian random variables with initial condition $\delta\mathbf{h}_0 \in \mathcal{N}(\mathbf{0}, \mathbf{P}_{\mathbf{h},0})$,
- the process noise is a zero-mean, Gaussian random vector $\boldsymbol{\omega}_k \in \mathcal{N}(\mathbf{0}, \mathbf{Q}_k)$. It is assumed to be independent of $\delta\mathbf{h}_0$,
- the measurement noise is a zero-mean random vector $\boldsymbol{\epsilon}_k \in \mathcal{N}(\mathbf{0}, \mathbf{R}_{\mathbf{r},k})$. It is assumed to be uncorrelated with $\boldsymbol{\omega}_k$ and $\delta\mathbf{h}_0$.

Our criterion of optimality for the estimate is the minimum Bayesian mean square error:

$$\min E\left\{\left(\delta\mathbf{h}_k - \delta\widehat{\mathbf{h}}_k\right)^T \left(\delta\mathbf{h}_k - \delta\widehat{\mathbf{h}}_k\right)\right\} \quad \forall\, k \geq 1 \tag{3.8}$$

We showed that the mmse estimate identifies with the mean of the posterior density:

$$\delta\widehat{\mathbf{h}}_k = E\left\{\delta\mathbf{h}_k | \{\,\delta\mathbf{y}\,\}_1^k\right\} \tag{3.9}$$

Vector $\delta\mathbf{h}_k$ depends linearly on the set of vectors $\{\delta\mathbf{h}_0, \boldsymbol{\omega}_1, \ldots, \boldsymbol{\omega}_k\}$ and vector $\delta\mathbf{y}_k$ depends linearly on $\delta\mathbf{h}_k$ and $\boldsymbol{\epsilon}_k$. By assumption, all the vectors in the set

$$\mathcal{S}_k \triangleq \{\delta\mathbf{h}_0, \boldsymbol{\omega}_1, \ldots, \boldsymbol{\omega}_k, \boldsymbol{\epsilon}_1, \ldots, \boldsymbol{\epsilon}_k\} \tag{3.10}$$

are Gaussian and independent of each other. As a result, any linear combination of them also produces a jointly Gaussian vector. In a batch framework, the estimate would write:

$$\delta\widehat{\mathbf{h}}_k = \mathbf{P}_{hy}\, \mathbf{P}_{yy}^{-1}\, \{\,\delta\mathbf{y}\,\}_1^k \tag{3.11}$$

To obtain a sequential estimator, we could split the estimation rule according to:

$$\delta\widehat{\mathbf{h}}_k = E\left\{\delta\mathbf{h}_k | \{\,\delta\mathbf{y}\,\}_1^{k-1}\right\} + E\left\{\delta\mathbf{h}_k | \delta\mathbf{y}_k\right\} \tag{3.12}$$

where the first term represents the estimate at time index k based on the $k-1$ previous data points and the second term represents the contribution of the most recent measurements to the estimate.

However, the additivity property (2.75) is valid only for independent data samples. Unfortunately, $\delta\mathbf{y}_k$ is correlated with previous data samples due to their dependence on the health condition, which is correlated in time through the transition model (3.6).

Let us consider the vector defined as the difference between the measurements obtained at time index k and the predicted measurements at time step k based on the estimate at

the previous time step, denoted $\delta\widehat{\mathbf{y}}_k^-$. This error is also known as the innovation or the residual $\mathbf{r}_k$:

$$\mathbf{r}_k \triangleq \delta\mathbf{y}_k - \delta\widehat{\mathbf{y}}_k^- \tag{3.13}$$

We now show that the residual is independent of the previous data *i.e.*, the residual is the part of the latest data sample that is uncorrelated with the previous ones. Inserting the measurement and transition equations into the expression of the residual, we obtain:

$$\begin{aligned}
\mathbf{r}_k &= (\mathbf{G}_{\mathbf{h},k}\ \delta\mathbf{h}_k + \boldsymbol{\epsilon}_k) - \mathbf{G}_{\mathbf{h},k}\ \delta\widehat{\mathbf{h}}_{k-1} \\
&= \mathbf{G}_{\mathbf{h},k}\ (\delta\mathbf{h}_{k-1} + \boldsymbol{\omega}_k) + \boldsymbol{\epsilon}_k - \mathbf{G}_{\mathbf{h},k}\ \delta\widehat{\mathbf{h}}_{k-1} \\
&= \mathbf{G}_{\mathbf{h},k}\ (\delta\mathbf{h}_{k-1} - \delta\widehat{\mathbf{h}}_{k-1}) + \mathbf{G}_{\mathbf{h},k}\ \boldsymbol{\omega}_k + \boldsymbol{\epsilon}_k \\
&= \mathbf{G}_{\mathbf{h},k}\ \mathbf{e}_{\mathbf{h},k-1} + \mathbf{G}_{\mathbf{h},k}\ \boldsymbol{\omega}_k + \boldsymbol{\epsilon}_k
\end{aligned} \tag{3.14}$$

First, the orthogonality principle tells us that the estimation error at the previous time index $\mathbf{e}_{\mathbf{h},k-1}$ is orthogonal to the sequence $\{\ \delta\mathbf{y}\ \}_1^{k-1}$. Second, this sequence of past data is made of linear combinations of the set of vectors $\mathcal{S}_{k-1}$ which are all independent of $\boldsymbol{\omega}_k$ and $\boldsymbol{\epsilon}_k$ by assumption. As a result, the residual $\mathbf{r}_k$ is uncorrelated with the previous data samples and the additivity property of the mmse estimator now applies to the whole sequence partitioned into $\left(\{\ \delta\mathbf{y}\ \}_1^{k-1}; \mathbf{r}_k\right)$:

$$\delta\widehat{\mathbf{h}}_k = E\left\{\delta\mathbf{h}_k | \{\ \delta\mathbf{y}\ \}_1^{k-1}\right\} + E\left\{\delta\mathbf{h}_k | \mathbf{r}_k\right\} \tag{3.15}$$

The first term in equation (3.15) represents the estimate of the health at time index k based on the the $k-1$ previous data points. For this reason, it is also called the a priori/prior estimate at time k and will be denoted $\delta\widehat{\mathbf{h}}_k^-$. The prior estimate is related to the estimate at time $k-1$ through the transition equation:

$$\begin{aligned}
\delta\widehat{\mathbf{h}}_k^- &\triangleq E\left\{\delta\mathbf{h}_k | \{\ \delta\mathbf{y}\ \}_1^{k-1}\right\} \\
&= E\left\{\delta\mathbf{h}_{k-1} + \boldsymbol{\omega}_k | \{\ \delta\mathbf{y}\ \}_1^{k-1}\right\} \\
&= E\left\{\delta\mathbf{h}_{k-1} | \{\ \delta\mathbf{y}\ \}_1^{k-1}\right\} + E\left\{\boldsymbol{\omega}_k | \{\ \delta\mathbf{y}\ \}_1^{k-1}\right\} \\
&= \delta\widehat{\mathbf{h}}_{k-1} + \mathbf{0}
\end{aligned} \tag{3.16}$$

where the second term on the third line is nought as the previous data are linear combination of vectors of the set $\mathcal{S}_{k-1}$ which are all independent of $\boldsymbol{\omega}_k$ by assumption.

The second term in equation (3.15) is the mmse estimate of $\delta\mathbf{h}_k$ based on the innovation and has therefore the following generic expression:

$$E\left\{\delta\mathbf{h}_k | \mathbf{r}_k\right\} = \mathbf{P}_{hr}\ \mathbf{P}_{rr}^{-1}\ \mathbf{r}_k = \mathbf{K}_k\ \mathbf{r}_k \tag{3.17}$$

As a reminder, the cross covariance matrix of the health parameters and the data is defined as:

$$\mathbf{P}_{hr} = E\left\{\delta\mathbf{h}_k\ \mathbf{r}_k^T\right\} \tag{3.18}$$

Before elaborating on the computation of the cross-covariance matrix, we note that the innovation is uncorrelated with the a priori estimate:

$$E\left\{\delta\widehat{\mathbf{h}}_k^-\ \mathbf{r}_k^T\right\} = \mathbf{0} \tag{3.19}$$

indeed, the a priori estimate is a linear combination of the past data and the innovation has been shown to be uncorrelated with these past data. Hence, the cross-covariance matrix can be computed as:

$$\begin{aligned}
\mathbf{P}_{hr} &= E\left\{(\delta\mathbf{h}_k - \delta\widehat{\mathbf{h}}_k^-)\ \mathbf{r}_k^T\right\} \\
&= E\left\{(\delta\mathbf{h}_k - \delta\widehat{\mathbf{h}}_k^-)\ (\delta\mathbf{y}_k - \mathbf{G}_{\mathbf{h},k}\ \delta\widehat{\mathbf{h}}_k^-)^T\right\} \\
&= E\left\{(\delta\mathbf{h}_k - \delta\widehat{\mathbf{h}}_k^-)\ (\mathbf{G}_{\mathbf{h},k}\ \delta\mathbf{h}_k + \boldsymbol{\epsilon}_k - \mathbf{G}_{\mathbf{h},k}\ \delta\widehat{\mathbf{h}}_k^-)^T\right\} \\
&= \mathbf{P}_{\mathbf{h},k}^-\ \mathbf{G}_{\mathbf{h},k}^T
\end{aligned} \tag{3.20}$$

where we use the fact that $\delta\mathbf{h}_k - \delta\widehat{\mathbf{h}}_k^-$ is a linear combination of $\{\delta\mathbf{h}_0, \boldsymbol{\omega}_1, \ldots, \boldsymbol{\omega}_k, \boldsymbol{\epsilon}_1, \ldots, \boldsymbol{\epsilon}_{k-1}\}$ which are all independent of $\boldsymbol{\epsilon}_k$ to step from line 3 to 4.

The covariance matrix of the residuals is readily computed as:

$$\begin{aligned}
\mathbf{P}_{rr} &= E\left\{\mathbf{r}_k\ \mathbf{r}_k^T\right\} \\
&= E\left\{[\mathbf{G}_{\mathbf{h},k}(\delta\mathbf{h}_k - \delta\widehat{\mathbf{h}}_k^-) + \boldsymbol{\epsilon}_k]\ [\mathbf{G}_{\mathbf{h},k}(\delta\mathbf{h}_k - \delta\widehat{\mathbf{h}}_k^-) + \boldsymbol{\epsilon}_k]^T\right\} \\
&= \mathbf{G}_{\mathbf{h},k}\ \mathbf{P}_{\mathbf{h},k}^-\ \mathbf{G}_{\mathbf{h},k}^T + \mathbf{R}_{\mathbf{r},k}
\end{aligned} \tag{3.21}$$

The gain matrix is defined as the product of the cross-covariance matrix and the inverse of the covariance of the residuals:

$$\mathbf{K}_k \triangleq \mathbf{P}_{hr}\ \mathbf{P}_{rr}^{-1} = \mathbf{P}_{\mathbf{h},k}^-\ \mathbf{G}_{\mathbf{h},k}^T\ \left(\mathbf{G}_{\mathbf{h},k}\ \mathbf{P}_{\mathbf{h},k}^-\ \mathbf{G}_{\mathbf{h},k}^T + \mathbf{R}_{\mathbf{r},k}\right)^{-1} \tag{3.22}$$

The gain matrix can be interpreted schematically as follows:

$$\text{Kalman gain} = \frac{\text{prior uncertainty}}{\text{prior uncertainty} + \text{measurement uncertainty}}$$

The matrix $\mathbf{G}_{\mathbf{h},k}\mathbf{P}_{\mathbf{h},k}^-\mathbf{G}_{\mathbf{h},k}^T$ represents the projection of the prior covariance of the health parameters onto the measurement space which allows its comparison to the measurement uncertainty represented by the noise covariance matrix $\mathbf{R}_{\mathbf{r},k}$. The gain matrix can therefore be seen as a weighting factor which trades off the information contained in the most recent measurements with that of the previous data, summarised by the prior knowledge.

If we have a great confidence in the prior belief ($\mathbf{P}_{\mathbf{h},k}^- \to \mathbf{0}$), the gain becomes tiny so as to avoid any update by the most recent measurements. On the other hand, if no prior is

available, $(\mathbf{P}^-_{\mathbf{h},k})^{-1} \rightarrow \mathbf{0}$ and the gain matrix tends to $(\mathbf{G}^T_{\mathbf{h},k}\ \mathbf{R}^{-1}_{\mathbf{r},\mathbf{k}}\ \mathbf{G}_{\mathbf{h},k})^{-1}\mathbf{G}^T_{\mathbf{h},k}\mathbf{R}^{-1}_{\mathbf{r},\mathbf{k}}$: the latest measurement sample is totally trusted since no other information is available (this comes down to a maximum likelihood estimate). The reference which determines if the prior is "high" or "low" is the measurement uncertainty represented by the measurement noise covariance matrix $\mathbf{R}_{\mathbf{r},\mathbf{k}}$.

To complete the description of the recursive mmse estimator, we shall provide relations for the computation of the covariance matrix of the estimated health parameters. The a priori covariance matrix is given by:

$$\begin{aligned}
\mathbf{P}^-_{\mathbf{h},k} &= E\left\{(\delta\mathbf{h}_k - \delta\widehat{\mathbf{h}}^-_k)\ (\delta\mathbf{h}_k - \delta\widehat{\mathbf{h}}^-_k)^T\right\} \\
&= E\left\{(\delta\mathbf{h}_{k-1} + \boldsymbol{\omega}_k - \delta\widehat{\mathbf{h}}_{k-1})\ (\delta\mathbf{h}_{k-1} + \boldsymbol{\omega}_k - \delta\widehat{\mathbf{h}}_{k-1})^T\right\} \\
&= \mathbf{P}_{\mathbf{h},k-1} + \mathbf{Q}_k
\end{aligned} \tag{3.23}$$

where we use the fact that $\delta\mathbf{h}_{k-1} - \delta\widehat{\mathbf{h}}_{k-1}$ is a linear combination of vectors in the set $\mathcal{S}_{k-1}$ which are all independent of $\boldsymbol{\omega}_k$ to step from line 2 to 3.

Noting that the (a posteriori) estimation error at time k can be written as:

$$\begin{aligned}
\mathbf{e}_{\mathbf{h},k} &= \delta\mathbf{h}_k - \delta\widehat{\mathbf{h}}_k \\
&= \delta\mathbf{h}_k - \delta\widehat{\mathbf{h}}^-_k - \mathbf{K}_k\ \mathbf{r}_k \\
&= \delta\mathbf{h}_k - \delta\widehat{\mathbf{h}}^-_k - \mathbf{K}_k\ [\mathbf{G}_{\mathbf{h},k}\ (\delta\mathbf{h}_k - \delta\widehat{\mathbf{h}}^-_k) + \boldsymbol{\epsilon}_k] \\
&= (\mathbf{I} - \mathbf{K}_k\ \mathbf{G}_{\mathbf{h},k})\ (\delta\mathbf{h}_k - \delta\widehat{\mathbf{h}}^-_k) - \mathbf{K}_k\ \boldsymbol{\epsilon}_k
\end{aligned} \tag{3.24}$$

the a posteriori covariance matrix is given by:

$$\begin{aligned}
\mathbf{P}_{\mathbf{h},k} &= E\left\{(\delta\mathbf{h}_k - \delta\widehat{\mathbf{h}}_k)\ (\delta\mathbf{h}_k - \delta\widehat{\mathbf{h}}_k)^T\right\} \\
&= (\mathbf{I} - \mathbf{K}_k\ \mathbf{G}_{\mathbf{h},k})\ \mathbf{P}^-_{\mathbf{h},k}\ (\mathbf{I} - \mathbf{K}_k\ \mathbf{G}_{\mathbf{h},k})^T + \mathbf{K}_k\ \mathbf{R}_{\mathbf{r},k}\ \mathbf{K}^T_k \qquad (3.25) \\
&= (\mathbf{I} - \mathbf{K}_k\ \mathbf{G}_{\mathbf{h},k})\ \mathbf{P}^-_{\mathbf{h},k} - (\mathbf{I} - \mathbf{K}_k\ \mathbf{G}_{\mathbf{h},k})\ \mathbf{P}^-_{\mathbf{h},k}\ \mathbf{G}^T_{\mathbf{h},k}\ \mathbf{K}^T_k + \mathbf{K}_k\ \mathbf{R}_{\mathbf{r},k}\ \mathbf{K}^T_k \\
&= (\mathbf{I} - \mathbf{K}_k\ \mathbf{G}_{\mathbf{h},k})\ \mathbf{P}^-_{\mathbf{h},k} - \mathbf{P}^-_{\mathbf{h},k}\ \mathbf{G}^T_{\mathbf{h},k}\ \mathbf{K}^T_k + \mathbf{K}_k\ (\mathbf{G}_{\mathbf{h},k}\ \mathbf{G}^T_{\mathbf{h},k} + \mathbf{R}_{\mathbf{r},k})\ \mathbf{K}^T_k \\
&= (\mathbf{I} - \mathbf{K}_k\ \mathbf{G}_{\mathbf{h},k})\ \mathbf{P}^-_{\mathbf{h},k} - \mathbf{P}^-_{\mathbf{h},k}\ \mathbf{G}^T_{\mathbf{h},k}\ \mathbf{K}^T_k + \mathbf{P}^-_{\mathbf{h},k}\ \mathbf{G}^T_{\mathbf{h},k}\ \mathbf{K}^T_k \\
&= (\mathbf{I} - \mathbf{K}_k\ \mathbf{G}_{\mathbf{h},k})\ \mathbf{P}^-_{\mathbf{h},k} \qquad (3.26)
\end{aligned}$$

Relation (3.25) is known as Joseph's form of covariance update and has the following advantage over the traditional form (3.26): the matrix subtraction is squared so that Joseph's form is better conditioned to guarantee positive definiteness and symmetry of $\mathbf{P}_{\mathbf{h},k}$ in the presence of round-off errors. It is therefore preferred in actual software implementation of the filter.

3.2.3 Implementation as a Kalman filter

The sequential mmse estimator derived above can be conveniently implemented in the form of a Kalman filter [Kalman, 1960]. Its architecture is summarised in pseudo-code style in algorithm 3.1.

Algorithm 3.1 Linear Kalman filter for health parameter estimation

Require: $\widehat{\mathbf{h}}_0 = \mathbf{h}^{\mathrm{hl}}$ and $\mathbf{P}_{\mathbf{h},0} = \mathbf{Q}_0$

1: **for** $k = 1$ to N **do**
2: $\quad \widehat{\mathbf{h}}_k^- = \widehat{\mathbf{h}}_{k-1}$
3: $\quad \mathbf{P}_{\mathbf{h},k}^- = \mathbf{P}_{\mathbf{h},k-1} + \mathbf{Q}_k$
4: $\quad \mathbf{r}_k = \mathbf{y}_k - \left(\widehat{\mathbf{y}}_k^{\mathrm{hl}} + \mathbf{G}_{\mathbf{h},k}\,(\widehat{\mathbf{h}}_k^- - \mathbf{h}^{\mathrm{hl}})\right)$
5: $\quad \mathbf{P}_{\mathbf{y},k} = \mathbf{G}_{\mathbf{h},k}\mathbf{P}_{\mathbf{h},k}^-\mathbf{G}_{\mathbf{h},k}^T + \mathbf{R}_{\mathbf{r},\mathbf{k}}$
6: $\quad \mathbf{K}_k = \mathbf{P}_{\mathbf{h},k}^-\mathbf{G}_{\mathbf{h},k}^T\,\mathbf{P}_{\mathbf{y},k}^{-1}$
7: $\quad \widehat{\mathbf{h}}_k = \widehat{\mathbf{h}}_k^- + \mathbf{K}_k\,\mathbf{r}_k$
8: $\quad \mathbf{P}_{\mathbf{h},k} = (\mathbf{I} - \mathbf{K}_k\,\mathbf{G}_{\mathbf{h},k})\,\mathbf{P}_{\mathbf{h},k}^-\,(\mathbf{I} - \mathbf{K}_k\,\mathbf{G}_{\mathbf{h},k})^T + \mathbf{K}_k\,\mathbf{R}_{\mathbf{r},k}\,\mathbf{K}_k^T$
9: **end for**

The Kalman filter involves only basic linear algebra operation, the most demanding one being a matrix inversion, and has a predictor-corrector structure. The block diagram in figure 3.2 completes the description of the algorithm.

1. **Prediction step** :
   - the estimated distribution of the health parameters at time $k-1$ is fed into the transition model (3.6) which gives a prior value $\widehat{\mathbf{h}}_k^-$ and a prior covariance matrix $\mathbf{P}_{\mathbf{h},k}^-$ (lines 2 and 3),
   - the a priori value of the health parameters is used in the engine model which delivers an estimation of the measurements $\widehat{\mathbf{y}}_k^-$; these predicted measurements are compared to the actual values measured on the engine to form the residual $\mathbf{r}_k = \mathbf{y}_k - \widehat{\mathbf{y}}_k^-$ (line 4),

2. **Correction step** :
   - the Kalman gain $\mathbf{K}_k$ is computed using equation (3.22) (line 6),
   - the a priori values of the health parameters are corrected with the residuals, weighted by the Kalman gain matrix $\mathbf{K}_k$ (line 7),
   - the covariance matrix is also updated, completing the description of the posterior distribution (line 8).

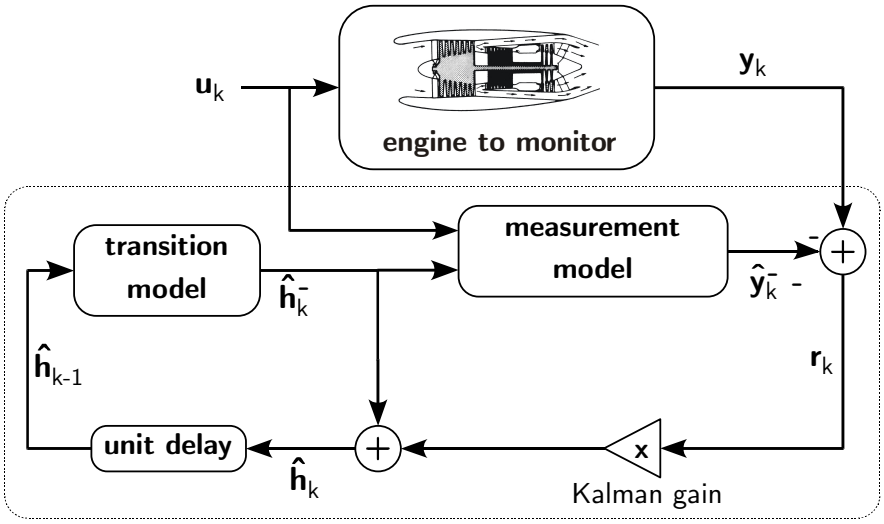


Figure 3.2: Health parameter update mechanism using a Kalman filter

For linear systems and white Gaussian noises, the Kalman filter is an optimal estimator in the sense that it minimises the bayesian mean square error. If the health parameters are moreover constant over time, the Kalman filter converges to the same estimate as the maximum a posteriori estimator (2.67), yet after a sequential processing of the data.

Due to its recursive structure, the Kalman filter has limited memory requirements. Indeed, the estimate at time k is computed on the basis of the estimate at time $k-1$ and the newest measurement sample. As the health parameters are assumed Gaussian, only the mean and the covariance matrix of the most recent estimate have to be stored. The sequential processing allows not only the tracking of a time-varying health condition, but also provides the user with an on-line health status, which leads to a more timely decision making than in the batch approach where the user would have to wait for the database to be built and processed.

The prediction step is an essential part of the algorithm. The more accurate the prediction model, the better the a priori estimate. The Kalman filter is driven by the innovation sequence, which is uncorrelated with previous data by construction.

The Kalman filter provides its own performance measure. The bayesian mean square error is computed as an integral part of the estimator through the covariance matrix $\mathbf{P}_{\mathbf{h},k}$. It can be noticed that the computation of the prior and posterior covariance matrices (lines 3 and 7) do not involve the measurements. Therefore these matrices can be computed beforehand if the sequence of matrices $\mathbf{G}_{\mathbf{h},k}$, $\mathbf{R}_{\mathbf{r},\mathbf{k}}$ and $\mathbf{Q}_k$ is known. From the analysis of algorithm 3.1, the mean square error increases at the prediction step (line 3) to take into account the uncertainty associated with the transition model. At the correction stage, this error is decreased proportionally to the Kalman gain as shown by equation (3.26), which translates the fact that some information has been extracted from the most recent data sample.

3.3 Extensions to non-linear models

3.3.1 Context

The Kalman filter is the mmse recursive estimator for stochastic linear models with Gaussian random variables. However, in our application, the engine performance model (2.25) relating the health parameters and the gas-path measurements is actually non-linear.

The particle filter [see *e.g.*, Arulampalam et al., 2002] was developed not so long ago as a means for solving non-linear and/or non-Gaussian estimation problems. This algorithm implements a sequential Monte-Carlo method in order to provide an estimate of the complete posterior distribution. However, the non-linearity of the engine performance model is mild enough so that non-linear (sub-optimal) extensions of the Kalman filter produce quite satisfactory results as we will see in chapter 4. For this reason, we do not investigate the use of the particle filter for module performance analysis in this document.

Taking a look back at algorithm 3.1, we see that the Kalman filter updates the first two moments (*i.e.*, mean and covariance) of the health parameter distribution. Such a practice has two advantages: first, it requires the maintenance of a constant and relatively small amount of information that is moreover generally sufficient for most kind of operational activities (*e.g.*, control, decision making). Second, the mean and covariance are linearly transformable quantities.

Obviously, we would like to replicate this simplicity in the non-linear case. If only the first two moments of a random variable are known, the most conservative assumption is to consider this variable as a Gaussian one. This is our first approximation: the health parameters are still considered as Gaussian random variables, even if gaussianity is lost through a non-linear mapping.

The second approximation is that we restrict to the class of linear estimators, for sake of convenience. In the Kalman filter, the update rule for the mean value of the health parameters writes conceptually:

$$\widehat{\mathbf{h}}_k = \text{prediction of } \mathbf{h}_k + \text{Gain matrix} \times (\mathbf{y}_k - \text{prediction of } \mathbf{y}_k) \tag{3.27}$$

Under the assumptions we used to derive the Kalman filter, an exact, closed form expression for each of the three terms in the right-hand side of equation (3.27) is available. In the non-linear case, they have to be approximated. As a result, the central problem in applying the Kalman filter technique to non-linear systems is basically the problem of propagating the mean and covariance estimates through non-linear transformations. We present below two popular approximation schemes: the extended Kalman filter (EKF) and the unscented Kalman filter (UKF) [Haykin, 2001].

3.3.2 The extended Kalman filter

The basic tenet behind the extended Kalman filter [see Jazwinski, 1970], summarised in algorithm 3.2, is to linearise the model at each time step around the most recent estimate $\widehat{\mathbf{h}}_k^-$ and then apply the standard Kalman update mechanism. The mean is propagated through the non-linear model while the covariance is propagated through the linearised model. For the trend monitoring problem, only the measurement equation is non-linear and the Jacobian matrix (aka. influence coefficient matrix) is assessed as:

$$\mathbf{G}_{\mathbf{h},k} = \left.\frac{\partial \mathcal{G}(\mathbf{u},\mathbf{h})}{\partial \mathbf{h}}\right|_{\mathbf{u}=\mathbf{u}_k;\mathbf{h}=\widehat{\mathbf{h}}_k^-} \tag{3.28}$$

Algorithm 3.2 Extended Kalman filter for health parameter estimation

Require: $\widehat{\mathbf{h}}_0 = \mathbf{h}^{\mathrm{hl}}$ and $\mathbf{P}_{\mathbf{h},0} = \mathbf{Q}_0$

1: **for** $k = 1$ to N **do**
2: $\quad\widehat{\mathbf{h}}_k^- = \widehat{\mathbf{h}}_{k-1}$
3: $\quad\mathbf{P}_{\mathbf{h},k}^- = \mathbf{P}_{\mathbf{h},k-1} + \mathbf{Q}_k$
4: $\quad\mathbf{r}_k = \mathbf{y}_k - \mathcal{G}(\mathbf{u}_k, \widehat{\mathbf{h}}_k^-)$
5: $\quad$Compute Jacobian matrix $\mathbf{G}_{\mathbf{h},k}$ as per equation (3.28)
6: $\quad\mathbf{P}_{\mathbf{y},k} = \mathbf{G}_{\mathbf{h},k}\mathbf{P}_{\mathbf{h},k}^-\mathbf{G}_{\mathbf{h},k}^T + \mathbf{R}_{\mathbf{r},\mathbf{k}}$
7: $\quad\mathbf{K}_k = \mathbf{P}_{\mathbf{h},k}^-\mathbf{G}_{\mathbf{h},k}^T\,\mathbf{P}_{\mathbf{y},k}^{-1}$
8: $\quad\widehat{\mathbf{h}}_k = \widehat{\mathbf{h}}_k^- + \mathbf{K}_k\,\mathbf{r}_k$
9: $\quad\mathbf{P}_{\mathbf{h},k} = (\mathbf{I} - \mathbf{K}_k\,\mathbf{G}_{\mathbf{h},k})\,\mathbf{P}_{\mathbf{h},k}^-\,(\mathbf{I} - \mathbf{K}_k\,\mathbf{G}_{\mathbf{h},k})^T + \mathbf{K}_k\,\mathbf{R}_{\mathbf{r},k}\,\mathbf{K}_k^T$
10: **end for**

The gain and covariance matrices have now to be computed on-line for they depend on the estimates via the influence coefficient matrix. Note also that the covariance matrix is only an approximation to the mean square error.

The evaluation of the influence coefficient matrix $\mathbf{G}_{\mathbf{h},k}$ involves at each time step n_h additional calls to the non-linear model in the case of computation by forward or backward differences and up to $2\,n_h$ for a central difference scheme. This increases substantially the computational load with respect to the linear Kalman filter. However, when small variations of the health parameters are considered, the Jacobian matrix $\mathbf{G}_{\mathbf{h},k}$ does not change significantly from time step to time step and can therefore be updated after a certain number of cycles, either predefined or depending on the estimated deterioration.

In addition, the Jacobian matrix $\mathbf{G}_{\mathbf{h},k}$ depends on the operating conditions $\mathbf{u}_k$ and should therefore still be evaluated at each time step. However, when processing data recorded in a neighbourhood of some average flight conditions, the variation of the matrix $\mathbf{G}_{\mathbf{h},k}$ is low

and the influence matrix can be computed for those average flight conditions as proposed by Qiu et al. [2007].

If this procedure obviously decreases the computational load of the EKF, it is also preferable for stability purposes. Indeed, the operating conditions $\mathbf{u}_k$ are also sensed and hence noisy. The assessment of the matrix $\mathbf{G}_{\mathbf{h},k}$ based on data corrupted by random errors can lead to results worse than with average flight conditions.

The EKF has no optimality properties and its performance is highly conditioned on the accuracy of the linearisation. Nonetheless, it has been widely used in many practical applications owing to its simplicity and suitability for real-time implementation. On the other hand, it has been recognised difficult to tune and proves reliable only for systems which are almost linear on the time scale of the update interval (sufficiently smooth non-linear mappings).

To try and address this issue, variations of the EKF have been devised. The well-named iterated extended Kalman filter [Gelb, 1974] iterates the EKF at each time step by re-defining the reference estimate and relinearising the model. The quadratic extended Kalman filter [Wang et al., 2006] considers a second-order representation (*i.e.*, accounts for quadratic terms) of the non-linear model. Both variations are reported to perform better than the basic EKF for highly non-linear systems, however at the price of a further increase in computational complexity which explains their limited practical usage.

Another drawback of the different versions of the EKF lies in the computation of the Jacobian matrix, a procedure that is often tedious and error prone, be it analytically or numerically. This was a major driver to come up with derivative-free approaches, such as the unscented Kalman filter presented below.

3.3.3 The unscented Kalman filter

The unscented Kalman filter addresses both the linear approximation and Jacobian computation issues of the EKF. In accordance with our simplifications, the health parameters are again represented by gaussian random variables. The technique is rooted in the so-called unscented transformation that is used to propagate the mean and covariance information through the non-linear model.

The unscented transformation

The unscented transformation, introduced by Julier and Uhlmann [1996], is a method for calculating the statistics of a random variable that undergoes a non-linear transformation. It relies on the bright intuition that "*with a fixed number of parameters, it should be easier to approximate a gaussian distribution than it is to approximate an arbitrary non-linear*

function". As a consequence, the unscented transformation does not approximate the non-linear function, but well the continuous probability density with a discrete one.

Basically, a minimal number of sample points are purposely chosen so as to capture the mean and covariance of the distribution. These points are then transformed through the non-linear mapping. The sample images capture the mean and covariance of the transformed distribution at least to the second order for any non-linearity, as proved in [Julier and Uhlmann, 1996] by a Taylor series expansion. This performance should be compared to the first-order accuracy achieved with a linearisation technique like the EKF. Figure 3.3 illustrates the concept for a hypothetical 2-D system. The left plot shows the true mean and covariance propagation using Monte-Carlo sampling, the centre plot shows the results using a linearisation approach, the right plot shows the behaviour of the unscented transform.

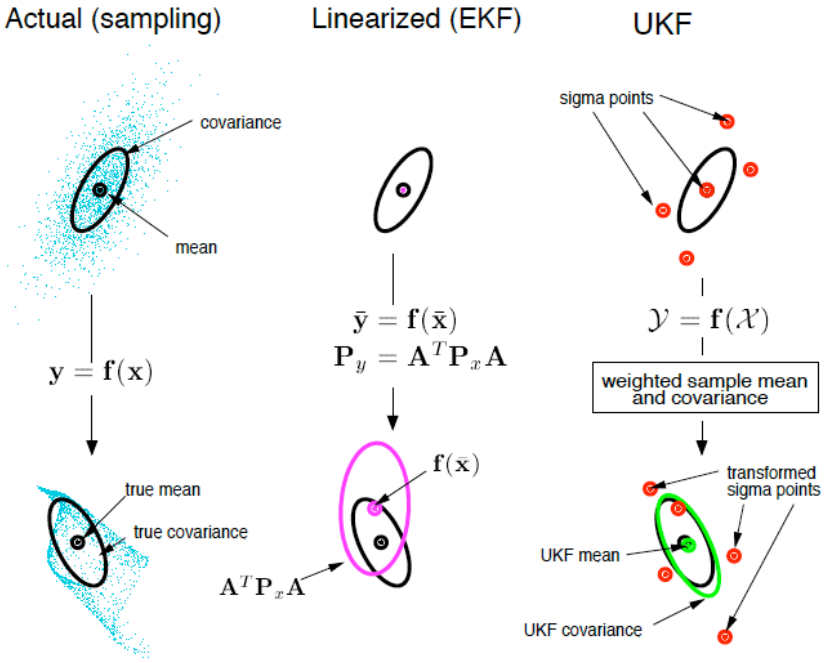


Figure 3.3: Comparison of the propagation of mean and covariance through a non-linear function. Left: actual, center: first-order linearisation, right: unscented transformation [Julier and Uhlmann, 1996]

Consider a parameter vector $\mathbf{x}$ of dimension n_x that is the realisation of a random variable with mean $\widehat{\mathbf{x}}$ and covariance $\mathbf{P}_x$ and a non-linear vector function defined by $\mathbf{y} = f(\mathbf{x})$. The unscented transformation provides the statistics of the variable $\mathbf{y}$ at the end of a three-stage process:

1. Generate the set of $2\,n_x + 1$ sigma points (vectors):

$$\begin{aligned}
\mathcal{X}_0 &= \widehat{\mathbf{x}}, \\
\mathcal{X}_i &= \widehat{\mathbf{x}} + \left(\gamma\sqrt{\mathbf{P}_x}\right)_i \qquad \text{for } i = 1, \ldots, n_x \\
\mathcal{X}_i &= \widehat{\mathbf{x}} - \left(\gamma\sqrt{\mathbf{P}_x}\right)_{i-n_x} \qquad \text{for } i = n_x + 1, \ldots, 2\,n_x
\end{aligned}$$

   where $\left(\sqrt{\mathbf{P}_x}\right)_i$ is the i^{th} row or column of the matrix square root of $\mathbf{P}_x$. The sigma points capture the same mean and covariance irrespective of the choice of the matrix square root. The Cholesky decomposition is therefore preferred due to its numerical efficiency and stability.

2. Propagate the sigma points through the non-linear function:

$$\mathcal{Y}_i = f(\mathcal{X}_i) \qquad \text{for } i = 0, \ldots, 2\,n_x \tag{3.29}$$

3. Compute the mean and covariance for $\mathbf{y}_k$ as a weighted sample mean and covariance of the image vectors $\mathcal{Y}_i$:

$$\begin{aligned}
\widehat{\mathbf{y}} &= \sum_{i=0}^{2\,n_x} W_i^{(m)} \mathcal{Y}_i \\
\mathbf{P}_y &= \sum_{i=0}^{2\,n_x} W_i^{(c)} (\mathcal{Y}_i - \widehat{\mathbf{y}})(\mathcal{Y}_i - \widehat{\mathbf{y}})^T
\end{aligned}$$

   with the weighting coefficients $W_i^{(\cdot)}$ given by:

$$\begin{aligned}
W_0^{(m)} &= \frac{\lambda}{n_x + \lambda}, \\
W_0^{(c)} &= \frac{\lambda}{n_x + \lambda} + 1 - \alpha^2 + \beta, \\
W_i^{(c)} &= W_i^{(m)} = \frac{1}{2(n_x + \lambda)}, \quad \forall\, i = 1, \ldots, 2\,n_x
\end{aligned}$$

In the above equations, $\gamma = \sqrt{n_x + \lambda}$ and $\lambda = \alpha^2(n_x + \kappa) - n_x$ are scaling parameters. The constant α determines the spread of the sigma points around $\widehat{\mathbf{x}}$ and is usually set to $10^{-4} \leq \alpha \leq 1$. The constant κ is a secondary scaling parameter which is heuristically set to $3 - n_x$. Finally, β is used to incorporate prior knowledge about the distribution of $\mathbf{x}$. For gaussian distributions, $\beta = 2$ is an optimal choice.

If the unscented transformation may at first resemble a Monte-Carlo type method, there is however a fundamental difference between the techniques. Indeed, the unscented transform does not draw the sample points randomly, but according to a deterministic rule such that they exhibit certain specific properties (*e.g.*, given mean and covariance).

The unscented Kalman filter

The unscented Kalman filter, detailed in algorithm 3.3, results from the integration of the unscented transformation in the recursive estimation procedure of algorithm 3.2. The UKF achieves an equal or better level of performance than the EKF at a comparable level of complexity. Next to the calls to the non-linear model, the most demanding task in the UKF is the computation of a matrix square root. Note that even if the number of calls to the non-linear model is the same, the UKF should not be interpreted as an EKF with central-difference computation of the Jacobian, as both filters handle the non-linearity totally differently.

Algorithm 3.3 Unscented Kalman filter for health parameter estimation

Require: $\widehat{\mathbf{h}}_0 = \mathbf{h}^{\mathrm{hl}}$ and $\mathbf{P}_{\mathbf{h},0} = \mathbf{Q}_0$

1: **for** $k = 1$ to N **do**
2: $\quad \widehat{\mathbf{h}}_k^- = \widehat{\mathbf{h}}_{k-1}$
3: $\quad \mathbf{P}_{\mathbf{h},k}^- = \mathbf{P}_{\mathbf{h},k-1} + \mathbf{Q}_k$
4: $\quad \mathcal{H}_{k-1} = \left[\widehat{\mathbf{h}}_k^- \quad \widehat{\mathbf{h}}_k^- + \gamma\sqrt{\mathbf{P}_{\mathbf{h},k}^-} \quad \widehat{\mathbf{h}}_k^- - \gamma\sqrt{\mathbf{P}_{\mathbf{h},k}^-}\right]$
5: $\quad \mathcal{Y}_{i,k} = \mathcal{G}(\mathbf{u}_k, \mathcal{H}_{i,k-1}) \qquad i = 0, 1, \ldots, 2\,n_h$
6: $\quad \widehat{\mathbf{y}}_k^- = \sum_{i=0}^{2\,n_h} W_i^{(m)} \mathcal{Y}_{i,k}$
7: $\quad \mathbf{r}_k = \mathbf{y}_k - \widehat{\mathbf{y}}_k^-$
8: $\quad \mathbf{P}_{\mathbf{y},k} = \sum_{i=0}^{2\,n_h} W_i^{(c)} (\mathcal{Y}_{i,k} - \widehat{\mathbf{y}}_k^-)(\mathcal{Y}_{i,k} - \widehat{\mathbf{y}}_k^-)^T + \mathbf{R}_{\mathbf{r},\mathbf{k}}$
9: $\quad \mathbf{P}_{\mathbf{hy},k} = \sum_{i=0}^{2n_h} W_i^{(c)} (\mathcal{H}_{i,k-1} - \widehat{\mathbf{h}}_k^-)(\mathcal{Y}_{i,k} - \widehat{\mathbf{y}}_k^-)^T$
10: $\quad \mathbf{K}_k = \mathbf{P}_{\mathbf{hy},k} \mathbf{P}_{\mathbf{y},k}^{-1}$
11: $\quad \widehat{\mathbf{h}}_k = \widehat{\mathbf{h}}_k^- + \mathbf{K}_k\, \mathbf{r}_k$
12: $\quad \mathbf{P}_{\mathbf{h},k} = \mathbf{P}_{\mathbf{h},k}^- - \mathbf{K}_k\, \mathbf{P}_{\mathbf{y},k}\, \mathbf{K}_k^T$
13: **end for**

In contrast to the EKF, no explicit Jacobian calculations are required for the UKF. This is a substantial advantage in the case of non-analytic models such as an engine performance deck. If needed, the influence coefficient matrix between the health parameters and the gas-path measurements can be approximately assessed by noting that:

$$\begin{aligned} \mathbf{P}_{\mathbf{hy},k} &\simeq \mathbf{P}_{\mathbf{h},k}^- \, \mathbf{G}_{\mathbf{h},k}^T \\ \Rightarrow \mathbf{G}_{\mathbf{h},k} &\simeq \left[(\mathbf{P}_{\mathbf{h},k}^-)^{-1} \, \mathbf{P}_{\mathbf{hy},k}\right]^T \end{aligned} \tag{3.30}$$

3.4 Fault detection module

3.4.1 Context

The Kalman filter presented in the previous section embeds a transition model that assumes a slow and continuous variation of the health parameters. As a result, the Kalman filter is very capable at tracking a gradual and distributed process like engine wear with a good accuracy. On the other hand, any short-time-scale (either rapid or abrupt) change in the engine condition (*e.g.*, a foreign object damage) is interpreted by the Kalman filter as a disturbance since it does not fit the health parameter dynamics. The transition model prevents a correct capture of this fast variation which translates into a sluggish response of the filter. The recognition of a faulty condition suffers therefore from a rather long delay that is incompatible with a proactive maintenance practice.

To overcome these deficiencies, Volponi [2003] proposed a strategy that combines a so-called single-fault isolator with a Kalman filter. The single-fault isolator consists in a bank of possible events that is examined whenever a residual is suspected to be representative of an abrupt fault. While very efficient, this solution implies the interaction between two different diagnosis tools. Moreover, the set-up of the fault library requires significant knowledge and experience.

A broader way to tackle this issue is to view it in the realm of *adaptive estimation* [Mehra, 1972]. The basic idea consists in increasing the mobility of the health parameters momentarily in order to recognise a rapid degradation. Mathematically, it comes down to an on-line tuning of the variance of the health parameters on the basis of a statistical processing of the residuals. Several techniques have been explored in the literature such as auto-correlation [Mehra, 1970, Odelson et al., 2006] or covariance matching [Jazwinski, 1969]. For that matter, Léonard et al. [2008] report the development of an adaptive monitoring tool that integrates such a covariance matching scheme.

Another attractive solution was proposed by Willsky and Jones [1974]. They acknowledge that short-time-scale events may occur, but they do occur infrequently. Moreover, the time of occurrence of the event, as well as its type (*i.e.*, the impacted component and the magnitude of the fault in our case) are typically unknown. The proposed technique rests on a modified transition model that account for possible "jumps" in the parameters and has been successfully applied, among others, in visual tracking [Bensalah and Chaumette, 1994].

In [Borguet and Léonard, 2009], we have transposed this framework to the particular case of engine performance monitoring. The resulting adaptive filter combines a Kalman filter, which relies on the assumption of a smooth variation of the health parameters, and a secondary system that monitors the residuals. This auxiliary component implements a generalised likelihood ratio (GLR) test in order to detect and estimate a rapid fault. In that way, the performance of the Kalman filter is maintained under normal operation

(long-time-scale deterioration) whereas the event detector provides a timely detection of sudden variations in the engine health. We showed the enhancement in terms of reactivity brought by this technique in [Borguet and Léonard, 2011] and compared it to the covariance matching approach in [Borguet and Léonard, 2010b].

3.4.2 An improved model for adaptive estimation

The key element of the adaptive filter is an enhanced transition model of the health parameters that accounts for possible abrupt events (also called "jumps" below). In the state-space model, equation (3.2) is replaced with equation (3.31):

$$\mathbf{h}_k = \mathbf{h}_{k-1} + \boldsymbol{\omega}_k + \boldsymbol{\Delta}_\mathbf{h}\, \delta_{\tau,k} \tag{3.31}$$

where

- $\boldsymbol{\Delta}_\mathbf{h}$ is a vector modelling the jump,
- $\delta_{\tau,k}$ is the Kronecker delta operator,
- τ is a positive integer that represents the time of occurrence of the jump.

Note that in the present framework, both $\boldsymbol{\Delta}_\mathbf{h}$ and τ are regarded as unknown, deterministic parameters and not as random variables, which means that no prior distribution is attached to them.

The strategy of adaptive estimation comes from viewing the modified state-space model according to two different hypotheses:

- H_0: no jump up to now ($\tau > k$)
- H_1: a jump has occurred ($\tau \leq k$)

Under assumption H_0 (no jump), the Kalman filter provides an optimal estimation of the health parameters in the least-square sense. Under assumption H_1, the residuals $\mathbf{r}_k$ and the estimated parameters $\widehat{\mathbf{h}}_k$ become a function of the jump characteristics τ and $\boldsymbol{\Delta}_\mathbf{h}$. Considering a first-order approximation of the measurement equation, the residuals $\mathbf{r}_k$ can be written as a sum of two terms:

$$\mathbf{r}_k = \mathbf{r}_{k,0} + \mathbf{H}_{k,\tau}\ \boldsymbol{\Delta}_\mathbf{h} \tag{3.32}$$

where $\mathbf{r}_{k,0}$ are the residuals in the no-jump case, distributed as $\mathcal{N}(\mathbf{0}, \mathbf{P}_{\mathbf{y},k})$ and the second term represents the contribution of a jump $\boldsymbol{\Delta}_\mathbf{h}$ that occurred at time τ on the residuals at time k.

Similarly, we can split the estimated health parameters into a sum of two terms:

$$\widehat{\mathbf{h}}_k = \widehat{\mathbf{h}}_{k,0} + \mathbf{F}_{k,\tau}\ \boldsymbol{\Delta}_{\mathbf{h}} \tag{3.33}$$

where $\widehat{\mathbf{h}}_{k,0}$ are the estimated values in the no-jump case and the second term represents the influence of a jump $\boldsymbol{\Delta}_{\mathbf{h}}$ that occurred at time τ on the estimated values at time k.

As will be shown in the next section, the detection of the jump and the estimation of its characteristics largely rely on the matrices $\mathbf{F}_{k,\tau}$ and $\mathbf{H}_{k,\tau}$. Explicit expressions for these matrices are obtained from algebraic manipulations of the state-space model and the equations of the Kalman filter. The final results are given below:

1. No jump has occurred yet ($\tau > k$)

$$\begin{aligned} \mathbf{F}_{k,\tau} &= \mathbf{0} \\ \mathbf{H}_{k,\tau} &= \mathbf{0} \end{aligned}$$

2. A jump has just occurred ($\tau = k$)

$$\begin{aligned} \mathbf{F}_{k,k} &= \mathbf{K}_k\ \mathbf{G}_{\mathbf{h},k} \\ \mathbf{H}_{k,k} &= \mathbf{G}_{\mathbf{h},k} \end{aligned}$$

3. A jump occurred in the past ($\tau < k$)

$$\begin{aligned} \mathbf{F}_{k,\tau} &= \sum_{j=\tau}^{k} \boldsymbol{\Theta}_{k,j}\ \mathbf{K}_j\ \mathbf{G}_{\mathbf{h},j} \\ \mathbf{H}_{k,\tau} &= \mathbf{G}_{\mathbf{h},k}\ (\mathbf{I} - \mathbf{F}_{k-1,\tau}) \end{aligned}$$

with $\boldsymbol{\Theta}_{k,j} = (\mathbf{I} - \mathbf{K}_k\ \mathbf{G}_{\mathbf{h},k})\ \boldsymbol{\Theta}_{k-1,j}$ and $\boldsymbol{\Theta}_{k,k} = \mathbf{I}$.

3.4.3 The generalised likelihood ratio test

In order to determine which hypothesis between H_0 and H_1 is true, we perform the celebrated likelihood ratio test (LRT). In short, it is a statistical test in which a ratio is computed between the maximum probability of a result under two different hypotheses, so that a decision can be made between them based on the value of this ratio [see van Trees, 1968].

In the present case, the likelihood ratio is equal to the ratio of the probabilities of observing the history of residuals under hypotheses H_1 and H_0 respectively:

$$\Lambda_k(\boldsymbol{\Delta}_{\mathbf{h}}, \tau) = \frac{p(\mathbf{r}_1, \ldots, \mathbf{r}_k | H_1, \boldsymbol{\Delta}_{\mathbf{h}}, \tau)}{p(\mathbf{r}_1, \ldots, \mathbf{r}_k | H_0)} \tag{3.34}$$

The numerator in equation (3.34) depends on the unknown $\mathbf{\Delta_h}$ and τ. As a workaround, van Trees [1968] proposes to use the generalised likelihood ratio (GLR). This procedure essentially consists in computing the maximum likelihood estimates $\widehat{\tau}$ and $\widehat{\mathbf{\Delta}}_\mathbf{h}$ assuming H_1 is true and then substituting these values into the usual likelihood ratio to finally decide whether a jump occurred.

The maximum likelihood estimates of the jump features are defined as the solutions of the following optimisation problem:

$$(\widehat{\mathbf{\Delta}}_\mathbf{h}, \widehat{\tau}) = \arg\max_{\mathbf{\Delta_h},\tau} \{p(\mathbf{r}_1, \ldots, \mathbf{r}_k | H_1, \mathbf{\Delta_h}, \tau)\} \tag{3.35}$$

According to equation (3.32) and our hypotheses, the residuals are a sequence of independent and normally distributed random variables:

$$p(\mathbf{r}_1, \ldots, \mathbf{r}_k | H_1, \mathbf{\Delta_h}, \tau) = \prod_{i=1}^{k} p(\mathbf{r}_i | H_1, \mathbf{\Delta_h}, \tau) \tag{3.36}$$

$$p(\mathbf{r}_i | H_1, \mathbf{\Delta_h}, \tau) = \mathcal{N}(\mathbf{H}_{i,\tau}\ \mathbf{\Delta_h}, \mathbf{P}_{\mathbf{y},i}) \tag{3.37}$$

We are obviously facing an estimation problem governed by similar equations as the one described in section 2.3. Applying the same solution technique, we first convert the maximisation problem into a minimisation one by taking the opposite of the natural logarithm of the right hand side of equation (3.35), which leads to the following objective function:

$$\mathcal{J}_{\text{glr}}(\mathbf{\Delta_h}, \tau) = -\sum_{i=1}^{k} (\mathbf{r}_i - \mathbf{H}_{i,\tau}\mathbf{\Delta_h})^T\ \mathbf{P}_{\mathbf{y},i}^{-1}\ (\mathbf{r}_i - \mathbf{H}_{i,\tau}\mathbf{\Delta_h}) \tag{3.38}$$

The maximum likelihood estimate of the jump is the value that cancels the derivatives with respect to $\mathbf{\Delta_h}$ of the objective function $\mathcal{J}_{\text{glr}}$ and can be written as:

$$\widehat{\mathbf{\Delta}}_\mathbf{h} = \mathbf{C}_{k,\tau}^{-1}\ \mathbf{d}_{k,\tau} \tag{3.39}$$

where matrix $\mathbf{C}_{k,\tau}$ is deterministic and does not depend on the data, whereas vector $\mathbf{d}_{k,\tau}$ is a linear combination of the residuals:

$$\mathbf{C}_{k,\tau} = \sum_{j=1}^{k} \mathbf{H}_{j,\tau}^T\ \mathbf{P}_{\mathbf{y},j}^{-1}\ \mathbf{H}_{j,\tau} \tag{3.40}$$

$$\mathbf{d}_{k,\tau} = \sum_{j=1}^{k} \mathbf{H}_{j,\tau}^T\ \mathbf{P}_{\mathbf{y},j}^{-1}\ \mathbf{r}_j \tag{3.41}$$

Equation (3.39) shows that the estimate of the jump $\widehat{\mathbf{\Delta}}_\mathbf{h}$ is a function of the unknown τ, whose maximum likelihood estimate is now determined as the value that maximises the

numerator of (3.34) – actually its natural logarithm for sake of simplicity. Its denominator is indeed independent of τ.

Acknowledging that both conditional probability densities in the likelihood ratio are gaussian, and replacing the jump structure with its maximum likelihood estimate (3.39), the log-likelihood ratio writes:

$$\begin{aligned} l_{k,\tau} &\triangleq 2\log\Lambda_k(\widehat{\mathbf{\Delta}}_{\mathbf{h},k},\tau) \\ &= -\sum_{i=1}^{k}\left(\mathbf{r}_i - \mathbf{H}_{i,\tau}\widehat{\mathbf{\Delta}}_{\mathbf{h},k}\right)^T \mathbf{P}_{\mathbf{y},i}^{-1}\left(\mathbf{r}_i - \mathbf{H}_{i,\tau}\widehat{\mathbf{\Delta}}_{\mathbf{h},k}\right) + \sum_{i=1}^{k}\mathbf{r}_i^T\,\mathbf{P}_{\mathbf{y},i}^{-1}\,\mathbf{r}_i \\ &= \mathbf{d}_{k,\tau}^T\,\mathbf{C}_{k,\tau}^{-1}\,\mathbf{d}_{k,\tau} \end{aligned} \tag{3.42}$$

Equations (3.40) and (3.41) show that the likelihood ratio (3.42) actually implements a matched filter *i.e.*, a correlation test between the variations in the residuals and the signature of a jump, represented by $\mathbf{H}_{k,\tau}$.

The value $\widehat{\tau}$ that maximises $l_{k,\tau}$ represents the most likely time at which a jump occurred since we began to monitor the system:

$$\widehat{\tau} = \arg\max_{\tau}\ \{l_{k,\tau}\} \tag{3.43}$$

The decision rule to choose between H_0 and H_1 simply compares the maximum value of the generalised likelihood ratio to a threshold η, chosen to provide a reasonable trade-off between false alarms and detection delay:

$$l_{k,\widehat{\tau}} \underset{H_0}{\overset{H_1}{\gtrless}} \eta \tag{3.44}$$

If hypothesis H_1 is verified at time step k, a detection flag is raised and a jump is declared to have occurred at the estimated time $\widehat{\tau}$. The estimated jump structure is given by:

$$\widehat{\mathbf{\Delta}}_{\mathbf{h},k} = \mathbf{C}_{k,\widehat{\tau}}^{-1}\,\mathbf{d}_{k,\widehat{\tau}} \tag{3.45}$$

The latter relation provides a maximum likelihood estimate of the jump $\mathbf{\Delta}_{\mathbf{h}}$ based on a guessed value $\widehat{\tau}$ of its time of occurrence and assuming that no prior information is available about the value of the jump. In that case, $\mathbf{C}_{k,\widehat{\tau}}^{-1}$ can be interpreted as the covariance matrix of the jump estimate (3.45).

3.4.4 Compensation of the jump

Besides an optimal decision rule for event detection, the framework provides useful information about the jump. The maximum likelihood estimates $\widehat{\tau}$ and $\widehat{\mathbf{\Delta}}_{\mathbf{h},k}$ can in fact be used to compensate the jump by a direct increment of the Kalman filter estimates:

$$\widehat{\mathbf{h}}_k = \widehat{\mathbf{h}}_{k,KF} + \underbrace{\left(\mathbf{I} - \mathbf{F}_{k,\widehat{\tau}}\right)\ \widehat{\mathbf{\Delta}}_{\mathbf{h},k}}_{\Delta\widehat{\mathbf{h}}_k} \tag{3.46}$$

where $\widehat{\mathbf{h}}_{k,KF}$ are the estimates provided by the Kalman filter before compensation, $\mathbf{I}\ \widehat{\mathbf{\Delta}}_{\mathbf{h},k}$ is the effect on the parameters at time k of a jump that occured at $\widehat{\tau}$ (the identity matrix should be seen here as the state transition matrix) and $\mathbf{F}_{k,\widehat{\tau}}\ \widehat{\mathbf{\Delta}}_{\mathbf{h},k}$ represents the response of the Kalman filter to the jump prior to its detection.

In some cases, the estimate of the jump may be quite inaccurate. To reflect this additional uncertainty in the estimate, caused by the jump, it is advised in Willsky and Jones [1976] to increase the covariance matrix of the health parameters accordingly. This rise in parameter covariance results in an increased Kalman gain *i.e.*, an increased bandwidth. The filter can improve its response to the jump and hence compensate for inaccuracies in $\widehat{\tau}$ and $\widehat{\mathbf{\Delta}}_{\mathbf{h},k}$. The covariance is adapted according to:

$$\mathbf{P}_{\mathbf{h},k} = \mathbf{P}_{\mathbf{h},k,KF} + \underbrace{\left(\mathbf{I} - \mathbf{F}_{k,\widehat{\tau}}\right)^T\ \mathbf{C}_{k,\widehat{\tau}}^{-1}\left(\mathbf{I} - \mathbf{F}_{k,\widehat{\tau}}\right)}_{\Delta\mathbf{P}_{\mathbf{h},k}} \tag{3.47}$$

where $\mathbf{P}_{\mathbf{h},k,KF}$ is the covariance matrix of the health parameters before compensation and $\Delta\mathbf{P}_{\mathbf{h},k}$ is the aforementioned compensation term for the covariance.

3.4.5 Practical implementation

The detection procedure as presented above implies that, at each time step k, the likelihood ratio $l_{k,\tau}$ is computed for $\tau = 1, \ldots, k$. We readily realise that such an implementation of the generalised likelihood ratio test requires a linear growth of storage and computing resources with time. To keep the problem tractable, the scrutinising is restricted to a sliding window that encompasses the latest $N_{wd} + 1$ residuals. This approximation is valid provided the window is sufficiently large to ensure detection of all faults of interest.

Practically, at time step k, the search for an event is constrained to the interval $k - N_{wd} \leq \tau \leq k$. In a sense, this turns the generalised likelihood ratio into a finite memory filter. It is worth noting that recursive relations are available for the update of the quantities involved in the likelihood ratio $l_{k,\tau}$ which is rather beneficial from a computational standpoint. The buffer is composed of a series of $3N_{wd} + 3$ matrices whose maximal size is $n_h \times n_h$ and $N_{wd} + 1$ vectors of dimension n_h too. The recursive relations are provided below:

$$
\begin{array}{lcl}
\multicolumn{3}{c}{\text{for } \tau \in [k-N_{wd}, k[} \\
\mathbf{C}_{k,\tau} & = & \mathbf{C}_{k-1,\tau} + \mathbf{H}_{k,\tau}^T \, \mathbf{P}_{\mathbf{y},k}^{-1} \, \mathbf{H}_{k,\tau} \\
\mathbf{d}_{k,\tau} & = & \mathbf{d}_{k-1,\tau} + \mathbf{H}_{k,\tau}^T \, \mathbf{P}_{\mathbf{y},k}^{-1} \, \mathbf{r}_k \\
\mathbf{F}_{k,\tau} & = & \mathbf{F}_{k-1,\tau} + \mathbf{K}_k \, \mathbf{H}_{k,\tau} \\
\mathbf{H}_{k,\tau} & = & \mathbf{G}_{\mathbf{h},k} \left(\mathbf{I} - \mathbf{F}_{k-1,\tau} \right)
\end{array}
\qquad
\begin{array}{lcl}
\multicolumn{3}{c}{\text{for } \tau = k} \\
\mathbf{C}_{k,k} & = & \mathbf{G}_{\mathbf{h},k}^T \, \mathbf{P}_{\mathbf{y},k}^{-1} \, \mathbf{G}_{\mathbf{h},k} \\
\mathbf{d}_{k,k} & = & \mathbf{G}_{\mathbf{h},k}^T \, \mathbf{P}_{\mathbf{y},k}^{-1} \, \mathbf{r}_k \\
\mathbf{F}_{k,k} & = & \mathbf{K}_k \, \mathbf{G}_{\mathbf{h},k} \\
\mathbf{H}_{k,k} & = & \mathbf{0}
\end{array}
$$

We have already mentioned several times that most practical cases of engine performance monitoring are characterised by a number of health parameters greater than the number of measurements. As a consequence, the system is only partially observable and each $n_h \times n_h$ matrix $\mathbf{C}_{k,\tau}$ is rank-deficient. The pseudo-inverse [see Golub and van Loan, 1996] advisedly replaces the common inverse in the computation of the likelihood ratio. To make a connexion with the observability notions introduced in section 2.5, the detection capability is therefore limited to events that lie – at least partially – in the observable subspace. In other words, any event totally located in the unobservable subspace is not detectable for it does not produce any signature on the residuals.

For the very same reason of ill-posedness, the jump estimate (3.45) should be computed either with the pseudo-inverse or with the map estimator (2.70). However, Akkaram et al. [2006] underline that both approaches lead to minimum ℓ_2-norm solutions which means that a localised fault is spread by the algorithm on several components. As a consequence, the compensation capability of the GLR technique is discarded in the implemented tool in favour of a dedicated module for fault isolation and quantification. This module is presented in section 3.5.

Figure 3.4 depicts the integration of the detection module, which comprises all the elements in the dashed box, with the Kalman filter. The performance model, actually

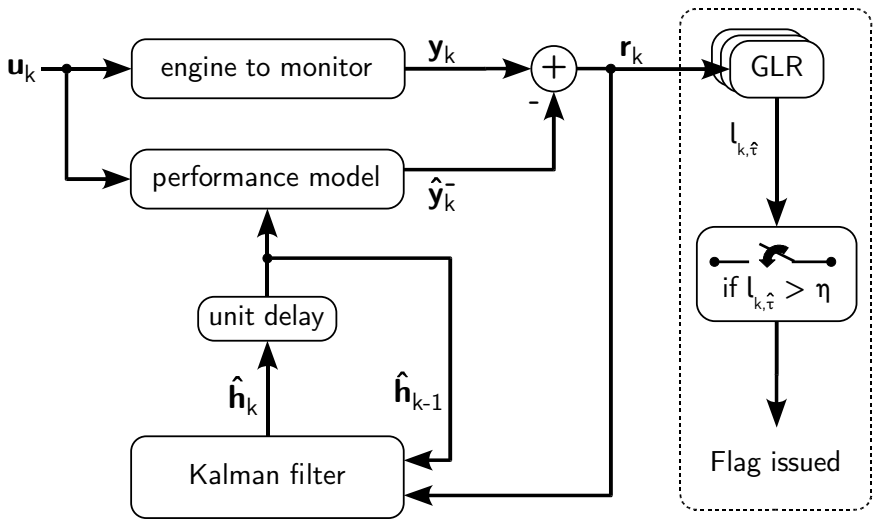


Figure 3.4: Integration of the detection module with the Kalman filter

embedded in the Kalman filter, is represented outside of it in order to underline its importance in this model-based algorithm. For sake of clarity, only the most relevant data streams are sketched in the diagram. Briefly explained, the adaptive component works as follows.

- The GLR box performs the following operations:
  1. update the quantities in the N_{wd}-sized buffer according to the relations given above,
  2. compute the likelihood ratio over the window by means of equation (3.42),
  3. search for the maximum value of the likelihood ratio.

  The outputs of the GLR box are the estimated time of occurrence of the jump $\widehat{\tau}$ (not represented in figure 3.4) and the maximum value of the likelihood ratio $l_{k,\widehat{\tau}}$.
- This value is compared to the threshold η in order to determine whether a jump has occurred.
- In case hypothesis H_1 is true, a detection flag is issued. This will serve as a trigger for the fault isolation module.

Once a jump is detected and subsequently identified, the buffer of residuals $\mathbf{d}_{k,\tau}$ is reset in order to avoid possible multiple detections of the same jump. With this modification, the detection module is moreover able to detect several successive jumps.

Two tuning parameters are available in the detection module, namely the width N_{wd} of the sliding window and the threshold η in the hypothesis testing (3.44). The selection of N_{wd} is dictated by a trade-off between accurate and fast detection of the events. The former implies to choose N_{wd} large enough while the latter advocates a small-sized buffer. On the other hand, the threshold η is directly related to the probability of false alarms P_F in jump detection through:

$$P_F = \int_{\eta}^{\infty} p(l_{k,\widehat{\tau}}|H_0)\ \mathrm{d}l \tag{3.48}$$

where $p(l_{k,\widehat{\tau}}|H_0)$ is the probability density of the likelihood ratio conditioned on H_0 which is a central Chi-squared density with n_h degrees of freedom, see van Trees [1968] for a proof. In practice, relation (3.48) is inverted numerically to obtain the threshold η for a prescribed false alarm rate P_F.

In terms of numerical operations, the once-per-time-step call to the detection module is roughly equivalent to $N_{wd}+1$ runs of a linear Kalman filter. The increase in computational burden is hence directly proportional to the length of the sliding window.

3.5 Fault isolation module

3.5.1 Context

Once the occurrence of an event is established, the next task is to localise it and to quantify its magnitude. Typically, accidental events impact at most one (or two) component(s) at a time so that a limited number of health parameters do actually participate in the fault. Isolation is therefore closely tied to the problem of subset selection.

In section 1.4, the review of methods for gas path analysis showed that artificial-intelligence based techniques deliver better performance for the isolation of accidental events essentially because of their classification nature that approximates the solution space with a limited number of cases. On the contrary, optimal estimation techniques are more suited to the assessment of distributed degradation, but tend to spread localised faults over several components.

This side-effect, sometimes termed *smearing* in the literature [see Provost, 1994], is due to the way the estimation problem is generally defined as we will see below. Briefly put, this usual formulation penalises more heavily large variations in a few parameters, which are typical of abrupt faults, than smaller variations in many parameters. This explains why we have discarded the compensation feature of the GLR detector presented in section 3.4 as it is rooted in the same framework.

The literature reports a few attempts aimed at reducing the smearing effect of optimal estimation techniques. Aretakis et al. [2003] performed the subset selection through a combinatorial approach. A least-square estimate is computed for each possible combination of $n_h^{'} \leq n_y$ health parameters. This solution suffers obviously from the curse of dimensionality when applied to real-world problems. Kobayashi and Simon [2003] applied a bank of estimators to isolate faults, each estimator being tailored to one particular fault type. Grodent and Navez [2003] tackled the deficiency of traditional least-square techniques at a deeper level by using a regularisation term that lowers the penalty on large deviations of the health parameters, hence improving fault isolation. Their elegant solution did not require to preset a library of possible faults.

We can take a further step to enhanced fault isolation by recognising that abrupt events involve only a reduced number of health parameters, which translates mathematically into the search for a *sparse* solution. Basically, the estimation problem is tweaked in order to favour a solution with many nought terms. Problems characterised by sparsity are encountered in various fields such as compressed sensing [Donoho, 2006], linear regression [Larsson and Selen, 2007] or source localisation [Malioutov et al., 2005] to name a few. Jokar and Pfetsch [2007] provide a good review of various algorithms looking for sparse solutions. Interestingly, Fuchs [2004b] shows that one flavour of the sparse estimation problem can be cast as a Quadratic Programming (QP) problem for which efficient solvers are available.

In [Borguet and Léonard, 2010a], we adopted this QP formulation to derive a fault isolation tool that promotes sparse solutions. The results indicated that the sparse estimator performs indeed better than the traditional regularised least-squares.

3.5.2 The cause of smearing in least-square estimation

As a first step towards improving the isolation capability of least-square-based estimators, let us try to figure out where its current deficiency comes from. For the task of fault isolation, we consider the statistical model (2.35) with its whitened noise distribution. Without loss of generality, we further assume that the columns of the matrix $\tilde{\mathbf{G}}_\mathbf{h}$ are normalised to one in Euclidian norm. This comes down to express the health parameters on a new scale. Moreover for sake of simplicity, we turn to the vocabulary from Optimisation Theory, instead of the vocabulary from Statistics used so far.

As presented in section 2.3, the least-squares estimates of the health parameters are obtained by minimising the squared Euclidian norm of the residuals:

$$\delta\widehat{\mathbf{h}}_{\mathrm{LS}} = \arg\min_{\delta\mathbf{h}} \left\{ \frac{1}{2} \left(\delta\tilde{\mathbf{y}} - \tilde{\mathbf{G}}_\mathbf{h}\, \delta\mathbf{h}\right)^T \left(\delta\tilde{\mathbf{y}} - \tilde{\mathbf{G}}_\mathbf{h}\, \delta\mathbf{h}\right) \right\} \tag{3.49}$$

The stationarity conditions are obtained by equating the first-order derivatives of the least-square criterion to zero, which gives the following system of n_h normal equations:

$$(\tilde{\mathbf{G}}_\mathbf{h}^T \tilde{\mathbf{G}}_\mathbf{h})\, \delta\widehat{\mathbf{h}}_{\mathrm{LS}} = \tilde{\mathbf{G}}_\mathbf{h}^T\, \delta\tilde{\mathbf{y}} \quad \Leftrightarrow \quad \mathbf{FIM}\, \delta\widehat{\mathbf{h}}_{\mathrm{LS}} = \mathbf{rhs} \tag{3.50}$$

As the number of health parameters n_h outweighs the number of measurements n_y, the rank of the $n_h \times n_h$ Fisher information matrix **FIM** is at best equal to n_y and consequently the matrix cannot be inverted. The problem is underdetermined and has an infinite number of solutions which can be written as:

$$\delta\widehat{\mathbf{h}}_{\mathrm{LS}} = \delta\widehat{\mathbf{h}}_{\mathrm{range}} + \delta\widehat{\mathbf{h}}_{\mathrm{null\ space}} \tag{3.51}$$

where $\delta\widehat{\mathbf{h}}_{\mathrm{range}}$ is unique as it lies in the range (observable subspace) of the matrix **FIM** whereas $\delta\widehat{\mathbf{h}}_{\mathrm{null\ space}}$ represents the undetermined part that lies in the null-space (unobservable subpsace) of the matrix **FIM** and therefore has no influence on the measurements as explained in section 2.5.

We can compute the minimum ℓ_2-norm solution of the underdetermined set of equations by means of the pseudo- (aka. Moore-Penrose) inverse. This strategy is applied by Akkaram et al. [2006] to the tuning of simulation models. The pseudo-inverse is a generalisation of the notion of inverse to the case of rectangular and/or singular matrices and is based on the SVD. Let us apply an SVD to the matrix **FIM**:

$$\mathbf{FIM} = \mathbf{U} \left[\begin{array}{c|c} \boldsymbol{\Sigma}_{11} & \mathbf{0}_{n_y \times (n_h - n_y)} \\ \hline \mathbf{0}_{(n_h - n_y) \times n_y} & \boldsymbol{\Sigma}_{22} \end{array}\right] \mathbf{V}^T \tag{3.52}$$

where $\boldsymbol{\Sigma}_{11}$ is a diagonal matrix containing the n_y non-zero singular values and $\boldsymbol{\Sigma}_{22}$ is a diagonal matrix containing the remaining $n_h - n_y$ null singular values. $\mathbf{U}$ and $\mathbf{V}$ are $n_h \times n_h$ orthonormal matrices.

The pseudo-inverse is defined according to [see Golub and van Loan, 1996]:

$$\mathbf{FIM}^{\dagger} \triangleq \mathbf{V} \left[\begin{array}{c|c} \boldsymbol{\Sigma}_{11}^{-1} & \mathbf{0}_{n_y \times (n_h - n_y)} \\ \hline \mathbf{0}_{(n_h - n_y) \times n_y} & \boldsymbol{\Sigma}_{22} \end{array}\right] \mathbf{U}^T \tag{3.53}$$

where $\boldsymbol{\Sigma}_{11}^{-1}$ is a diagonal matrix containing the inverse of the non-zero singular values. Loosely speaking, the pseudo-inverse inverts the "invertible part" of the matrix and discards the contribution of the null space. Therefore it leads to the minimum ℓ_2-norm solution:

$$\begin{aligned} \delta\widehat{\mathbf{h}}_{\mathrm{LS}} &= \mathbf{FIM}^{\dagger}\ \mathbf{rhs} \\ &= \left[\begin{array}{c|c} \mathbf{V}_1 & \mathbf{V}_2 \end{array}\right] \left[\begin{array}{c} \boldsymbol{\Sigma}_1^{-1} \\ \hline \boldsymbol{\Sigma}_2^{-1} \end{array}\right] \mathbf{U}^T\ \mathbf{rhs} \\ &= \underbrace{\mathbf{V}_1\ \boldsymbol{\Sigma}_1^{-1}\ \mathbf{U}^T\ \mathbf{rhs}}_{\delta\widehat{\mathbf{h}}_{\text{range}}} + \underbrace{\mathbf{V}_2\ \boldsymbol{\Sigma}_2^{-1}\ \mathbf{U}^T\ \mathbf{rhs}}_{\delta\widehat{\mathbf{h}}_{\text{null space}}} \end{aligned} \tag{3.54}$$

as $\delta\widehat{\mathbf{h}}_{\text{null space}} = \mathbf{0}$ by construction of the pseudo-inverse.

Another possibility to solve the underdetermined set of equations is to make use of regularisation. We add a set of soft constraints on the health parameters, which was termed a priori knowledge in the probabilistic framework of section 2.4. The most common regularisation scheme for underdetermined least-squares problems consists in adding a quadratic penalisation on the deviations of the health parameters. As a result, the algorithm is driven towards a solution that has the minimum weighted (by matrix $\mathbf{Q}_0$) ℓ_2-norm as we can realise from the mathematical expression of the problem:

$$\delta\widehat{\mathbf{h}}_{\mathrm{RLS}} = \arg\min_{\boldsymbol{\delta h}} \left\{ \frac{1}{2}\ (\delta\tilde{\mathbf{y}} - \tilde{\mathbf{G}}_{\mathbf{h}}\ \delta\mathbf{h})^T\ (\delta\tilde{\mathbf{y}} - \tilde{\mathbf{G}}_{\mathbf{h}}\ \delta\mathbf{h}) + \frac{1}{2}\ \delta\mathbf{h}^T\ \mathbf{Q_0}^{-1}\ \delta\mathbf{h} \right\} \tag{3.55}$$

The choice of a quadratic function for the regularisation term allows an analytic relation to be worked out which is the major reason of its popularity.

These developments explain the smearing effect that is observed with the usual least-square techniques. Indeed, due to the quadratic penalisation on the parameter deviations, be it either implicit or explicit, solutions involving small variations in numerous health parameters have a lower cost than solutions characterised by large deviations in a few health parameters. As a result, the algorithm has the tendency to spread the effect of localised faults over several components.

3.5.3 Favouring sparsity in the estimation process

As underlined in section 2.4, regularisation helps solving underdetermined problems by artificially improving their mathematical conditioning. An adequate selection of the regularisation term is hence of paramount importance and should be made in accordance with the structure of the expected solution.

Short timescale faults impact only a limited number of health parameters, which means that many elements in the vector of health parameter deviations are equal to zero. Such a pattern is termed sparse. It should therefore be possible to enhance the concentration capability of a least-square-based algorithm by introducing a regularisation term that favours sparsity.

A first possible formulation of the sparse estimation problem is to look for the solution having the smallest number of non-zero components [Fuchs, 2004a], which can be thought of as an ideal measure of sparsity, this writes:

$$\min_{\delta\mathbf{h}} \left\{ \frac{1}{2}(\delta\tilde{\mathbf{y}} - \tilde{\mathbf{G}}_{\mathbf{h}}\, \delta\mathbf{h})^T\, (\delta\tilde{\mathbf{y}} - \tilde{\mathbf{G}}_{\mathbf{h}}\, \delta\mathbf{h}) + \lambda\, \|\delta\mathbf{h}\|_0 \right\} \text{ with } \lambda > 0 \tag{3.56}$$

where $\|\delta\mathbf{h}\|_0$ is the ℓ_0-quasi-norm, defined as the number of non-zero elements in vector $\delta\mathbf{h}$. The scalar parameter λ sets the trade-off between least-square fit of the data and sparsity in the solution.

Note that other definitions of sparse estimation are conceivable. For instance we could look for the solution of a given sparsity p that minimises the data misfit. This formulation would also involve the ℓ_0-quasi-norm, but now as an equality constraint, in the optimisation problem:

$$\min_{\delta\mathbf{h}} \left\{ \frac{1}{2}(\delta\tilde{\mathbf{y}} - \tilde{\mathbf{G}}_{\mathbf{h}}\, \delta\mathbf{h})^T\, (\delta\tilde{\mathbf{y}} - \tilde{\mathbf{G}}_{\mathbf{h}}\, \delta\mathbf{h}) \right\} \text{ subject to } \|\delta\mathbf{h}\|_0 = p \tag{3.57}$$

While their mathematical expression looks incredibly simple, the sparse estimation problems (3.56-3.57) are quite burdensome to solve in practice. Indeed, the term $\|\delta\mathbf{h}\|_0$ makes them of combinatorial nature. The optimal solution can only be obtained by means of an exhaustive search, which rapidly becomes unmanageable in industrial applications.

3.5.4 A convex approximation to sparse estimation

Different strategies have been developed in order to approximate the optimisation problem (3.56), see for instance [Tropp, 2004] for a short review, and make its computation tractable. Among them, the so-called relaxation approach has enjoyed a reasonable level of success and is elected here. Basically, the idea is to replace the ℓ_0-quasi-norm with a

suitable ℓ_p-norm[1]. Any ℓ_p-(pseudo-)norm promotes sparsity provided $p \leq 1$ essentially because the slope of the penalisation function decreases as the parameter deviation increases. Among this subfamily, the ℓ_1-norm has the considerable advantage to turn the resulting optimisation problem into a convex one, which admits a global optimum.

Figure 3.5 shows the penalty induced by three different norms for a positive deviation of a scalar parameter. For the ℓ_0 quasi-norm, the penalty has a unit value as soon as the parameter deviation differs from zero. In that way, large deviations have no higher cost than tiny ones. It can be seen that the ℓ_1-norm is the convex function of lowest order that places a unit penalty on unit coefficients and sets the cost to zero for nought coefficients. The ℓ_2-norm used in the classical regularised least-squares provides the same costs as the ℓ_1-norm for zero and unit deviations, but penalises much more large deviations and somewhat less small deviations. This sketch shows clearly that the ℓ_1-norm provides the natural convex relaxation of the ℓ_0-quasi-norm. On the contrary, the ℓ_2-norm is totally inappropriate for favouring sparsity.

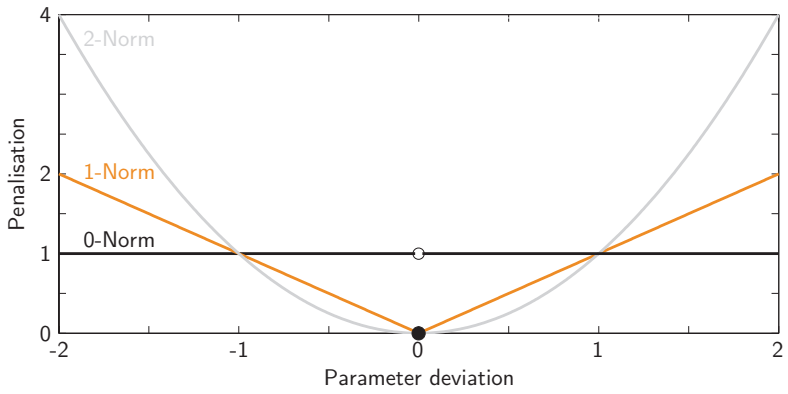


Figure 3.5: Comparison of the penalty induced by different norms

Thus, the convex relaxation to (3.56) is:

$$\min_{\delta\mathbf{h}} \left\{ \frac{1}{2}(\delta\tilde{\mathbf{y}} - \tilde{\mathbf{G}}_{\mathbf{h}}\, \delta\mathbf{h})^T\, (\delta\tilde{\mathbf{y}} - \tilde{\mathbf{G}}_{\mathbf{h}}\, \delta\mathbf{h}) + \lambda\, \|\delta\mathbf{h}\|_1 \right\} \text{ with } \lambda > 0 \tag{3.58}$$

where the ℓ_1-norm of the parameter deviations replaces the sparsity measure $\|\delta\mathbf{h}\|_0$.

We can translate the convex approximation to sparse estimation into Bayesian language as done by Granai and Vandergheynst [2005] for a closely-related problem. Leaving the derivation to the interested reader, we state that the optimisation problem (3.58) has the same mathematical expression as the search for the maximum a posteriori estimate where each health parameter is considered as an independent random variable that follows a Laplace distribution. Roughly speaking, we have moved from a Gaussian to a Laplace prior with respect to the developments of section 2.4. The Laplace distribution has a fatter tail than the Gaussian one, therefore large deviations in the health parameters

[1] the family of ℓ_p-norms is defined in appendix A.2

are more likely. The assumption of independence between the parameters might be a downside of the present approach because we cannot easily induce some coupling between health parameters of a same component as we did when using a Gaussian prior.

The scalar λ in the convex relaxation (3.58) plays a similar role as in (3.56) in the sense that it balances the least-square data fit and the relaxed-sparsity term. There are three remarkable values for this parameter [see Fuchs, 2004a]:

- if $\lambda = 0$, the optimisation problem (3.58) boils down to the classical underdetermined least-squares and the set of solutions is convex,
- if $\lambda = 0^+$, that is λ positive and arbitrarily close to zero, the solution is attained at the point(s) in the previous set having least ℓ_1-norm,
- if $\lambda \geq \left\| \tilde{\mathbf{G}}_{\mathbf{h}}^T \, \delta\tilde{\mathbf{y}} \right\|_\infty$, the optimum is $\delta\widehat{\mathbf{h}} = \mathbf{0}$ (see the dual formulation below).

As λ increases, the quality of the least-squares fit degrades whereas the solution becomes sparser. We shall consequently tune this parameter to reach the best performance. To better understand the influence of λ on the solution, it is worth taking a look at the dual formulation of (3.58) [Fuchs, 2001].

First, we convert the optimisation problem (3.58) into a QP problem [see appendix A.3 or Fletcher, 2000] by replacing the vectors $\delta\mathbf{h}$ and $|\delta\mathbf{h}|$ with their positive $\delta\mathbf{h}^+$ and negative $\delta\mathbf{h}^-$ parts as shown by Fuchs [2004b]:

$$\left\{ \begin{array}{r} \delta\mathbf{h} = \delta\mathbf{h}^+ - \delta\mathbf{h}^- \\ |\delta\mathbf{h}| = \delta\mathbf{h}^+ + \delta\mathbf{h}^- \end{array} \right. \quad \text{with} \quad \left\{ \begin{array}{l} \delta\mathbf{h}^+ \triangleq \max(\delta\mathbf{h}, \mathbf{0}) \\ \delta\mathbf{h}^- \triangleq \max(-\delta\mathbf{h}, \mathbf{0}) \end{array} \right. \tag{3.59}$$

where the operator $\max(\mathbf{v}, \mathbf{0})$ is applied element-wise.

This rather simple change of variables doubles the number of optimisation variables, but on the other hand it leads to the following quasi-unconstrained quadratic program:

$$\min_{\delta\mathbf{h}^+, \delta\mathbf{h}^-} \left\{ \frac{1}{2} \left\| \delta\tilde{\mathbf{y}} - \tilde{\mathbf{G}}_{\mathbf{h}} \left(\delta\mathbf{h}^+ - \delta\mathbf{h}^- \right) \right\|_2^2 + \lambda \, \mathbf{1}_{n_h}^T \left(\delta\mathbf{h}^+ + \delta\mathbf{h}^- \right) \right\} \text{ subject to } \left\{ \begin{array}{l} \delta\mathbf{h}^+ \geq \mathbf{0} \\ \delta\mathbf{h}^- \geq \mathbf{0} \end{array} \right. \tag{3.60}$$

where $\mathbf{1}_{n_h}$ is a column vector of length n_h with each element equal to one.

Let us stack the unknowns into one single vector $\delta\mathbf{h}_s^T = [\delta\mathbf{h}^{+T} \; \delta\mathbf{h}^{-T}]$, the quadratic program (3.60) becomes:

$$\min_{\delta\mathbf{h}_s} \left\{ \frac{1}{2} \left\| \delta\tilde{\mathbf{y}} - \tilde{\mathbf{G}}_{\mathbf{h},s} \, \delta\mathbf{h}_s \right\|_2^2 + \lambda \, \mathbf{1}_{2n_h}^T \, \delta\mathbf{h}_s \right\} \text{ subject to } \delta\mathbf{h}_s \geq \mathbf{0} \tag{3.61}$$

with matrix $\tilde{\mathbf{G}}_{\mathbf{h},s} = [\tilde{\mathbf{G}}_{\mathbf{h}} \mid -\tilde{\mathbf{G}}_{\mathbf{h}}]$

We form the Lagrangian function associated to the constrained estimation problem (3.61):

$$\mathcal{L}(\delta\mathbf{h}_s, \boldsymbol{\mu}) = \frac{1}{2} \left\| \delta\tilde{\mathbf{y}} - \tilde{\mathbf{G}}_{\mathbf{h},s}\, \delta\mathbf{h}_s \right\|_2^2 + \lambda\, \mathbf{1}_{2n_h}^T\, \delta\mathbf{h}_s - \boldsymbol{\mu}^T\, \delta\mathbf{h}_s \tag{3.62}$$

where $\boldsymbol{\mu}$ is the vector of the Lagrangian multipliers.

The stationarity conditions of the Lagrangian with respect to $\delta\mathbf{h}_s$ writes:

$$-\tilde{\mathbf{G}}_{\mathbf{h},s}^T \left(\delta\tilde{\mathbf{y}} - \tilde{\mathbf{G}}_{\mathbf{h},s}\, \delta\mathbf{h}_s\right) + \lambda\, \mathbf{1}_{2n_h} - \boldsymbol{\mu} = \mathbf{0} \tag{3.63}$$

Left-multiplying the above relation with $\delta\mathbf{h}_s^T$ and rearranging the terms gives:

$$\lambda\, \delta\mathbf{h}_s^T\, \mathbf{1}_{2n_h} - \delta\mathbf{h}_s^T\, \boldsymbol{\mu} = -\delta\mathbf{h}_s^T\, \tilde{\mathbf{G}}_{\mathbf{h},s}^T \left(\delta\tilde{\mathbf{y}} - \tilde{\mathbf{G}}_{\mathbf{h},s}\, \delta\mathbf{h}_s\right) \tag{3.64}$$

Introducing (3.64) into (3.62), the Lagrangian function writes:

$$\begin{aligned} \mathcal{L}(\delta\mathbf{h}_s, \boldsymbol{\mu}) &= \frac{1}{2} \left\| \delta\tilde{\mathbf{y}} - \tilde{\mathbf{G}}_{\mathbf{h},s}\, \delta\mathbf{h}_s \right\|_2^2 + \delta\mathbf{h}_s^T\, \tilde{\mathbf{G}}_{\mathbf{h},s}^T \left(\delta\tilde{\mathbf{y}} - \tilde{\mathbf{G}}_{\mathbf{h},s}\, \delta\mathbf{h}_s\right) \\ &= -\frac{1}{2} \left\{ \left\| \tilde{\mathbf{G}}_{\mathbf{h},s} \delta\mathbf{h}_s \right\|_2^2 + \|\delta\tilde{\mathbf{y}}\|_2^2 \right\} \end{aligned} \tag{3.65}$$

The dual optimisation problem consists in maximising the Lagrangian function with respect to $\boldsymbol{\mu}$ under the constraint that each component in $\boldsymbol{\mu}$ is positive. An equivalent formulation is the following one:

$$\min_{\delta\mathbf{h}_s} \left\{ \frac{1}{2} \left\| \tilde{\mathbf{G}}_{\mathbf{h},s}\, \delta\mathbf{h}_s \right\|_2^2 \right\} \text{ subject to } \begin{cases} \boldsymbol{\mu} \geq \mathbf{0} \\ \tilde{\mathbf{G}}_{\mathbf{h},s}^T \left(\delta\tilde{\mathbf{y}} - \tilde{\mathbf{G}}_{\mathbf{h},s}\, \delta\mathbf{h}_s\right) = \lambda\, \mathbf{1}_{2n_h} - \boldsymbol{\mu} \end{cases} \tag{3.66}$$

As the Lagrangian multipliers do not appear explicitly anymore in the objective function of (3.66), the set of constraints can be alternately rewritten as:

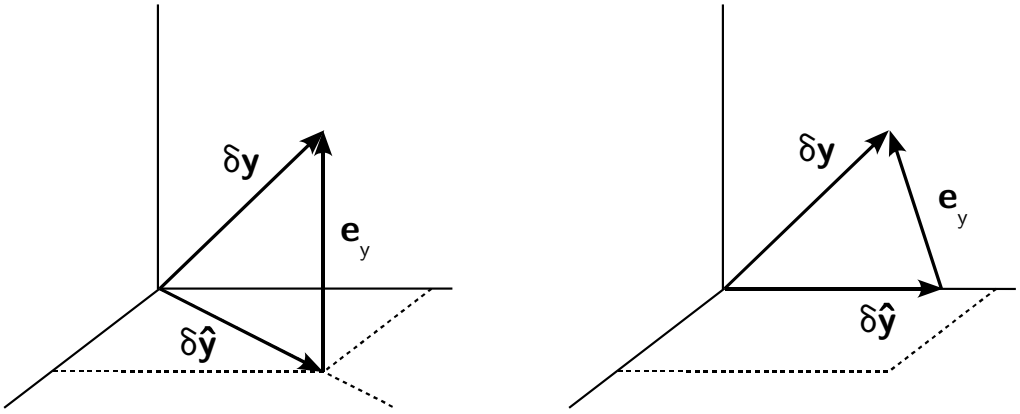
$$\tilde{\mathbf{G}}_{\mathbf{h},s}^T \left(\delta\tilde{\mathbf{y}} - \tilde{\mathbf{G}}_{\mathbf{h},s}\, \delta\mathbf{h}_s\right) \leq \lambda\, \mathbf{1}_{2n_h} \tag{3.67}$$

Finally, we revert back to the original unknowns *i.e.*, the deviations in the health parameters $\delta\mathbf{h}$. The dual problem (3.66) is then expressed as:

$$\min_{\delta\mathbf{h}} \left\{ \|\tilde{\mathbf{G}}_{\mathbf{h}}\, \delta\mathbf{h}\|_2^2 \right\} \text{ subject to } \|\tilde{\mathbf{G}}_{\mathbf{h}}^T\, (\delta\tilde{\mathbf{y}} - \tilde{\mathbf{G}}_{\mathbf{h}}\, \delta\mathbf{h})\|_\infty \leq \lambda \tag{3.68}$$

In the dual formulation, the scalar inequality constraint states that, at the solution, the vector of residuals $\mathbf{e}_y = \delta\tilde{\mathbf{y}} - \tilde{\mathbf{G}}_\mathbf{h}\,\delta\widehat{\mathbf{h}}$ is such that its correlation with any of the columns of $\tilde{\mathbf{G}}_\mathbf{h}$ never exceeds λ. As a matter of fact, λ allows to tune the maximum magnitude of these correlations. When processing noisy data, it is advised in [Fuchs, 2004a] to set λ to the standard deviation of the noise, which is equal to one for the scaled system.

We can interpret geometrically this constraint and further put it in perspective with the orthogonality principle (2.53) that characterises the traditional least-square solution. This least-square solution, depicted in subfigure 3.6(a), is obtained through an orthogonal projection of the measurement vector $\delta\tilde{\mathbf{y}}$ onto the subspace (a plane in our 2-D sketch) spanned by the columns of the matrix $\tilde{\mathbf{G}}_\mathbf{h}$, each column giving the contribution of one particular parameter. Consequently, the residual vector $\mathbf{e}_y$ is perpendicular to the plane. On the other hand, the constraint associated with the dual formulation of the sparse estimation framework (subfigure 3.6(b)) can be seen as a relaxed version of the orthogonality condition. The sparse solution is obtained here through an oblique projection whose deviation from orthogonality is controlled via parameter λ. The Euclidian norm of the residual vector $\mathbf{e}_y$ is larger than in the traditional least-square case, but the reconstructed measurement vector $\delta\widehat{\mathbf{y}}$ uses a smaller number of health parameters.



(a) Traditional least-squares (b) Sparse estimation

Figure 3.6: Geometric interpretation of the solutions

3.5.5 Practical implementation

The optimisation problem (3.61) is solved by means of the package MINQ by Neumaier [1998]. A specificity of this solver is that it can handle quadratic programs with semi-positive-definite Hessian. It consists of a combination of coordinate searches and subspace minimisation steps and uses rank 1 updates to keep the computation cheap. In every computation dedicated to fault isolation and quantification, we use an updated influence coefficient matrix $\tilde{\mathbf{G}}_\mathbf{h}$ evaluated at the latest estimate of the health parameters.

One of the main objectives of the isolation module is to provide the maintenance staff with an automated assessment of the faulty component(s). The vector of health parameter deviations obtained after solving the optimisation problem (3.60) is however in a not so convenient form for that purpose. Therefore, we apply a relatively simple isolation logic, inspired by the one devised by Aretakis et al. [2003] in order to determine the faulty component. This isolation logic assumes that only one component is faulty at a time. First, we compute the magnitude of each fault type as the ℓ_2-norm of the parameter deviations – typically one or two – associated to the said component fault. The component with the largest fault magnitude is then deemed as the faulty one.

Due to the presence of noise in the data, the best we can expect is that the sparse solution has its non-zero elements at the right locations and with the right signs [Fuchs, 2006]. We have however no guarantee that the amplitudes are accurately assessed. The deviations in the health parameters of the suspected faulty component are then evaluated via a maximum likelihood approach limited to the subset of parameters associated to the faulty components. To sum up, the isolation and quantification tasks are carried out in a two-pass process. Firstly we perform a subset selection with the sparse estimation tool and secondly we evaluate the magnitude of the fault with the traditional least-square criterion.

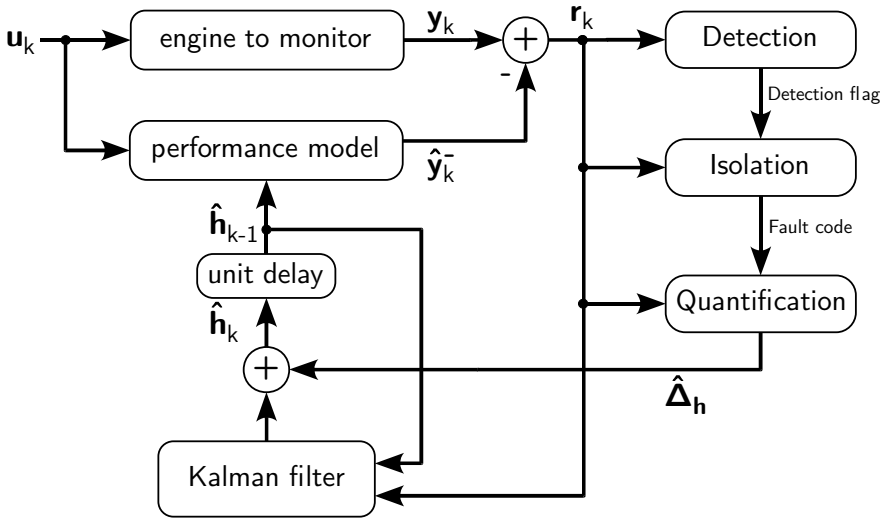


Figure 3.7: Sketch of the proposed EHM solution

Figure 3.7 shows the integration of the isolation and quantification modules with the adaptive Kalman filter discussed in section 3.4. We restrict the presentation to the main data stream for sake of clarity. The outputs of the detection and isolation modules, respectively a detection flag and a fault code together with their timestamps, could be archived in a file associated to the engine. In order to improve the signal-to-noise ratio for the isolation and quantification tasks, the actual implementation follows the flowchart below:

1. once a detection flag is raised by the detection module, both the Kalman filter and the GLR detector are put at rest and residuals based on the latest estimate of the health parameters are gathered for the N_{wi} next cycles. N_{wi} should be chosen small enough in order to minimise the delay between detection and isolation while at the same time providing some reduction in the noise level,

2. the sparse estimation problem is solved for the N_{wi} vectors of residuals in the buffer. The fault isolation logic determines the faulty component and passes its associated code to the quantification unit,

3. the fault magnitude is assessed as previously explained. The latest Kalman filter estimate is then compensated with the estimated fault. The buffers of the detection and isolation modules are emptied and the adaptive Kalman filter is relaunched until the next detection.

Chapter 4

Assessment of the diagnosis tool

Contents

4.1 Foreword

This chapter is dedicated to the evaluation of the engine health monitoring tool developed in the previous chapter. The assessment is carried out on the basis of simulated test-cases, also called scenarios in the sequel. It is indeed almost impossible to have access to real operational data and to a performance deck of a real engine given the highly proprietary nature of such elements. Although simulated data will ever be only a more or less faithful image of actual field recordings, it has some advantages for developers of health management solutions such as: i) the ability to simulate a large range of fault conditions, ii) a low cost to generate these data and iii) the knowledge of the true solution, thereby allowing the quantification of the performance of the monitoring tool.

The chapter is organised as follows: in section 4.2, we take a look at the engine performance model that is used to generate the scenarios as well as in the diagnosis tool. Section 4.3 describes the main features of the scenarios and their generation process. In section 4.4, we select a number of relevant metrics to assess quantitatively the performance of the diagnosis tool. The results obtained on the batch of scenarios are presented and discussed in section 4.5. Finally, section 4.6 comes back to the topic of sensor–parameter suites selection and analysis and proposes an application of the tools developed in section 2.5.

4.2 The engine performance model

The engine model was developed in the frame of the OBIDICOTE project[1] and is detailed in [Stamatis et al., 2001, Ruiz, 2001]. It simulates a two-spool, high bypass ratio, mixed flow turbofan engine in the 25 klbf ($\simeq$ 110 kN) thrust class. While not derived from an existing engine, it is representative of the power plants of single–aisle airliners such as the Boeing 737 and Airbus A320. Such aircrafts are part of the short–haul fleet of virtually all legacy airlines and are the workhorses of the low–cost carriers which may fly them as much as a dozen times a day. Figure 4.1 depicts a schematic of the engine where the location of the health parameters and the numbering of major inter–component stations are also indicated.

As seen in section 2.1.1, the operating point of an engine is defined through the control and environmental variables. The OBIDICOTE model features three control variables, of which the main one is the fuel flow WF injected in the combustor. In a real engine, this fuel flow is set by the control system in order to reach the thrust setting parameter ordered by the pilot or the flight management computer. Without loss of generality, we simply impose the fuel flow directly thereby disregarding the effect of any deterioration on thrust. The other two control parameters drive the variable geometry devices on the low– and high–pressure compressors to improve operability at part speed and during transients.

[1] A Brite/Euram project concerned with On-Board Identification Diagnosis and Control in Turbine Engines.

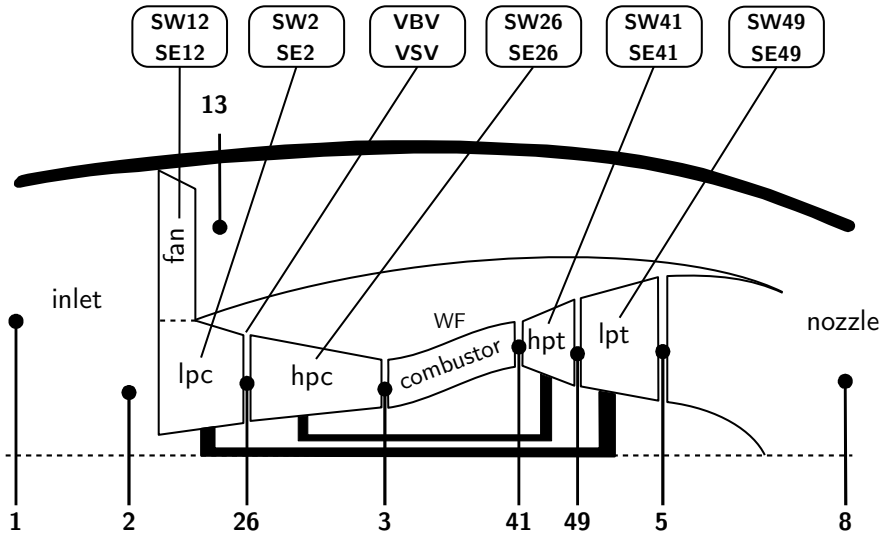


Figure 4.1: Turbofan layout with station numbering and health parameters location

Multivariable control having not yet made its way to commercial engines, these actuators are open-loop scheduled with respect to a regime indicator such as the corrected spool speeds [Spang and Brown, 1999, Linke-Diesinger, 2008].

The environmental variables define the thermodynamical state of the air surrounding the engine. As shown in figure 2.2, the triplet made of ambient pressure, ambient temperature and flight velocity (true air speed) is a possible choice. At the scenario generation stage, we prefer the alternate triplet formed by pressure altitude ALT, flight Mach number MN and the offset in temperature with respect to the International Standard Atmosphere $\mathsf{DT}_{\mathsf{ISA}}$. This is motivated by the fact that the first two quantities are effectively input by the pilot into the aircraft autopilot and the third one is quite convenient to define cold or hot days. These variables are however not readily available on board so yet another triplet of environmental variables is fed into the diagnosis tool. These are the inlet (ram) total pressure P2 and temperature T2 and the ambient pressure Pamb.

The health condition of the gas–path components of the engine is modelled by means of twelve parameters. Ten of them are traditional correction factors that modify the efficiency (SEi) and the flow capacity (SWi) of each turbomachinery component with respect to a reference condition as described in Mathioudakis et al. [2002]. In accordance with the discussion of section 1.2, no health parameter is attached to the combustor. The last two health parameters represent deviations relative to the nominal schedule of the variable geometry devices, namely variable stator vanes (VSV) and blow–off valves (VBV). They model either a fault on the sensed actuator position or a fault on the actuator itself (*e.g.*, mechanical failure).

The model generates a large number of outputs that are either measurable or not on a flying engine. On the measurable side, we have total temperatures and pressures at various inter-component planes and the spool speeds. The non–measurable family includes air

mass flows and performance indicators such as thrust, stall margins and efficiencies. The specific instrumentation used to perform the diagnosis will be presented later on.

4.3 The scenario generator

To generate the so–called scenarios, we have developed a tool inspired from the Engine Fleet Simulator reported in [Simon et al., 2008]. As sketched in figure 4.2, each scenario represents the operational life of an engine in a given fleet and is defined by a unique operating history and a unique health degradation profile. These inputs are processed with the engine performance model to generate a database of sensed parameters as if collected once per flight on a real operating engine. The length of a scenario is arbitrarily set to 3000 flights. More details about the input and output files are provided below.

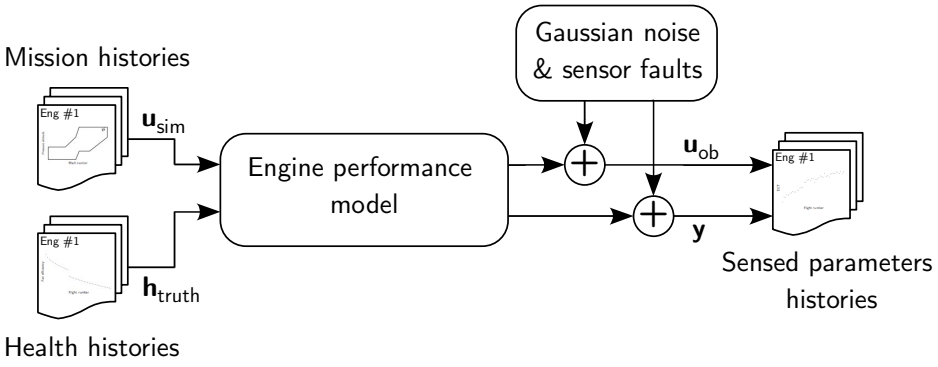


Figure 4.2: Schematic of the scenario generator

4.3.1 Mission histories

We assume that the virtual data collection is performed only during cruise given that it is the most likely part of the flight at which (almost perfect) steady–state operation can be expected. As a result, each operating point $\mathbf{u}_{sim}$ of each scenario is randomly selected in the envelope defined in table 4.1 which is representative of the spread observed in actual cruise conditions.

Name	Units	Lower bound	Upper bound
WF	kg/s	0.380	0.400
ALT	m	10058	11278
MN	–	0.78	0.82
DT_{ISA}	K	−10	+10

Table 4.1: Operating envelope

4.3.2 Health histories

The degradation profile $\mathbf{h}_{\mathsf{truth}}$ combines gradual wear of the engine modules and a single, abrupt fault picked from a predefined library. The abrupt event is superimposed to the gradual deterioration profile at a random flight in the window between flight 1000 and flight 2000. Complementary details about each category are given below.

Gradual deterioration

Progressive deterioration, representing engine wear, is modelled by altering the efficiency and flow correcting factors of the modules (fan, lpc, hpc, hpt, lpt). Despite the decrease in performance it causes, we do not regard gradual deterioration as a faulty condition in this work, but rather as a normal mechanism induced by engine usage. This point of view is shared by several authors *e.g.*, Jaw [2005] or Simon et al. [2008].

In the late seventies/early eighties, a number of studies, luckily published in the open literature [Sallee, 1978, Wulf, 1980, Sasahara, 1985], established typical performance deterioration profiles from historical data collected on first–generation large commercial turbofans. As stated by Simon [2008], a linear+exponential law captures the essential of these deterioration profiles. The four main parameters of this wear model are depicted in figure 4.3. The out–of–factory (*i.e.*, at flight 0) health condition $\mathbf{h}_0$ is assumed to be normally distributed with a mean value of one for each health parameter and a variance quoted in the third column of table 4.2, as per Roth et al. [2005]. The deterioration after 3000 flights $\mathbf{h}_{N_{fl}}$ is uniformly distributed in an interval ranging between 70% and 130% of the mean value quoted in the fourth column of table 4.2. Note that this fleet–average final deterioration is expressed in percent change with respect to the initial health condition.

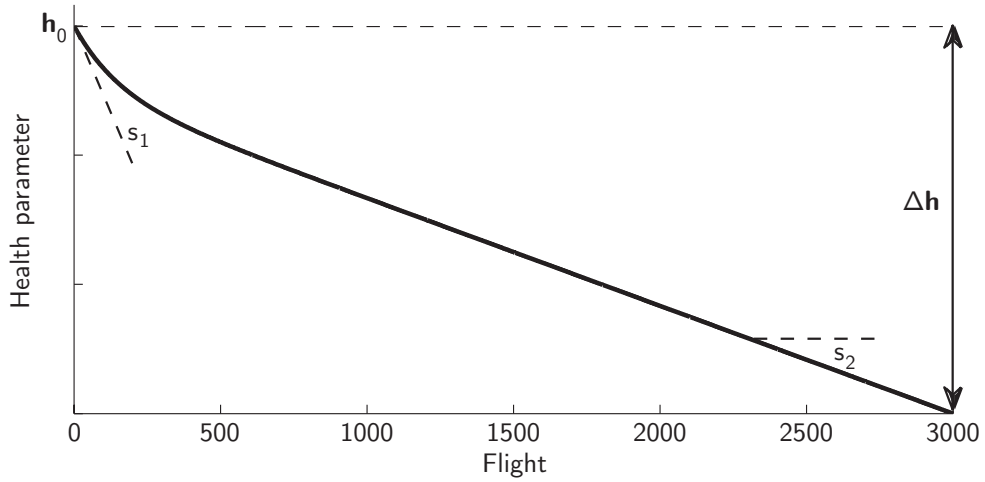


Figure 4.3: Main parameters of the wear model

As sketched in figure 4.3, the deterioration profile is mostly exponential for the first flights, this is representative of the usual running–in period of the engine, and mostly linear later on. The scalars s_1 and s_2 control the slope of the exponential and linear part respectively. Both parameters are uniformly distributed in an interval whose mean value and width are given respectively in the second and third columns of table 4.3, as per Simon [2008]. The randomness in all these four parameters of the wear model accounts for various effects such as engine–to–engine manufacturing variations and more/less rapid and severe deterioration of each module.

	Flight 0: $\mathbf{h}_0$		**Flight 3000: $\Delta\mathbf{h}$**	
Parameter	Mean	Std	Mean	Range
SW12	1.0	0.4%	−1.82%	±30%
SE12	1.0	1.0%	−1.50%	±30%
SW2	1.0	1.0%	−1.00%	±30%
SE2	1.0	0.5%	−0.65%	±30%
SW26	1.0	1.0%	−3.22%	±30%
SE26	1.0	1.0%	−2.71%	±30%
SW41	1.0	0.7%	2.28%	±30%
SE41	1.0	1.0%	−2.30%	±30%
SW49	1.0	0.2%	0.25%	±30%
SE49	1.0	0.6%	−0.47%	±30%

Table 4.2: Distribution of initial health and deterioration after 3000 flights

Parameter	Mean	Range
s_1	1/600	±10%
s_2	1/150	±10%

Table 4.3: Distribution of the shape parameters of the linear+exponential fit

Abrupt faults

The abrupt faults are picked up from the library summarised in table 4.4. Each fault type impacts only one entity which can be a component (FC's 1–5), a system (FC's 6–7) or a sensor (FC's 8–18). The no–fault case (FC 0) completes the list.

Turbomachinery module faults involve alterations in both the efficiency and flow correcting factors. During the generation of a scenario, two related quantities, namely the magnitude f_m and the coupling factor f_c, are used. They are uniformly distributed in the intervals quoted in table 4.4. The fault magnitude f_m is defined as the Euclidian norm of the efficiency and flow *variations*. The coupling factor f_c is defined as the ratio between the change in flow capacity and the change in efficiency, see equations (4.1).

FC	Description	Magnitude f_m	Coupling f_c
0	no fault	n/a	n/a
1	fan fault	1 to 5 %	+ (0.5 to 2.0)
2	lpc fault	1 to 5 %	+ (0.5 to 2.0)
3	hpc fault	1 to 5 %	+ (0.5 to 2.0)
4	hpt fault	1 to 5 %	− (0.5 to 2.0)
5	lpt fault	1 to 5 %	− (0.5 to 2.0)
6	vbv fault	1 to 5	n/a
7	vsv fault	1 to 5	n/a
8	P13 bias	± 1 to 5 σ	n/a
9	T26 bias	± 1 to 5 σ	n/a
10	P3 bias	± 1 to 5 σ	n/a
11	T3 bias	± 1 to 5 σ	n/a
12	NF bias	± 1 to 5 σ	n/a
13	NC bias	± 1 to 5 σ	n/a
14	T49 bias	± 1 to 5 σ	n/a
15	WF bias	± 1 to 5 σ	n/a
16	Pamb bias	± 1 to 5 σ	n/a
17	P2 bias	± 1 to 5 σ	n/a
18	T2 bias	± 1 to 5 σ	n/a

Table 4.4: Library of abrupt faults – FC stands for fault code

$$\begin{cases} f_m \triangleq \sqrt{\Delta\mathsf{SEi}^2 + \Delta\mathsf{SWi}^2} \\ f_c \triangleq \dfrac{\Delta\mathsf{SWi}}{\Delta\mathsf{SEi}} \end{cases} \quad \Rightarrow \quad \begin{cases} \Delta\mathsf{SEi} = \dfrac{f_m}{\sqrt{1+f_c^2}} \\ \Delta\mathsf{SWi} = f_c \cdot \Delta\mathsf{SEi} \end{cases} \tag{4.1}$$

Once randomly drawn, f_m and f_c are converted back to deviations in the health parameters according to equations (4.1). For compressors ΔSWi and ΔSEi have the same sign, while for turbines they are of opposite sign. This point of view is shared by several authors see *e.g.*, [Ogaji et al., 2002] and [Volponi et al., 2003].

The vbv and vsv system faults are implemented as true off–schedule deviations. The uniformly distributed magnitude for these fault types is reported here as some kind of severity index, for sake of simplicity. A unit value corresponds to a small modification with respect to the nominal setting (*e.g.,* only a slight mistuning of the vsv), while a value of 5 hints at a deep malfunction (*e.g.*, fully open vbv).

Instrumentation faults are modelled as biased readings from one sensor, either in the flow path or for the operating conditions. The magnitude, expressed in table 4.4 in units of sensor standard deviation σ, and the sign of the bias are both randomly selected.

4.3.3 Sensed parameter histories

The database of sensed parameters, processed by the diagnosis tool, contains engine data collected once per flight at cruise conditions. Data acquisition and management comes at a certain cost and the historical trend has been to record a snapshot of sensor readings *i.e.*, one single, instantaneous view of these quantities. With the continuous improvement in computing power, data transmission and storage capability, more advanced algorithms are used to identify steady-state operation and to extract averaged values over a window of time from the continuous data stream [Simon and Litt, 2010]. This greatly enhances the quality of the archived data. In an attempt to mimic such a system, we generate the snapshots in the following way:

1. select a random operating condition from the distribution specified in table 4.1 and read the deterioration relative to the current flight,
2. run the engine model to generate 30 data samples for these inputs, this number corresponds to a recording window of 3 seconds at a sample frequency of 10 Hz, which are common values [see Simon and Simon, 2003],
3. add Gaussian noise, whose magnitude is specified in table 4.5, to the noise–free simulated measurements. In case of a sensor fault, also add the bias to the faulty sensor (see figure 4.2),
4. average the readings and store them in the database. This last step provides a first reduction in the noise level of the measurements.

The sensor suite $\mathbf{y}$ selected to perform the diagnosis is similar to the instrumentation installed on contemporary turbofan engines and is detailed in table 4.5 where the nominal accuracy of each sensor is also quoted. These values are adapted from [Curnock, 2000] and [Roth et al., 2005]. The table is complemented with the on–board sensors $\mathbf{u}_{\mathsf{ob}}$ that define the operating conditions of the engine.

Label	Description	Uncertainty
P13	fan outlet total pressure	±150 Pa
T26	lpc outlet total temperature	±2 K
P3	hpc outlet total pressure	±5000 Pa
T3	hpc outlet total temperature	±4 K
NF	fan speed	±3 rpm
NC	core speed	±6 rpm
T49	lpt inlet total temperature	±8 K
WF	fuel flow	±5 g/s
Pamb	ambient pressure	±100 Pa
P2	fan inlet total pressure	±100 Pa
T2	fan inlet total temperature	±2 K

Table 4.5: Gas-path and operating point instrumentation

4.4 Selected metrics for evaluation

Implementation of a condition–based maintenance strategy depends on the expected savings that it will generate. The quantification of these savings, preferably expressed in terms of currency units, goes through the definition of metrics that can be broadly divided into two groups [Orsagh et al., 2002]: performance metrics and effectiveness metrics. On one hand, effectiveness metrics assess the complexity of the algorithm (*e.g.*, number of tuning parameters, length of the code) and the cost to deploy and maintain it [Sambasivan et al., 2007]. Such metrics are to a large extent operator specific. On the other hand, performance metrics quantify mainly the detection and isolation capability, as well as their sensitivity with respect to signal–to–noise ratio and tuning parameters [Depold et al., 2006, Butler et al., 2006]. The mapping between these metrics, such as false alarm rate or misclassifications, and their financial impact is again quite application dependent [DePold et al., 2004]. Following the review by Davison and Bird [2008] we focus here on a few performance metrics that are relevant to gauge our diagnosis tool.

4.4.1 Metrics for estimation

In addition to performing fault detection and isolation, our diagnosis tool provides also an estimate of the engine health. To assess the accuracy of this estimate, we rely on the well–known *Root Mean Square (estimation) Error* as defined in equation (4.2) for the ith health parameter. This scalar is an image of the average distance between the estimated and true health condition and captures both the bias and the variance components of the estimation error.

$$RMSE_i = \sqrt{\frac{1}{N_{scn}\ N_{fl}} \sum_{j=1}^{N_{scn}} \sum_{k=1}^{N_{fl}} \left(\widehat{\mathbf{h}}_k^j(i) - \mathbf{h}_k^j(i) \right)^2} \tag{4.2}$$

where

- N_{scn} is the number of scenarios over which the RMSE is evaluated,
- N_{fl} is the length of a scenario *i.e.*, $N_{fl} = 3000$ here,
- $\widehat{\mathbf{h}}_k^j(i)$ is the estimated value of the ith parameter at flight k of scenario j,
- $\mathbf{h}_k^j(i)$ is the true value of the ith parameter at flight k of scenario j.

We use the RMSE metric to gauge the tracking performance of the diagnosis tool. We evaluate it only on no–fault scenarios (FC 0), since possible detection delays, missed detections or misclassifications that might occur in faulty scenarios are likely to blur the picture.

4.4.2 Metrics for detection

We have selected two metrics to assess the detection capability of the diagnosis tool. The first metric is the *Detection Decision Matrix* (DDM). As shown in table 4.6, it is a square matrix of dimension two. The elements on the main diagonal reflect correct predictions (the predicted and true states are the same). False negatives are cases where an existing fault is not detected by the algorithm. For this reason, they are also called *missed detections*. False positives are cases where the algorithm detects erroneously a non existing fault. This situation is known as a *false alarm*. It is obviously desirable to maintain the number of missed detections as low as possible, but it is evenly desirable to achieve the lowest possible false alarm rate. A too high false alarm rate would indeed reduce the user's trust in the algorithm, which could ultimately cause actual faults to be ignored.

		Predicted state	
		Fault	No fault
True state	Fault	True Positives	False Negatives
	No fault	False Positives	True Negatives

Table 4.6: Sample detection decision matrix (DDM)

As by–products of this matrix, the True Positive Rate (TPR) and False Positive Rate (FPR) are given by

$$\text{TPR} = \frac{\text{TP}}{\text{TP+FN}} \quad \text{and} \quad \text{FPR} = \frac{\text{FP}}{\text{FP+TN}} \tag{4.3}$$

The second metric is the *Detection Delay*, which is defined as the time required to detect a fault after its initiation. It is desirable to have a minimum detection delay in order to take corrective actions as soon as possible.

4.4.3 Metrics for isolation

We have also selected two metrics to assess the isolation capability of the diagnosis tool. The first metric is the *Classification Confusion Matrix* (CCM). It is a square matrix, whose dimension is equal to the number of fault types N_{fc}. In the most general form, the no–fault type can be included as well. It is not the case here as we emphasise on fault isolation, fault detection performance being assessed with the metrics mentioned above. The elements on the main diagonal of the CCM reflect correct classifications. Each row gives an overview of how the true occurrences of a given fault type (*e.g.*, lpc) are affected to the various fault types by the algorithm. As a by–product of this matrix, the *Percent Correctly Classified* (PCC) for a given fault type is defined as the ratio of the number of correct classifications to the total number of occurrences for the said type.

		Predicted state			
		Fault 1	Fault 2	...	Fault N_{fc}
True state	Fault 1	$\mathbf{CCM_{1,1}}$	$\mathsf{CCM}_{1,2}$	...	$\mathsf{CCM}_{1,N_{fc}}$
	Fault 2	$\mathsf{CCM}_{2,1}$	$\mathbf{CCM_{2,2}}$	...	$\mathsf{CCM}_{2,N_{fc}}$
	⋮	⋮	⋮	⋱	⋮
	Fault N_{fc}	$\mathsf{CCM}_{N_{fc},1}$	$\mathsf{CCM}_{N_{fc},2}$	...	$\mathbf{CCM_{N_{fc},N_{fc}}}$

Table 4.7: Sample classification confusion matrix (CCM)

The second metric is the *Kappa Coefficient* κ, defined in equation (4.4). It conveniently summarises the content of the confusion matrix into a single scalar and can be interpreted as a measure of an algorithm's ability to correctly classify a fault, which takes into account the expected number of correct classifications occurring by chance, i.e., if the classifications were totally random. The upper bound on κ is one, which means that the algorithm achieves perfect classification.

$$\kappa = \frac{N_{cc} - N_{ec}}{N_{tot} - N_{ec}} \tag{4.4}$$

where

- N_{cc} is the number of correctly classified cases,
- N_{tot} is the total number of cases,
- N_{ec} is the number of cases expected correct by chance,

$$\begin{aligned} N_{cc} &= \sum_{i=1}^{N_{fc}} \mathsf{CCM}_{i,i} \qquad N_{tot} = \sum_{i=1}^{N_{fc}} \sum_{j=1}^{N_{fc}} \mathsf{CCM}_{i,j} \\ N_{ec} &= \sum_{i=1}^{N_{fc}} \left\{ \left(\sum_{j=1}^{N_{fc}} \frac{\mathsf{CCM}_{i,j}}{N_{tot}} \right) \cdot \sum_{j=1}^{N_{fc}} \mathsf{CCM}_{i,j} \right\} \end{aligned}$$

4.5 Results and discussion

We have generated 100 scenarios for each of the 18 fault classes listed in table 4.4. Such a number allows a fair coverage of the fault pattern for module faults (both in magnitude and coupling factor), as well as for system and sensor faults. Moreover, we have created a complementary batch of 1800 no–fault scenarios (FC 0) in order to process an equal number of faulty and non–faulty scenarios. We remind the reader that we do not consider gradual deterioration as a faulty condition due to its continuous nature.

We configure our diagnosis tool such that the Kalman filter estimates only the changes in the efficiency SEi and flow SWi correcting factors of the turbomachinery components. Indeed, the main purpose of the Kalman filter is to track the gradual deterioration of the engine. Neither the systems nor the sensors are expected to exhibit this long–time–scale degradation trend.

4.5.1 Comparison of various Kalman filters

We begin the analysis of the results with the comparison of the estimation accuracy achieved by different flavours of the Kalman filter. More specifically, we compare the semi-linear Kalman filter with the extended and unscented Kalman filters presented in chapter 3. The semi-linear Kalman filter (SLKF) is a simplified version of the extended Kalman filter in which the influence coefficient matrix is updated each fixed number of flights (as specified by the user), around an average operating condition and the latest estimate of the health condition. In the present study, we consider four different SLKFs, whose update rate of the Jacobian matrix is listed in table 4.8.

Name	Jacobian updating
SLKF-1	never
SLKF-2	every 1000 flights
SLKF-3	every 100 flights
SLKF-4	every 10 flights

Table 4.8: Tested configurations of the semi-linear Kalman filter (SLKF)

On a side note, we evaluate the Jacobian matrix through a centred–difference scheme, hence the EKF makes the same number of calls to the non–linear engine model as the UKF at each time step.

The health parameter distribution is initialised with the mean and standard deviation vectors given in table 4.2. The health parameters of a brand new engine are assumed independent to reflect the randomness in the manufacturing tolerances. The covariance matrix of the process noise is set–up based on the average final deterioration reported in table 4.2 and integrates coupling factors between the health parameters of a given module, as explained in section 2.4.2.

Table 4.9 reports the RMSE, as defined in equation (4.2), evaluated over the batch of 1800 no–fault scenarios and expressed in percent of the average baseline value (equal to one here) for each of the ten module health parameters and the six filter configurations. The next–to–last row simply sums up the previous ones. The last row gives the mean computing time to process one scenario *i.e.*, 3000 flights. Note that more than the absolute figures, which are highly dependent on the computer platform, the relative increase in computing time from one configuration to another one is the main information to extract from this last row.

RMSE (%)	SLKF-1	SLKF-2	SLKF-3	SLKF-4	EKF	UKF
SW12	0.133	0.127	0.123	0.123	0.035	0.035
SE12	0.263	0.255	0.250	0.252	0.043	0.043
SW2	0.025	0.032	0.036	0.037	0.039	0.039
SE2	0.217	0.216	0.215	0.216	0.164	0.162
SW26	0.218	0.213	0.210	0.211	0.168	0.166
SE26	0.123	0.120	0.119	0.119	0.094	0.093
SW41	0.020	0.020	0.020	0.020	0.020	0.020
SE41	0.085	0.084	0.084	0.083	0.066	0.066
SW49	0.139	0.138	0.137	0.137	0.108	0.108
SE49	0.167	0.164	0.161	0.163	0.081	0.081
Sum	**1.390**	**1.370**	**1.355**	**1.360**	**0.817**	**0.813**
TCPU (s)	2.76	4.77	16.88	33.88	187.62	178.33

Table 4.9: Performance of different filters in terms of percent RMSE and mean computing time per scenario

The analysis of table 4.9 reveals that the basic SLKF without any Jacobian update *i.e.*, the SLKF-1 configuration is a viable solution for trending the engine health fairly accurately at a really modest computational cost. We note a slight improvement in the SLKF's estimation accuracy as the Jacobian update rate increases. Turning to the fully non-linear filters leads to a reduction of roughly 70% in the global RMSE with respect to the SLKF-1 configuration. The EKF and UKF achieve virtually the same RMSE performance, which we can explain by the mild non–linearity level of the engine performance model. Unlike the UKF, the EKF provides readily the influence coefficient matrix which is used abundantly in the fault detection and isolation modules. For this reason, and given the comparable level of accuracy, it seems wiser to prefer the EKF over the UKF. From the last line of table 4.9, we note a continuous increase in computing time as the number of Jacobian updates increases. The EKF and UKF have a similar computing time due to the fact that the Jacobian matrix is evaluated with a centred–difference scheme. They are however about two orders of magnitude slower than the SLKF-1 configuration. The above comments are in line with the findings of other researchers who compared different Kalman filters in various applications such as motion tracking [Orderud, 2005], estimation of aircraft aerodynamic parameters [Chowdhary and Jategaonkar, 2010] or, interestingly, jet engine performance trending [Simon, 2008].

Figure 4.4 depicts the tracking of gradual deterioration for one of the 1800 no–fault scenario processed with the SLKF-1 algorithm. The subplots show the true (black line) and estimated (orange line) health parameters, expressed in terms of percent deviation with respect to a reference value (equal to one for all parameters). On the abscissa, time is expressed in terms of flights. The non–zero deviation at flight zero is due to the engine–to–engine variations in manufacturing. On the qualitative side, we can state that the trend in each health parameter is well captured. In particular, the estimated values look in fairly good agreement with the true ones for the lpc, hpc and hpt.

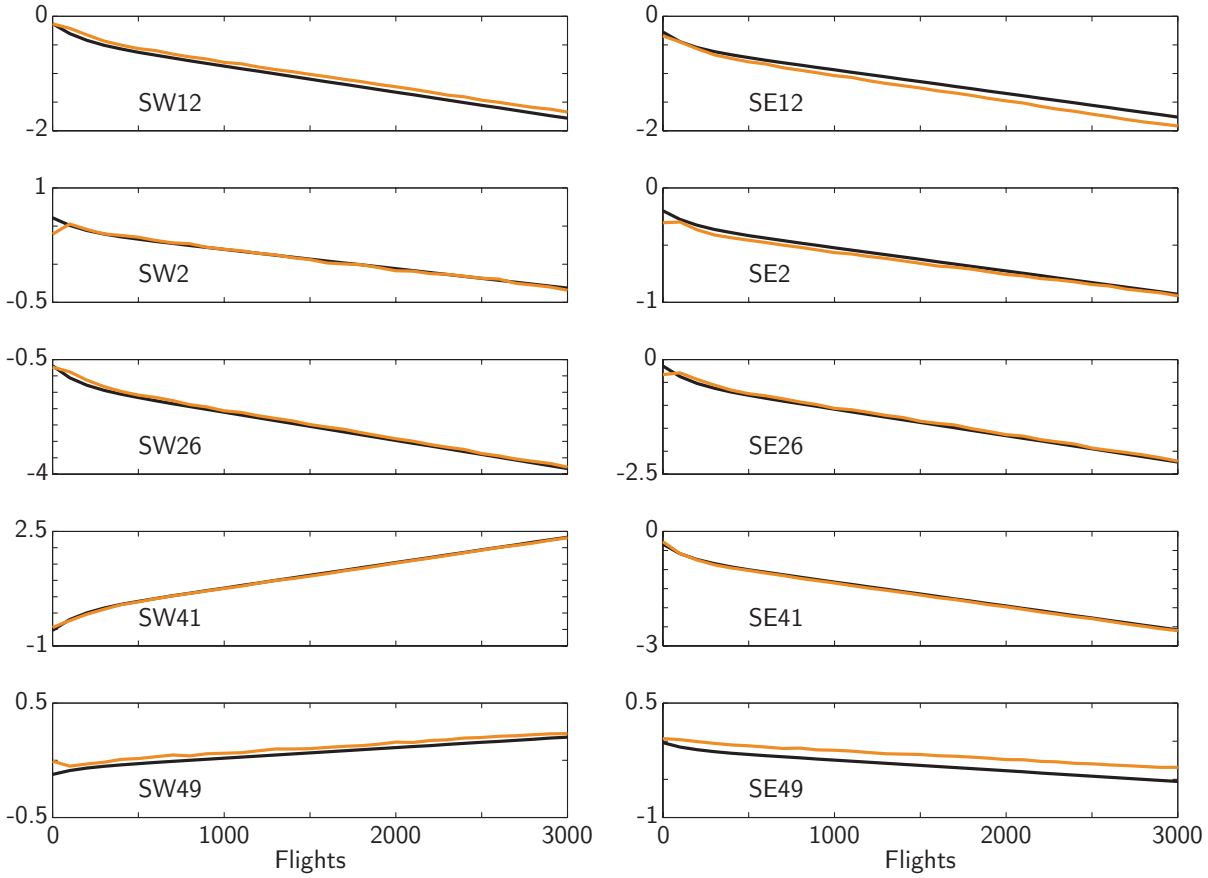


Figure 4.4: True (black line) and estimated (orange line) health parameters of a non–faulty scenario – Health parameters are expressed as percent deviation from the reference value – SLKF-1 algorithm

Figure 4.5 complements nicely the previous one by allowing a more quantitative assessment. Indeed, each subplot shows the evolution of the absolute value of the estimation error (orange line) and the 3σ confidence bound (black line). This threshold is computed here as three times the square root of the diagonal terms of the covariance matrix of the health parameters. Both quantities are expressed in percent. On the abscissa, time is again expressed in terms of flights. The estimation error is globally smaller than 0.1% for all health parameters except the fan and lpt efficiencies (resp. SE12 and SE49). The observability analysis carried out in section 4.6.3 will provide an explanation to this fact. This accuracy level is fairly acceptable except for the lpt whose deterioration is, at about 0.5%, much lower than for the other components. The estimation error is moreover always below the 3σ confidence bound *i.e.*, the estimation error is inside the uncertainty region caused by the measurement noise and the negative redundancy. Therefore, we can say the solution is statistically meaningful. We observe a continuous increase in the 3σ confidence bound for the health parameters of the fan, hpc and lpt. This increase is due on one hand to the process noise needed to track the time evolving degradation and on the other hand to the underdetermined nature of the estimation problem that makes some health parameters only partially observable, as will be shown in section 4.6.3.

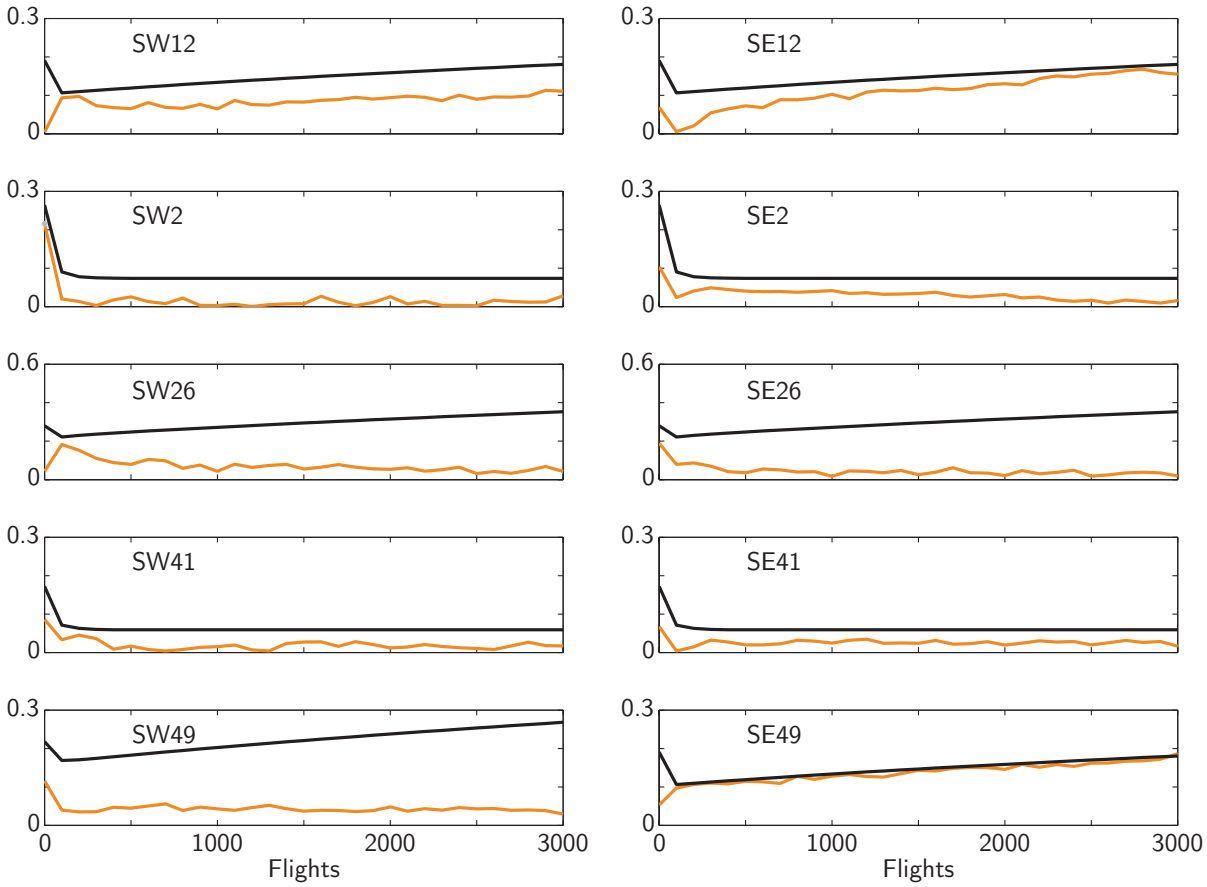


Figure 4.5: Absolute estimation error (orange line) and 3σ confidence bound (black line) for the scenario plotted in figure 4.4 – Values expressed in percent – SLKF-1 algorithm

4.5.2 Fault detection performance

In this section, we focus on the fault detection capability of the diagnosis tool. First of all, we study the influence of the two tuning parameters of the detection module which are i) the length of the sliding window N_{wd} over which the GLR is computed and ii) the detection threshold which is indirectly set through the expected probability of false alarm P_F, see equation (3.48). We consider three window lengths of 5, 10 and 20 flights (one flight being here the time unit) and five threshold values corresponding to target probabilities of false alarm of 10^{-1}, 10^{-2}, 10^{-3}, 10^{-4} and 10^{-5}.

For each of the fifteen configurations, we process the database of 3600 scenarios and compute the TPR and FPR metrics. We have arbitrarily decided to evaluate the FPR aka. the false alarm rate only from the no–fault scenarios. Any false alarm that might occur on a faulty scenario is not accounted for in the FPR. Doing so, we implicitly assume that the false alarm rate is the same for faulty scenarios as for no–fault ones. The TPR aka. the correct detection rate is obviously evaluated from the faulty scenarios. Nonetheless, we require the detection delay to be less than ten flights to increment the TPR counter.

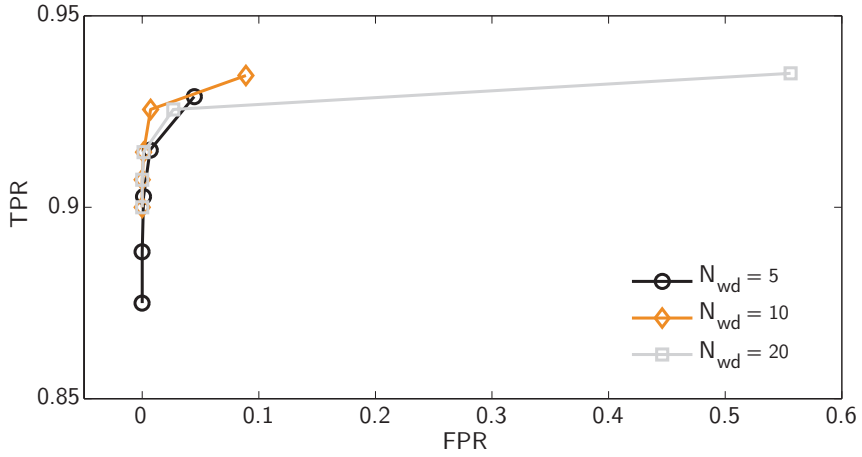


Figure 4.6: Partial receiver operating characteristic curves of the detection module for three different window length N_{wd}

Figure 4.6 depicts portions of the so–called receiver operating characteristic (ROC) curves for each of the three window lengths. These curves show the evolution of the TPR versus the FPR as the detection threshold is varied. The axes are normalised so that the FPR and TPR range between zero and one, we can therefore refer to them respectively as the probability of false alarm and the probability of (correct) detection.

The curves in figure 4.6 have the typical shape of ROC curves see *e.g.*, [DePold et al., 2004]: a decrease in the detection threshold causes a beneficial increase in the TPR and a simultaneous (but undesired) increase in the FPR. The worst TPR over the fifteen examined configurations is achieved for the couple $(N_{wd}, P_F) = (5, 10^{-5})$ and reaches a rather encouraging value of roughly 87%. On the other hand, the best TPR tops the 93% mark for the couple $(N_{wd}, P_F) = (10, 10^{-1})$ but this comes with a false alarm rate of about 10%. These relatively high values of TPR are due to a favourable signal to noise ratio of most of the faults as well as to their abrupt occurrence. Considering the FPR, table 4.10 shows that the computed values are in fairly good agreement with the P_F values that set indirectly the detection threshold.

P_F (%)	10^{-3}	10^{-2}	0.1	1.0	10.0
FPR (%)	0.0	0.0	0.1	0.7	8.9

Table 4.10: Comparison of expected and achieved false alarm rates – $N_{wd} = 10$

Figure 4.6 also shows that, for a fixed value of the threshold, an increase in the length of the sliding window N_{wd} leads to an increase in both TPR and FPR. This can be explained by the fact that a wider window enhances the resolution of the detection module which therefore allows more faults to be detected. However, the enhanced resolution applies also to outliers hence the increase in FPR. We further note that TPRs are almost unaltered when moving from a 10– to a 20–flight window width. We suspect this is caused by the fact that the additional correct detections actually occur after a delay larger than the

10–flight threshold discussed previously. As a result, those additional detections do not contribute to the TPR.

On the basis of figure 4.6, we select the settings $(N_{wd},P_F) = (10,10^{-3})$ for a deeper analysis of the results. This configuration appears as a good trade–off in terms of TPR/FPR as well as in terms of computational load. The detection decision matrix in table 4.11 summarises the global detection performance of the diagnosis algorithm and is a good quantitative complement to the corresponding point on the ROC curve. A mere two false alarms were issued over the 1800 no–fault scenarios, leading to a false positive rate of 0.1% which we consider acceptable. Of course, other users might have more stringent requirement on the FPR, which could be achieved at the price of a lower TPR as shown by the ROC curves. Looking at the first row of the DDM, the number of missed detections amounts to 154 out of 1800 faulty scenarios. This translates in a score of 91.4% for the true positive rate. The next step is to analyse the results for each fault type separately.

		Predicted state		
		Fault	No fault	
True	Fault	**1646**	154	TPR = 91.4%
state	No fault	2	**1798**	FPR = 0.1%

Table 4.11: Detection decision matrix

Table 4.12 provides an overview of the detection capability for the different fault types. The third column gives the percentage of detected events for the said type. The mean detection delay is reported in the fourth column. This average value is obviously computed from the detection delay, as defined in section 4.4, of the detected cases. Finally, the fifth column gives the so-called "Span", defined as the difference between the maximum and minimum detection delays for a specific fault type. This quantity provides somehow a measure of the variability in the detection delay.

The PCD reaches 100% for all module faults (FC's 1–5), meaning that for the range of magnitudes and coupling factors considered in this study, any abrupt deterioration of a component will be captured by the detection module. Detection of system faults (FC's 6–7) is also successful over their whole range of severity index as confirmed by the PCDs of 100%. The smal–valued MDDs and "Spans" hint at a rapid detection of the fault, actually one or two flights after its initiation for all module and system faults. As stated above, this level of performance results from the combination of a good signal to noise ratio of all these fault types and of their abrupt occurrence.

Figure 4.7 shows the norm of the residuals, scaled by each sensor's standard deviation, for the lowest fault level considered in the study *i.e.*, a 1% magnitude for module faults and a unit severity index for system faults. We see that lpc and vbv faults have the smallest overall signature on the gas–path measurements, which provides the explanation to the fact that a handful of such faults are detected two flights after their initiation. On the other hand, hpt and vsv faults cause the largest deviations in the gas–path measurements and are accordingly the easiest to detect.

FC	Description	PCD	MDD	Span
1	fan fault	100	1.00	0
2	lpc fault	100	1.05	1
3	hpc fault	100	1.00	0
4	hpt fault	100	1.00	0
5	lpt fault	100	1.00	0
6	vbv fault	100	1.01	1
7	vsv fault	100	1.00	0
8	P13 bias	98	2.07	7
9	T26 bias	100	1.50	4
10	P3 bias	100	1.78	5
11	T3 bias	100	1.62	3
12	NF bias	82	3.17	9
13	NC bias	70	2.53	8
14	T49 bias	100	1.51	4
15	WF bias	100	1.50	3
16	Pamb bias	**0**	n/a	n/a
17	P2 bias	96	2.34	9
18	T2 bias	100	1.60	5

PCD: Percent Correctly Detected
MDD: Mean Detection Delay
Span: Difference between minimum and maximum detection delays

Table 4.12: Detection performance for the considered fault types

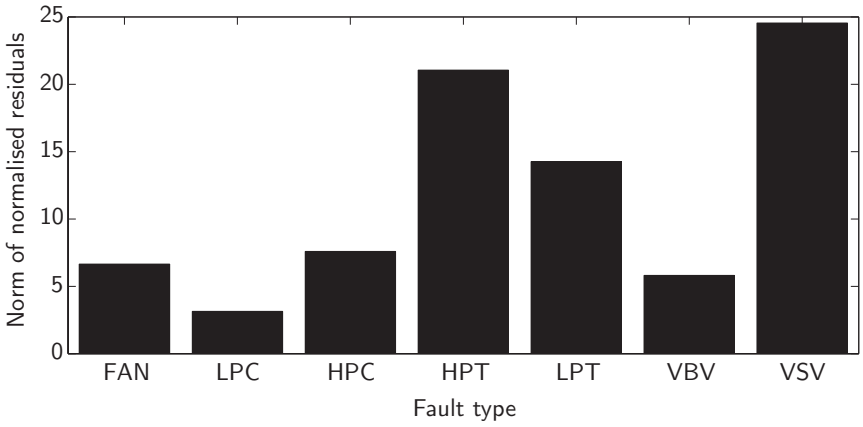


Figure 4.7: Norm of residuals for component and system faults of lowest magnitude

Back to the analysis of table 4.12, the detection of gas-path instrumentation faults (FC's 8–14) is again satisfactory as far as PCD is concerned. Surprisingly, the PCD associated to low– and high– spool speed faults reaches a mere 70% and 82% respectively while the tachometers are among the most precise sensors (see table 4.5). However, we should not forget that the diagnosis tool uses a covariance matrix for the measurement noise that also accounts for the uncertainty in the sensed operating conditions. Figure 4.8 shows that both spool speeds are the most sensitive to uncertainty in the operating conditions therefore their effective standard deviation, as given in equation (2.24), is larger than the one of the "bare" sensor. As a result, tachometer biases of low magnitudes are less detectable, leading to lower PCD values.

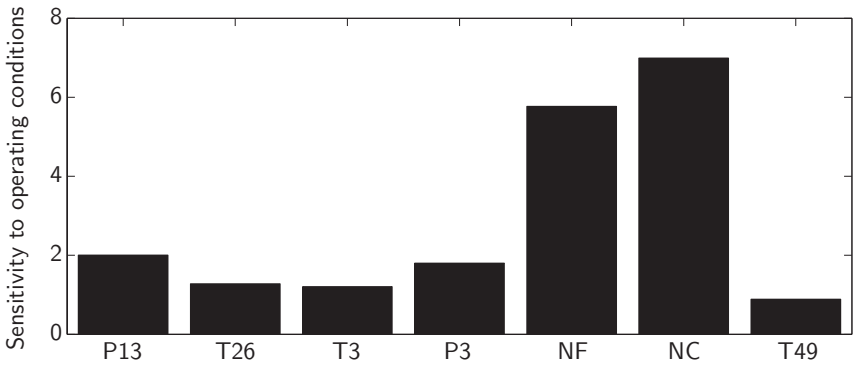


Figure 4.8: Sensitivity of the gas–path measurements to uncertainty in the operating conditions

Looking at the two other metrics for the gas–path sensor biases, the MDDs rise up to 3.17 flights while at the same time the "Spans" range between three and nine flights. As a conclusion, the detection delay of a gas–path sensor fault is dependent not only on its magnitude, but also on the faulty sensor itself. We suspect that the variability in detection delay finds its roots in the very formulation of the detection module. In fact, taking a look at equation (3.42), we see that the likelihood ratio is evaluated in the space of the health parameters tracked by the Kalman filter. It means that before being detected, any fault, and in particular a sensor fault, is first converted inside the algorithm into a change in the module health parameters SWi and SEi. This transformation may modify the signal to noise ratio of the original fault and therefore increase the detection delay.

With the notable exception of a biased Pamb sensor, detection of operating condition instrumentation faults (FC's 15–18) is of similar quality as for gas–path sensor faults. Again the detection delay depends to some extent on the magnitude of the fault for the same reason as gas–path sensor faults do. More intriguing is the case of Pamb. For the 15 configurations of the detection module considered in the present study, the PCD of that fault type is equal to zero. Figure 4.9 provides the explanation of such a miserable score at a glance. It shows the evolution of the norm of the residuals, scaled by the sensor's standard deviation, with respect to the sensor fault magnitude for Pamb (black line) and for P2 (orange line). We clearly see that a faulty Pamb has no influence on the gas path measurements for the range of sensor magnitudes considered here. The reason behind

this is that the nozzle is choked at the cruise regimes considered here. The Mach number in the exhaust section is hence equal to one, meaning that the engine *simulation model* is unaffected by downstream conditions *i.e.*, Pamb (see [Cumpsty, 2000] for further details). One solution to detect Pamb faults at such operating points would be to rely on a thrust measurement, which is unfortunately unavailable in flight.

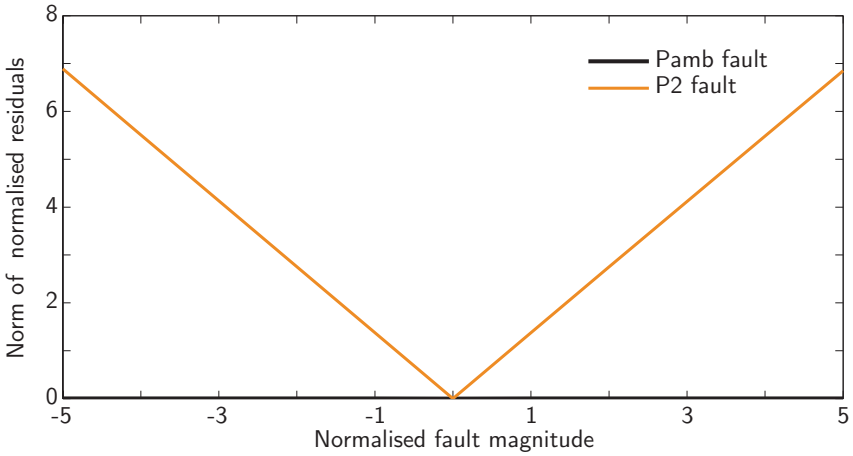


Figure 4.9: Norm of normalised residuals for Pamb (black line) and P2 (orange line) sensor faults – magnitudes expressed in number of standard deviations

To conclude the analysis of the detection capability, table 4.13 reports the minimum level of sensor bias (negative ones in the third column, positive ones in the fourth column) that was successfully detected by the algorithm. The values quoted in the table are normalised with the respective standard deviation of the sensor and are obtained from the processing of the batch of scenarios previously defined.

Sensor	PCC	Negative bias	Positive bias
P13	98	-1.22	1.09
T26	100	-1.05	1.08
P3	100	-1.15	1.04
T3	100	-1.00	1.00
NF	**82**	**-1.95**	**1.71**
NC	**70**	**-2.23**	**2.15**
T49	100	-1.03	1.18
WF	100	-1.12	1.02
Pamb	**0**	n/a	n/a
P2	96	-1.25	1.03
T2	100	-1.04	1.02

Table 4.13: Minimum levels of detectable sensor bias (expressed in number of standard deviations)

Sensor faults for which a 100% PCD is achieved show bounds very close to one standard deviation (in absolute value), which is the minimum fault magnitude considered in the study. The values reported in table 4.13 are actually the lowest magnitudes for each fault type in the processed batch of scenarios. On the other hand, sensors for which missed detections occur (P13, NF, NC and P2) have a negative bound larger (in absolute value) than the positive one (*e.g.*, -1.25σ versus 1.03σ for a biased P2). Tachometer biases should at least reach a magnitude of around two standard deviations *i.e.*, 2 rpm for fan speed and 4 rpm for compressor speed, to be detectable by the algorithm.

4.5.3 Fault isolation performance

In this section, we focus on the performance of the fault isolation module *i.e.*, the combination of the sparse estimator *and* the isolation logic described in section 3.5. We begin with a parametric study of the influence of the regularisation parameter λ in (3.60) on global performance factors of the isolation module. We process the batch of 1800 faulty scenarios (even if we know that Pamb faults are not detectable) with the diagnosis tool for four different values of the parameter λ: 10^{-4}, 10^{-1}, 1.0 and 10.0. On the basis of the CCM, we evaluate i) the overall percentage of correctly classified cases, defined as the ratio of the total number of classifications to the total number of examined cases, ii) the number of non–zero off–diagonal terms in the CCM, which provides a direct measure of the randomness in the classification process and iii) the kappa coefficient. Going back to its definition (4.4), the kappa coefficient is somehow related to the other two metrics.

λ	PCC (%)	$N_{n,o}$	κ
10^{-4}	88.1	50	0.873
10^{-1}	92.6	46	0.921
1.0	**93.6**	**41**	**0.931**
10.0	85.4	44	0.845

Table 4.14: Influence of the parameter λ on the performance of the isolation module

These three metrics, respectively abbreviated PCC, $N_{n,o}$ and κ, are reported in table 4.14 for the four λ values mentioned above. Setting λ to one leads to the best performance. In this case, the overall PCC reaches an encouraging value of 93.6%, which means that 1540 fault cases are correctly classified out of the 1646 scenarios that are declared faulty by the diagnosis tool. The number of non–zero off–diagonal terms $N_{n,o}$ equals 41, so roughly fifteen percent of the non–diagonal terms are different from zero. Choosing $\lambda = 0.1$ causes a slight drop (about 1 point) in the global PCC and κ, as well as an increase in $N_{n,o}$.

Selecting a much smaller value for λ, here 10^{-4}, we note a stronger deterioration in isolation performance: the PCC goes down to 88.1% while $N_{n,o}$ rises to 50. We can explain this behaviour by the fact that the lower the λ value, the weaker the sparsity constraint in the estimator. As a result, computed solutions contain spurious changes in

several health parameters. Such spoiled estimates may then fool the isolation logic. Large λ values have the same detrimental effect on the isolation capability. In fact, for $\lambda = 10.0$ we achieve a PCC of 85% and 44 off–diagonal terms are non–zero. This time, the loss in PCC is due to a too important weight of the sparsity constraint which reduces both the number of non–zero terms in the sparse solution and their magnitudes. This translates directly in misclassification of low–magnitude faults.

Table 4.15 reports the classification confusion matrix for $\lambda = 1.0$. The true and predicted fault classes are respectively on the vertical and horizontal axes, as in the sample CCM shown in table 4.7. For sake of completeness, we kept the Pamb fault class in the matrix, although that fault type is not detectable. The percentage of correctly classified (PCC) cases for each fault type is quoted in the last column of the table. The horizontal and vertical lines separate component/system faults from sensor faults. In an attempt to maintain good legibility, only the non–zero elements in the CCM are displayed.

From the PCC figures of table 4.15, the isolation capability of all component/system faults but the lpc and the hpc is nearly perfect, with PCCs in excess of 98%. The discrimination of lpc and hpc faults is somewhat worse, with PCCs of respectively 85% and 77%. The isolability of sensor faults is almost as good, with PCCs greater than 95% except for P13, NF, NC and P2. On a side note, those four sensors were already not perfectly detectable. Nonetheless, the worst PCC nearly tops the 80% mark, which is a decent performance. In conclusion, the sparse estimation approach allows a rather accurate isolation over a wide range of fault classes and magnitudes.

Looking in more details at the CCM in table 4.15, 14 percent of lpc faults are erroneously assigned to a vsv malfunction. This percentage rises to 23 in the case of hpc faults. On the other hand, one lonesome vsv fault is reported as a compressor module fault. In some way, those results are not so surprising. Indeed, the border is quite thin between the impact of a compressor module fault and a system fault, especially for the duet hpc–vsv as exposed below. Figure 4.10 depicts the situation for one of the erroneously isolated hpc faults. The health parameter values, on the vertical axis, are expressed as percent deviations from the condition before the occurrence of the fault for the efficiency and flow factors and as the severity index for the system faults. Orange bars are related to the solution provided by the sparse algorithm, black ones to the true fault.

We realise at a glance that the computed solution is in fact a sparse one as only two parameters out of 12 show non–zero deviations. If the alteration of the hpc efficiency SE26 is indeed captured with a reasonable accuracy, the change in hpc flow capacity SW26 is believed to be a vsv problem by the sparse algorithm. Given that the deviation on the vsv is larger than the combined one on the hpc, the isolation logic declares the vsv as the faulty entity. Figure 4.11 provides an explanation to the confusion made by the sparse algorithm between the hpc flow factor SW26 and the vsv parameter. It shows the variation in each measurement, quoted in number of sensor's standard deviation, for a 1% change in SW26 (black bars) and a vsv malfunction of severity 0.5 (orange bars). The value of the severity index for the vsv malfunction is chosen so that the maximal deviation is of comparable magnitude for both parameters.

	fan	lpc	hpc	hpt	lpt	vbv	vsv	P13	T26	P3	T3	NF	NC	T49	WF	Pamb	P2	T2	**PCC (%)**
fan	99											1				–			**99.0**
lpc		85				1	14									–			**85.0**
hpc			77				23									–			**77.0**
hpt				100												–			**100.0**
lpt					99							1				–			**99.0**
vbv						98			2							–			**98.0**
vsv			1				99									–			**99.0**
P13	4							90		1						–	3		**81.6**
T26		1			1				98							–			**98.0**
P3	2			1				1		95		1				–			**95.0**
T3			2		2						96					–			**96.0**
NF	1	1	1		1		1					65				–	12		**79.3**
NC			2				3		1		1	2	61			–			**87.1**
T49		1		1	1									97		–			**97.0**
WF	2		1												97	–			**97.0**
Pamb	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
P2		1		2		1	1	4								–	87		**90.6**
T2	2	1														–		97	**97.0**

Table 4.15: Classification confusion matrix and PCC

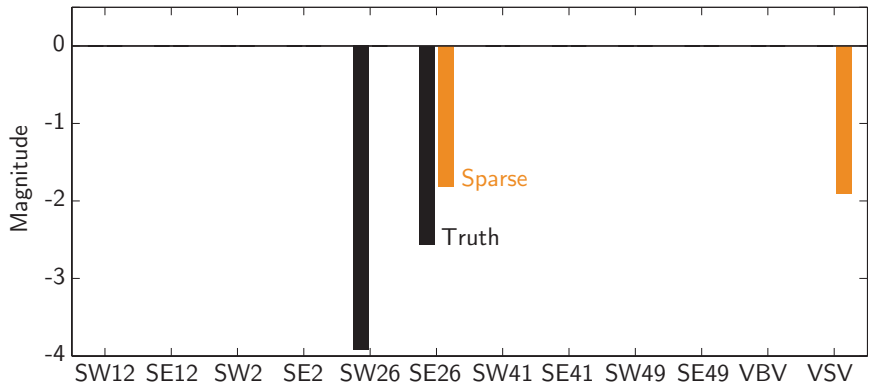


Figure 4.10: Sparse estimate of a misclassified hpc fault

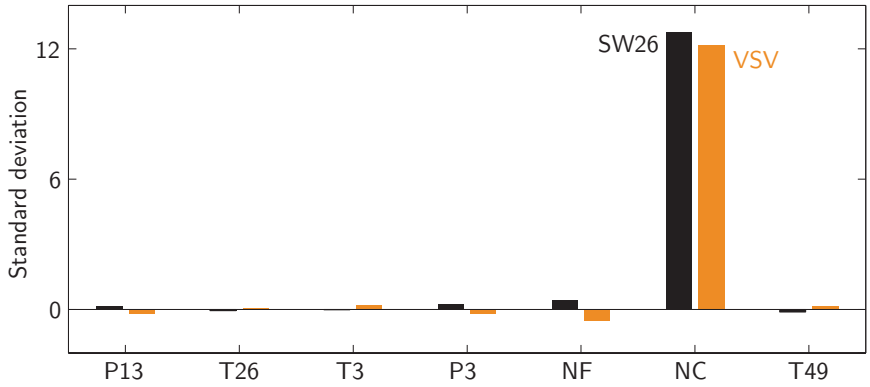


Figure 4.11: Signatures of SW26 and VSV on the sensors

From the graph, it is obvious that the signature of SW26 is highly similar to the one of vsv for that level of fault magnitude. In both cases, only one sensor reading, namely the core speed NC, undergoes a significant deviation. Consequently, it is very hard to distinguish between these two parameters in the case of faults of small magnitude. The true loss in SW26 is preferentially affected to vsv as the same deviation in the sensors is achieved for a fault severity of half the value it would have in the case of SW26. Part of the criterion of the sparse algorithm penalises in fact the l_1-norm of the parameter deviations.

As far as the isolation of sensor faults is concerned, we notice that the bottom left quadrant of the CCM is pretty densely populated, with not less than 25 non–zero terms. It means that the isolation module attributes most of the misclassified sensor faults are to component/system faults in a rather random manner. To complement the discussion, figure 4.12 summarises the classification of sensor faults. The abscissa is the true sensor fault magnitude, expressed in number of standard deviations. Grey dots indicate correctly isolated cases while black crosses show the misclassifications. We see that for all sensors but the quadruplet P13, NF, NC and P2, misclassifications occur for low fault magnitudes (typically smaller than 1.5σ). For the other four sensors, misclassifications can occur at larger fault magnitudes.

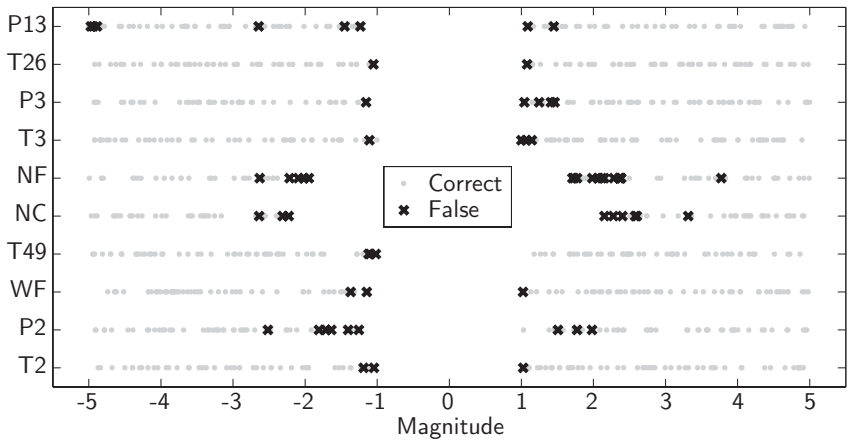


Figure 4.12: Graphical summary of sensor fault isolation

To conclude the analysis of the isolation capability, we compare the sparse estimation tool (3.60) to the classical least-squares estimator (3.55). The results presented below are extracted from [Borguet and Léonard, 2010]. This paper was oriented towards the fault isolation task only and the set-up was different in three aspects from the one considered so far: i) the fault library was limited to component and system faults (*i.e.,* FC 1-7 of table 4.12); ii) we used a slightly different set of seven gas-path sensors; iii) we processed 1000 occurrences for each of the seven fault types. Tables 4.16 and 4.17 report the classification confusion matrix for the sparse algorithm and for the usual least-squares method respectively. Both tables are organised in the same way as table 4.15.

	fan	lpc	hpc	hpt	lpt	vbv	vsv	**PCC (%)**
fan	1000							**100**
lpc		870					130	**87**
hpc		2	716				282	**71.6**
hpt				1000				**100**
lpt					1000			**100**
vbv						1000		**100**
vsv		1	7				992	**99.2**

Table 4.16: Classification confusion matrix and PCC – Sparse estimation

Proportionally, the numbers in table 4.16 are almost identical to their counter-parts in table 4.15, again underlining the quite good overall isolation capability of the sparse estimation tool. On the other hand, the usual least-squares method has a much worse isolation capability, as can be seen from table 4.17. Due to their good observability, fan and hpt faults enjoy a correct classification rate of 100 %. The PCCs for the other turbomachinery components are between ten and twenty percents lower than their corresponding values obtained with the sparse algorithm. The situation of system faults is even worse, with only roughly one occurrence over two correctly isolated. Moreover, we notice a much larger number of non-zero off-diagonals terms in table 4.17.

	fan	lpc	hpc	hpt	lpt	vbv	vsv	**PCC (%)**
fan	1000							**100**
lpc	40	688	161		111			**68.8**
hpc	4	284	602				110	**60.2**
hpt				1000				**100**
lpt	3			220	777			**77.7**
vbv	152	113		229		506		**50.6**
vsv	0	181	331				488	**48.8**

Table 4.17: Classification confusion matrix and PCC – Least-squares estimation

The kappa coefficients for the sparse tool and for the classical least-squares method have a value of 0.930 and 0.677 respectively. The lower value of the least-squares estimator comes from the lower number of total correct classifications (sum of the diagonal terms of the matrix), but also from the population of terms outside the main diagonal. Indeed, in a given column of the CCM, the larger the number of non-zero terms, the more random the fault is associated to a class by the algorithm. This effect is taken into account in the computation of the kappa coefficient.

Finally, figure 4.13 compares the solutions obtained with the traditional least-squares method (3.55), abbreviated LS2 below, and with the sparse algorithm for an lpc fault. The first thing to point out is the radically different pattern of the LS2 and sparse solutions. The sparse solution and the true fault have their non-zero components at the same locations and with the same signs, so that the fault is clearly concentrated on the lpc. The estimated magnitudes are slightly lower than the actual ones. This results from the effect of the measurement noise, but also illustrates the trade-off between fitting the data and achieving a sparse solution. The solution computed by equation (3.55) shows the typical spread (aka. smearing effect) over multiple parameters of a localised fault. Among all the parasitic deviations, the one on SW26 is of the same order of magnitude as the one on the two parameters of the lpc. As a consequence, a combined degradation of both the lpc and the hpc can be suspected.

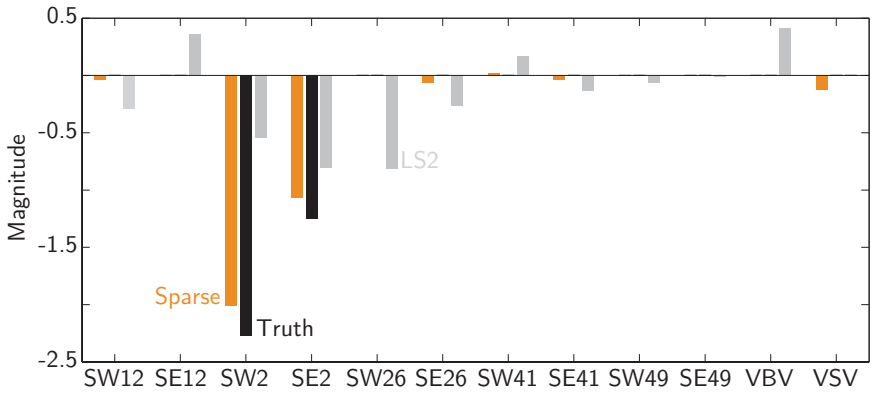


Figure 4.13: Comparison of the sparse and usual least-squares solutions for an lpc fault

4.6 Sensor set selection and analysis

In this section, we apply the tools developed in section 2.5 for sensor set selection and analysis. For illustrative purpose, we consider an onboard engine performance monitoring tool that shall track the deterioration of the turbomachinery components from cruise flight data recordings.

4.6.1 Optimisation of the sensor configuration

The sensor selection problem is formulated for the operating point specified in table 4.18, which is the centroid of the envelope mentioned in table 4.1. The set of health parameters is restricted to the ten modification factors on the efficiency and flow capacity of the rotating components.

Label	WF	ALT	MN	DT_{ISA}
Value	0.390 kg/s	10668 m	0.80	0.0 K

Table 4.18: Mean operating conditions for the optimal selection problem

Sensors that may be fitted on the engine are listed with their associated uncertainty (noise level) in table 4.19. We do not consider sensors that are not available in flight (*e.g.*, thrust, air mass flow) nor advanced instrumentation (*e.g.*, torque, hpt inlet temperature, emissions) that is still at the development level [Simon et al., 2004]. Therefore, the catalogue mostly consists of total temperatures and pressures at various inter–component planes as well as both spool speeds.

Label	Description	Uncertainty
P13	fan outlet total pressure	±150 Pa
T13	fan outlet total temperature	±2 K
P26	lpc outlet total pressure	±500 Pa
T26	lpc outlet total temperature	±2 K
P3	hpc outlet total pressure	±5000 Pa
T3	hpc outlet total temperature	±4 K
NF	fan speed	±3 rpm
NC	core speed	±6 rpm
P49	hpt outlet total pressure	±500 Pa
T49	hpt outlet total temperature	±8 K
P5	lpt outlet total pressure	±300 Pa
T5	lpt outlet total temperature	±5 K

Table 4.19: Available sensors for the optimal selection problem

In a nutshell, we shall determine the n_y gas path sensors, among the twelve listed above, that maximise the figure of merit of equation (2.88) for the set of $n_h = 10$ health parameters associated to the turbomachinery components. The number of sensors that are installed on the engine n_y is not strictly speaking a free design parameter as it should ideally be as small as possible. Nonetheless, a number of $n_y = 7$ sensors seems a reasonable choice for contemporary commercial turbofans according to the information reported in Jaw and Garg [2005].

The optimisation problem is basically of combinatorial nature as any sensor is either part or not of the optimal instrumentation. Although some algorithms are especially tailored to that kind of problem, we opt for a naive and brute force technique: we evaluate the figure of merit for every of the 792 (12 choose 7 binomial coefficient) possible combinations. We apply unit weights to each component of the FOM ($W_i = 1.0,\ i = 1, 2, 3$) in equation (2.88).

Table 4.20 reports the first ten optimal configurations based on the figure of merit defined in equation (2.88). The bottom row in the table gives the frequency of each sensor in the ten best sensor sets. To complete the picture, the first five optimal configurations according to each component of the FOM are given in tables 4.21, 4.22 and 4.23 (respectively the condition number, the trace and the determinant). The optimal configuration according to the full FOM (*i.e.*, first row of table 4.20) is ranked respectively fourth, first and second according to the condition number, the trace and the determinant criterion. Hence, the aggregated metric elegantly merges the different objectives of the optimisation problem.

Looking at the last row of each table, we deduce that T13, T26 and P5 are certainly no good candidate to the optimal instrumentation whatever the considered metric. On the contrary, P13, P26, P49 and the spool speeds appear to be mandatory in the optimised configuration. This point will be discussed in more details subsequently. As underlined by Roth et al. [2005], installation of an inter–turbine pressure sensor becomes more difficult on modern engines. Consequently, we remove P49 from the list of available sensors and carry out the search process a second time over the mere 330 possible configurations.

Table 4.24 lists the four sensor sets selected for further investigations. Set **A** is the standard instrumentation defined in table 4.5 and used to validate our diagnosis tool in the preceding sections. It serves here as the baseline for comparison, set **B** is the optimal configuration when considering P49 as available whereas set **C** is the optimal one when discarding P49. Finally, set **D** is the worst sensor set. The notion of optimality is of course relative to the metric in equation (2.88).

Upon examination of table 4.24, we notice that P13, both tachometers and P3 are common to the sets **A**, **B** and **C**. P13 is indeed required to assess correctly fan performance. Concerning the spool speeds and P3, these sensors are mandatory for control purpose. It is thus a good point that they appear in the optimal sets. Furthermore, spool speeds are generally measured with a high accuracy, which is one of the reasons of their presence in the optimal configurations. The main characteristic of set **B** is obviously the presence of P49. This sensor is in fact the only one in the list that allows full discrimination between

	P13	T13	P26	T26	P3	T3	NF	NC	P49	T49	P5	T5
1	✓		✓		✓		✓	✓	✓	✓		
2	✓		✓			✓	✓	✓	✓	✓		
3	✓		✓		✓	✓	✓	✓	✓			
4	✓		✓		✓		✓	✓	✓			✓
5	✓		✓			✓	✓	✓	✓			✓
6	✓				✓	✓	✓	✓	✓	✓		
7	✓		✓		✓	✓		✓	✓	✓		
8	✓		✓		✓	✓	✓		✓	✓		
9	✓		✓		✓	✓	✓		✓	✓		
10			✓		✓	✓	✓	✓	✓	✓		
	9	0	9	0	8	8	9	8	10	7	0	2

Table 4.20: Top 10 configurations based on the **aggregated metric**

	P13	T13	P26	T26	P3	T3	NF	NC	P49	T49	P5	T5
1	✓		✓		✓	✓	✓	✓	✓			
2	✓		✓			✓	✓	✓	✓	✓		
3	✓		✓			✓	✓	✓	✓			✓
4	✓		✓		✓		✓	✓	✓	✓		
5	✓		✓		✓		✓	✓	✓			✓
	5	0	5	0	3	3	5	5	5	2	0	2

Table 4.21: Top 5 configurations based on the **condition number** of the FIM

	P13	T13	P26	T26	P3	T3	NF	NC	P49	T49	P5	T5
1	✓		✓		✓		✓	✓	✓	✓		
2	✓		✓				✓	✓	✓	✓		✓
3			✓		✓		✓	✓	✓	✓		✓
4	✓		✓			✓	✓	✓	✓	✓		
5	✓				✓		✓	✓	✓	✓		✓
	4	0	4	0	3	1	5	5	5	5	0	3

Table 4.22: Top 5 configurations based on the **trace** of the FIM

	P13	T13	P26	T26	P3	T3	NF	NC	P49	T49	P5	T5
1	✓		✓			✓	✓	✓	✓	✓		
2	✓		✓		✓		✓	✓	✓	✓		
3	✓				✓	✓	✓	✓	✓	✓		
4	✓		✓			✓	✓	✓	✓			✓
5	✓		✓		✓	✓	✓	✓	✓			
	5	0	4	0	3	4	5	5	5	3	0	1

Table 4.23: Top 5 configurations based on the **determinant** of the FIM

set **A**	P13	T26	P3	T3	NF	NC	T49
set **B**	P13	P26	P3	NF	NC	P49	T49
set **C**	P13	P26	P3	T3	NF	NC	T5
set **D**	P13	T13	T26	T3	T49	T5	P5

Table 4.24: Selected sensors sets for further investigation

hpt and lpt degradation. It is also worth noting that the Exhaust Gas Temperature (defined as T49 or T5, depending on the manufacturer) is part of these three sensor sets. An EGT sensor is already fitted on every engine as this quantity serves nowadays as a crude indicator of engine health in the cockpit (through the so–called EGT margin).

We see that the optimal sets **B** and **C** have the largest number of pressure sensors, respectively four and three out of seven. This is explained by the fact that pressure sensors are more informative about the engine condition than temperature sensors. From this statement, the absence of the lpt exhaust pressure P5 from the optimal instrumentation may appear strange. However, in a mixed–flow turbofan, this quantity is quite correlated with the fan outlet pressure P13. Consequently selecting P13 discards P5 and vice–versa. Set **A** is made of two pressure and three temperature sensors, relatively well distributed across the engine. On the contrary, set **D** is truly the black sheep among the possible instrumentation sets. It combines the five available temperatures, two almost redundant pressures (namely P13 and P5) and no tachometer.

4.6.2 Evaluation of the sensor sets

In order to illustrate the benefit brought by optimal instrumentation, we follow a procedure similar to the one used for the comparison of various flavours of the Kalman filter in section 4.5.1. We generate a series of 200 scenarios for each of the four sensor sets listed in table 4.24. As we formulated the selection of optimal instrumentation to estimate the ten health parameters of the turbomachinery components, these scenarios contain only gradual deterioration. The progressive change in engine condition is tracked by means of the SLKF-1 type of Kalman filter.

Table 4.25 reports the RMSE, as defined in equation (4.2), evaluated over the batch of 200 no–fault scenarios and expressed in percent of the average baseline value (equal to one here) for each of the ten module health parameters and the four sensor sets. The last row sums up the previous ones.

The last row of table 4.9 shows clearly the influence of the sensor configuration on the estimation accuracy. The optimal set including P49 *i.e.*, set **B** leads to the lowest RMSE, closely followed by the optimal set **C** (without P49). The RMSE achieved with the baseline instrumentation is slightly larger than the one of the optimal sensor sets. Among the 330 possible configurations without P49, set **A** is ranked number 10, indeed not so

RMSE (%)	Set **A**	Set **B**	Set **C**	Set **D**
SW12	0.133	0.102	0.083	0.656
SE12	0.263	0.174	0.163	0.075
SW2	0.025	0.206	0.235	0.198
SE2	0.217	0.146	0.153	1.172
SW26	0.218	0.198	0.177	2.121
SE26	0.123	0.211	0.109	0.736
SW41	0.020	0.050	0.079	1.102
SE41	0.085	0.107	0.083	0.240
SW49	0.139	0.020	0.115	0.710
SE49	0.167	0.122	0.151	0.346
Sum	**1.390**	**1.337**	**1.346**	**7.356**

Table 4.25: Performance of the selected sensor sets in terms of RMSE

far from the best one, set **C**. On the other hand table 4.9 confirms that setbf D is a true poor performer. Indeed, its RMSE is 5.5 larger than the one of set **B**.

It seems worth pointing out that the criterion supporting our sensor selection is mainly related to the variance in the estimated parameters as it is based on the Fisher information matrix. On the other hand, the RMSE is a measure of the global estimation error that combines both the bias and variance parts. The validation procedure tends to show that the optimal sets according to the figure of merit (2.88) are also optimal in the RMSE sense.

4.6.3 Observability analysis of the sensor sets

We conclude this chapter with an observabilty analysis as per the concepts defined in section 2.5.4. Our goal is to obtain a coarse indication of how accurately a health parameter can be estimated on one hand and how responsive a sensor is to a change in the health condition on the other hand. This information is given respectively by the observability and sensitivity indices.

Figure 4.14 depicts the observability indices of the health parameters for the four sensors sets. Concerning set **A**, the observability index of SW2 and SW49 reaches one so these two parameters are perfectly observable (*i.e.*, unbiased estimates on average) The observability levels of SW12, SW26, SE26 and SE41 are quite high. The least observable parameters are SE12 and SW49. The optimal set **B** makes SW49 perfectly observable due to the presence of the inter–turbine pressure sensor P49. On the other hand, the observability of the health parameters of the booster and SE26 is degraded with respect to set **A** because we have one less sensor in the neighbourhood of these components. This kind of "transfer of observability" between the different parameters that depends on the instrumentation

is a result of the underdetermined nature of the estimation problem. The optimal set **C** enhances the observability of the fan, the hpc and SE2 with respect to the baseline instrumentation. This comes at the price of a reduced observability of SW2. The situation of set **D** is as expected much worse. Except SE12 that shows a perfect observability due to the duet of sensors P13 and T13, all health parameters exhibit observability levels that are well below those of sensor set **A**. This is particularly obvious for SW26 that is almost unobservable and, to a lesser extent, for the health parameters of the booster.

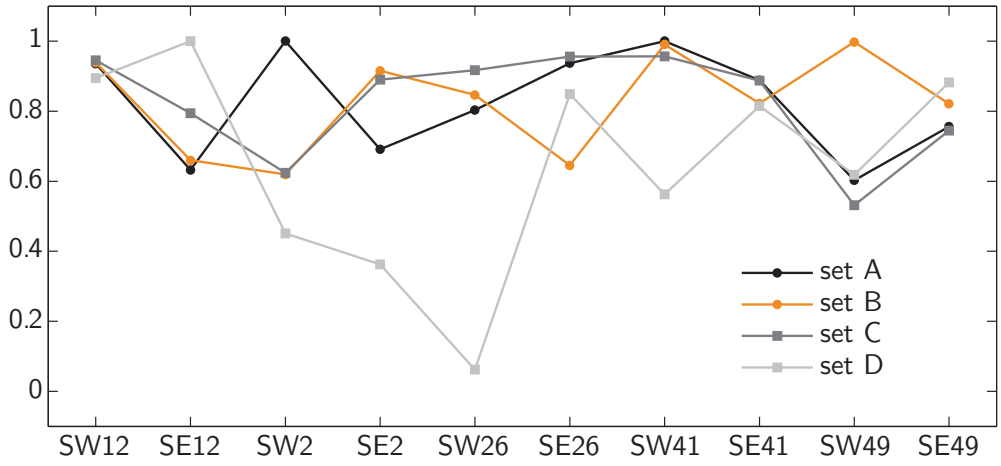


Figure 4.14: Comparison of the observability indices OI

Figure 4.15 shows the sensitivity indices of the sensor sets as computed with equation (2.90). As a matter of fact, the bars in each subfigure sum to unity. Looking at the three well–designed instrumentation sets (**A**, **B** and **C**), the spool speeds are the most sensitive to a change in the health condition followed by the pressures and finally the temperatures as mentioned above. Table 4.26 completes the picture by reporting the global sensitivity of each sensor set *i.e.*, the denominator in equation (2.90). The optimal sets **B** and **C** have obviously the highest sensitivity. Nonetheless, the gap between set **A** and **B** is narrow. As a further argument of its poor nature, the overall sensitivity of set **D** is more than one order of magnitude lower than for the three others.

set **A**	set **B**	set **C**	set **D**
192.52	231.90	201.83	16.37

Table 4.26: Total sensitivity of the selected sensor sets

Finally, figure 4.16 shows the unobservable directions for sensor set **A**. Each direction is presented as a cobweb plot [see Butler et al., 2006] that we find convenient to plot a 10-dimensional vector. Each spoke of the cobweb plot is associated to a health parameter, the sign in front of the parameter's name specifying the sense of the change. The circles are contours of deviation magnitude and range between zero and one since the unobservable directions are vector of unit norms. Indeed the unobservable directions are the base

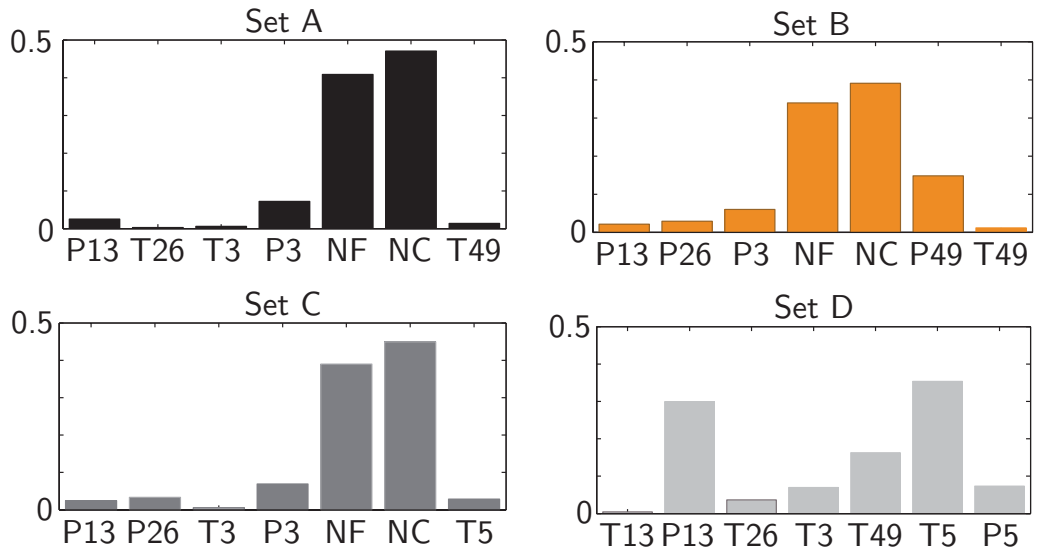


Figure 4.15: Comparison of the sensitivity indices SI

vectors of the null space of the influence coefficient matrix. As we consider a well–designed configuration with ten health parameters and seven sensors, we have three unobservable directions in the space of the health parameters.

Roughly speaking, the unobservable direction in the left subfigure involves the triplet SE2, SW26 and SE26. The unobservable direction in the central subfigure impacts the triplet SE41, SW49 and SE49 and is a well-known issue in gas–path analysis: standard instrumentation does not allow perfect discrimination between hpt and lpt degradation. The unobservable direction in the right subfigure mixes SW12, SE12 and SE49. This last subfigure explains the increasing estimation errors on these three parameters that we observe in figure 4.5. On the other hand, the fact that SW2 and SW41 are perfectly observable and that we introduce correlations between the flow and efficiency factors on each component minimises the effects of the other two unobservable directions.

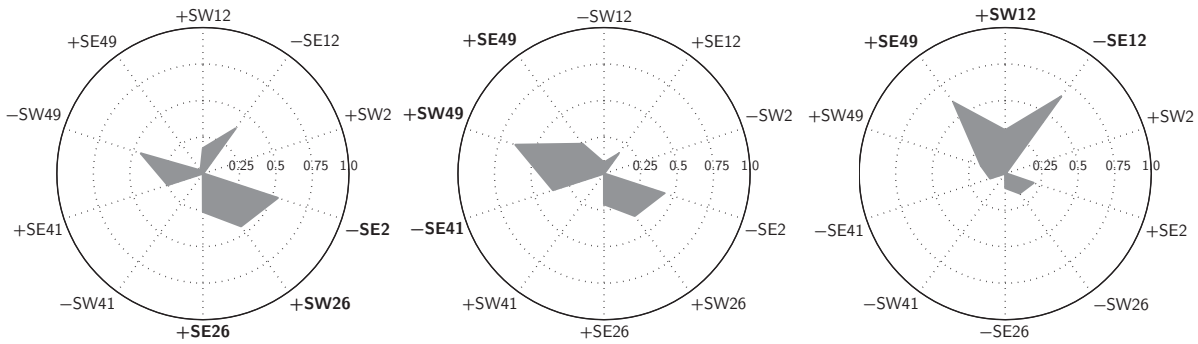


Figure 4.16: Unobservable directions for sensor set **A**

Chapter 5

Diagnostics in transient operation

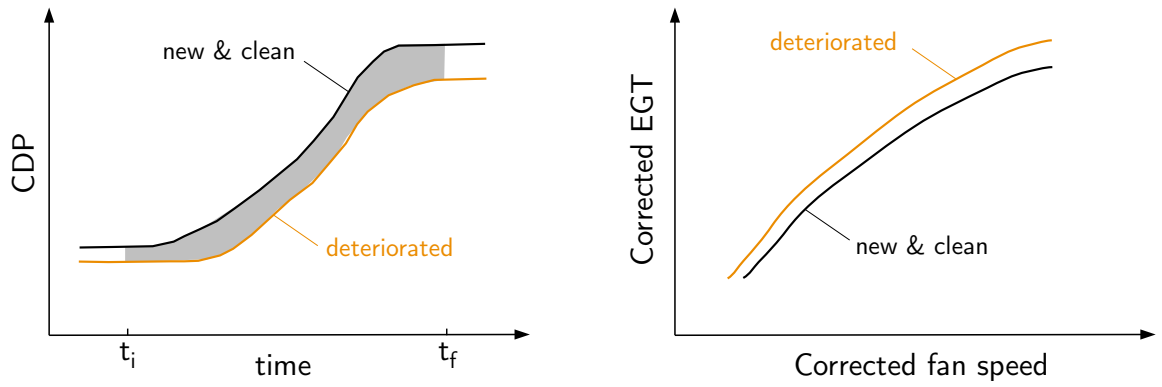
Contents

5.1 Context and motivations

MODULE performance analysis was initially formulated for a steady-state operation of the engine, mainly for simplicity and computational cost reasons. However, we have seen previously that the major downside associated with the processing of data collected at a single steady-state operating point is that the resulting estimation problem is under-determined. One of the remedies proposed in the community is to extend the analysis to multiple steady-state operating points [Gulati et al., 2001, Grönstedt, 2002]. This solution takes advantage of the non-linearity of the engine with respect to its operating point to increase the informational content of the data.

In practice, collecting steady-state data of good quality can be tedious. A commercial aircraft operates in (quasi-)steady-state at cruise and possibly at the end of the take-off phase. A military aircraft on the other hand spends up to 70% of the time in transient operation. In fact, stabilising an engine at a number of steady-state points turns out to be quite a lengthy – read costly – procedure that is best achieved outside regular operations, preferably in a test cell.

Consequently the idea emerged in the late 1980's to try and assess the engine condition directly from transient recordings. Much of the early research was targeted to military applications and was of *qualitative* nature *i.e.,* concerned with the detection of a change in the engine condition. The first efforts [Henry, 1988, Cue and Muir, 1991, Eustace et al., 1994] were dedicated to the derivation of so-called signatures, in both healthy (new and clean) and deteriorated conditions, from transient manoeuvres representative of take-off accelerations. The implanted fault types were mainly of mechanical nature, such as biased variable guide vanes, biased nozzle area and customer bleed leak. Malfunction of various sensors was also investigated. The signatures take the form of a time trace of the gas-path measurements (figure 5.1(a)) or a cross-plot of two gas-path measurements (figure 5.1(b)). This research also scrutinised the factors that impact the repeatability of the data, such as ambient conditions and rate of change of the throttle lever.



(a) Time trace

(b) Correlation between gas-path sensors

Figure 5.1: Signature of deterioration on the transient response of an engine

Soon enough, trend monitoring based on transient data was formulated as an estimation problem. In a first step, the goal was to estimate the parameters of transfer functions that model the dynamic response of the engine [Merrington, 1988]. Equation (5.1) shows such a transfer function that approximates the response of the fan speed with respect to the fuel flow as a first order dynamic system. Deterioration modifies that relationship which translates into changes in the static gain K and the time constant τ. These changes in the dynamic characteristics can then be correlated to particular engine faults. Additional studies [Merrington et al., 1991, Merrington, 1994] investigated the influence of various effects such as heat soakage, under-modelling and linearisation errors on the effectiveness of the estimation scheme.

$$\delta\mathsf{NF} = \frac{K}{1+\tau s}\,\delta\mathsf{WF} \tag{5.1}$$

The most important findings that were extracted from these studies are that:

- some faults appear only during transient operation but could seriously degrade the operability of the engine in its whole envelope *e.g.,* misscheduling of the variable nozzle,
- transient records provide significant diagnostic content,
- transient data is more sensitive to faults thereby facilitating their detection, localisation and quantification.

The estimation problem turned then to a more *quantitative* formulation with the use of state variable models incorporating health parameters similar to those found in steady-state MPA (*i.e.,* correcting factors on the flow capacity and efficiency of the modules). The earliest publications [Luppold et al., 1989, Kerr et al., 1992] seem to favour the Kalman filter to estimate the health parameters. Later, various techniques used in steady-state analysis made their way to the transient realm. These include artificial neural networks [Li, 2003, Sampath et al., 2003], genetic algorithms [Ogaji et al., 2003] and traditional least-squares [Grönstedt, 2005]. These batch methods rely on a so-called accumulated deviation, defined in equation (5.2) to evaluate the engine condition. The accumulated deviation can be visualised by the shaded area in figure 5.1(a). Next to these batch approaches, further studies on the effectiveness of a sequential treatment of the data by means of the Kalman filter are also reported in the recent literature [Dewallef and Léonard, 2003, Kobayashi et al., 2005].

$$\Delta\mathbf{y} = \int_{t_i}^{t_f} \left(\mathbf{y}(t) - \widehat{\mathbf{y}}^{\mathrm{hl}}(t)\right)\,\mathrm{d}t \tag{5.2}$$

Transient diagnostics was advocated as a palliative to the underdetermined nature of the estimation problem. As paradoxical as it can be, most of the references cited above consider "well-posed" estimation problems, *i.e.,* configurations for which the number of gas-path measurements is at least equal to the number of health parameters. Nonetheless, a common outcome of these studies consists in an improvement in the accuracy of the health parameter estimates with respect to the results obtained with steady-state data.

Most authors relate this improvement to a better information content of transient data than steady-state one, yet it seems relatively difficult to provide a generic proof that supports this observation. In a recent contribution [Borguet et al., 2011], we proposed a theoretical framework that formalises the benefit of transient analysis for a particular class of input signals.

Another important issue concerns the quality of the performance deck. If steady-state models of high fidelity can be obtained relatively easily, transient models on the other hand are recognised to suffer from modelling errors. The influence of these errors on the estimation accuracy as well as ways to cope with them have been the topic of several studies as we will see later.

The remainder of this chapter reports our contributions to module performance analysis based on transient data. In section 5.2 we point out the differences between steady-state and transient analysis. Then we extend the performance monitoring tool based on a Kalman filter to the processing of transient data. Section 5.3 addresses the issue of model-plant-mismatch, *i.e.,* using a model that is not perfectly faithful. We propose an original approach that mixes the physics based, imperfect model with a black-box component to improve the robustness of the diagnosis tool with respect to these modelling errors. Finally, these theoretical developments are assessed on simulated scenarios. The results obtained on an underdetermined configuration are presented and discussed in section 5.4.

5.2 Changes with respect to the steady-state case

5.2.1 The engine model

We saw in section 2.1.2 that the application of the equations of conservation of mass, momentum and energy to the components of the engine flow path leads naturally to a non-linear state-space model in continuous time:

$$\dot{\mathbf{x}}(t) = f_c\left(\mathbf{x}(t), \mathbf{u}(t), \mathbf{h}(t)\right) \tag{5.3}$$

$$\mathbf{y}(t) = g\left(\mathbf{x}(t), \mathbf{u}(t), \mathbf{h}(t)\right) \tag{5.4}$$

where $\mathbf{x}$ are the dynamic state variables associated to the transient phenomena (*e.g.*, spool speeds for the shaft inertia, metal temperatures for the heat soakage), $\dot{\mathbf{x}}$ stands for the time derivatives of the states, $\mathbf{u}$ are the input signals, $\mathbf{h}$ are the health parameters and $\mathbf{y}$ are the gas path measurements.

The continuous-time nature of the dynamic equation (5.3) is not quite suited for the sequential processing of data on a digital computer. Hence, the equation is integrated over one time step and under the zero-order hold assumption (*i.e.,* the value of the input signals is constant over the time step). As no analytical formulation of the vector function f_c is available, it is integrated numerically to lead to its discrete time counter-part (5.5).

On the other hand, the static equation (5.4) generating the measurements is valid in both time settings.

$$\mathbf{x}_{k+1} = \mathcal{F}(\mathbf{x}_k, \mathbf{u}_k, \mathbf{h}_k) \tag{5.5}$$
$$\mathbf{y}_k = \mathcal{G}(\mathbf{x}_k, \mathbf{u}_k, \mathbf{h}_k) \tag{5.6}$$

Finally, following the example of the statistical model used in steady-state, we acknowledge the fact that the transient model contains some inaccuracies and that the measurements are corrupted by noise. Moreover, we complement the description of the system with a random walk for the temporal evolution of the health parameters which leads to the following set of equations:

$$\mathbf{h}_{k+1} = \mathbf{h}_k + \boldsymbol{\omega}_k \qquad \text{with } \boldsymbol{\omega}_k \in \mathcal{N}(\mathbf{0}, \mathbf{Q}_{\mathbf{h},k}) \tag{5.7}$$
$$\mathbf{x}_{k+1} = \mathcal{F}(\mathbf{x}_k, \mathbf{u}_k, \mathbf{h}_k) + \boldsymbol{\nu}_k \qquad \text{with } \boldsymbol{\nu}_k \in \mathcal{N}(\mathbf{0}, \mathbf{Q}_{\mathbf{x},k}) \tag{5.8}$$
$$\mathbf{y}_k = \mathcal{G}(\mathbf{x}_k, \mathbf{u}_k, \mathbf{h}_k) + \boldsymbol{\epsilon}_k \qquad \text{with } \boldsymbol{\epsilon}_k \in \mathcal{N}(\mathbf{0}, \mathbf{R}_{\mathbf{y},k}) \tag{5.9}$$

where the random vectors $\boldsymbol{\omega}_k$, $\boldsymbol{\nu}_k$ and $\boldsymbol{\epsilon}_k$ are respectively the process noise for the health parameters, the process noise for the state variables and the measurement noise. They are assumed to be zero-mean, white, Gaussian and independent of each other.

5.2.2 The dual estimation problem

Performance trending based on transient data can be cast as an estimation problem. Let us again define the residual $\mathbf{r}_k$ as the difference between the actual measurements $\mathbf{y}_k$ taken on the engine and some prediction $\widehat{\mathbf{y}}_k^-$ from the simulation deck. For transient operation, the residual is a function of both the health parameters $\mathbf{h}_k$ and the state variables $\mathbf{x}_k$:

$$\mathbf{r}_k \triangleq \mathbf{y}_k - \widehat{\mathbf{y}}_k^- \equiv f(\mathbf{x}_k, \mathbf{h}_k) \tag{5.10}$$

In fact the prediction of measurements with the transient simulation deck (*i.e.,* the deterministic part of equation (5.9)) requires that values for the health parameters $\mathbf{h}_k$ and the state variables $\mathbf{x}_k$ be provided in addition to the control inputs $\mathbf{u}_k$:

$$\widehat{\mathbf{y}}_k^- \triangleq \mathcal{G}(\mathbf{x}_k, \mathbf{u}_k, \mathbf{h}_k) \tag{5.11}$$

Consequently, two tasks must be simultaneously carried out during transient performance monitoring: i) the estimation of the state variables $\mathbf{x}_k$ associated to the engine dynamics, and ii) the estimation of the health parameters $\mathbf{h}_k$ reflecting the degradation of the engine.

Such a situation where both the model parameters and the state variables must be determined from the same stream of noisy measurements is known as the dual (read double) estimation problem [Wan and Nelson, 2001]. Estimates of these quantities are obtained by minimising the distance between the actual and predicted measurements, according to some metric specified by the scalar-valued function ρ_k:

$$\left(\widehat{\mathbf{h}}_k, \widehat{\mathbf{x}}_k\right) = \arg\min_{\mathbf{h}_k, \mathbf{x}_k} \left\{\rho_k(\mathbf{y}_k - \widehat{\mathbf{y}}_k^-)\right\} \tag{5.12}$$

For sake of completeness, let us mention that the dual estimation problem can be simplified in the case of a batch processing of the data. Assuming that the health condition does not change during the collection of the data and that the engine is initially at steady-state, the problem reduces to the estimation of the constant health condition **h**. This can be solved with proven techniques, see [Loisy et al., 1992].

These simplifications do not match our requirements of a sequential processing of the data under a possibly time-varying health condition. As a result, we have to stick to the dual estimation framework that is illustrated schematically in figure 5.2. The "dual estimation procedure" box hides the resolution of the minimisation problem (5.12). If we assume that the states and the health parameters are Gaussian random vectors and that the system is linear, the choice of the mean-square error as the cost function ρ_k leads to the Kalman filter as we saw in chapter 3.

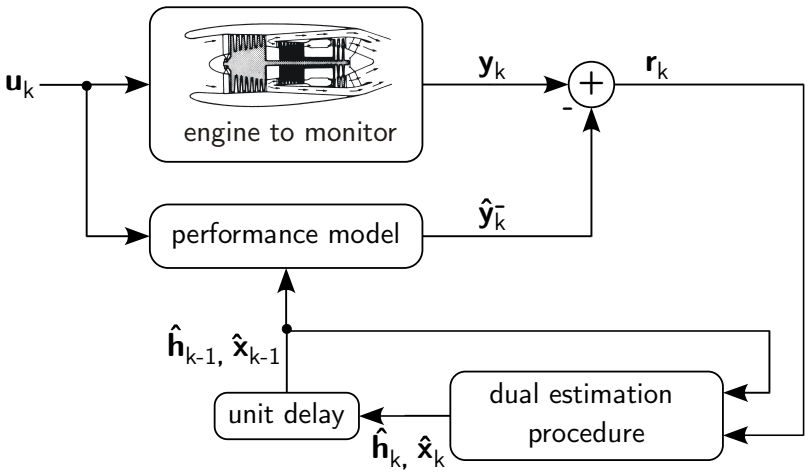


Figure 5.2: The dual estimation problem

The Kalman filter can be applied in two ways to the dual estimation problem. In the *joint estimation* approach, the states and health parameters are concatenated into one augmented state vector that is estimated with a single Kalman filter. Although it was successfully applied to engine health monitoring [Kobayashi and Simon, 2003, Kobayashi et al., 2005], the joint approach was also reported to suffer from instability problems in other applications [Wan and Nelson, 1996].

The alternative is to handle the dual estimation task by running two Kalman filters concurrently, one dealing with the estimation of the health parameters and the other, with the estimation of the state variables. We prefer this *dual Kalman filter* approach to the joint one given the different time-scales associated with the engine dynamics (rather fast) and with the health parameters (much slower) [Borguet et al., 2005]. Moreover, this implementation allows to tune each estimation procedure to its characteristics. The next section describes in more details the principle of operation of the dual Kalman filter.

5.2.3 The dual Kalman filter

The dual Kalman filter was successfully applied in numerous fields as various as speech processing [Wan and Nelson, 1997], training of neural networks [Ciocoiu, 2002], aerodynamic characterisation of spacecraft [VanDyke et al., 2004] and geophysics modelling [Gove and Hollinger, 2006]. All these systems are genuinely non-linear and require *ad hoc* extensions of the Kalman filter to solve the dual estimation problem. We also use non-linear filters in the case of transient performance monitoring. Indeed, typical manoeuvres span the whole power range of the engine at several points in its flight envelope.

Figure 5.3 illustrates the aforementioned structure of the dual Kalman filter that decomposes the dual estimation problem into separate state estimation and parameter estimation blocks. At time step k, both filters are fed with the most recent estimates of the states $\widehat{\mathbf{x}}_{k-1}$ and of the parameters $\widehat{\mathbf{h}}_{k-1}$ as well as with the latest measurements $\mathbf{y}_k$ and inputs $\mathbf{u}_k$. Basically, each filter updates one of the two vectors of interest assuming the other one is known. This is akin to optimising a multidimensional objective function by optimising one argument at a time while the others are fixed. As we have only access to *estimates* of the states and parameters, we should take into account the influence of their respective uncertainty on the other quantity that we estimate [Nelson, 2000]. Such an error-coupling strategy allows to account, at least approximately, for the cross-dependencies between the states and the parameters that appear explicitly in the joint approach.

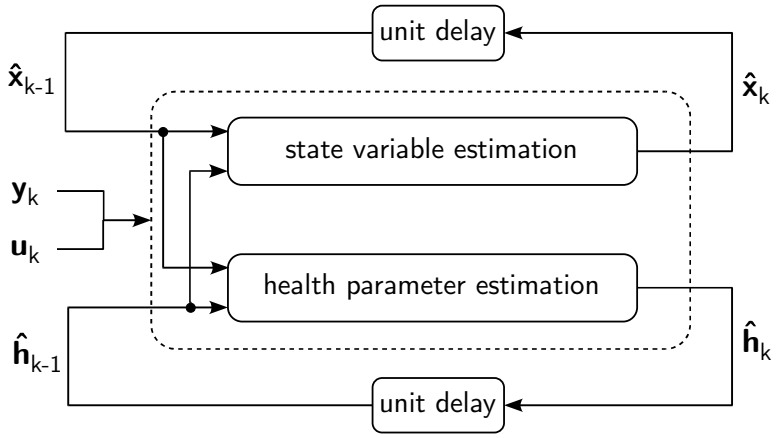


Figure 5.3: Structure of the dual Kalman filter

It is worth mentioning that in contrast to batch methods, the present technique does not provide the actual trajectory of the system in the state space due to the estimation error on the health parameters. Nonetheless, as the estimated health parameters converge toward their actual values, the estimated state trajectory gets closer to the true one. It is not a drawback *per se* since our foremost interest is in the health parameter values.

We have implemented both the extended and unscented Kalman filters for our transient performance monitoring tool. These algorithms were derived in chapter 3 under the assumption of a steady-state operation. In the following, we focus on the modifications brought to deal with the peculiarities of a processing of transient data.

Algorithm 5.1 Extended Kalman filter for state variable estimation

Require: $\widehat{\mathbf{x}}_0$ and $\mathbf{P}_{\mathbf{x},0}$

for $k = 1$ to N **do**

$\widehat{\mathbf{x}}_k^- = \mathcal{F}(\widehat{\mathbf{x}}_{k-1}, \mathbf{u}_k, \widehat{\mathbf{h}}_{k-1})$

Compute Jacobian matrix $\mathbf{F}_{\mathbf{x},k}$ as per equation (5.15)

$\mathbf{P}_{\mathbf{x},k}^- = \mathbf{F}_{\mathbf{x},k}\mathbf{P}_{\mathbf{x},k-1}\mathbf{F}_{\mathbf{x},k}^T + \mathbf{Q}_{\mathbf{x},k} + \mathbf{P}_{\mathbf{xhx},k-1}$

$\mathbf{r}_k = \mathbf{y}_k - \mathcal{G}(\widehat{\mathbf{x}}_k^-, \mathbf{u}_k, \widehat{\mathbf{h}}_{k-1})$

Compute Jacobian matrix $\mathbf{G}_{\mathbf{x},k}$ as per equation (5.16)

$\mathbf{P}_{\mathbf{y},k} = \mathbf{G}_{\mathbf{x},k}\mathbf{P}_{\mathbf{x},k}^-\mathbf{G}_{\mathbf{x},k}^T + \mathbf{R}_{\mathbf{y},k} + \mathbf{P}_{\mathbf{yhy},k-1}$

$\mathbf{K}_k = \mathbf{P}_{\mathbf{x},k}^-\mathbf{G}_{\mathbf{x},k}^T\ \mathbf{P}_{\mathbf{y},k}^{-1}$

$\widehat{\mathbf{x}}_k = \widehat{\mathbf{x}}_k^- + \mathbf{K}_k\ \mathbf{r}_k$

$\mathbf{P}_{\mathbf{x},k} = (\mathbf{I} - \mathbf{K}_k\ \mathbf{G}_{\mathbf{x},k})\ \mathbf{P}_{\mathbf{x},k}^-\ (\mathbf{I} - \mathbf{K}_k\ \mathbf{G}_{\mathbf{x},k})^T + \mathbf{K}_k\ \mathbf{R}_{\mathbf{y},k}\ \mathbf{K}_k^T$

end for

Algorithm 5.2 Unscented Kalman filter for state variable estimation

Require: $\widehat{\mathbf{x}}_0$ and $\mathbf{P}_{\mathbf{x},0}$

for $k = 1$ to N **do**

$\mathcal{X}_{k-1} = \left[\widehat{\mathbf{x}}_{k-1} \quad \widehat{\mathbf{x}}_{k-1} + \gamma\sqrt{\mathbf{P}_{\mathbf{x},k-1}} \quad \widehat{\mathbf{x}}_{k-1} - \gamma\sqrt{\mathbf{P}_{\mathbf{x},k-1}}\right]$

$\mathcal{X}_{i,k}^- = \mathcal{F}(\mathcal{X}_{i,k-1}, \mathbf{u}_k, \widehat{\mathbf{h}}_{k-1}) \quad i = 0, 1, \ldots, 2\,n_x$

$$\widehat{\mathbf{x}}_k^- = \sum_{i=0}^{2\,n_x} W_i^{(m)}\mathcal{X}_{i,k}^-$$

$$\mathbf{P}_{\mathbf{x},k}^- = \sum_{i=0}^{2\,n_x} W_i^{(c)}(\mathcal{X}_{i,k}^- - \widehat{\mathbf{x}}_k^-)(\mathcal{X}_{i,k}^- - \widehat{\mathbf{x}}_k^-)^T + \mathbf{Q}_{\mathbf{x},k} + \mathbf{P}_{\mathbf{xhx},k-1}$$

$\mathcal{X}_k = \left[\widehat{\mathbf{x}}_k^- \quad \widehat{\mathbf{x}}_k^- + \gamma\sqrt{\mathbf{P}_{\mathbf{x},k}^-} \quad \widehat{\mathbf{x}}_k^- - \gamma\sqrt{\mathbf{P}_{\mathbf{x},k}^-}\right]$

$\mathcal{Y}_{i,k} = \mathcal{G}(\mathcal{X}_{i,k}, \mathbf{u}_k, \widehat{\mathbf{h}}_{k-1}) \quad i = 0, 1, \ldots, 2\,n_x$

$$\widehat{\mathbf{y}}_k^- = \sum_{i=0}^{2\,n_x} W_i^{(m)}\mathcal{Y}_{i,k}$$

$\mathbf{r}_k = \mathbf{y}_k - \widehat{\mathbf{y}}_k^-$

$$\mathbf{P}_{\mathbf{y},k} = \sum_{i=0}^{2\,n_x} W_i^{(c)}(\mathcal{Y}_{i,k} - \widehat{\mathbf{y}}_k^-)(\mathcal{Y}_{i,k} - \widehat{\mathbf{y}}_k^-)^T + \mathbf{R}_{\mathbf{y},k} + \mathbf{P}_{\mathbf{yhy},k-1}$$

$$\mathbf{P}_{\mathbf{xy},k} = \sum_{i=0}^{2\,n_x} W_i^{(c)}(\mathcal{X}_{i,k} - \widehat{\mathbf{x}}_k^-)(\mathcal{Y}_{i,k} - \widehat{\mathbf{y}}_k^-)^T$$

$\mathbf{K}_k = \mathbf{P}_{\mathbf{xy},k}\,\mathbf{P}_{\mathbf{y},k}^{-1}$

$\widehat{\mathbf{x}}_k = \widehat{\mathbf{x}}_k^- + \mathbf{K}_k\,\mathbf{r}_k$

$\mathbf{P}_{\mathbf{x},k} = \mathbf{P}_{\mathbf{x},k}^- - \mathbf{K}_k\,\mathbf{P}_{\mathbf{y},k}\,\mathbf{K}_k^T$

end for

Estimation of the state variables

In the framework of module performance analysis, some states (*e.g.*, the spool speeds) are measured while others (*e.g.,* metal temperatures) are not measurable. It is in fact quite difficult to define a representative carcass temperature for the heat soakage from a reasonable number of thermocouples.

The estimation of the states of a dynamic system was the original scope of Kalman's work in the early 60's [Kalman, 1960]. As stated above, we estimate the state variables at time step k assuming that the health parameters are fixed at their latest available estimate $\widehat{\mathbf{h}}_{k-1}$. Consequently, we rely on the following state-space model:

$$\mathbf{x}_k = \mathcal{F}(\mathbf{x}_{k-1}, \mathbf{u}_k, \widehat{\mathbf{h}}_{k-1}) + \boldsymbol{\nu}_k \tag{5.13}$$

$$\mathbf{y}_k = \mathcal{G}(\mathbf{x}_k, \mathbf{u}_k, \widehat{\mathbf{h}}_{k-1}) + \boldsymbol{\epsilon}_k \tag{5.14}$$

Algorithm 5.1 describes in pseudo-code style the operations involved within the extended Kalman filter. As both the state transition and the measurement equations are non-linear, we have to compute two Jacobian matrices in order to assess the Kalman gain. These Jacobian matrices (5.15) and (5.16) give respectively the sensitivity of the updated states and of the measurements with respect to the states.

$$\mathbf{F}_{\mathbf{x},k} = \left.\frac{\partial \mathcal{F}(\mathbf{x},\mathbf{u},\mathbf{h})}{\partial \mathbf{x}}\right|_{\mathbf{u}=\mathbf{u}_k;\,\mathbf{h}=\widehat{\mathbf{h}}_{k-1};\,\mathbf{x}=\widehat{\mathbf{x}}_{k-1}} \tag{5.15}$$

$$\mathbf{G}_{\mathbf{x},k} = \left.\frac{\partial \mathcal{G}(\mathbf{x},\mathbf{u},\mathbf{h})}{\partial \mathbf{x}}\right|_{\mathbf{u}=\mathbf{u}_k;\,\mathbf{h}_k=\widehat{\mathbf{h}}_{k-1};\,\mathbf{x}=\widehat{\mathbf{x}}_k^-} \tag{5.16}$$

Algorithm 5.2 describes in pseudo-code style the operations involved within the unscented Kalman filter. We notice that the unscented transform (see section 3.3.3) is called twice due, again, to the non-linearity of both equations in the state-space model (5.13-5.14). The state transition equation is applied to sigma points sampled from the distribution of the previous state estimate while the measurement equation is applied to sigma points sampled from the distribution of the prior state estimate, in the manner of the evaluation of the Jacobian matrices $\mathbf{F}_{\mathbf{x},k}$ and $\mathbf{G}_{\mathbf{x},k}$ respectively at $\widehat{\mathbf{x}}_{k-1}$ and $\widehat{\mathbf{x}}_k^-$.

The matrices $\mathbf{P}_{\mathbf{xhx},k-1}$ and $\mathbf{P}_{\mathbf{yhy},k-1}$ appearing at lines 4 and 7 in the extended Kalman filter and at lines 5 and 10 in the unscented Kalman filter embed the uncertainty in the health parameters into the state estimation process. We discuss them in more details in the section dedicated to the error-coupling strategy.

Both filters are quite equivalent from the viewpoint of computational cost if the Jacobian matrices in the EKF are evaluated *via* a centred-difference scheme. The number of calls to the non-linear performance model amounts in that case to $2\,(2\,n_x + 1)$ at each time step.

Algorithm 5.3 Extended Kalman filter for health parameter estimation

Require: $\widehat{\mathbf{h}}_0$ and $\mathbf{P}_{\mathbf{h},0}$

for $k = 1$ to N **do**

$\widehat{\mathbf{h}}_k^- = \widehat{\mathbf{h}}_{k-1}$

$\mathbf{P}_{\mathbf{h},k}^- = \mathbf{P}_{\mathbf{h},k-1} + \mathbf{Q}_{\mathbf{h},k}$

$\mathbf{r}_k = \mathbf{y}_k - \mathcal{M}(\widehat{\mathbf{x}}_{k-1}, \mathbf{u}_k, \widehat{\mathbf{h}}_k^-)$

Compute Jacobian matrix $\mathbf{M}_{\mathbf{h},k}$ as per equation (5.20)

$\mathbf{P}_{\mathbf{y},k} = \mathbf{M}_{\mathbf{h},k}\mathbf{P}_{\mathbf{h},k}^-\mathbf{M}_{\mathbf{h},k}^T + \mathbf{R}_{\mathbf{y},\mathbf{k}} + \mathbf{P}_{\mathbf{yxy},k-1}$

$\mathbf{K}_k = \mathbf{P}_{\mathbf{h},k}^-\mathbf{M}_{\mathbf{h},k}^T\ \mathbf{P}_{\mathbf{y},k}^{-1}$

$\widehat{\mathbf{h}}_k = \widehat{\mathbf{h}}_k^- + \mathbf{K}_k\ \mathbf{r}_k$

$\mathbf{P}_{\mathbf{h},k} = (\mathbf{I} - \mathbf{K}_k\ \mathbf{M}_{\mathbf{h},k})\ \mathbf{P}_{\mathbf{h},k}^-\ (\mathbf{I} - \mathbf{K}_k\ \mathbf{M}_{\mathbf{h},k})^T + \mathbf{K}_k\ \mathbf{R}_{\mathbf{y},k}\ \mathbf{K}_k^T$

end for

Algorithm 5.4 Unscented Kalman filter for health parameter estimation

Require: $\widehat{\mathbf{h}}_0$ and $\mathbf{P}_{\mathbf{h},0}$

for $k = 1$ to N **do**

$\widehat{\mathbf{h}}_k^- = \widehat{\mathbf{h}}_{k-1}$

$\mathbf{P}_{\mathbf{h},k}^- = \mathbf{P}_{\mathbf{h},k-1} + \mathbf{Q}_{\mathbf{h},k}$

$\mathcal{H}_{k-1} = \left[\widehat{\mathbf{h}}_k^- \quad \widehat{\mathbf{h}}_k^- + \gamma\sqrt{\mathbf{P}_{\mathbf{h},k}^-} \quad \widehat{\mathbf{h}}_k^- - \gamma\sqrt{\mathbf{P}_{\mathbf{h},k}^-}\right]$

$\mathcal{Y}_{i,k} = \mathcal{M}(\widehat{\mathbf{x}}_{k-1}, \mathbf{u}_k, \mathcal{H}_{i,k-1}) \quad i = 0, 1, \ldots, 2\,n_h$

$\widehat{\mathbf{y}}_k^- = \sum_{i=0}^{2\,n_h} W_i^{(m)}\mathcal{Y}_{i,k}$

$\mathbf{r}_k = \mathbf{y}_k - \widehat{\mathbf{y}}_k^-$

$\mathbf{P}_{\mathbf{y},k} = \sum_{i=0}^{2\,n_h} W_i^{(c)}(\mathcal{Y}_{i,k} - \widehat{\mathbf{y}}_k^-)(\mathcal{Y}_{i,k} - \widehat{\mathbf{y}}_k^-)^T + \mathbf{R}_{\mathbf{y},\mathbf{k}} + \mathbf{P}_{\mathbf{yxy},k-1}$

$\mathbf{P}_{\mathbf{hy},k} = \sum_{i=0}^{2n_h} W_i^{(c)}(\mathcal{H}_{i,k-1} - \widehat{\mathbf{h}}_k^-)(\mathcal{Y}_{i,k} - \widehat{\mathbf{y}}_k^-)^T$

$\mathbf{K}_k = \mathbf{P}_{\mathbf{hy},k}\mathbf{P}_{\mathbf{y},k}^{-1}$

$\widehat{\mathbf{h}}_k = \widehat{\mathbf{h}}_k^- + \mathbf{K}_k\ \mathbf{r}_k$

$\mathbf{P}_{\mathbf{h},k} = \mathbf{P}_{\mathbf{h},k}^- - \mathbf{K}_k\ \mathbf{P}_{\mathbf{y},k}\ \mathbf{K}_k^T$

end for

Estimation of the health parameters

The estimation of the health parameters from a transient data stream is conceptually identical to its steady-state counter-part that we considered in chapter 3. The major difference lies in the fact that the measurements depend also on the states in the present case. Furthermore, the state vector is itself a function of the health parameters. Consequently, the state-space model relies on a modified measurement equation that makes this dependence explicit:

$$\mathbf{h}_k = \mathbf{h}_{k-1} + \boldsymbol{\omega}_k \tag{5.17}$$

$$\mathbf{y}_k = \mathcal{M}(\widehat{\mathbf{x}}_{k-1}, \mathbf{u}_k, \mathbf{h}_k) + \boldsymbol{\epsilon}_k \tag{5.18}$$

The non-linear function $\mathcal{M}$ results from the aggregation of the state transition equation (5.5) and the measurement equation (5.6) and is readily available from a performance deck. Interestingly, this transformation allows the measurements $\mathbf{y}_k$ to be predicted based on the state estimate at the previous time step $\widehat{\mathbf{x}}_{k-1}$ which fits into our formulation of the dual estimation problem.

$$\left.\begin{array}{ll} \mathbf{x}_k & = \mathcal{F}(\mathbf{x}_{k-1}, \mathbf{u}_k, \mathbf{h}_k) \\ \mathbf{y}_k & = \mathcal{G}(\mathbf{x}_k, \mathbf{u}_k, \mathbf{h}_k) \end{array}\right\} \Rightarrow \mathbf{y}_k = \mathcal{M}(\mathbf{x}_{k-1}, \mathbf{u}_k, \mathbf{h}_k) \tag{5.19}$$

Algorithm 5.3 describes in pseudo-code style the operations involved within the extended Kalman filter. The similarities with the extended Kalman filter for steady-state data (see algorithm 3.2) are obvious. The Jacobian matrix (5.20) gives the sensitivity of the aggregated measurement model with respect to the health parameters. It is equivalent to the computation of the *total* derivative of equation (5.6) with respect to the parameters.

$$\mathbf{M}_{\mathbf{h},k} = \left.\frac{\partial \mathcal{M}(\mathbf{x}, \mathbf{u}, \mathbf{h})}{\partial \mathbf{h}}\right|_{\mathbf{u}=\mathbf{u}_k;\, \mathbf{h}=\widehat{\mathbf{h}}_k^-\,;\, \mathbf{x}=\widehat{\mathbf{x}}_{k-1}} \tag{5.20}$$

Algorithm 5.4 describes in pseudo-code style the operations involved within the unscented Kalman filter. Due to the linearity of the transition model for the health parameters, we perform the unscented transformation only on the measurement equation. This limits the number of calls to the non-linear performance model to $(2\,n_h + 1)$ at each time step.

The matrix $\mathbf{P}_{\mathbf{yxy},k-1}$ appearing at line 6 in the extended Kalman filter and at line 8 in the unscented Kalman filter embeds the uncertainty in the state variables into the parameter estimation process. We discuss it in more details below.

Error-coupling strategy

As stated above, the tenet behind the dual Kalman filter is to split the dual estimation task into two blocks. For sake of clarity, let us consider the state estimation problem.

The discussion below can obviously be transposed to the health parameter estimation problem *mutatis mutandis*. The Kalman filter estimates the state vector while the health parameters (the other unknown in our dual estimation problem) are fixed at their latest estimates. In the algorithms (5.1-5.2), we call the performance model with the mean value of this estimate. However, in the Kalman filter framework, the health parameter estimates are not deterministic variables, but Gaussian random variables. We should therefore incorporate the available information about the error in this estimate, which is conveyed by its covariance matrix.

The approach taken here to feedback the uncertainty in one quantity (the health parameters) into the filter estimating the other quantity (the state vector) is identical to the way we dealt with the influence of the noise on the operating conditions over the residuals (see section 2.1.3). As a reminder, we increased the uncertainty in the prediction error by using a modified measurement noise signal whose variance was larger than the nominal one.

Transposing the technique to the state estimation problem, we first realise that the health parameters appear in both the state transition (5.13) and the measurement (5.14) equations. As a result, the variance of both the process noise $\boldsymbol{\nu}_k$ and the measurement noise $\boldsymbol{\epsilon}_k$ are complemented respectively with the matrices $\mathbf{P}_{\mathbf{xhx},k}$ and $\mathbf{P}_{\mathbf{yhy},k}$ which basically project the uncertainty in the health parameters onto the state and measurement spaces. In the EKF framework, these two matrices are evaluated through:

$$\mathbf{P}_{\mathbf{xhx},k} = \mathbf{M}^{\star}_{\mathbf{h},k}\,\mathbf{P}_{\mathbf{h},k}\,\mathbf{M}^{\star,T}_{\mathbf{h},k} \tag{5.21}$$

$$\mathbf{P}_{\mathbf{yhy},k} = \mathbf{M}_{\mathbf{h},k}\,\mathbf{P}_{\mathbf{h},k}\,\mathbf{M}^{T}_{\mathbf{h},k} \tag{5.22}$$

with

$$\mathbf{M}^{\star}_{\mathbf{h},k} = \left.\frac{\partial \mathcal{M}^{\star}(\mathbf{x},\mathbf{u},\mathbf{h})}{\partial \mathbf{h}}\right|_{\mathbf{u}=\mathbf{u}_k\,;\,\mathbf{h}=\widehat{\mathbf{h}}^{-}_{k}\,;\,\mathbf{x}=\widehat{\mathbf{x}}_{k-1}} \tag{5.23}$$

where the function $\mathcal{M}^{\star}$ stands for the aggregated model $\mathcal{M}$ that outputs the states instead of the measurements.

The case of the health parameter estimation problem is somewhat simpler given that the states appear in the measurement equation (5.18) only. The variance of the measurement noise is complemented with the matrix $\mathbf{P}_{\mathbf{yxy},k-1}$ that projects the uncertainty in the states onto the measurement space. In the EKF framework, the matrix is computed through:

$$\mathbf{P}_{\mathbf{yxy},k} = \mathbf{G}_{\mathbf{x},k}\,\mathbf{P}_{\mathbf{x},k}\,\mathbf{G}^{T}_{\mathbf{x},k} \tag{5.24}$$

In our implementation of the dual Kalman filter, the matrices $\mathbf{P}_{\mathbf{xhx},k}$ and $\mathbf{P}_{\mathbf{yhy},k}$ are evaluated inside the health parameter filter while the matrix $\mathbf{P}_{\mathbf{yxy},k}$ is evaluated in the state filter. The matrices are exchanged between the filters with a one-step delay, as indicated by the time index $k-1$ in algorithms (5.1-5.4). A direct benefit of this approximation is that no additional call to the performance model is required to assess these matrices. As a side note, the error-coupling strategy extends to the UKF framework, where the unscented transform replaces the Jacobian matrices to project the uncertainty.

5.3 Improving robustness to model-plant-mismatch

5.3.1 Context

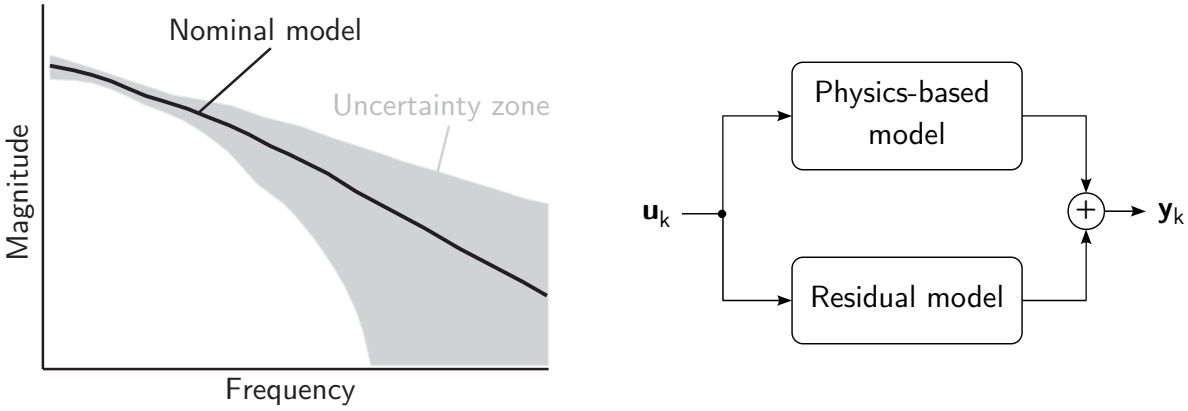
The vast majority of the literature about module performance analysis based on transient data assumes the availability of a perfectly faithful simulation model. Nonetheless, this hypothesis is rarely met in practice. In fact, complex phenomena such as heat transfer, volume dynamics, clearances, secondary airflows and power off-takes to name a few are poorly (or even not at all) modelled in current state-of-the-art aero-thermodynamic models [RTO, 2002, Nielsen et al., 2005]. Consequently, the measurements predicted with the simulation model are biased with respect to the actual measurements taken on the engine. Such biases are generally called model-plant-mismatch or modelling errors.

As reported by Volponi [2005], model-plant-mismatch has a detrimental effect on the accuracy of the estimated health parameters. We recall that the principle of operation of the Kalman filter consists, loosely speaking, in modifying the health parameters so as to drive the residuals to zero on average. If the residuals contain modelling errors in addition to the signature induced by engine deterioration, the health parameters lose their useful meaning for diagnosis and become mere *tuners* that capture the modelling errors. In this respect, Muske and Badgwell [2002] adopted such a solution in the framework of robust adaptive control. They introduced dummy parameters in their physics-based model of a distillation column to cancel the model-plant-mismatch.

In order to maintain a good diagnosis capability we have to take the model-plant-mismatch into account in our estimation procedure. Reinelt et al. [2002] review different methods used in robust identification to model modelling errors. Three approaches have emerged:

1. stochastic embedding is a frequency domain method where the non-modelled dynamics are represented by a non-stationary stochastic process whose variance increases with frequency. This comes down to encircle the nominal transfer function with an uncertainty tube whose size grows with increasing frequencies (see figure 5.4(a)),
2. model-error modelling consists in building a model of the residuals (*i.e.,* the difference between the gas turbine response and its physics-based model) by means of system identification techniques. The combination of the physics-based model and the residual model enhances the fidelity of the predictions (see figure 5.4(b)),
3. set membership provides a set of feasible models for a given batch of input-output data under the assumption of an unknown-but-bounded modelling error.

To the best of the author's knowledge, only the first two approaches made their way to module performance analysis. On one hand, Yedavalli and Li [2007] developed a robust algorithm for fault detection based on adaptive thresholding. The dynamics of the threshold is computed from the uncertainty associated to the engine model. The



(a) Stochastic embedding

(b) Model-error modelling

Figure 5.4: Modelling model-plant-mismatch

technique was found to decrease significantly the false alarm rate caused by modelling errors. Unfortunately, this work tackled only the detection part of the diagnosis task due to the qualitative nature of the stochastic embedding approach.

On the other hand, Volponi and his coworkers followed the model-error modelling approach to design so-called hybrid models. Basically, a hybrid model combines a traditional aero-thermodynamic model (or a linearised version thereof) and an empirical (black-box) component. In [Volponi, 2005], the empirical component consists in a neural network whose purpose is to reproduce the modelling errors. The neural network is trained from a database of residuals collected on an engine supposed in a reference health condition. The work reported by the team covers many issues encompassing the modelling of model biases from mild transients [Brotherton et al., 2003], the coverage of the whole flight envelope [Volponi and Brotherton, 2003] and the implementation on airworthy computers [Volponi et al., 2007]. Used in a Kalman filter framework, the hybrid model improved dramatically the accuracy of the health parameter estimates in contrast to the physics-based model alone.

Due to the finer description of model biases it provides, the hybrid model approach enhances the fidelity of the predicted measurements and leads therefore to more quantitative results than the stochastic embedding approach. For sake of completeness, we shall mention that Bird and Schwartz [1994] already advocated the concept of hybrid model as a means to achieve robust fault detection and isolation in transient operation.

Part of our research [Borguet et al., 2008] also addressed the use of hybrid models to deal with model-plant-mismatch. In that publication, we selected a response surface model as the empirical component and witnessed a noticeable improvement in the estimation accuracy for a simplified diagnosis application. In the next section, we present our methodology for taking into account the model biases in the estimation procedure and we discuss the integration of our black-box component, called the bias compensation module (BCM), with the dual Kalman filter.

5.3.2 Dealing with model biases in the estimation procedure

Going back to the state-space description, the model-plant-mismatch results from approximations in both the state transition equation (5.8) and the measurement equation (5.9) with respect to the actual behaviour of the gas turbine. A possible mathematical translation of this fact is to consider that neither $\boldsymbol{\nu}_k$ nor $\boldsymbol{\epsilon}_k$ are zero-mean, Gaussian random vectors. However, only the input/output data streams (*i.e.,* the sequence of controls $\mathbf{u}_k$ and gas-path measurements $\mathbf{y}_k$) are available in practice. From this external viewpoint, model biases look similar to sensor malfunctions.

Consequently, we propose to gather all the effects caused by model biases in the noise term of equation (5.9). The Gaussian random vector $\boldsymbol{\epsilon}_k$ has now time-varying properties and combines the sensor inaccuracy and the model biases. As can be seen from equation (5.25), both the mean and the covariance matrix of $\boldsymbol{\epsilon}_k$ are adapted. The mean value $\mathbf{b}_k$ obviously represents the average model bias for the operating point at time step k. The covariance matrix $\mathbf{R_{b,k}}$ accounts not only for the sensor uncertainty (similarly to $\mathbf{R_{y,k}}$), but also for a possible local inaccuracy of the mean bias value.

$$\boldsymbol{\epsilon}_k \in \mathcal{N}(\mathbf{0}, \mathbf{R_{y,k}}) \quad \rightarrow \quad \boldsymbol{\epsilon}_k \in \mathcal{N}(\mathbf{b}_k, \mathbf{R_{b,k}}) \tag{5.25}$$

This revised definition of the noise term $\boldsymbol{\epsilon}_k$ leads to two changes in the dual Kalman filter, one impacting the mean value and the other, the covariance of the residuals. Considering the mean value, relation (5.26) indicates that the mean bias $\mathbf{b}_k$ is added to the predicted measurements $\widehat{\mathbf{y}}_k^-$ to form the outputs of the hybrid model. Considering the covariance, the matrix $\mathbf{R_{b,k}}$ replaces the matrix $\mathbf{R_{y,k}}$ used so far.

$$\mathbf{r}_k = \mathbf{y}_k - \widehat{\mathbf{y}}_k^- \quad \rightarrow \quad \mathbf{r}_k = \mathbf{y}_k - (\widehat{\mathbf{y}}_k^- + \mathbf{b}_k) \tag{5.26}$$

$$\mathbf{R_{y,k}} \quad \rightarrow \quad \mathbf{R_{b,k}} \tag{5.27}$$

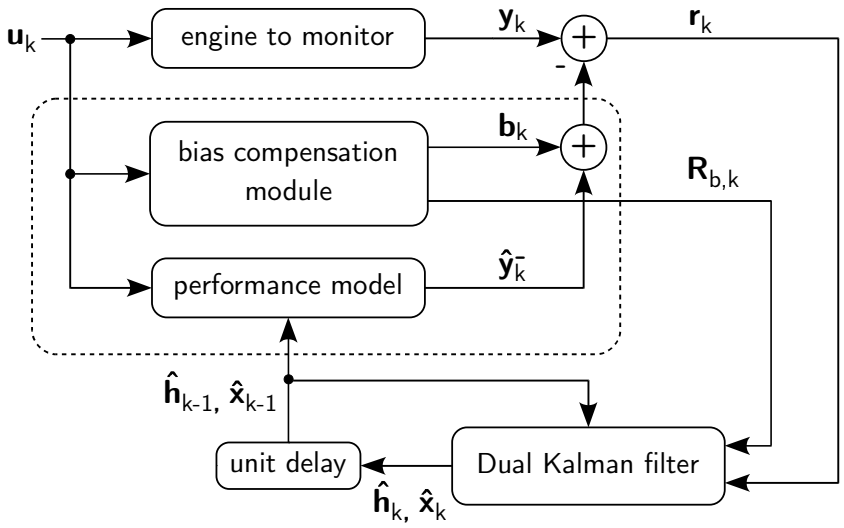


Figure 5.5: Integration of the bias compensation module with the dual Kalman filter

Figure 5.5 shows the block diagram of the modified diagnosis algorithm. Conceptually, the only change with respect to the basic algorithm of figure 5.2 resides in the addition of the bias compensation module (BCM) to the biased physics-based model to form the so-called hybrid model delimited by the dashed perimeter. Given that the performance deck alone incorporates state variables and health parameters, the bias compensation module is called outside the dual estimation procedure and its outputs $\mathbf{b}_k$ and $\mathbf{R}_{\mathbf{b},\mathbf{k}}$ are handed to each Kalman filter in the practical implementation. In the next section, we detail the structure of the bias compensation module.

5.3.3 Inside the bias compensation module

As stated above, model-plant-mismatch resembles sensor errors from an operator's point of view. However, unlike sensor errors, which are totally random, model biases have a much more predictable nature. Consequently, they can be studied beforehand, for example during the acceptance tests that every engine undergoes before it is put into service.

During this learning phase, outputs of the physics-based model are compared to the observed measurements without estimating the health parameters which are assumed to be at their reference level. As the model is not perfect, the residuals $\mathbf{r}_k$ identify with the model-plant-mismatch. The next step of the learning phase consists in characterising the mean $\mathbf{b}_k$ and the covariance matrix $\mathbf{R}_{\mathbf{b},\mathbf{k}}$ of the observed biases.

Modelling the mean bias $\mathbf{b}_k$

Modelling the mean bias $\mathbf{b}_k$ is a so-called system identification exercise. Numerous textbooks address this problem in an exhaustive way, see for instance Ljung [1999]. Basically, system identification consists in selecting a class of models $\mathcal{H}$ and determining parameters $\mathbf{p}$ of the model to fit a batch of noisy input-output data. The model for the mean bias takes the generic form of equation (5.28). The mean bias at time step k, $\mathbf{b}_k$ depends on the controls at time step k, $\mathbf{u}_k$, as well as on lagged values of itself and of the controls. In fact, the model of the mean bias is identified at the baseline health condition of the engine. The random vector $\boldsymbol{\zeta}_k$ includes the measurement noise and other possible perturbations.

$$\mathbf{b}_k = \mathcal{H}(\mathbf{u}_k, \mathbf{p}) + \boldsymbol{\zeta}_k \tag{5.28}$$

This formulation suffers from the fact that $\mathbf{b}_k$ represents the difference between two nonlinear systems that are respectively the real engine and the biased aero-thermodynamic model. To find a class of models suitable for representing these differences turns out to be quite tedious. As a palliative, we decide to identify a black-box model of the real engine instead, which comes with some advantages. First, the real engine is a nonlinear system that obeys physical laws, much like any other engineering system (*e.g.,* a satellite, an electrical engine). Second, such black-box models of the engine are used throughout

the development and operational stages of an engine's life. Third, this approach allows to bring evolutions in the performance deck without having to re-identify the bias model. On the other hand, the major flaw lies in the introduction of additional model-plant mismatch in case of a not-accurate-enough black-box model.

The identification of dynamic models of jet engines has been widely addressed both at the local (linear) and global (nonlinear) levels [Arkov et al., 2000, Evans et al., 2001, Boaghe et al., 2002, Chiras et al., 2002] due to their abundant use at the control system department of engine manufacturers as previously said. We have selected the class of Wiener models to capture the behaviour of the engine. Wiener models combine a Linear-Time-Invariant dynamic system in series with a static output nonlinearity, as sketched in figure 5.6 and have already been applied successfully to the modelling of gas turbine by other authors [Chiras et al., 2001, Kulikov and Thompson, 2004].

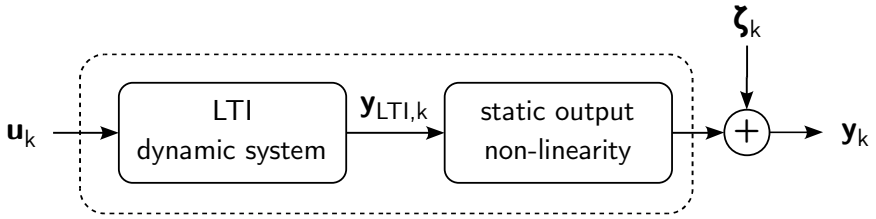


Figure 5.6: Sketch of a generic Wiener model

In the application reported in section 5.4, the dynamic part of the model is represented by a rational transfer function $H(z)$ and the static nonlinearity $h(\cdot)$ is a piecewise linear function. The parameters of the model are therefore the location and number of the poles and zeros of the transfer function as well as the location and number of breakpoints of the piecewise linear function. Mathematically, our Wiener model writes:

$$\mathbf{y}_{LTI,k} = H(z)\,\mathbf{u}_k \tag{5.29}$$

$$\mathbf{y}_k = h\left(\mathbf{y}_{LTI,k}\right) + \boldsymbol{\zeta}_k \tag{5.30}$$

where z is the discrete time shift operator

Although the engine is a true multi-input, multi-output system, we set up one Wiener model for each gas-path sensor. Considering the inputs, we mentioned earlier that the main control variable is the fuel flow, variable geometry devices such as stator vanes and bleed valves being scheduled on engine parameters such as corrected speeds. To further simplify the identification procedure, we reasonably suppose that the model biases are collected at a single atmospheric condition, the one prevailing during the pass-off test. As a result, we identify n_y single-input (the fuel flow), single-output (each of the gas-path sensors) Wiener models. As the development of algorithms dedicated to the identification of Wiener models is beyond the scope of the present work, we rely on the System Identification toolbox built in Matlab to perform the task.

Computation of the covariance matrix $\mathbf{R_{b,k}}$

Once the model of the mean bias is available, we can assess its covariance matrix $\mathbf{R_{b,k}}$. Let us recall that $\mathbf{R_{b,k}}$ is intended to capture both the uncertainty due to the measurement noise and possible inaccuracies in the model of the mean bias. Similarly to the stochastic embedding approach, we assume that the fidelity of the mean bias model decreases when the "intensity" of the transient increases. We motivate this assumption by the fact that it is relatively easier to obtain an accurate aero-thermodynamic model of (quasi-)steady-state operation than of severe transient operation (*e.g.,* slam or bodie manoeuvres). To carry on the comparison with the stochastic embedding, the confidence region around the mean bias model, represented by the covariance matrix $\mathbf{R_{b,k}}$ is here scheduled with an indicator of the "intensity" of the transient in lieu of the frequency.

For sake of effectiveness, the intensity indicator shall be a scalar quantity easy to compute from the various data streams involved in the bias compensation module. We define the transient index TI as:

$$\mathsf{TI}_k = \frac{1}{n_x} \sum_{i=1}^{n_x} \frac{\dot{\widehat{\mathbf{x}}}_k(i)}{\mathbf{x}^{ref}(i)} \tag{5.31}$$

where n_x is the number of state variables of the on-board model, $\dot{\widehat{\mathbf{x}}}_k(i)$ is the derivative of state variable i at time index k and $\mathbf{x}^{ref}(i)$ is the reference value of the ith state variable (*e.g.,* at take-off rating). The units of TI are s^{-1}. The normalisation by a reference state value is required given the different orders of magnitude of the state variables. The transient index is zero in steady-state operation, positive when the engine is accelerating and negative otherwise.

Equation (5.31) shows that the transient index is computed from the state derivatives provided by the biased performance deck rather than from measurements taken on the real engine. Three reasons dictate this choice: first, the engine model produces noise-free signals and outputs directly the state derivatives. Second, not all state variables are measurable on the real engine (*e.g.,* the metal temperatures involved in the heat transfers), but are available in the engine model. Third, it avoids the computation of numerical derivative, a process that is generally very sensitive to the noise level. It would be illusory to evaluate the functional dependence of $\mathbf{R_{b,k}}$ with respect to TI. Therefore, we chop the TI axis in a small number of segments (*e.g.,* sharp deceleration, mild deceleration, quasi-steady, mild acceleration, sharp acceleration) and evaluate one covariance matrix per interval.

The procedure to compute $\mathbf{R_{b,k}}$ over a generic segment $[\mathsf{TI}^{min}, \mathsf{TI}^{MAX}]$ is given in algorithm 5.5. As a first step, the gap between each data point and the mean bias is computed for all data points belonging to the particular segment (lines 2 to 5, note that each vector $\mathbf{e}_i$ is a $ny \times 1$ vector). Finally, each element of the symmetric covariance matrix is obtained using the well-known definition of the covariance of random variables (line 6).

Algorithm 5.5 Covariance matrix computation

1: **set** $N = 0$
2: **for all** k such that $\mathsf{TI}^{min} < \mathsf{TI}_k \leq \mathsf{TI}^{MAX}$ **do**
3: $N = N + 1$
4: $\mathbf{e}_N = \mathbf{r}_k - \mathbf{b}_k$
5: **end for**
6: $\mathbf{R_b} = \dfrac{1}{N-1} \sum_{i=1}^{N} \left\{ \mathbf{e}_i \ \mathbf{e}_i^T \right\}$

It should not be surprising that some off-diagonal terms of the matrix $\mathbf{R_{b,k}}$ are non-zero. This simply indicates that the model biases are inter-dependent as the modelling errors introduce some relationships between the measurements. For instance the bias on the EGT sensor measurement is linked to that of low pressure spool speed, since temperature recovery factor for thermocouples varies with mass flow.

Extension of the BCM to other flight conditions

The bias model is built from a database collected at one point in the flight envelope and is therefore valid at that particular point. However, the engine will obviously be operated at several other points in its flight domain during its life. This raises the question of the extension of the bias model to the whole operating envelope.

A first solution would be to collect additional mismatch data at a number of atmospheric conditions, but we already mentioned the difficulty to achieve it in practice. Another solution consists in having recourse to the so-called corrected parameters which rely on similarity laws and first order approximation of the aero-thermodynamic processes. The idea behind similarity laws is to define dimensionless groups of parameters that are associated with the flow field in the engine. These corrected parameters allow comparison, generally on a Mach number basis, of the performance of the engine operating under different atmospheric conditions. As a reminder, the general expression to correct a parameter X is given by:

$$\mathsf{X}^{co} = \frac{\mathsf{X}}{\theta^a \ \delta^b} \tag{5.32}$$

where $\theta = \mathsf{T2}/288.15$ K and $\delta = \mathsf{P2}/101325$ Pa.

Parameter correction is a well-established practice and the theoretical values of exponents a and b for the variables of interest can be found in many references, see for instance [Walsh and Fletcher, 1998] or [Volponi, 1999]. Yet, additional effects, such as the modification of the thermo-physical properties of the working fluid, Reynolds number effects or geometrical effects (*e.g.,* clearance, blade untwist) make the engine behaviour deviate from the assumptions of the Mach number similarity. As a remedy, the a and b exponents for each parameter involved in the BCM can be fine-tuned as explained by Kurzke [2003] to further improve the accuracy.

Practical implementation

Figure 5.7 shows the different components and main data streams of the bias compensation module. As a first step, the main input (fuel flow) is corrected to ISA-SLS conditions before entering the core of the BCM. Note that all variables appearing inside the dashed perimeter are corrected quantities.

The corrected fuel flow feeds on one hand the Wiener models that generate corrected outputs $\widehat{\mathbf{y}}_{W,k}$ representative of the real engine at baseline health condition. On the other hand, the biased performance deck (the same that is used in the dual Kalman filter) also receives the corrected fuel flow on input and outputs gas path measurements $\widehat{\mathbf{y}}_k^{hl}$ for ambient conditions set to ISA-SLS and health condition fixed at its reference value. The corrected mean bias is computed as the difference between the Wiener models and the performance deck outputs. It is then "un-corrected" back to the current atmospheric conditions to provide the actual mean bias $\mathbf{b}_k$ that is used in the computation of the residuals, see equation (5.26).

The biased physics-based model also outputs the state derivatives $\widehat{\dot{\mathbf{x}}}_k^{hl}$ that are converted into the transient index TI_k defined previously. Finally, the adequate covariance matrix $\mathbf{R}_{\mathbf{b},\mathbf{k}}$ is selected from a predefined bank based on the transient index value.

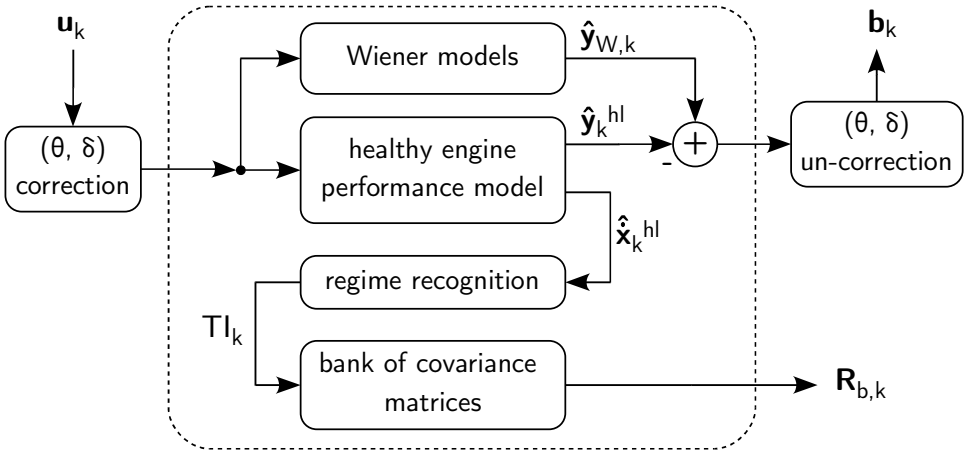


Figure 5.7: Structure of the bias compensation module

As can be noted from figure 5.7, the mean bias $\mathbf{b}_\mathbf{k}$ is "un-corrected", but not its associated covariance matrix $\mathbf{R}_{\mathbf{b},\mathbf{k}}$. Doing so, we implicitly assume that the uncertainty in the mean bias is constant throughout the flight domain. It means that the highest contribution to the uncertainty is the measurement noise (the magnitude of which is here supposed independent of the operating conditions) rather than the inaccuracy in the mean bias model.

5.4 Application of the methodology

This section is dedicated to the evaluation of the potential benefit that can be achieved from the processing of transient data by means of the dual Kalman filter presented above. For the very same reasons as in chapter 4, the assessment is again carried out on the basis of simulated test-cases. In section 5.4.1, we describe the transient capabilities offered by the OBIDICOTE model and present the test-cases. Section 5.4.2 provides a comparison of the estimation accuracy of the transient method with respect to the steady-state one in the case of a perfect performance deck. Finally, section 5.4.3 illustrates the enhancement in estimation accuracy offered by the hybrid model approach in the case of a biased performance deck.

5.4.1 Setting up the engine model and the test-cases

The OBIDICOTE model is used both at the test-case generation stage, mimicking a real engine, and at the data processing stage, as a classical performance deck. For sake of clarity and with some abuse of language, we will therefore refer to the OBIDICOTE model as the "real engine" when it is used for data generation whereas we will name it the "on-board model" when it is used for data processing.

For the creation of the transient test-cases, we add a basic control system to the real engine. The percent corrected fan speed PCNF serves as the thrust setting parameter. It is indeed well correlated with the fan airflow and, as such, with the delivered thrust given that the fan is responsible for more than 80% of the total thrust in high by-pass ratio turbofan engines. The control system and the PCNF scale are presented in appendix B. The engine control system translates the PCNF request into a fuel flow demand. The variable geometry devices (vbv and vsv) still follow a static schedule based on the corrected spool speeds.

At the test-case generation stage, the atmospheric conditions are again specified through the triplet formed by pressure altitude ALT, flight Mach number MN and the temperature offset with respect to a standard day DT_{ISA}. On the other hand, the on-board model is fed with the inlet total pressure P2 and temperature T2 and the ambient pressure Pamb. As the collection of transient data can theoretically take place anywhere in the flight envelope, a complete description of the thermodynamical state of the ambient air should include an information about its moisture content. In fact, the humidity level of a mass of air modifies its density and consequently affects the engine performance. Although neglected in the simulated test-cases, the effect of humidity could quite easily be accounted for as we will see below.

The deterioration level of the engine is described through ten health parameters which are correction factors on the efficiency (SEi) and flow capacity (SWi) of each rotating component as illustrated in figure 5.8. Unlike in the steady-state test-cases, we do not

consider here parameters related to the working condition of the variable geometry devices. Let us take the case of the bleed valves. When functioning nominally, they are open at low power (idle regime) and progressively closed as the engine reaches full power (take-off regime). In the case of a failed bleed valve that is stuck open, it means that a parameter attached to its health condition will vary depending on the operating point. For sake of simplicity, we restrict to health parameters that are constant throughout the test-case.

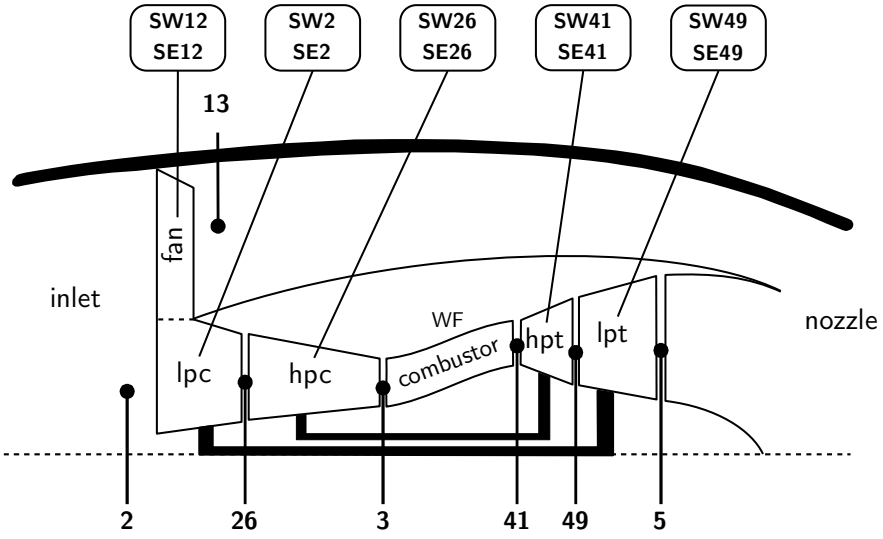


Figure 5.8: Turbofan layout with station numbering and health parameters location

Although a version of the model including gas dynamics is available, we discard its use for two reasons. First it requires tiny integration steps (less than a millisecond), which increases dramatically the computing load. Second, gas dynamics effects are negligible for normal operation of the engine. Consequently, the transient effects taken into account in the model are the inertia of the low- and high-pressure spools and the heat soakage in the components of the gas generator (*i.e.,* hpc, combustor and hpt) leading to the set of seven state variables listed in table 5.1.

Label	Description	Units
NF	fan speed	rpm
NC	core speed	rpm
TM3b	hpc blade temperature	K
TM3c	hpc casing temperature	K
TM4c	burner casing temperature	K
TM41b	hpt blade temperature	K
TM41c	hpt casing temperature	K

Table 5.1: State variables

The instrumentation selected to perform the diagnosis is the same as in the steady-state test-cases and is recalled in table 5.2, the first seven sensors are located in the gas-path and

the last four define the operating point. The noise level of the gas-path sensors is supposed to be independent of the operating point which causes lower signal-to-noise ratios at low regimes. The operating point is assumed to be measured without any uncertainty.

Label	Description	Uncertainty
P13	fan outlet total pressure	±150 Pa
T26	lpc outlet total temperature	±2 K
P3	hpc outlet total pressure	±5000 Pa
T3	hpc outlet total temperature	±4 K
NF	fan speed	±3 rpm
NC	core speed	±6 rpm
T49	lpt inlet total temperature	±8 K
WF	fuel flow	±0 g/s
Pamb	ambient pressure	±0 Pa
P2	fan inlet total pressure	±0 Pa
T2	fan inlet total temperature	±0 K

Table 5.2: Gas-path and operating point instrumentation.

Figure 5.9 sketches the procedure to generate the scenarios. Each test-case combines the same transient manoeuvre, namely a typical take-off sequence, and a unique health condition that is constant over the whole duration of the scenario. We add Gaussian noise, whose magnitude is specified in table 5.2, to the clean, simulated outputs to make them closer to real data. We set the sampling frequency to 50 Hz which is the usual update rate of most of the current digital engine control systems. Additionally, the associated discrete time step (20 milliseconds) allows a stable temporal integration of the dynamic model with any explicit integration method.

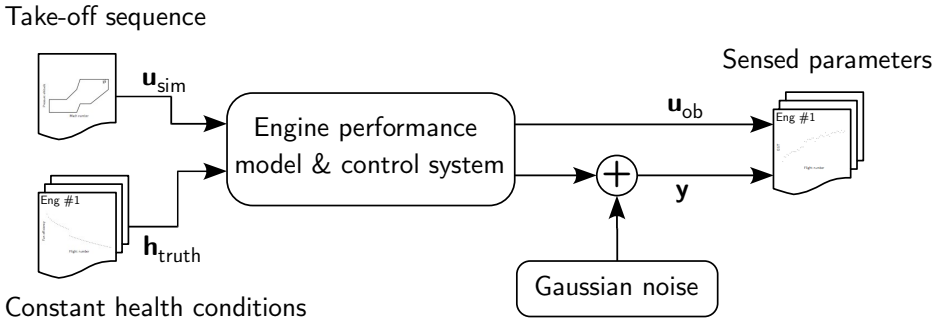


Figure 5.9: Schematic of the scenario generator

The transient manoeuvre is representative of a take-off sequence on a standard day ($\mathsf{DT_{ISA}}$= 0 K) from an airport located at sea level. Figure 5.10 depicts the evolution of the requested PCNF, the altitude and the aircraft speed (in terms of Mach number) with respect to time. The engine is initially at the Min Idle regime (PCNF = 40%). At

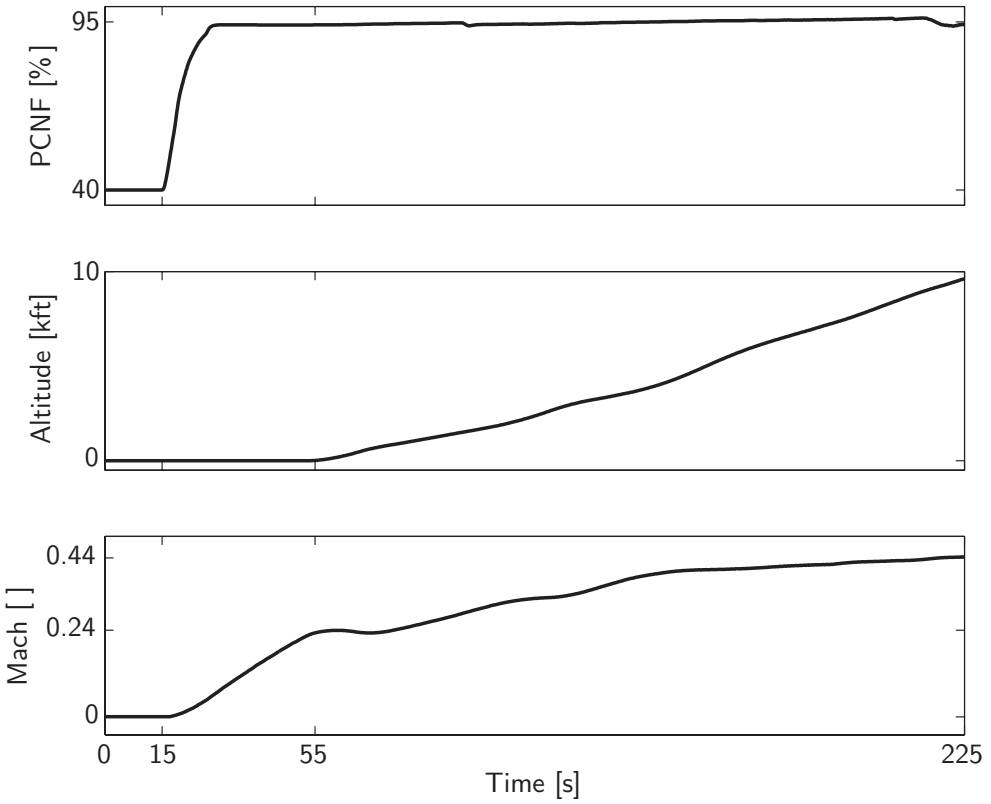


Figure 5.10: Typical take-off sequence from an airport at sea level, standard day

time t = 15 s, the pilot activates the take-off mode which causes the engine to progressively accelerate to PCNF = 95%. This value depends on many factors such as aircraft weight, ambient air temperature and runway properties and is nowadays computed by the flight management computer of the aircraft. At time t = 55 s, the aircraft has gained a sufficient speed so that it is able to lift off. The scenario terminates arbitrarily at time t = 225 s when the aircraft reaches an altitude of 10 kft (3048 m).

On a side note, let us mention that the effect of humidity could be taken into account relatively easily for the take-off roll part of the test-case, that is up to time t = 55 s. In fact, this portion of the scenario happens in a narrow geographic area and at a constant altitude so that the humidity content of the surrounding mass of air can be regarded as uniform. Although neither the engine, nor the aircraft are fitted with a hygrometer, this information is available in the weather reports issued at the airport and could therefore be used as an additional input to the on-board model.

As mentioned above, each scenario is characterised by a unique health condition that is constant throughout the complete take-off sequence. In fact, the duration of the test-case is far too short for engine wear to change appreciably. Moreover, we suppose that no accidental event, for instance the ingestion of a foreign object on the runway, occurs during the first 225 seconds of the flight. At the test-case generation stage, the health condition is specified by means of a deterioration level. As shown in figure 5.11, we have

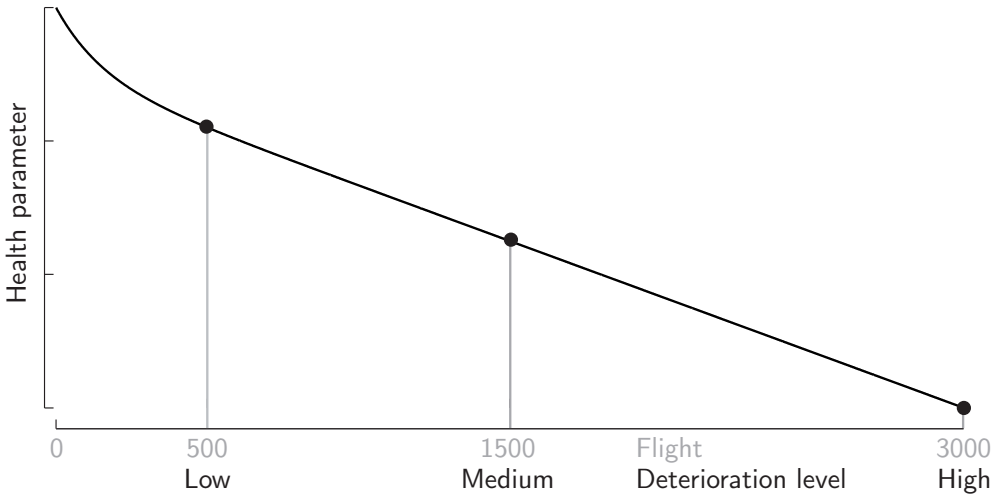


Figure 5.11: Definition of the three deterioration levels from the wear model

defined three deterioration levels along the wear model presented in section 4.3.2: low, medium and high corresponding to the ageing of an engine after respectively 500, 1500 and 3000 cycles. Given that all parameters defining the wear model are random variables, we effectively have an envelope of possible values for each health parameter at each of the three deterioration levels.

In the specific scope of performance trending, the crucial piece of information that we are interested in is the health condition. On the contrary, we have little to no interest in the state variables, which we see as a secondary quantity that we are however forced to estimate due to the formulation of the problem. Consequently, we assess the performance of the transient methodology only from the health parameter side, which allows furthermore a comparison with the results obtained from steady-state data.

To gauge the accuracy of the health parameter estimates, we rely on the *Root Mean Square (estimation) Error* as defined in equation (5.33) for the ith health parameter. This scalar is a measure of the average gap between the estimated and true health condition and captures both the bias and the variance components of the estimation error.

$$RMSE_i(i_{end}) = \sqrt{\frac{1}{N_{scn} \cdot N_w} \sum_{j=1}^{N_{scn}} \sum_{k=i_{end}-N_w+1}^{i_{end}} \left(\widehat{\mathbf{h}}_k^j(i) - \mathbf{h}_k^j(i)\right)^2} \tag{5.33}$$

where

- N_{scn} is the number of scenarios,
- N_w is the length of the window, here $N_w = 100$,
- i_{end} is the discrete time index at which the evaluation window ends,
- $\widehat{\mathbf{h}}_k^j(i)$ is the estimated value of the ith parameter at time step k of scenario j,
- $\mathbf{h}^j(i)$ is the true value of the ith parameter over the whole jth scenario.

As can be seen from equation (5.33), the RMSE is evaluated at a certain moment, for instance at lift-off, represented by its discrete time index i_{end}. The computation is carried out for a certain number of scenarios N_{scn} and over a window spanning the $N_w = 100$ latest data points (corresponding to a time interval of two seconds) in order to average possible fluctuations in the estimates.

To assess the performance of the transient methodology, we generate 10 instances of each of the three deterioration levels. For each of the 30 health conditions, we consider 10 different realisations of the measurement noise. This leads to a grand total of 300 take-off scenarios. In order to have a fair basis of comparison between the steady-state and transient approaches, we generate an equivalent number of steady-state scenarios based on the same batch of health conditions. Each steady-state scenario consists of data simulated with the steady-state model to which Gaussian noise is added. The operating point is the same throughout all scenarios. It is taken as the mean cruise condition defined in table 4.18. A steady-state scenario contains the same number of data samples as the take-off sequence, that is 11251 points.

5.4.2 Comparison of Kalman filters

In this section, we consider that a perfect model of the engine is available. We process the batch of 300 transient scenarios with the dual Kalman filter based either on two extended or two unscented filters. We do not consider mixed configurations (*e.g.,* EKF for the parameters, UKF for the states).

For all scenarios, we initialise the health parameter distribution in the following way: the mean value is set at the fleet average baseline value, that is a value of one for each health parameter. The diagonal covariance matrix is based on an a priori assumed standard deviation of one percent for each parameter. The covariance matrix of the process noise was set to the identity matrix multiplied by a tiny factor of 10^{-12}. This provides a mobility of roughly one percent over the full duration of the scenario.

On the other hand, we initialise the state distribution as follows: the mean value is set at the steady-state value predicted by the on-board model for the initial value of the health parameters. Therefore, we totally overlook the fact that two of the states, namely the spool speeds, are actually measured and could be initialised more accurately. The selection of the covariance matrix turns out to be much more difficult, especially in the case of the UKF. Indeed, setting a too large value results in a non-convergence of the on-board model during the unscented projection. After some trials, we end up with a covariance matrix equal to the identity matrix multiplied by a factor 100. The covariance of the associated process noise is then set to one percent of the initial covariance matrix of the health parameters.

Considering the steady-state scenarios, we process them with the monitoring tool presented in chapter 3 in which we disable the fault detection and isolation modules given

that the health condition is constant throughout the test-case. We initialise the health parameter distribution in the exact same way as for the transient scenarios.

Tables 5.3 and 5.4 report the RMSE, as defined in equation (5.33), expressed in percent of the fleet average baseline value achieved respectively with the extended and unscented filters. The last row sums up the ten previous ones to provide a global measure over the full set of health parameters. Column-wise, both tables are organised in three main blocks, corresponding to the three deterioration levels defined previously (low, medium, high). Each block is itself subdivided into three columns corresponding to the evaluation of the RMSE at the three moments defined below:

TR1	time of lift-off in the transient scenarios (t = 55 s)
TR2	final time in the transient scenarios (t = 225 s)
SS	final sample in the steady-state scenarios

The first valuable result to extract from table 5.3 is the neat superiority of the transient approach with respect to the steady-state one in terms of estimation accuracy. Ratios between the RMSE evaluated at the end of the transient sequence (TR2 columns) and at the end of the steady-state test-case (SS columns) range between two and ten after the processing of an equal number of data samples. We explain this dramatic improvement by the wealthier amount of information provided by the unsteady data sequence, that covers a substantial part of the operating regimes of the engine, compared to the one provided by steady-state data captured at a single cruise condition.

Focusing more specifically on the transient results, the estimates at the end of the test-case (TR2 columns) are more accurate than the estimates at lift-off (TR1 columns) by a factor of roughly 1.55, whatever the deterioration level is. The reason behind this improvement is basically due to the larger number of data samples processed over the complete test-case than on part of it which allows a better filtering of the measurement noise. The reduction in RMSE between the moments TR1 and TR2 is mainly noticeable on the flow capacity factors SWi. The trend is less obvious for the lpt flow factor given its relatively lower deterioration magnitude with respect to the other rotating components.

Table 5.3 shows that at a given moment (TR1, TR2 or SS), the RMSE increases with stronger deterioration level. This loss in estimation accuracy is due to the fact that the health parameter distribution is initialised independently of the deterioration level. As a result, the initial estimation error grows with stronger deterioration level. Moreover, the bandwidth of the filter estimating the health parameters, which is essentially governed by the initial covariance of the parameters and by the covariance of the process noise (given that the noise level is constant), could possibly be too low to capture the larger changes in the health parameters within the duration of the test-case. We can indeed observe this reduction in bandwidth in table 5.3: the overall RMSE at time TR2 for the medium deterioration level is almost equal to the overall RMSE at time TR1 for the low deterioration level. In practice, we can expect a better accuracy at the medium and high deterioration levels given that the mean value of the health parameters would be initialised to the most recent available estimate of the engine condition.

	Low			**Medium**			**High**		
	TR1	TR2	SS	TR1	TR2	SS	TR1	TR2	SS
SW12	0.110	0.054	0.224	0.132	0.065	0.389	0.240	0.105	0.529
SE12	0.092	0.084	0.431	0.114	0.104	0.767	0.224	0.182	1.327
SW2	0.047	0.014	0.033	0.096	0.015	0.033	0.088	0.017	0.035
SE2	0.038	0.033	0.331	0.038	0.055	0.369	0.053	0.061	0.542
SW26	0.053	0.026	0.269	0.120	0.049	0.302	0.132	0.054	0.450
SE26	0.022	0.016	0.159	0.028	0.025	0.178	0.039	0.029	0.260
SW41	0.050	0.007	0.022	0.098	0.009	0.022	0.113	0.011	0.025
SE41	0.055	0.045	0.114	0.112	0.084	0.120	0.138	0.106	0.163
SW49	0.098	0.077	0.205	0.125	0.142	0.217	0.204	0.177	0.298
SE49	0.095	0.065	0.275	0.167	0.128	0.502	0.218	0.143	0.913
Sum	**0.659**	**0.421**	**2.063**	**1.031**	**0.676**	**2.899**	**1.448**	**0.883**	**4.542**

Table 5.3: Performance of the extended Kalman filter in terms of percent RMSE

	Low			Medium			High		
	TR1	TR2	SS	TR1	TR2	SS	TR1	TR2	SS
SW12	0.146	0.078	0.223	0.188	0.108	0.398	0.386	0.244	0.535
SE12	0.123	0.114	0.430	0.176	0.168	0.784	0.399	0.395	1.340
SW2	0.043	0.018	0.030	0.063	0.021	0.034	0.124	0.023	0.034
SE2	0.088	0.090	0.319	0.082	0.087	0.370	0.123	0.148	0.541
SW26	0.064	0.071	0.258	0.053	0.064	0.305	0.094	0.112	0.448
SE26	0.099	0.038	0.153	0.200	0.041	0.180	0.274	0.072	0.262
SW41	0.028	0.010	0.023	0.038	0.010	0.023	0.076	0.009	0.024
SE41	0.055	0.046	0.115	0.120	0.075	0.128	0.135	0.111	0.175
SW49	0.102	0.081	0.204	0.125	0.128	0.234	0.204	0.200	0.317
SE49	0.109	0.083	0.276	0.188	0.161	0.503	0.265	0.246	0.916
Sum	**0.863**	**0.629**	**2.033**	**1.232**	**0.862**	**2.959**	**2.081**	**1.560**	**4.592**

Table 5.4: Performance of the uscented Kalman filter in terms of percent RMSE

Taking a look at table 5.4, we realise that similar comments apply to the results obtained with the unscented Kalman filter. Contrary to the expectations, the RMSE's of the UKF are slightly higher than those of the EKF, especially on the efficiency factors of the compressors (SE12, SE2, SE26). The last row of the table shows that the global RMSE is 0.2% higher for the low and medium deterioration levels and up to 0.7% higher for the high deterioration level. We explain this lower performance of the UKF by a very touchy behaviour of the performance deck during the unscented projection.

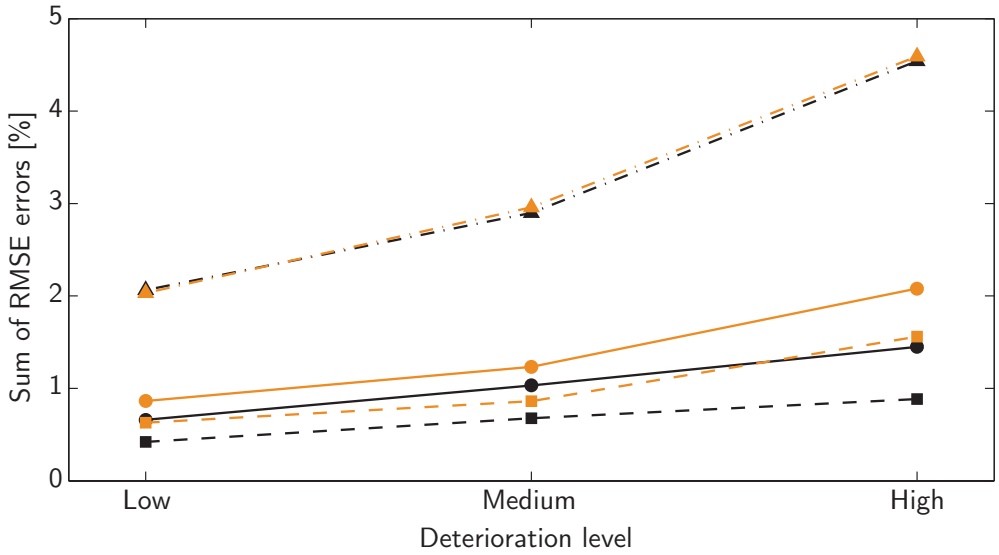


Figure 5.12: Comparison of the overall estimation accuracy between transient and steady-state approaches in terms of global percent RMSE. Black lines are for the EKF, orange lines are for the UKF. Plain lines are for evaluation at time TR1, dashed lines are for evaluation at time TR2, Dash-dotted lines are for evaluation at time SS

Figure 5.12 gathers the information contained in the last row of tables 5.3 and 5.4. Black lines are related to the extended filters and orange lines are related to the unscented filters. There are three families of broken lines related to the three moments at which the RMSE is evaluated: plain lines for time TR1, dashed lines for time TR2 and dash-dotted lines for time SS. For steady-state data, the performance of the EKF and UKF is essentially the same. This is in line with the results obtained in section 4.5.1. For transient data, the gap between the EKF and the UKF is larger for the reason explained above. The improvement in estimation accuracy between time TR1 and TR2, should be traded-off with the additional computational cost and data management.

Furthermore, we note that the slope of the broken lines related to transient test-cases is smaller than those related to steady-state test-cases, especially considering the segment between the medium and high deterioration levels. We can conclude that with the same initialisation procedure of the health parameter distribution, the transient methodology has better exploration properties than the steady-state one. This is a consequence of the fact that the transient algorithm processes data representative of a myriad of operating points of a non-linear system.

Figure 5.13 illustrates the results that are obtained by running the dual extended Kalman filter on one of the test-cases at the high deterioration level. Each subplot depicts the evolution of the absolute value of the estimation error (orange line) and the 3σ confidence bound (black line) for a given health parameter. The confidence bound is computed as three times the square root of the diagonal terms of the covariance matrix of the health parameters. Both quantities are expressed in percent.

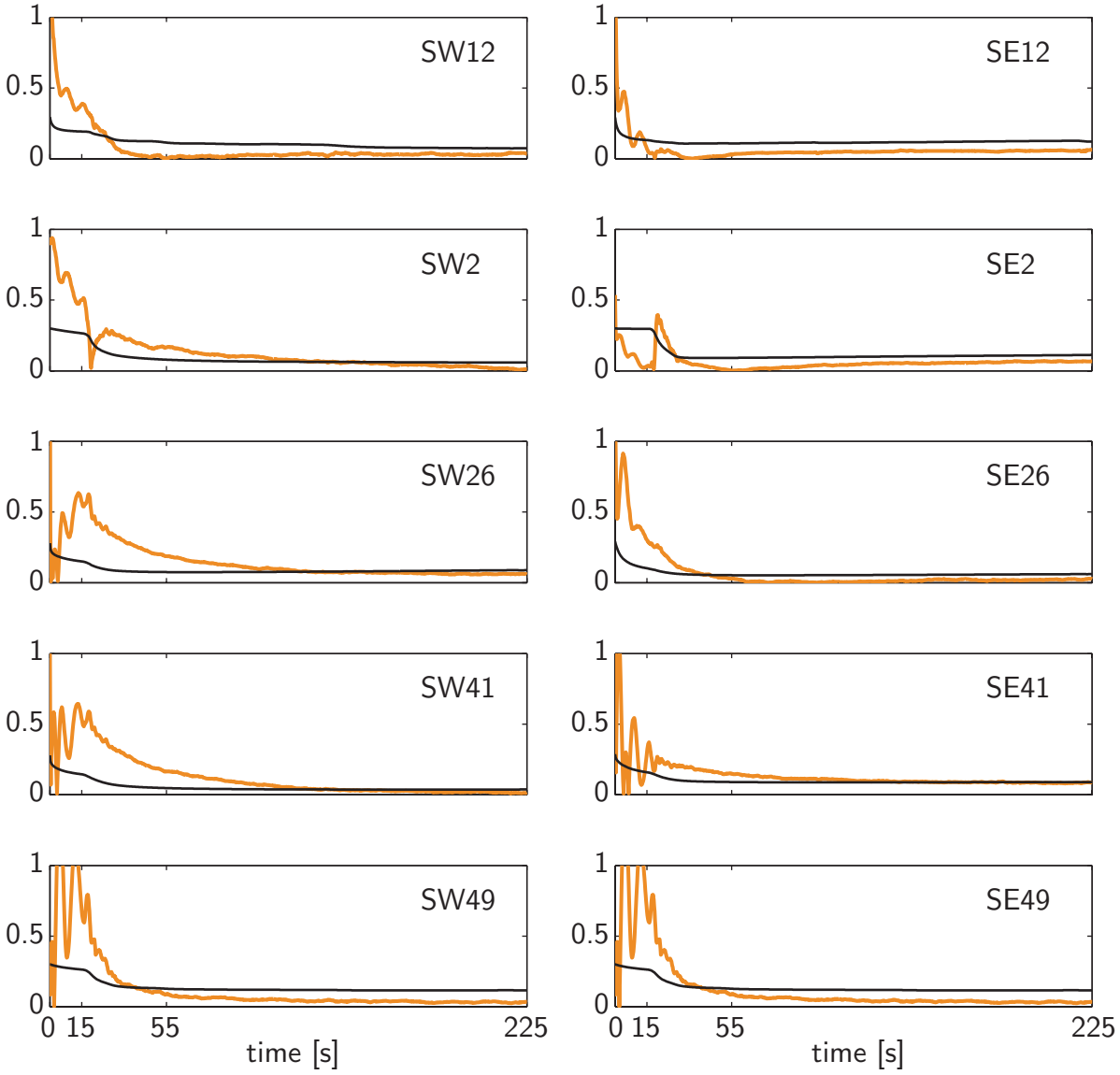


Figure 5.13: Absolute estimation error (orange line) and 3σ confidence bound (black line) for the health parameters – Values expressed in percent – Dual EKF algorithm – high deterioration level

Focusing on the first 15 seconds of the test-cases *i.e.,* before any transient manoeuvre is initiated, the convergence of the estimates toward the actual health condition is relatively poor. Although the estimation error indeed decreases for the health parameters related to the fan and the lpc, we observe rather severe oscillations for the parameters linked to

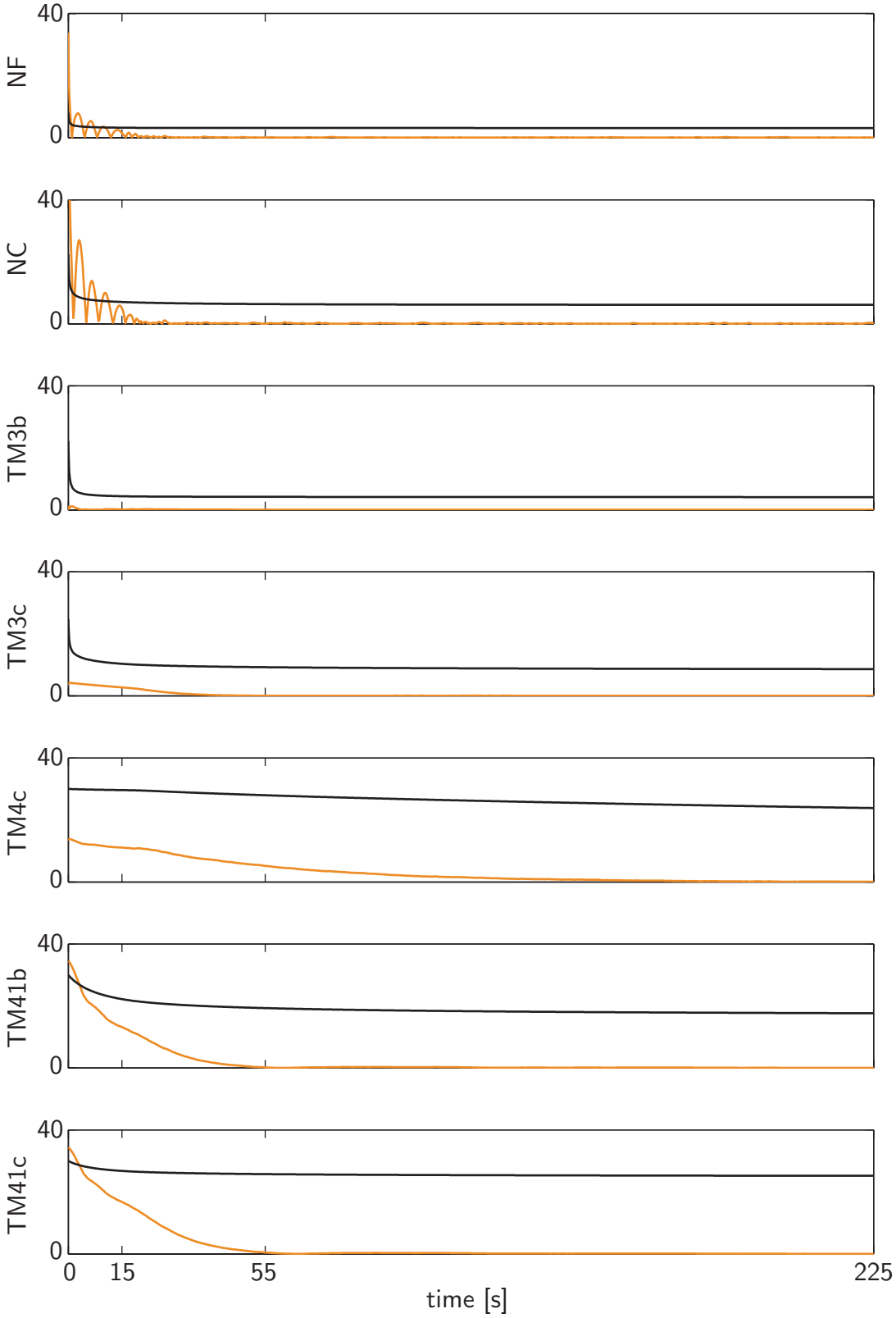


Figure 5.14: Absolute estimation error (orange line) and 3σ confidence bound (black line) for the state variables – Values expressed in percent – Dual EKF algorithm – high deterioration level

both turbines. Besides, the estimation error increases on the health parameters associated to the hpc. Shortly after the take-off mode is engaged (around t = 17 s), we see a dramatic improvement in the convergence of the health parameters toward the solution. This illustrates the (probably obvious) statement that the transient methodology is most effective when processing transient data. The parameters with the lowest convergence speed are SW26, SW41 and SE41, which is not surprising given that they are the parameters undergoing the largest change in magnitude.

At the time of lift-off the estimation error on most health parameters except SW2, SW26, SW41 and SE41 are within the confidence bound which explains the slight improvement in overall RMSE between times TR1 and TR2. By the end of the test-case, the estimation error of all the parameters is inside the 3σ region. Therefore, we can say the solution is statistically meaningful.

To complete the analysis of that specific test-case, figure 5.14 shows the results from the state filter. Each subplot depicts the evolution of the absolute value of the estimation error (orange line) and the 3σ confidence bound (black line) for a given state variable. The confidence bound is computed as three times the square root of the diagonal terms of the covariance matrix of the state variables. Both quantities are expressed in actual units *i.e.,* rpm for the spool speeds and Kelvin for the metal temperatures.

The initial estimation error is relatively large on the spool speeds and on the hpt metal temperatures. The spool speeds converge quite rapidly to their true values, even during the first 15 seconds of steady-state operation of the scenario, given that they are measured. We notice a very small estimation error on the metal temperatures of the hpc, which seems to indicate that this particular case of deterioration does not affect appreciably the temperatures in the cold section of the engine. The convergence speed of the metal temperatures correlates with the characteristic time of the heat transfers on each component of the gas generator: the hpc has the fastest dynamics, due to its relatively low mass, the hpt ranks at the second position and the combustor has the largest thermal inertia due to its large volume.

Considering the uncertainty in the state estimates, the spool speeds achieve the lowest level, since they are measured. Then come the metal temperatures of the hpc, the hpt and the combustor in that particular order. The converged values of the confidence bound are related to the observability of the different state variables.

For indicative purpose, table 5.5 reports the average computing time for one scenario (11251 data samples) in the transient and steady-state cases. More than the absolute values, which are highly dependent on hardware as well as software characteristics, the relative change in computational load is the main information to extract from this table.

In the steady-state case, both the EKF and the UKF have similar computing demand as was already witnessed in table 4.9. In the transient case, the UKF is almost 50 % slower than the EKF although the performance deck is called the same number of times in both algorithms. Unlike the computation of the Jacobian matrices, which involves

tiny perturbation steps, the unscented transformation literally challenges the internal solver of the OBIDICOTE model with the larger perturbations in both states and health parameters it uses.

	EKF	UKF
Steady-state	157	195
Transient	551	816

Table 5.5: Average computing time expressed in seconds per scenario (11251 data samples)

Comparing the computing times of the EKF's in the steady-state and transient cases, we notice that the average time to process 11251 data points rises by a factor of about 3.5. Most of this increase comes from the higher number of calls to the performance deck for each data sample, which amounts to 21 in steady-state and 51 in transient operation. In the algorithms, the most computationally demanding task is in fact running the performance deck.

5.4.3 Robustness against modelling errors

In this section, we investigate the ability of the hybrid model approach to provide accurate estimate of the engine condition in the presence of modelling errors. As we work with simulated data, the first step is to introduce model-plant-mismatch in the on-board model so that its response is different from the real engine. Second, we briefly discuss some issues related to the building of the bias compensation module. Finally, we report some results obtained on the transient scenarios described previously.

Introducing bias in the on-board model

So far, we hypothesised that the on-board model could predict perfectly *i.e.,* without any substantial bias, the response of the real engine. This situation is however never met in practice despite the continuous efforts deployed by the manufacturers to improve the fidelity of their decks. To study the effect of model-plant mismatch on the estimation accuracy of the dual Kalman filter with simulated data, we therefore have to introduce modelling errors in the on-board model.

As far as the dynamic effects are concerned, the on-board model incorporates the rotating inertia of both spools and a partial modelling of the heat soakage processes that is confined to the hpt. This leads to the set of three state variables listed in table 5.6. Heat transfer processes are the slowest dynamics in a turbine engine and they influence quite strongly its transient response (see Nielsen et al. [2005]). In the present case, we account for the heat transfer in the hpt since it is the hottest part of the engine and therefore the thermal effects are expected to be more important than in the other components.

Label	Description
NF	fan speed
NC	core speed
TM41	bulk hpt metal temperature

Table 5.6: State variables of the on-board model

The incomplete modelling of the heat transfers generates model-plant-mismatch only during transient operation of the gas turbine. To spoil the steady-state predictions (and as a side effect the transient behaviour), we introduce on one hand a power off-take on the high pressure spool, that could result from an incorrect bookkeeping of the power consumption of the auxiliary equipment (such as fuel and oil pumps). On the other hand, we introduce a bleed leak at the outlet of the hpc that could result from neglecting some bleed air demands such as cabin climate control or anti-ice systems for instance.

Building and validation of the bias compensation module

To build the empirical component of the hybrid model, previously termed the bias compensation module, we first create a learning set of data. We assume that the data collection occurs during the acceptance testing, that is when the real engine is in its "new and clean" condition. The testing takes place on a standard day at sea level.

One of the most critical issues in system identification concerns the choice of the input signal, whose content should be sufficiently rich to allow all major features of the engine response to be identified. For non-linear systems, this is still an open question whose solution is well beyond the scope of this work. Therefore, we consider the manoeuvre sketched in figure 5.15 as the input sequence. The first part of the manoeuvre (up to t = 350 s) is a sharply shortened version of a typical cycle used in endurance testing. The plateau at 100% PCNF is representative of the take-off, climb and cruise portions of a flight, the plateau at 45% PCNF is representative of the descent and approach phases, the short hump at 75% PCNF accounts for thrust reversing. A staircase-shaped power sweep complements the sequence. The database contains ten realisations of the input sequence in order to reduce the effect of the measurement noise.

For each of the seven gas-path sensors we identify a single-input (fuel flow) single-output (said sensor) Wiener model of the response of the real engine. If this approximation simplifies greatly the identification task, it unfortunately comes at a price: we loose the correlation between the response of the various sensors that embeds some of the physics of the turbine engine. We do not automate the procedure to determine the optimal level of complexity of the Wiener models (*i.e.,* number of poles and zeros of the transfer function, number of segments for the piece-wise linear static non-linearity). Instead we start with the lowest possible model complexity and progressively increase the number of parameters. Once the goodness of fit does not appreciably vary, we stop the process. The most complex model we end up with has three poles, two zeros and an eight-segment

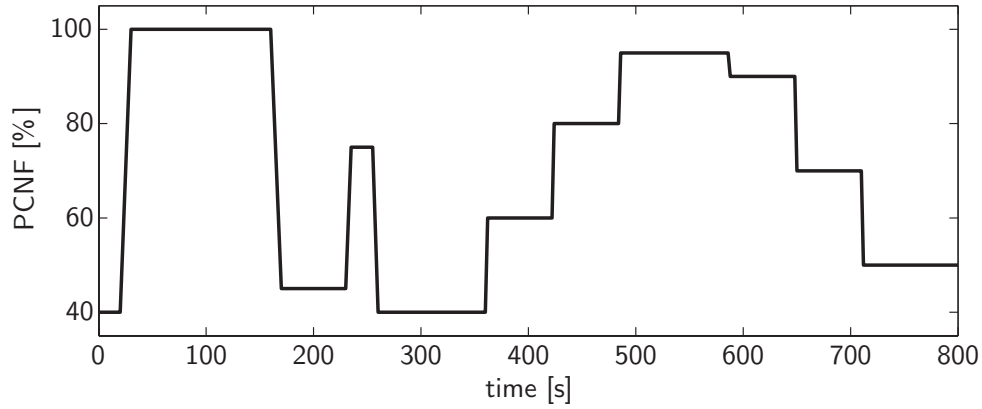


Figure 5.15: Sequence of requested PCNF used to build the bias compensation module

non-linearity. The next step consists in creating the bank of covariance matrices from the residuals between the Wiener models and the actual recordings. We split the TI axis in five portions corresponding to fierce deceleration, mild deceleration, quasi-steady-state, mild acceleration and fierce acceleration.

As a validation procedure of the bias compensation module, we compute the residuals between the real engine and the hybrid model for the take-off sequence at sea level and compare them to the residuals obtained with the biased model alone. During this validation step, we assume that the real engine is still in its "new and clean" condition so that the residuals represent actually the modelling errors.

Figure 5.16 depicts the normalised RMS error integrated over the whole take-off sequence and expressed in number of nominal standard deviations[1] for each sensor. Note that the vertical axis is in logarithmic scale. Let us consider the plain lines for now. The black line links the errors obtained with the biased model, the orange line is related to the hybrid model. The fidelity of the biased model is obviously very poor, except for T26 and T3. Indeed, the error for a perfect model would be equal to one as the only contributor would be the random measurement noise. In the case of the biased model, the largest error, occurring on NC reaches about seventy standard deviations while the lowest one, occurring on T26 and T3, is about three standard deviations. The addition of the bias compensation module to the biased model, leading to the hybrid model, improves considerably the quality of the predictions. The reduction factor on the various errors ranges between a factor two for T3 to thirty for NC. We can therefore expect a much more accurate estimate of the engine condition when using the hybrid model even if it still contains some modelling errors.

We built the bias compensation module from data collected at ISA-SLS conditions and validated it on a take-off sequence starting from the same ambient conditions. We now would like to gauge the suitability of the corrected parameters when using the bias compensation module at other atmospheric conditions. To this end, we devise a new take-off

[1] that is one third of the values quoted in table 5.2

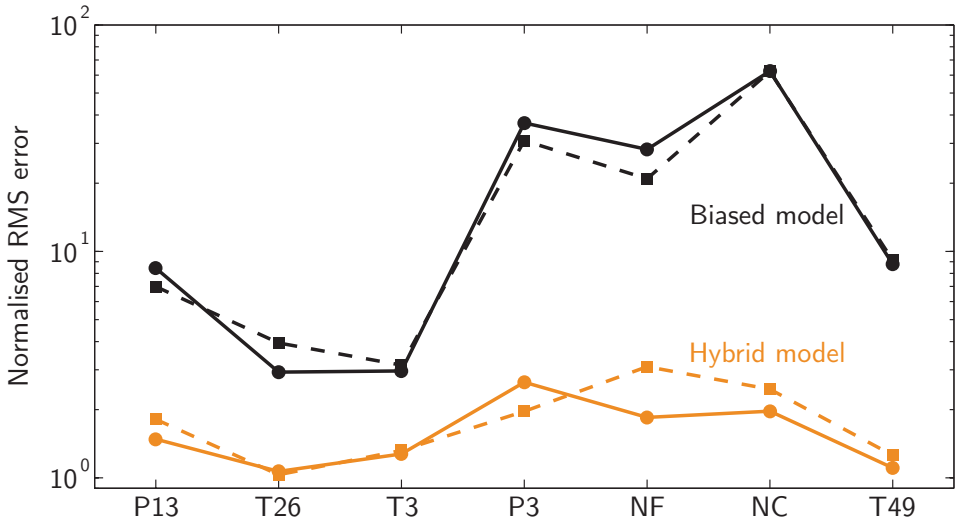


Figure 5.16: Comparison of the prediction error on each sensor with the biased and hybrid models – plain lines are for the ISA, sea-level take-off, dashed lines are for the hot 'n high take-off

sequence from an airport at an elevation of 7316 feet (2230 m) and a temperature of ISA + 25°C which are so-called hot 'n high conditions. Obviously, the ambient temperature and pressure are quite different from those in the standard conditions. The results are illustrated in figure 5.16 through the dashed lines, with the same colour code as for the take-off at sea level. We notice that the errors made with the biased model are shockingly close in magnitude to those achieved for the nominal take-off. Furthermore, we observe the same trend for the dashed orange line, meaning that the bias compensation module formulated in corrected variables can effectively take care of changing ambient conditions.

Results

We evaluate the gains in estimation accuracy brought by the hybrid model from the original batch of 300 transient scenarios. We process these test-cases with the dual extended Kalman filter only as it turned out to be more accurate than its unscented cousin.

Table 5.7 reports the RMSE, as defined in equation (5.33), expressed in percent of the fleet average baseline value evaluated at the time of lift-off (t = 55 s). The last row sums up the ten previous ones to provide a global measure over the full set of health parameters. Column-wise, both tables are organised in three main blocks, corresponding to the three deterioration levels defined previously (low, medium, high). Each block is itself subdivided into three columns corresponding to a different version of the on-board model used in the dual Kalman filter. Next to the results obtained with the biased and hybrid model, those achieved with a perfect model are recalled as target values.

	Low			**Medium**			**High**		
	Biased	Hybrid	Perfect	Biased	Hybrid	Perfect	Biased	Hybrid	Perfect
SW12	0.115	0.129	0.110	0.141	0.156	0.132	0.228	0.281	0.240
SE12	1.033	0.097	0.092	1.042	0.126	0.114	1.200	0.370	0.224
SW2	0.097	0.545	0.047	0.176	0.967	0.096	0.200	1.788	0.087
SE2	2.045	0.404	0.038	2.053	0.727	0.038	2.083	1.439	0.053
SW26	0.276	0.136	0.053	0.184	0.177	0.120	0.162	0.217	0.132
SE26	3.279	0.113	0.022	3.281	0.146	0.028	3.313	0.099	0.039
SW41	0.078	0.070	0.049	0.142	0.120	0.098	0.186	0.257	0.113
SE41	0.462	0.064	0.055	0.423	0.085	0.112	0.458	0.205	0.138
SW49	0.633	0.114	0.098	0.604	0.155	0.125	0.487	0.311	0.204
SE49	0.634	0.104	0.095	0.608	0.174	0.167	0.480	0.167	0.218
Sum	**8.653**	**1.777**	**0.659**	**8.653**	**2.833**	**1.031**	**8.797**	**5.136**	**1.448**

Table 5.7: Effect of the model-plant mismatch on the estimation accuracy in terms of percent RMSE

The results obtained with the biased model indicate that the model-plant-mismatch impacts primarily the efficiency factors SEi and the flow capacity factor of the lpt. We do not think it is worth attempting to explain this phenomenon for it is entirely dependent on the modelling errors. Adding the bias compensation module in the loop improves unmistakably the quality of the estimates. Most of the errors are in the same ballpark as those obtained with the perfect model, except for the health parameters of the booster and to a lesser extent for those of the hpc. This is most likely caused by the remaining modelling errors on the hpc outlet pressure and spool speeds that we highlighted in figure 5.16.

Figure 5.17 depicts the information contained in the last row of table 5.7. For sake of completeness, the results obtained from steady-state test-cases are added to the plot. First, we notice that the overall estimation error achieved with the biased model is almost independent of the deterioration level. In fact, going back to table 5.7 teaches us that the combined error on the efficiency factors of the compressors is nearly constant with respect to the deterioration level and accounts for not less than 73% of the total RMSE. As previously mentioned, the use of the hybrid model reduces appreciably the RMSE, however it stays at a higher level than what would be achieved in a perfect world. The key information to grab from the graph is that the global RMSE obtained with the hybrid model is of comparable magnitude to that achieved with steady-state data collected at a single operating point.

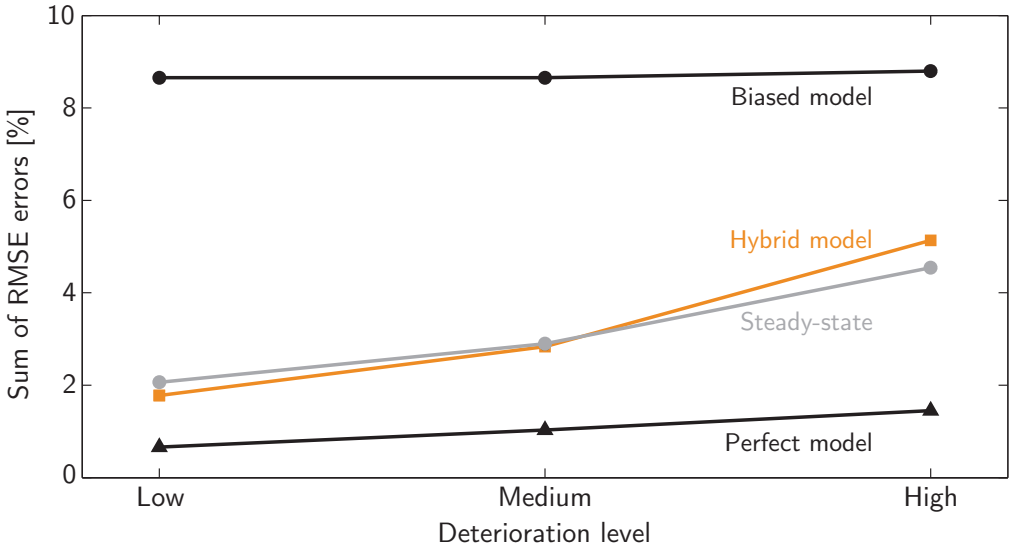


Figure 5.17: Comparison of the estimation accuracy in the presence of model-plant-mismatch

Chapter 6

Conclusions and perspectives

Contents

6.1 Conclusions

The objective of this thesis was to bring additional capabilities to an existing performance monitoring tool. The theoretical developments were assessed on a large panel of test-cases such that the results are sufficiently representative of the performance of the enhanced algorithm. Nonetheless, we remind the reader that the validation was carried out with simulated data, which are always "better looking" than true operational data.

Even if we considered an aircraft engine application in this thesis, we would like to underline that the developed algorithms can be applied to any gas turbine configuration. In the following, we summarise the essential findings for the four topics listed in section 1.5 to which our work has contributed.

Tools for sensor set selection and analysis

We revisited the problem of the optimal selection of the sensor suite from the viewpoint of Information Theory. From sound mathematical arguments, the Fisher Information Matrix appeared to be a relevant quantity for the problem of sensor selection. Although the Fisher Information Matrix is related to the variance in the estimates, simulation results showed that sensor sets selected according to the FIM-based figure of merit led to lower root mean square estimation errors over the full set of parameters. It would be interesting to complement the FIM-based metric with a term that explicitly accounts for the bias part of the estimation error, similarly to research by Simon and Garg [2009].

The framework developed in this thesis can be extended to variations of the sensor selection problem such as the selection of the best sensors to add to an existing set (for instance imposed by control system requirements) for a given set of health parameters. Another theme is the selection of the best sensor configuration to estimate engine performance parameters that are currently not measurable in flight such as thrust, stall margin and turbine entry temperature, see *e.g.*, [Litt, 2008]. Additional considerations such as the cost and the reliability of the sensors could be embedded in the formulation of the sensor selection problem. Finally the developed methodology could also be applied in more prospective studies to assess the benefit of new sensor technology (*e.g.*, emissions, turbine inlet temperature, airflow).

Improved fault detection

We improved the reactivity of the diagnosis tool with respect to accidental faults by introducing a transition model that explicitly accounts for such events. The resulting algorithm combines a traditional Kalman filter, which provides accurate estimation of the health condition for long-time-scale deterioration (such as engine wear), and a detection module based on a hypothesis testing that monitors the residuals and looks for abrupt

changes in the health condition. We carried out a parametric study on the influence of the two tuning parameters of the detection module on its performance.

High correct detection rates and low detection latency could be achieved on abrupt component faults while keeping the false alarm rate at an acceptable level. The detection capability of sensor faults was slightly worse, this is probably due to the fact that the hypothesis testing is formulated in terms of signatures in the health parameter space and as such any sensor fault is projected into that space prior to its detection.

We have identified at least one area of improvement of the detection module, namely its robustness to outliers in the data. Actual recordings can indeed contain spikes which would increase the false alarm rate. The addition of a persistency logic to the detection module should take care of this issue.

Improved fault isolation

We improved the fault isolation capability of the diagnosis tool by using a regularisation term that better matches the expected pattern of a fault *i.e.,* many nought terms in the solution vector. The resulting fault isolation module implements an estimator that combines least-squares fit of the data and sparsity in the solution. Moreover, it is efficiently formulated as a quadratic programming problem. The isolation module has one tuning parameter whose selection can be done according to sound guidelines. An interesting feature of the fault isolation tool is that it does not require the set-up of a pre-defined bank of accidental faults, but makes better use of prior knowledge about the structure of localised faults to enhance its ability to concentrate them.

High correct classification rates were achieved on most of the fault types considered in the library of scenarios. On the other hand, a non-negligible number of high pressure compressor faults were misclassified as variable stator vane faults. After investigation, we found the misclassification was due to highly similar signatures of the high pressure compressor flow capacity and variable stator vane health parameters. Nonetheless, the results also illustrated the superiority of the sparse estimation approach versus the legacy least-squares estimator as far as fault isolation is concerned.

Our latest work [Borguet and Léonard, 2011] consisted in adding constraints to the sparse estimation problem. These constraints translate additional knowledge about the parameters such as the fact that efficiencies do not improve when a fault occurs, or that a component fault involves a simultaneous change in efficiency and flow capacity. The constrained framework helped resolve the confusion issue between high pressure compressor and variable stator vane faults.

Diagnosis from transient data and robustness with respect to modelling errors

In a sequential framework, module performance analysis from data recorded during transient operation requires two vectors to be estimated, namely the heath parameters and the state variables associated to the dynamic behaviour of the engine. Our solution is based on two Kalman filters running concurrently. Such a dual approach allows, if necessary, the tuning of each filter to the quantity it estimates. Results obtained on data representative of a take-off sequence showed that the processing of transient data leads to a more accurate estimation of the health condition than the processing of single point, steady-state data. This improvement comes from the wealthier information contained in the large number of operating points visited during a transient manoeuvre. The bottleneck of the dual Kalman filter lies in the computational demand associated to the numerous calls to the engine performance deck at each time step. It currently prevents running the algorithm on on-board hardware.

In a second step, we saw that modelling errors in the performance deck have a disastrous effect on the estimation procedure. We applied the concept of hybrid model to cope with these modelling errors. The hybrid model combines the biased performance deck, based on first principles, and a black-box component called the bias compensation module in this work. The BCM relies on Wiener models and is identified from data collected when the engine is in a "new and clean" condition. Results obtained on take-off roll scenarios showed a neat improvement in estimation accuracy when the hybrid model replaced the biased performance deck. However, the level of fidelity achieved with the hybrid model led to estimation errors of the same order of magnitude as those obtained from single point steady-state data, which warrants further refinements of the BCM.

Complementary developments to the transient algorithm include its extension to system faults (such as variable blade valve and variable stator vane) and sensor faults. Another question of practical interest is the possibility to decrease the sampling frequency of data down to around one Hertz, which would reduce the memory requirements and might allow real-time execution on on-board hardware.

6.2 Perspectives

Next to the complementary developments proposed in the previous section, which are directly related to the performance monitoring tool designed in the scope of this thesis, we would like to mention three directions for future research efforts in the same perimeter.

The first direction concerns the validation of the monitoring tool. In this thesis, we evaluated its performance on the basis of simulated data. The next logical step would be to check its behaviour against real data. It is however particularly difficult to be granted access to such data due to proprietary reasons. As a workaround, NASA set up

the ProDiMES initiative, see [Simon et al., 2008], whose purpose is to provide developers of health monitoring solutions with a benchmark problem. Even if it still consists of simulated measurements, the ProDiMES data is released as a blind-test designed by representatives of industries and research agencies.

Secondly, the topic of "more intelligent engines" [RTO, 2009] is gaining momentum in the community. One of the objectives is to bring additional capabilities in the engine control system. On one hand, the control laws would adapt to the actual condition of the engine (*e.g,* the acceleration schedule would be modified in the case where the high pressure compressor is degraded). On the other hand, the regulation would be based on so-called virtual sensors. For instance, the thrust predicted by an on-board model would replace current thrust setting parameters such as the fan speed or engine pressure ratio. Both of these strategies require accurate estimates of the health condition and of non-measurable quantities which could be handled by tools similar to the one designed in this thesis.

The third area for future research that we propose addresses a number of questions related to modelling. Firstly, we remember that the Kalman filter tracks gradual deterioration on the basis of a very coarse model for the temporal evolution of the health parameters. It mainly serves to dampen the measurement noise, but does not take into account the operational history of the engine. However, the wear profile depends to a large extent on engine usage. The development of ageing models based on operational factors, see *e.g.*, [Spieler et al., 2007, Wensky et al., 2010], is an emerging field that will increase the level of information available to track performance changes.

Another axis of research that deserves further work is the introduction of zooming in the performance deck. Basically, zooming consists in replacing 0-D models of some components – mainly compressors and turbines – with higher order representations – mainly 1-D not to penalise excessively the computational load, see for instance [Aretakis et al., 2010, Pilet et al., 2011]. The major advantage offered by zooming would be to describe the health of rotating modules with more physical parameters than the legacy correction factors applied to component maps. This is especially important in the case of multi-point (either steady-state or transient) estimation for which the validity of uniform health indices throughout the operating range may be questioned.

Appendix A

The mathematics toolbox

A.1 The Gaussian distribution

The Gaussian distribution is a probability density function (pdf) that is very frequently used to model the statistical behaviour of a random variable. For a scalar random variable x, it writes:

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x-\hat{x})^2}{2\sigma^2}\right\} \tag{A.1}$$

The scalar Gaussian distribution is completely specified by its mean value $\hat{x}$ and its variance σ^2. These two parameters are defined as:

$$\hat{x} = \int_{-\infty}^{+\infty} x\, p(x)\, \mathrm{d}x = E[x] \tag{A.2}$$

$$\sigma^2 = \int_{-\infty}^{+\infty} (x-\hat{x})^2\, p(x)\, \mathrm{d}x = E[(x-\hat{x})^2] \tag{A.3}$$

where $E[\cdot]$ stands for the mathematical expectation operator. The square root of the variance is called the standard deviation σ.

Figure A.1 illustrates the celebrated bell-shaped allure of the scalar Gaussian pdf that has a mean value of zero and a unit variance. The distribution is symmetric about the mean value, which is the most likely one. The variance provides a measure of the dispersion about the mean value.

The Gaussian distribution can be generalised to a random vector $\mathbf{x}$ of dimension n. In this case, the pdf writes:

$$p(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^n\, |\mathbf{\Sigma}|}} \exp\left[-\frac{1}{2}(\mathbf{x}-\widehat{\mathbf{x}})^T \mathbf{\Sigma}^{-1} (\mathbf{x}-\widehat{\mathbf{x}})\right] \tag{A.4}$$

where $\widehat{\mathbf{x}}$ is a n-dimensional column vector that contains the mean value of each element of the vector, $\mathbf{\Sigma}$ a symmetric, positive definite and square matrix of dimension n and $|\mathbf{\Sigma}|$

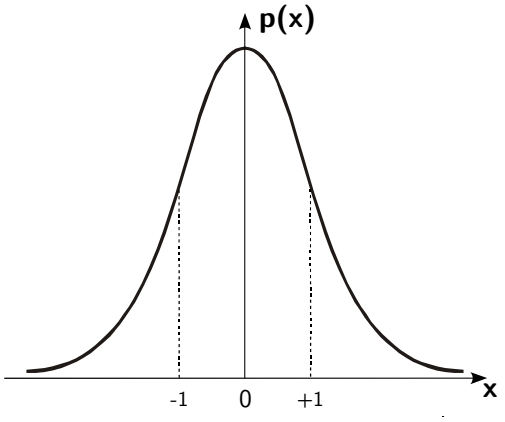


Figure A.1: The scalar Gaussian pdf with zero mean and unit variance

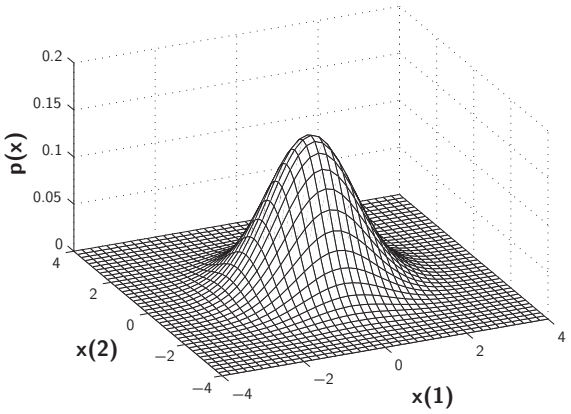


Figure A.2: The 2D Gaussian pdf with zero mean and an identity covariance matrix

denotes the determinant of $\mathbf{\Sigma}$. The mean value $\hat{\mathbf{x}}$ and covariance matrix $\mathbf{\Sigma}$ are defined as:

$$\widehat{\mathbf{x}} = E[\mathbf{x}] \tag{A.5}$$

$$\mathbf{\Sigma} = E[(\mathbf{x} - \widehat{\mathbf{x}})(\mathbf{x} - \widehat{\mathbf{x}})^T] \tag{A.6}$$

Figure A.2 depicts the Gaussian pdf of a random vector of dimension two whose mean is the nought vector and whose covariance matrix is the identity matrix. In such cases where the covariance matrix is diagonal, the components of the random vector are statistically independent.

As shown in figure A.3 for the two-dimensional case, the surface of constant probability density are hyper-ellipsoids defined through:

$$(\mathbf{x} - \widehat{\mathbf{x}})^T \mathbf{\Sigma}^{-1} (\mathbf{x} - \widehat{\mathbf{x}}) = C^2 \tag{A.7}$$

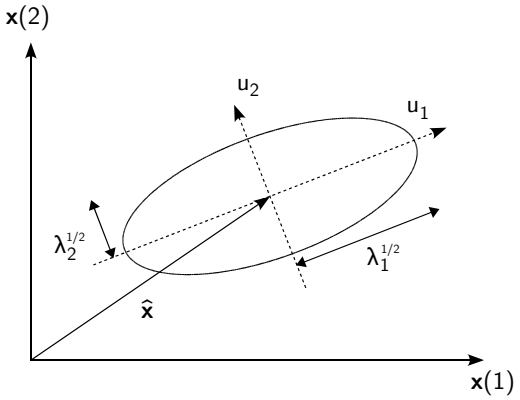


Figure A.3: Principal directions of a covariance matrix.

The axes of the hyper-ellipsoids are given by the eigenvectors of $\mathbf{\Sigma}$ which satisfy:

$$\mathbf{\Sigma}\,\mathbf{u}_i = \lambda_i\,\mathbf{u}_i \tag{A.8}$$

and the corresponding eigenvalues λ_i give the variances along the respective principal directions. The eigenvectors of $\mathbf{\Sigma}$ actually define an auxiliary coordinate system in which the variables are statistically independent.

The Gaussian distribution benefits from a number of very handy properties that explain its popularity in many scientific fields.

1. It has a relatively simple analytical form allowing many closed-form solutions to be obtained.
2. The central limit theorem states that, under rather general circumstances, the mean of n independent random variables tends to be asymptotically Gaussian *i.e.,* follows a Gaussian probability density as n tends to infinity. A common application is to the sum of a set of variables drawn independently from the same distribution. In practice, convergence tends to be very fast, so that for values as small as 10 the approximation to a Gaussian distribution can be very good. We can reasonably expect that measurements of naturally occurring phenomena have several components, leading to a distribution which is close to Gaussian.
3. Any non-singular linear transformation changes a Gaussian distribution into another Gaussian distribution characterised by different mean and covariance parameters,
4. The marginal probabilities of a Gaussian probability, obtained by integrating out some of the variables, are themselves Gaussian. Similarly, the conditional probabilities obtained by setting some variables to fixed values, are also Gaussian.

A.2 The ℓ_p norm family

For any real number $p \geq 1$ and a real-valued column vector $\mathbf{x}$ with n components, the ℓ_p-norm is defined as:

$$\|\mathbf{x}\|_p = \left(\sum_{i=1}^{n} |\mathbf{x}_i|^p\right)^{1/p} \tag{A.9}$$

In particular, the ℓ_2-norm is the well-known Euclidian norm

$$\|\mathbf{x}\|_2 = \sqrt{\sum_{i=1}^{n} |\mathbf{x}_i|^2} = \sqrt{\mathbf{x}^T\mathbf{x}} \tag{A.10}$$

and the ℓ_1-norm can be expressed as

$$\|\mathbf{x}\|_1 = \left(\sum_{i=1}^{n} |\mathbf{x}_i| \right) = \mathbf{1}^T |\mathbf{x}| \tag{A.11}$$

where $\mathbf{1}$ is a column vector of appropriate dimensions with all elements equal to one.

The formula is also valid for $0 < p < 1$, but the resulting function does not define a norm strictly speaking because it violates the triangle inequality. Taking the limit $p \to \infty$ yields the uniform norm and taking the limit $p \to 0$ yields the ℓ_0 quasi-norm.

A.3 The Quadratic Programming problem

A Quadratic Programming problem is an optimisation problem in which the objective function is quadratic and the constraint functions are linear. The problem is to find a solution vector $\mathbf{x}_{opt}$ to:

$$\begin{aligned} \min_{\mathbf{x}} f(\mathbf{x}) &= \frac{1}{2}\, \mathbf{x}^T\, \mathbf{H}\, \mathbf{x} + \mathbf{g}^T\, \mathbf{x} \\ \text{subject to} \quad & \begin{cases} \mathbf{a}_i^T\, \mathbf{x} = b_i, & i \in E \\ \mathbf{c}_j^T\, \mathbf{x} \geq d_j, & j \in I \end{cases} \end{aligned} \tag{A.12}$$

where E and I are respectively the sets of equality and inequality constraints.

If the symmetric Hessian matrix $\mathbf{H}$ is positive semi-definite then $f(\mathbf{x})$ is a convex function. In this case, the optimisation problem has a global optimiser if there exists some feasible vector $\mathbf{x}_{opt}$ satisfying the constraints and if f is bounded below on the feasible domain. If the Hessian is positive definite and the problem has a feasible solution, then the global optimiser is moreover unique.

Appendix B

A basic engine control system

B.1 Definition of the 100% PCNF regime

For the transient test-cases, the engine is piloted in terms of percent corrected fan speed PCNF, rather than directly through the fuel flow. The reference point on the fan speed scale, namely the 100% corrected fan speed in Sea Level Static, standard conditions (ISA-SLS) is defined in the table below:

Label	Description	Value
WF	fuel flow	1.238 kg/s
Alt	pressure altitude	0 m
Mach	flight Mach number	0
DT_{ISA}	offset in temperature	0 K
NF	fan speed	4114 rpm
NC	core speed	12401 rpm
FN	net thrust	127.8 kN
Wair	total air mass flow	421 kg/s

Table B.1: Definition of the 100% corrected fan speed regime at ISA-SLS conditions

The OBIDICOTE model is limited to simulation of above-idle operations. Moreover, in order to ensure the convergence of the model in the range of considered deterioration, the minimum regime (Min Idle) is set to 40% PCNF.

B.2 About the controller

The controller used in the transient test-cases is derived from an existing control system provided with the OBIDICOTE model. It combines a feed-forward action to maintain the current fan speed with three Proportional-Integral feedback loops that handle transient manoeuvres. One of the PI loop tracks the PCNF value set by the user, the other two

loops limit the acceleration/deceleration rate of the high-pressure spool in order to avoid compressor surge. This basic system controls only the fuel flow. The other two actuators, namely the blow-off valves at the LPC outlet and the variables stator vanes on the HPC follow a pre-defined schedule, much as in the current engine control system actually. A single set of gains is used throughout the operating envelope of the engine. As a result the control system is far from delivering optimal performance, it is however deemed to be sufficiently representative for the application at hand.

Figure B.1 shows the closed-loop engine response to a fast acceleration-deceleration manoeuvre between the Min Idle and 100% regimes. The upper left graph gives the evolution of the fuel flow, the upper right graph plots the requested (grey line) and effective (black line) percent corrected fan speed. The lower left graph shows the relative acceleration of the high pressure spool and the lower right graph gives the active loop (1: NF tracking, 2: limiter on HP spool acceleration, 3: limiter on HP spool deceleration).

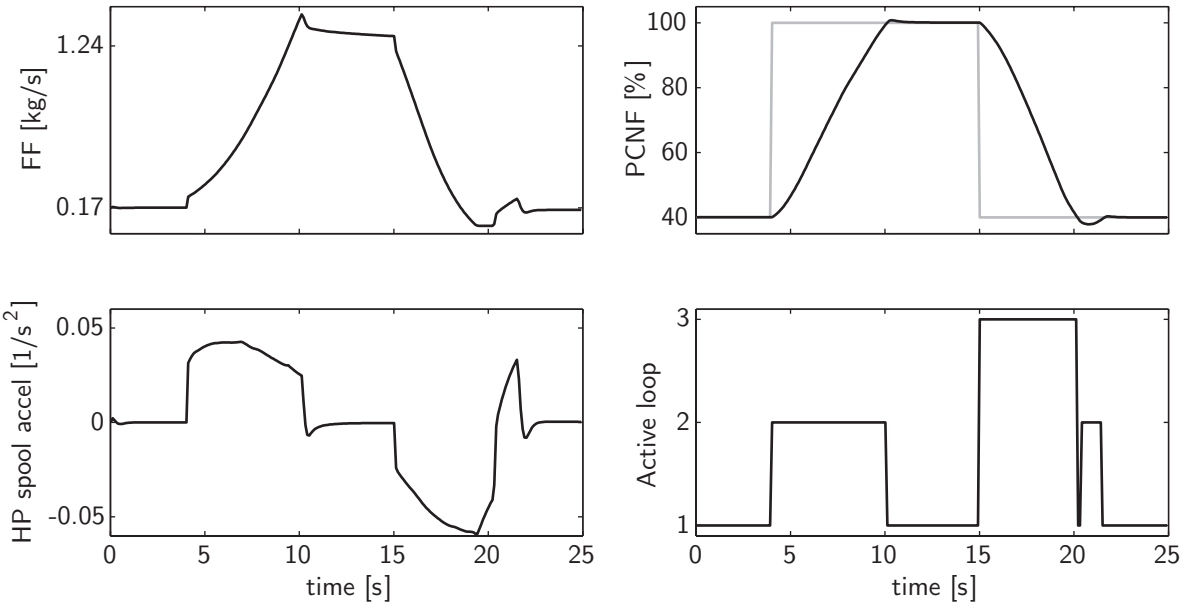


Figure B.1: Closed-loop response of the engine to a typical acceleration-deceleration manoeuvre

Bibliography

Chapter 1

AGARD. *Erosion, Corrosion and Foreign Object Damage Effects in Gas Turbines*, volume CP-558. Advisory Group for Aerospace Research and Development, France, 1994.

J. Barbosa, G. Matuck, C. Bringhenti, and I. Lima. Gas turbine fault detection and isolation using MLP artificial neural network. In *ASME Turbo Expo*, number GT2007-27987, Montréal QC, Canada, 2007.

S. Borguet and O. Léonard. Coupling principal component analysis and Kalman filter algorithms for on-line aircraft engine diagnostics. *Control Engineering Practice*, 17(4): 494–502, 2009.

O. Cordoba. Gas path analysis study for overhaul engines. In *ASME Turbo Expo*, number GT2005-68137, Reno-Tahoe NV, USA, 2005.

S. Danielsson. Gas path analysis to pre and post overhaul testing of JT9D turbofan engine. In *SAE Aerospace Meeting*, number 770993, Los Angeles CA, USA, 1977.

S. de Smet. Vibration monitoring systems for aero gas turbines. In *Engine System Symposium*, Cranfield, UK, 2005. Cranfield University.

P. Dempsey and A. Afjeh. Integrating oil debris and vibration gear damage detection technologies using fuzzy logic. Technical Memorandum TM-2002-211126, NASA, 2002.

H. DePold and F. Gass. The application of expert systems and neural networks to gas turbine prognostics and diagnostics. *ASME J. of Eng. for Gas Turbine and Power*, 121:607–612, 1999.

P. Dewallef. *Application of the Kalman Filter to Health Monitoring of Gas Turbine Engines: A Sequential Approach to Robust Diagnosis*. PhD thesis, University of Liège, 2005.

P. Dewallef, K. Mathioudakis, and O. Léonard. On-line aircraft engine diagnostic using a soft-constrained Kalman filter. In *ASME Turbo Expo*, number GT2004-53539, Vienna, Austria, 2004.

P. Dewallef, C. Romessis, O. Léonard, and K. Mathioudakis. Combining classification techniques with Kalman filters for aircraft engine diagnostics. *ASME J. of Eng. for Gas Turbines and Power*, 128(2):281–287, 2006.

I. S. Diakunchak. Performance deterioration in industrial gas turbines. *ASME J. of Eng. for Gas Turbines and Power*, 114:161–168, 1992.

D. L. Doel. The role for expert systems in commercial gas turbine engine monitoring. In *ASME Turbo Expo*, number 90-GT-374, Brussels, Belgium, 1990.

D. L. Doel. Temper: gas path analysis tool for commercial jet engines. *ASME J. of Eng. for Gas Turbines and Power*, 116:82–89, 1994a.

D. L. Doel. An assessment of weighted-least-squares-based gas path analysis. *ASME J. of Eng. for Gas Turbines and Power*, 116:366–373, 1994b.

M. Fast, M. Assadi, and S. De. Condition based maintenance of gas turbines using simulation data and artificial neural network: a demonstration of feasability. In *ASME Turbo Expo*, number GT2008-50768, Berlin, Germany, 2008.

A. Green. Oil system monitoring. In *Engine System Symposium*, Cranfield, UK, 2005. Cranfield University.

T. Grönstedt. Identifiability in multi-point gas turbine parameter estimation problems. In *ASME Turbo Expo*, number GT2002-30020, Amsterdam, Netherlands, 2002.

A. Gulati, D. Taylor, and R. Singh. Multiple operating point analysis using genetic algorithm optimisation for gas turbine diagnostics. In *15th International Symposium on Air Breathing Engines*, number 1139, Bangalore, India, 2001.

L. Jaw. Recent advancements in aircraft engine health management technologies and recommandations for the next step. In *ASME Turbo Expo*, number GT2005-68625, Reno-Tahoe NV, USA, 2005.

Ph. Kamboukos and K. Mathioudakis. Comparison of linear and non-linear gas turbine performance diagnostics. *ASME J. of Eng. for Gas Turbine and Power*, 127(1):49–56, 2005. doi: 10.1115/1.1788688.

S-H. Kim, M.A. Cohen, and S. Netessine. Performance contracting in after-sales service supply chains. *Management science*, 53(12):1843–1858, 2007. doi: 10.1287/mnsc.1070.0741.

T. Kobayashi and D. L. Simon. A hybrid neural network-genetic algorithm technique for aircraft engine performance diagnostics. In *37th AIAA/ASME/SAE/ASEE Joint Propulsion Conference*, number AIAA-2001-3763, Salt Lake City UT, USA, 2001.

M. Koehl. On-board engine life usage monitoring by real time computation. In *21st symposium of the international committee on aeronautical fatigue*, volume 2, Toulouse, France, 2001.

R. Kurz, K. Brun, and M. Wollie. Degradation effects on industrial gas turbines. *ASME J. of Eng. for Gas Turbines and Power*, 131(6):062401, 2009. doi: 10.1115/1.3097135.

A. Kyriasis, A. Tsalavoutas, and K. Mathioudakis. Gas turbine fault identification by fusing vibration trending and gas path analysis. In *ASME Turbo Expo*, number GT2009-59942, Orlando FL, USA, 2009.

Y. G. Li. Performance-analysis-based gas turbine diagnostics: A review. *IMechE J. of Power and Energy*, 216(5):363–377, 2002.

L. Marinai, D. Probert, and R. Singh. Prospects for aero gas-turbine diagnostics: a review. *Applied Energy*, 79:109–126, 2004.

K. Mathioudakis, Ph. Kamboukos, and A. Stamatis. Turbofan performance deterioration tracking using non-linear models and optimization techniques. *ASME J. of Turbomachinery*, 124:580–587, 2002.

S. Menon, Ö. Uluyol, K. Kim, and E. Nwadiogbu. Incipient fault detection and diagnosis in turbine engines using hidden markov models. In *ASME Turbo Expo*, number GT2003-38589, Atlanta GA, USA, 2003. doi: 10.1115/GT2003-38589.

G. Merrington. Fault diagnosis of gas turbine engines from transient data. *ASME J. of Eng. for Gas Turbines and Power*, 111:237–243, 1989.

J. Mukkherjee, V. Kini, S. Menon, and L. Eswara. Gas turbine fault detection and diagnosis using nonlinear feature extraction methods. In *ASME Turbo Expo*, number GT2005-68802, Reno-Tahoe NV, USA, 2005.

S. Ogaji, S. Sampath, R. Singh, and S. Probert. Parameter selection for diagnosing a gas-turbine's performance deterioration. *Applied Energy*, 73(1):25–46, 2002. doi: 10.1016/S0306-2619(02)00042-9.

S.O.T Ogaji and R. Singh. Advanced engine diagnostics using artificial neural networks. In *IEEE International Conf. on Artificial Intelligence Systems (ICAIS)*, pages 236–241, Divnomorskoe, Russia, 2002.

M. J. Provost. *The use of optimal estimation techniques in the analysis of gas turbines.* PhD thesis, Cranfield University, 1994.

R. Rajamani, J. Wang, and K.Y Jeong. Condition-based maintenance for aircraft engines. In *ASME Turbo Expo*, number GT2004-54127, Vienna, Austria, 2004.

Rolls-Royce. *The Jet Engine.* Rolls-Royce, 1999.

C. Romessis and K. Mathioudakis. Setting up of a probabilistic neural network for sensor fault detection including operation with component faults. *ASME J. of Eng. for Gas Turbine and Power*, 125(3):634–641, 2003. doi: 10.1115/1.1582493.

C. Romessis and K. Mathioudakis. Bayesian network approach for gas path fault diagnosis. *ASME J. of Eng. for Gas Turbines and Power*, 128(1):64–72, 2006.

C. Romessis, A. Kyriazis, and K. Mathioudakis. Fusion of gas turbines diagnostic inference – the Dempster-Schafer approach. In *ASME Turbo Expo*, number GT2007-27043, Montréal QC, Canada, 2007.

G. P. Sallee. Performance deterioration based on existing (historical) data - JT9D jet engine diagnostics program. Contractor Report CR-135448, NASA, 1978.

A. Stamatis, K. Mathioudakis, and K. Papailou. Jet engine fault detection with discrete operating points gas path analysis. *AIAA J. of Propulsion and Power*, 7:1043–1048, 1991.

A. Stamatis, K. Mathioudakis, and K. Papailou. Optimal measurements and health indices selection for gas turbine performance status and fault diagnosis. *ASME J. of Eng. for Gas Turbines and Power*, 114(2):209–216, 1992.

I. Tumer and A. Bajwa. A survey of aircraft engine health monitoring system. In *35th AIAA/ASME/SAE/ASEE Joint Propulsion Conference*, number AIAA-99-2528, Los Angeles CA, USA, 1999.

J. Turso and J. S. Litt. A foreign object damage event detector data fusion system for turbofan engines. Technical Memorandum TM-2004-213192, NASA, 2004.

L. A. Urban. Gas path analysis applied to engine condition monitoring. In *8th Joint Propulsion Specialist Conference*, number AIAA 72-1082, New-Orleans LA, USA, 1972.

A. J. Volponi. Foundation of gas path analysis (part I and II). In *von Karman Institute Lecture Series*, number 01 in Gas Turbine Condition Monitoring and Fault Diagnosis, Brussels, Belgium, 2003.

A. J. Volponi. Data fusion for enhanced aircraft engine prognostics and health management. Contractor Report CR-2005-214055, NASA, 2005.

A. J. Volponi, H. DePold, R. Ganguli, and C. Daguang. The use of Kalman filter and neural network methodologies in gas turbine performance diagnostic: A comparative study. *ASME J. of Eng. for Gas Turbines and Power*, 125:917–924, 2003.

M. Wallin and T. Grönstedt. A comparative study of genetic algorithms and gradient methods for RM12 turbofan engine diagnostics and performance estimation. In *ASME Turbo Expo*, number GT2004-53591, Vienna, Austria, 2004.

R. Wulf. Engine diagnostics program CF6-50 engine performance deterioration. Contractor Report CR-159867, NASA, 1980.

A. V. Zaita, G. Buley, and G. Karlsons. Performance deterioration modeling in aircraft gas turbine engines. *ASME J. of Eng. for Gas Turbines and Power*, 120:344–349, 1998.

Chapter 2

G. Bechini, G. Ameyugo, L. Marinai, and R. Singh. Gas path diagnostics: the importance of measurement selection in the monitoring process. In *17th International Symposium on Air Breathing Engines*, number ISABE-2005-1281, München, Germany, 2005.

S. Borguet and O. Léonard. The Fisher information matrix as a relevant tool for sensor selection in engine health monitoring. *International Journal of Rotating Machinery*, 2008:784749, 2008.

R. G. Brown. Not just observable, but how observable. In *National Electronics Conference*, volume 22, pages 709–714, 1966.

E. Castillo, A. Conejo, R. Pruneda, and C. Solares. State estimation observability based on the null space of the measurement jacobian matrix. *IEEE Trans. on Power Systems*, 20(3):1656–1658, 2005.

D. L. Doel. Interpretation of weighted-least-squares gas path analysis results. *ASME J. of Eng. for Gas Turbines and Power*, 125(3):624–633, 2003.

E.H. Dudgeon, editor. *Guide to the measurement of the transient performance of aircraft turbine engines and components*, volume AR-320. Advisory Group for Aerospace Research and Development, France, Neuilly-sur-Seine, France, 1994.

G. Golub and C. van Loan. *Matrix computations*. The John Hopkins University Press, Baltimore MD, USA, 3rd edition, 1996.

M. Härefors. *A study in jet engine control*. PhD thesis, Chalmers University of Technology, Göteborg, Sweden, 1999.

M.S. Horobin. Cycle-match engine models used in functional engine design - an overview. In *AGARD RTO-MP-8: Design Principles and Methods for Aircraft Gas Turbine Engines*, number 44, Toulouse, France, 1999.

R. Isermann. Supervision, fault-detection and fault-diagnosis methods – an introduction. *Control Engineering Practice*, 5(5):639–652, 1997.

Ph. Kamboukos, P. Oikonomou, A. Stamatis, and K. Mathioudakis. Optimizing effectiveness of mixed turbofans by means of adaptive modelling and choice of appropriate monitoring parameters. In *AGARD RTO-MP-79: Ageing Mechanisms and Control*, number 9, Manchester, UK, 2001.

S. M. Kay. *Fundamentals of Statistical Signal Processing: Estimation Theory*. Prentice Hall, London, 1993.

P. H. Kirkegaard and R. Brincker. On the optimal locations of sensors for parametric identification of linear structural systems. *Mechanical Systems and Signal Processing*, 8:639–647, 1994.

A. Kreiner and K. Lietzau. The use of onboard real-time models for jet engine control. In *von Karman Institute Lecture Series*, number 01 in Gas Turbine Condition Monitoring and Fault Diagnosis, Brussels, Belgium, 2003.

J. S. Litt. An optimal orthogonal decomposition method for Kalman filter-based turbofan engine thrust estimation. *ASME J. of Eng. for Gas Turbines and Power*, 130(1):011601, 2008.

L. Ljung. *System Identification, Theory for the User*. Prentice Hall Information and System Sciences Series, 2nd edition, 1999.

R. Mushini and D. Simon. On optimization of sensor selection for aircraft gas turbine engines. In *18th International Conference on Systems Engineering*, pages 9–14, Las Vegas NV, USA, 2005.

C. Papadimitriou. Optimal sensor placement methodology for parametric identification of structural systems. *Journal of Sound and Vibration*, 278:923–947, 2004.

M. Pinelli and P.R. Spina. Gas turbine field performance determination: sources of uncertainties. *ASME J. of Eng. for Gas Turbine and Power*, 124:155–160, 2002.

M. J. Provost. *The use of optimal estimation techniques in the analysis of gas turbines*. PhD thesis, Cranfield University, 1994.

C. R. Rao. *Linear Statistical Inference and Its Applications*. John Wiley & Sons, New-York, 2nd edition, 1973.

RTO. Performance prediction and simulation of gas turbine engine operation. Technical Report 44, Research and Technology Organisation, 2002.

H.I.H. Saravanamuttoo, editor. *Recommended practices for measurement of gas path pressures and temperatures for performance assessment of aircraft turbine engines and components*, volume AR-245. Advisory Group for Aerospace Research and Development, Neuilly-sur-Seine, France, 1993.

D. L. Simon and S. Garg. Optimal tuner selection for Kalman filter based aircraft engine performance estimation. In *ASME Turbo Expo*, number GT2009-59684, Orlando FL, USA, 2009.

T. Sowers, G. Kopasakis, and D. L. Simon. Application of the systematic sensor selection strategy for tubofan engine diagnostics. In *ASME Turbo Expo*, number GT2008-50525, Berlin, Germany, 2008.

T. Sowers, J. Fittje, G. Kopasakis, and D. L. Simon. Expanded application of the systematic sensor selection strategy for turbofan engine diagnostics. In *ASME Turbo Expo*, number GT2009-59251, Orlando FL, USA, 2009.

A. Stamatis, K. Mathioudakis, and K. Papailou. Adaptive simulation of gas turbine performance. *ASME J. of Eng. for Gas Turbine and Power*, 112:168–175, 1990.

F. van den Berg, H. Hoefsloot, H. Boelens, and A. Smilde. Selection of optimal sensor position in a tubular reactor using robust degree of observability criteria. *Chemical Engineering Science*, 55:827–837, 2000.

H. L. van Trees. *Detection, Estimation and Modulation Theory.* John Wiley & Sons, New-York, 1968.

A. Vande Wouwer, N. Point, S. Porteman, and M. Remy. An approach to the selection of optimal sensor locations in distributed parameters systems. *Journal of Process Control*, 10:291–300, 2000.

V. Venkatasubramanian, R. Rengaswamy, K. Yin, and S. Kavuri. A review of fault detection and diagnosis part 1: Quantitative model-based methods. *Computers and Chemical Engineering*, 27:293–311, 2003.

P.P. Walsh and P. Fletcher. *Gas Turbine Performance.* Blackwell Science, New-York, 1998.

Chapter 3

S. Akkaram, D. Beeson, H. Agarwal, and G. Wiggs. Inverse modeling techniques for parameter estimation in engineering simulation models. In *ASME Turbo Expo*, number GT2006-90058, Barcelona, Spain, 2006.

N. Aretakis, K. Mathioudakis, and A. Stamatis. Non-linear engine component fault diagnosis from a limited number of measurements using a combinatorial approach. *ASME J. of Eng. for Gas Turbines and Power*, 125(3):642–650, 2003.

S. Arulampalam, S. Maskell, N. Gordon, and T. Clapp. A tutorial on particle filters for on-line non-linear/non-gaussian bayesian tracking. *IEEE Trans. on Signal Processing*, 50(2):174–188, 2002.

F. Bensalah and F. Chaumette. Real-time visual tracking using abrupt changes detection. Technical Report 2425, INRIA, 1994.

S. Borguet and O. Léonard. A generalized likelihood ratio test for adaptive gas turbine performance monitoring. *ASME J. of Eng. for Gas Turbines and Power*, 131:011601, 2009.

S. Borguet and O. Léonard. A sparse estimation approach to fault isolation. *ASME J. of Eng. for Gas Turbines and Power*, 132(2):021601, 2010a. doi: 10.1115/1.3156815.

S. Borguet and O. Léonard. Assessment of an anomaly detector for jet engine health monitoring. *International Journal of Rotating Machinery*, 2011:942576, 2011. doi: 10.1155/2011/942576.

S. Borguet and O. Léonard. Comparison of adaptive filtering schemes for gas turbine performance diagnostics. *Journal of Computational and Applied Mechanics*, 234(7): 2202–2212, 2010b. doi: 10.1016/j.cam.2009.08.075.

P. Dewallef. *Application of the Kalman Filter to Health Monitoring of Gas Turbine Engines: A Sequential Approach to Robust Diagnosis*. PhD thesis, University of Liège, 2005.

D. Donoho. Compressed sensing. *IEEE Trans. on Information Theory*, 52(4):1289–1306, 2006.

R. Fletcher. *Practical Methods of Optimization*. John Wiley & Sons, New-York, 2000.

J. J. Fuchs. On the application of the global matched filter to doa estimation with uniform circular arrays. *IEEE Trans. on Signal Processing*, 49(4):702–709, 2001.

J. J. Fuchs. On sparse representations in arbitrary redundant basis. *IEEE Trans. on Information Theory*, 50:1341–1344, 2004a.

J. J. Fuchs. Recovery conditions of sparse representations in the presence of noise. In *IEEE International Conf. on Acoustics, Speech and Signal Processing*, volume 3, pages 337–340, Toulouse, France, 2006.

J. J. Fuchs. Recovery of exact sparse representations in the presence of noise. In *IEEE International Conf. on Acoustics, Speech and Signal Processing*, volume 2, pages 533–536, 2004b.

A. Gelb. *Applied Optimal Estimation*. The MIT Press, Cambride MA, USA, 1st edition, 1974.

G. Golub and C. van Loan. *Matrix computations*. The John Hopkins University Press, Baltimore MD, USA, 3rd edition, 1996.

L. Granai and P. Vandergheynst. Sparse approximation by linear programming using an L1 data fidelity term. In *Proc. of the Workshop on Signal Processing with Adaptive Sparse Structured Representations*, Rennes, France, 2005.

M. Grodent and A. Navez. A robust parameter identification approach for engine health monitoring and diagnosis. In *von Karman Institute Lecture Series*, number 01 in Gas Turbine Condition Monitoring and Fault Diagnosis, Brussels, Belgium, 2003.

S. Haykin, editor. *Kalman filtering and neural networks*. Wiley series on adaptive and learning systems for signal processing, communications and control. John Wiley & Sons, New-York, 2001.

A.H. Jazwinski. Adaptive filtering. *Automatica*, 5(4):475–485, 1969.

A.H. Jazwinski. *Stochastic processes and filtering theory*. Academic Press, New-York, 1970.

S. Jokar and M. Pfetsch. Exact and approximate sparse solutions of underdetermined linear equations. ZIB-Report 07-05, Zentrum für Informationstechnik, Berlin, March 2007.

S.J. Julier and J.K. Uhlmann. A general method for approximating nonlinear transformations of probability distributions. Technical report, RRG, Department of engineering science, University of Oxford, 1996.

R.E. Kalman. A new approach to linear filtering and prediction problems. *Trans. ASME, series D, J. of Basic Engineering*, 82(1):35–45, 1960.

T. Kobayashi and D. L. Simon. Application of a bank of Kalman filters for aircraft engine fault diagnostics. In *ASME Turbo Expo*, number GT2003-38550, Atlanta GA, USA, 2003.

E. Larsson and Y. Selen. Linear regression with a sparse parameter vector. *IEEE Trans. on Signal Processing*, 55(2):451–460, 2007.

O. Léonard, S. Borguet, and P. Dewallef. An adaptive estimation algorithm for aircraft engine performance monitoring. *AIAA J. of Propulsion and Power*, 28(4):763–769, 2008.

D. Malioutov, M. Cetin, and A. Willsky. A sparse signal reconstruction perspective for source localization with sensor arrays. *IEEE Trans. on Signal Processing*, 53(8):3010–3022, 2005.

R. K. Mehra. On the identification of variances and adaptive Kalman filtering. *IEEE Trans. on Automatic Control*, 15(2):175–184, 1970.

R. K. Mehra. Approaches to adaptive filtering. *IEEE Trans. on Automatic Control*, 17 (5):693–698, 1972.

A. Neumaier. MINQ - General definite and bound constrained indefinite quadratic programming, 1998. URL `http://www.mat.univie.ac.at/~neum/software/minq`.

B. Odelson, M. Rajamani, and J. Rawlings. A new autocovariance least-squares method for estimating noise covariances. *Automatica*, 42:303–308, 2006.

M. J. Provost. *The use of optimal estimation techniques in the analysis of gas turbines.* PhD thesis, Cranfield University, 1994.

H. Qiu, N. Eklund, W. Yan, P. Bonissone, F. Xue, and K. Goebel. Estimating deterioration level of aircraft engines. In *ASME Turbo Expo*, number GT2007-27519, Montréal QC, Canada, 2007.

D. L. Simon. An integrated architecture for on-board aircraft engine performance trend monitoring and gas-path fault diagnostics. Technical Memorandum TM-2010-216358, NASA, 2010.

J. Tropp. Just relax: convex programming methods for subset selection and sparse approximation. ICES Report 0404, Univeristy of Texas, February 2004.

H. L. van Trees. *Detection, Estimation and Modulation Theory*. John Wiley & Sons, New-York, 1968.

A. J. Volponi. Extending gas path analysis coverage for other fault conditions. In *von Karman Institute Lecture Series*, number 01 in Gas Turbine Condition Monitoring and Fault Diagnosis, Brussels, Belgium, 2003.

W. Wang, Z. Liu, and R. Xie. Quadratic extended Kalman filter approach for GPS/INS integration. *Aerospace Science and Technology*, 10(8):709–713, 2006.

A. Willsky and H. Jones. A generalized likelihood ratio approach to state estimation in linear systems subject to abrupt changes. In *1974 IEEE Conf. Decision and Control*, pages 846–853, 1974.

A. Willsky and H. Jones. A generalized likelihood ratio approach to the detection and estimation of jumps in linear systems. *IEEE Trans. on Automatic Control*, 21(1):108–112, 1976.

Chapter 4

S. Borguet and O. Léonard. A sparse estimation approach to fault isolation. *ASME J. of Eng. for Gas Turbines and Power*, 132(2):021601, 2010. doi: 10.1115/1.3156815.

S. Butler, K. Pattipati, A. J. Volponi, J. Hull, R. Rajamani, and J. Siegel. An assessment methodology for data-driven and model-based techniques for engine health monitoring. In *ASME Turbo Expo*, number GT2006-91096, Barcelona, Spain, 2006.

G. Chowdhary and R. Jategaonkar. Aerodynamic parameter estimation from flight data applying extended and unscented Kalman filter. *Aerospace Science and Technology*, 14: 106–117, 2010.

N. Cumpsty. *Jet Propulsion*. Cambridge University Press, Cambridge, UK, 2000.

B. Curnock. Obidicote project - work package 4: Steady-state test cases. Technical Report DNS62433, Rolls-Royce PLc, 2000.

C. Davison and J. Bird. Review of metrics and assignment of confidence intervals for health management of gas turbine engines. In *ASME Turbo Expo*, number GT2008-50849, Berlin, Germany, 2008.

H. DePold, J. Siegel, and J. Hull. Metrics for evaluating the accuracy of diagnostic fault detection systems. In *ASME Turbo Expo*, number GT2004-54144, Vienna, Austria, 2004.

H. Depold, R. Rajamani, W. Morrison, and K. Pattipati. A unified metric for fault detection and isolation in engines. In *ASME Turbo Expo*, number GT2006-91095, Barcelona, Spain, 2006.

L. Jaw. Recent advancements in aircraft engine health management technologies and recommandations for the next step. In *ASME Turbo Expo*, number GT2005-68625, Reno-Tahoe NV, USA, 2005.

L. Jaw and S. Garg. Propulsion control technology development in the United States (a historical perspective). Technical Memorandum TM-2005-213978, NASA, 2005.

A. Linke-Diesinger. *Systems of commercial turbofan engines*. Springer, Berlin, 2008.

K. Mathioudakis, Ph. Kamboukos, and A. Stamatis. Turbofan performance deterioration tracking using non-linear models and optimization techniques. *ASME J. of Turbomachinery*, 124:580–587, 2002.

S. Ogaji, S. Sampath, R. Singh, and S. Probert. Parameter selection for diagnosing a gas-turbine's performance deterioration. *Applied Energy*, 73(1):25–46, 2002. doi: 10.1016/S0306-2619(02)00042-9.

F. Orderud. Comparison of Kalman filter estimation approaches for state space models with nonlinear measurements. In *46th Scandinavian Conference on Simulation and Modelling*, Trondheim, Norway, 2005.

R. Orsagh, M. Roemer, C. Savage, and M. Lebold. Development of performance and effectiveness metrics for gas turbine diagnostic technologies. In *IEEE Aerospace Conference*, volume 6, pages 2825–2834, 2002.

B. A. Roth, D. L. Doel, and J. J. Cissell. Probabilistic matching of turbofan performance models to test data. In *ASME Turbo Expo*, number GT2005-68201, Reno-Tahoe NV, USA, 2005.

J. Ruiz. Obidicote model v4.0 user's manual. Technical report, SNECMA, 2001.

G. P. Sallee. Performance deterioration based on existing (historical) data - JT9D jet engine diagnostics program. Contractor Report CR-135448, NASA, 1978.

L. Sambasivan, D. Mylaraswamy, and J. Mukherjee. Benchmarking diagnostic algorithms. In *ASME Turbo Expo*, number GT2007-28194, Montréal QC, Canada, 2007.

O. Sasahara. JT9D engine/module performance deterioration results from back to back testing. In *Proceedings of the 7th International Symposium on Air Breathing Engines*, pages 528–535, Beijing, China, September 1985.

D. Simon. A comparison of filtering approaches for aircraft engine health estimation. *Aerospace Science and Technology*, 12(4):276–284, 2008. doi: 10.1016/j.ast.2007.06.002.

D. Simon and D. L. Simon. Aircraft turbofan engine health estimation using constrained Kalman filtering. In *ASME Turbo Expo*, number GT2003-38584, Atlanta GA, USA, 2003.

D. L. Simon and J. S. Litt. A data filter for identifying steady-state operating points in engine flight data for condition monitoring applications. In *ASME Turbo Expo*, number GT2010-22818, Glasgow, UK, 2010.

D. L. Simon, S. Garg, G. Hunter, T. Guo, and K. Semega. Sensor need for control and health management of intelligent aircraft engines. In *ASME Turbo Expo*, number GT2004-54324, Vienna, Austria, 2004.

D. L. Simon, J. Bird, C. Davison, A. J. Volponi, and R. E. Iverson. Benchmarking gas path diagnostic methods: a public approach. In *ASME Turbo Expo*, number GT2008-51360, Berlin, Germany, 2008.

H. Spang and H. Brown. Control of jet engines. *Control Eng. Practice*, 7:1043–1059, 1999.

A. Stamatis, K. Mathioudakis, J. Ruiz, and B. Curnock. Real-time engine model implementation for adaptive control and performance monitoring of large civil turbofans. In *ASME Turbo Expo*, number 2001-GT-0362, New-Orleans LA, USA, 2001.

A. J. Volponi, H. DePold, R. Ganguli, and C. Daguang. The use of Kalman filter and neural network methodologies in gas turbine performance diagnostic: A comparative study. *ASME J. of Eng. for Gas Turbines and Power*, 125:917–924, 2003.

R. Wulf. Engine diagnostics program CF6-50 engine performance deterioration. Contractor Report CR-159867, NASA, 1980.

Chapter 5

V. Arkov, C. Evans, P.J. Fleming, D.C. Hill, J.P. Norton, I. Pratt, D. Rees, and K. Rodriguez-Vazquez. System identification strategies applied to aircraft gas turbine engines. *Annual review in control*, 24:67–81, 2000.

J. Bird and H. Schwartz. Diagnosis of turbine engine transient performance with model-based parameter estimation techniques. In *ASME Turbo Expo*, number 94-GT-317, The Hague, Netherlands, 1994.

O. M. Boaghe, S. A. Billings, L. M. Li, P. J. Fleming, and J. Liu. Time and frequency domain identification and analysis of a gas turbine engine. *Control Engineering Practice*, 10(12):1347–1356, 2002. doi: 10.1016/S0967-0661(02)00100-4.

S. Borguet, P. Dewallef, and O. Léonard. On-line transient engine diagnostics in a Kalman filtering framework. In *ASME Turbo Expo*, number GT2005-68013, Reno-Tahoe NV, USA, 2005.

S. Borguet, P. Dewallef, and O. Léonard. A way to deal with model-plant mismatch for a reliable diagnosis in transient operation. *ASME J. of Eng. for Gas Turbines and Power*, 130:031601, 2008.

S. Borguet, M. Henriksson, T. McKelvey, and T. Grönstedt. A study on engine health monitoring in the frequency domain. *ASME J. of Eng. for Gas Turbines and Power*, 133(8):081604, 2011. doi: 10.1115/1.4002832.

T. Brotherton, A. J. Volponi, R. Luppold, and D.L. Simon. eSTORM: Enhanced selft tuning on-board real-time engine model. In *2003 IEEE Aerospace Conference*, number 1023, Big Sky, MT, USA, 2003.

N. Chiras, C. Evans, and D. Rees. Nonlinear modelling of gas turbine dynamics using a Wiener structure. In University of Glamorgan, editor, *2nd International Cloolquium on Electronics*, pages 25–29, Cardiff, UK, 2001.

N. Chiras, C. Evans, and D. Rees. Global nonlinear modeling of gas turbine dynamics using NARMAX structures. *ASME J. of Eng. for Gas Turbines and Power*, 124:817–826, 2002.

I. Ciocoiu. RBF networks training using a dual extended Kalman filter. *Neurocomputing*, 48:609–612, 2002.

R.W. Cue and D.E. Muir. Engine performance monitoring and troubleshooting techniques for the CF-18 aircraft. *ASME J. of Eng. for Gas Turbines and Power*, 113(1):11–19, 1991.

P. Dewallef and O. Léonard. On-line performance monitoring and engine diagnostic using robust kalman filtering techniques. In *ASME Turbo Expo*, number GT2003-38379, Atlanta GA, USA, 2003.

R. Eustace, B. Woodyatt, G. Merrington, and A. Runacres. Fault signatures obtained from fault implant tests on an F404 engine. *ASME J. of Eng. for Gas Turbines and Power*, 116:178–183, 1994.

C. Evans, P.J. Fleming, D.C. Hill, J.P. Norton, I. Pratt, D. Rees, and K. Rodriguez-Vazquez. Application of system identification techniques to aircraft gas turbine engines. *Control Engineering Practice*, 9:135–148, 2001.

J.H. Gove and D.Y. Hollinger. Application of a dual unscented Kalman filter for simultaneous state and parameter estimation in problems of surface-atmosphere exchange. *Journal of Geophysical research*, 111(D08S07), 2006. doi: 10.1029/2005JD006021.

T. Grönstedt. Identifiability in multi-point gas turbine parameter estimation problems. In *ASME Turbo Expo*, number GT2002-30020, Amsterdam, Netherlands, 2002.

T. Grönstedt. Least squares based transient nonlinear gas path analysis. In *ASME Turbo Expo*, number GT2005-68717, Reno-Tahoe NV, USA, 2005.

A. Gulati, D. Taylor, and R. Singh. Multiple operating point analysis using genetic algorithm optimisation for gas turbine diagnostics. In *15th International Symposium on Air Breathing Engines*, number 1139, Bangalore, India, 2001.

J.R. Henry. CF-18/F404 transient performance trending. In *AGARD-CP-448, Engine Condition Monitoring - Technology and Experience*, number 37, 1988.

R.E. Kalman. A new approach to linear filtering and prediction problems. *Trans. ASME, series D, J. of Basic Engineering*, 82(1):35–45, 1960.

L.J. Kerr, T.S. Nemec, and G.W. Gallops. Real-time estimation of gas turbine engine damage using a control-based kalman filter algorithm. *ASME J. of Eng. for Gas Turbines and Power*, 114:187–195, 1992.

T. Kobayashi and D. L. Simon. Application of a bank of Kalman filters for aircraft engine fault diagnostics. In *ASME Turbo Expo*, number GT2003-38550, Atlanta GA, USA, 2003.

T. Kobayashi, D. L. Simon, and J. S. Litt. Application of a constant gain extended Kalman filter for in-flight estimation of aircraft engine performance parameters. In *ASME Turbo Expo*, number GT2005-68494, Reno-Tahoe NV, USA, 2005.

G. Kulikov and H. Thompson, editors. *Dynamic modelling of gas turbines: Identification, simulation, condition monitoring and optimal control*. Advances in industrial control. Springer-Verlag, London, 1st edition, 2004.

J. Kurzke. Model based gas turbine parameter corrections. In *ASME Turbo Expo*, number GT2003-38234, Atlanta GA, USA, 2003.

Y.G. Li. A gas turbine diagnostic approach with transient measurements. *IMechE J. of Power and Energy*, 217(2):169–177, 2003.

L. Ljung. *System Identification, Theory for the User*. Prentice Hall Information and System Sciences Series, 2nd edition, 1999.

J. Loisy, J.P. Duponchel, and R. Carrillo. Steady and transient performance calculation method for prediction, analysis and identification. In *AGARD Lecture Series 183*, number 3, 1992.

R.H. Luppold, G.W. Gallops, L.J. Kerr, and J.R. Roman. Estimating in-flight engine performance variations using Kalman filter concepts. In *25th AIAA/ASME/SAE/ASEE Joint Propulsion Conference*, number AIAA-89-2584, Monterey CA, USA, 1989.

G. Merrington. Identification of dynamic characteristics for fault isolation purposes in a gas turbine using closed-loop measurements. In *AGARD-CP-448, Engine Condition Monitoring - Technology and Experience*, number 36, 1988.

G. Merrington. Fault diagnosis in gas turbines using a model-based technique. *ASME J. of Eng. for Gas Turbines and Power*, 116:374–380, 1994.

G. Merrington, O. Kwon, G. Goodwin, and B. Carlsson. Fault detection and diagnosis in gas turbines. *ASME J. of Eng. for Gas Turbines and Power*, 113:276–282, 1991.

K. Muske and T. Badgwell. Disturbance modeling for offset-free linear model predictive control. *Journal of Process Control*, 12:617–632, 2002.

A.T. Nelson. *Nonlinear Estimation and Modeling of Noisy Time-Series by Dual Kalman Filtering Methods*. PhD thesis, Oregon Graduate Institute of Technology, 2000.

A. E. Nielsen, C. W. Moll, and S. Staudacher. Modeling and validation of the thermal effects on gas turbine transients. *ASME J. of Eng. for Gas Turbines and Power*, 127: 564–572, 2005.

S. Ogaji, Y.G. Li, S. Sampath, and R. Singh. Gas path fault diagnosis of a turbofan engine from transient data using artificial neural networks. In *ASME Turbo Expo*, number GT2003-38423, Atlanta GA, USA, 2003.

W. Reinelt, A. Garulli, and L .Ljung. Comparing different approaches to model error modeling in robust identification. *Automatica*, 38:787–803, 2002.

RTO. Performance prediction and simulation of gas turbine engine operation. Technical Report 44, Research and Technology Organisation, 2002.

S. Sampath, Y.G. Li, S.O.T Ogaji, and R. Singh. Fault diagnosis of a two spool turbofan engine using transient data: A genetic algorithm approach. In *ASME Turbo Expo*, number GT2003-38300, Atlanta GA, USA, 2003.

M. VanDyke, J. Schwartz, and C. Hall. Unscented Kalman filtering for spacecraft attitude state and parameter estimation. In *AAS/AIAA Space flight mechanics conference*, number AAS-04-115, Maui HI, USA, 2004.

A. J. Volponi. Gas turbine parameter corrections. *ASME J. of Eng. for Gas Turbines and Power*, 121:613–621, 1999.

A. J. Volponi. Use of hybrid engine modeling for on-board module performance tracking. In *ASME Turbo Expo*, number GT2005-68169, Reno-Tahoe NV, USA, 2005.

A. J. Volponi and T. Brotherton. A bootstrap data methodology for sequential hybrid engine model building. In *2003 IEEE Aerospace Conference*, number 1061, 2003.

A. J. Volponi, T. Brotherton, and R. Luppold. Empirical tuning of an on-board gas turbine engine model for real-time module performance estimation. In *ASME Turbo Expo*, number GT2007-27535, Montréal QC, Canada, 2007.

P.P. Walsh and P. Fletcher. *Gas Turbine Performance*. Blackwell Science, New-York, 1998.

E. Wan and A.T. Nelson. *Kalman filtering and neural networks*, chapter 5 Dual extended Kalman filter methods. Wiley series on adaptive and learning systems for signal processing, communications and control. John Wiley & Sons, New-York, 2001.

E. Wan and A.T. Nelson. Dual Kalman filtering methods for nonlinear prediction, smoothing and estimation. In *Advances in Neural Information Processing Systems*, pages 793–799, Denver, CO, USA, 1996.

E. Wan and A.T. Nelson. Neural dual extended Kalman filtering: applications in speech enhancement and monaural blind signal separation. In *IEEE Workshop on neural networks for signal processing*, pages 466–475, Amelia Island, FL, USA, 1997.

R.K. Yedavalli and W. Li. Aircraft engine fault detection using dynamic/adaptive threshold approach. In *ASME Turbo Expo*, number GT2007-28180, Montréal QC, Canada, 2007.

Chapter 6

N. Aretakis, I. Roumeliotis, and K. Mathioudakis. Performance model zooming for in-depth component fault diagnosis. In *ASME Turbo Expo*, number GT2010-23262, Glasgow, UK, 2010.

S. Borguet and O. Léonard. Constrained sparse estimation for improved fault isolation. *ASME J. of Eng. for Gas Turbines and Power*, 133(12):121602, 2011. doi: 10.1115/1.4004013.

J. S. Litt. An optimal orthogonal decomposition method for Kalman filter-based turbofan engine thrust estimation. *ASME J. of Eng. for Gas Turbines and Power*, 130(1):011601, 2008.

M. Morini, M. Pinelli, P.R. Spina, and M. Venturini. Influence of blade deterioration on compressor and turbine performance. *ASME J. of Eng. for Gas Turbines and Power*, 132:032401, 2010. doi: 10.1115/1.4000248.

J. Pilet, J-L. Lecordix, N. Garcia-Rosa, R. Barènes, and G. Lavergne. Towards a fully coupled component zooming approach in engine performance simulation. In *ASME Turbo Expo*, number GT2011-46320, Vancouver BC, Canada, 2011.

RTO. More intelligent gas turbine engines. Technical Report AVT-128, Research and Technology Organisation, 2009.

D. L. Simon and S. Garg. A systematic approach to sensor selection for aircraft engine health estimation. In *19th International Symposium on Air Breathing Engines*, number ISABE-2009-1125, Montréal QC, Canada, 2009.

D. L. Simon, J. Bird, C. Davison, A. J. Volponi, and R. E. Iverson. Benchmarking gas path diagnostic methods: a public approach. In *ASME Turbo Expo*, number GT2008-51360, Berlin, Germany, 2008.

S. Spieler, S. Staudacher, R. Fiola, P. Sahm, and M. Weisschuh. Probabilistic engine performance scatter and deterioration modeling. In *ASME Turbo Expo*, number GT2007-27051, Montréal QC, Canada, 2007.

T. Wensky, L. Winkler, and J. Friedrichs. Environmental influences on engine performance degradation. In *ASME Turbo Expo*, number GT2010-22748, Glasgow, UK, 2010.

List of publications

Conference proceedings

1. S. Borguet, P. Dewallef, O. Léonard. *On-Line Transient Engine Diagnostics in a Kalman Filtering Framework*, in Proceedings of ASME Turbo Expo 2005, paper GT2005-68013, 2005, Reno NV, USA.

2. S. Borguet, V. Kelner, O. Léonard. *Cycle Optimisation of a Turbine Engine: an Approach Based on Genetic Algorithms*, in Proceedings of NCTAM06, 2006, Mons, Belgium.

3. M. Henriksson, S. Borguet, O.Léonard, T. Grönstedt. *On Inverse Problems in Turbine Engine Parameter Estimation*, in Proceedings of ASME Turbo Expo 2007, paper GT2007-27756, 2007, Montréal QC, Canada.

4. O. Léonard, F.Denis, J.-P. Thomas, S. Borguet. *From manual to model-based control of a small jet engine*, in Proceedings of XIX International Symposium on Air Breathing Engines (ISABE), paper ISABE-2009-1302, 2009, Montréal QC, Canada.

Peer-reviewed journals

1. S. Borguet, P. Dewallef, O. Léonard. *A Way to Deal with Model-Plant Mismatch for a Reliable Diagnosis in Transient Operation*, in ASME J. of Eng. for Gas Turbine and Power 130, 031601, 2008, doi: 10.1115/1.2833491.

2. S. Borguet, O.Léonard. *A Sensor-Fault-Tolerant Diagnosis Tool based on a Quadratic Programming Approach*, in ASME J. of Eng. for Gas Turbine and Power 130, 021605, 2008, doi: 10.1115/1.2772637.

3. O. Léonard, S. Borguet, P. Dewallef. *An Adaptive Estimation Algorithm for Aircraft Engine Performance Monitoring*, in AIAA Journal of Propulsion and Power, 28:4, pp 763-769, 2008, doi: 10.2514/1.34320.

4. S. Borguet, O. Léonard. *The Fisher Information Matrix as a Relevant Tool for Sensor Selection in Engine Health Monitoring*, in International Journal of Rotating Machinery, vol. 2008, article ID 784749, 2008, doi: 10.1155/2008/784749.

5. S. Borguet, O. Léonard. *A Generalized Likelihood Ratio Test for Adaptive Gas Turbine Performance Monitoring*, in ASME J. of Eng. for Gas Turbine and Power, 131, 011601, 2009, doi: 10.1115/1.2967493.

6. O. Léonard, J. P. Thomas, S. Borguet. *Ten Years of Experience With a Small Jet Engine as a Support for Education*, in ASME J. of Eng. for Gas Turbine and Power, 131, 012303, 2009, doi: 10.1115/1.2967487.

7. S. Borguet, O. Léonard. *Coupling principal component analysis and Kalman filtering algorithms for on-line aircraft engine diagnostics*, in Control Engineering Practice, 17:4, pp 494-502, 2009, doi: 10.1016/j.conengprac.2008.09.008.

8. S. Borguet, O. Léonard. *A Sparse Estimation Approach to Fault Isolation*, in ASME J. of Eng. for Gas Turbine and Power, 132(2), 021601, 2010, doi: 10.1115/1.3156815.

9. S. Borguet, O. Léonard. *Comparison of Adaptive Filtering Schemes for Gas Turbine Performance Diagnostics*, in Journal of Computational and Applied Mathematics, 2010, doi: 10.1016/j.cam.2009.08.075.

10. S. Borguet, O. Léonard. *Assessment of an Anomaly Detector for Jet Engine Health Monitoring*, in International Journal of Rotating Machinery, vol. 2011, article ID 942576, 2011, doi: 10.1155/2011/942576.

11. S. Borguet, M. Henriksson, T. McKelvey, O. Léonard. *A Study on Engine Health Monitoring in the Frequency Domain*, in ASME J. of Eng. for Gas Turbine and Power, 133(8), 081604, 2011, doi: 10.1115/1.4002832

12. S. Borguet, O. Léonard. *Constrained Sparse Estimation for Improved Fault Isolation*, in ASME J. of Eng. for Gas Turbine and Power, 133(12), 121602, 2011, doi: 10.1115/1.3156815.

Book chapters

1. S. Borguet, O. Léonard. A Quadratic Programming Framework for Constrained and Robust Jet Engine Health Monitoring, in *Progress in Propulsion Physics, vol. 1*, Eucass Book Series – Advances in Aerospace Sciences, EDP Sciences & Torus Press, pp. 669-692, 2009, ISBN 978-2-7598-0411-5.

University of Liège
Turbomachinery Group
Campus du Sart-Tilman, B52/3
B-4000 Liège Belgium
Tel: +32 43 66 91 87
Web: www.ulg.ac.be/turbo