Optimisation of a Diagnostic Test for a Truck Engine

Master's thesis
performed in **Vehicular Systems**

by
Petter Haraldsson

Reg nr: LiTH-ISY-EX-3183-2002

9th September 2002

Optimisation of a Diagnostic Test for a Truck Engine

Master's thesis

performed in **Vehicular Systems**,
Dept. of Electrical Engineering
at **Linköpings universitet**

by **Petter Haraldsson**

Reg nr: LiTH-ISY-EX-3183-2002

Supervisor: **Dr Mattias Nyberg**
SCANIA
Dr Erik Frisk
LiU

Examiner: **Assistant professor Erik Frisk**
Linköpings Universitet

Linköping, 9th September 2002

Avdelning, Institution
Division, Department

Vehicular Systems,
Dept. of Electrical Engineering
581 83 Linköping

Datum
Date

9th September 2002

Språk
Language
☐ Svenska/Swedish
☒ Engelska/English
☐ _____

Rapporttyp
Report category
☐ Licentiatavhandling
☒ Examensarbete
☐ C-uppsats
☐ D-uppsats
☐ Övrig rapport
☐ _____

ISBN
—

ISRN
LITH-ISY-EX-3183-2002

Serietitel och serienummer **ISSN**
Title of series, numbering —

URL för elektronisk version
http://www.vehicular.isy.liu.se
http://www.ep.liu.se/exjobb/isy/2002/3183/

Titel Optimering av ett diagnostest för en lastbilsmotor

Title Optimisation of a Diagnostic Test for a Truck Engine

Författare Petter Haraldsson
Author

Sammanfattning
Abstract

Diagnostic systems become more and more an important within the field of vehicle systems. This is much because new rules and regulation forcing the manufacturer of heavy duty trucks to survey the emission process in its engines during the whole lifetime of the truck. To do this a diagnostic system has to be implemented which always survey the process and check that the thresholds of the emissions set by the government not are exceeded. There is also a demand that this system should be reliable, i.e. not producing false alarms or missed detection. One way of producing such a system is to use model based diagnosis system where thresholds has to be set deciding if the system is corrupt or not. There is a lot of difficulties involved in this. Firstly, there is no way of knowing if the signals logged are corrupt or not. This is because faults in these signals should be detected. Secondly, because of strict demand of reliability the thresholds has to be set where there is very low probability of finding values while driving. In this thesis a methodology is proposed for setting thresholds in a diagnosis system in an experimental test engine at Scania. Measurement data has been logged over 20 hours of effective driving by two individuals of the same engine. It is shown that the result is improved significantly by using this method and the threshold can be set so smaller faults in the system reliably can be detected.

Nyckelord
Keywords threshold levels, OBD, residual, statistic, model, outliers, filtering

Abstract

Diagnostic systems become more and more an important within the field of vehicle systems. This is much because new rules and regulation forcing the manufacturer of heavy duty trucks to survey the emission process in its engines during the whole lifetime of the truck. To do this a diagnostic system has to be implemented which always survey the process and check that the thresholds of the emissions set by the government not are exceeded. There is also a demand that this system should be reliable, i.e. not producing false alarms or missed detection. One way of producing such a system is to use model based diagnosis system where thresholds has to be set deciding if the system is corrupt or not. There is a lot of difficulties involved in this. Firstly, there is no way of knowing if the signals logged are corrupt or not. This is because faults in these signals should be detected. Secondly, because of strict demand of reliability the thresholds has to be set where there is very low probability of finding values while driving. In this thesis a methodology is proposed for setting thresholds in a diagnosis system in an experimental test engine at Scania. Measurement data has been logged over 20 hours of effective driving by two individuals of the same engine. It is shown that the result is improved significantly by using this method and the threshold can be set so smaller faults in the system reliably can be detected.

Keywords: threshold levels, OBD, residual, statistic, model, outliers, filtering

Acknowledgement

This work has been carried out with cooperation with Scania AB. First, I want to thank my two supervisors Erik Frisk and Mattias Nyberg. I want to thank you for all the discussions and that you took time with all my questions. You have really helped me in the work of producing this thesis.

I want to thank all the people at Scania which helped me get the informations that I needed. I also want to thank all the people at the division of Vehicle System at Linköping Universitet for all their support and help.

Petter Haraldsson
Södertälje, 2002

Contents

Chapter 1

Introduction

This master's thesis has been performed for Scania in Södertälje, Sweden, both at the Linköping University at the department of Vehicle Systems and in Scania, Södertälje. Scania is a multinational company and a world leading truck manufacturer.

Background

The emission requirements for heavy truck engines have over the years become more and more strict, both in the US and Europe. In Europe there are legislation rules forcing the manufacturer to meet the needs of EURO 4 in 2005. Included in demands of EURO 4 there is both restrictions on pollution and demands of an on-board diagnosis (OBD) system. The purpose of such a system is to make sure that the requirements on emissions are kept, not only when the truck is new but also during the truck's whole operative life. The requirement is when a fault that will increase emissions appear, it should be detected. Faults would typically be due to wear or malfunction. An example of how a fault could influence emission is if the intake mass flow sensor has a bias fault by some percent and always show a higher value than it should. The fuel injection would then be affected and emission increases.

There are also demands, not from the government but from Scania not to produce any false alarms from the OBD system during the lifetime of a truck.

One way to construct a diagnosis system is to utilise model based diagnosis. This approach is, as the name implies, based on having a model of the engine and then on-line compare the measured signals from the engine with the output from the model. When the measured signals from the engine are sufficiently separated from the output from the model, then a fault has occurred. This is done by constructing a

test quantity and when this test quantity exceeds a certain threshold, a fault has occurred. This is done using statistical methods.

In an early development stage it is not economically realistic to base the threshold on measured data but on statistic assumptions. The reason for this is because it would require such a huge amount of data to base the thresholds on measured data, that it would be very expensive to collect this amount of data. In this thesis an algorithm will be proposed for constructing a test quantity and furthermore a methodology for thresholding this test quantity.

Objectives

Work has been put into developing a diagnosis system. In this system, a test quantity has to be constructed which will be thresholded. This threshold has to be set correctly to avoid missed detection and false alarms. There is though stringent condition for both false alarm rate and missed detection rate. With these stringent conditions a huge amount of data would have to be logged if not using statistical methods and this would be economical unrealistic. By using statistical methods much less data has to be logged and this is the reason why using statistical methods here.

The objective in this thesis is to develop an algorithm for producing a test quantity and correctly set the threshold which fulfill the stringent condition of false alarm rate and missed detection rate. The algorithm consists of several distinct "blocks" which can be changed and/or replaced when further development is done in the diagnosis system.

Methods

All signal processing has been evaluated in the Matlab/Simulink environment. The signals are taken from a measurement system installed in different heavy trucks. These measurement has been logged on a, by Scania ordained, test course.

Specifications

The problem of which fault that increases the emission is hard to specify. The reason for this is because there is not at the moment a proper understanding of which faults that are actually increasing the emissions. Because this lack of knowledge some assumptions has to be made about the faults. Assumptions that has to be verified lately on. In this thesis, bias and gain fault in some sensors are considered and assumed to increase the emissions. The reason for choosing to consider these faults is both, that there is a limit for how much a thesis may

contain, and bias and gain faults is two common faults in the diagnosis system.

Reader's Guide

Some fundamental mathematics and control theory are assumed to be known by the reader. This would not be a problem for undergraduate and graduate engineers specialised in signal processing. Not much knowledge is needed in vehicle systems, though it will make it easier to read with such a knowledge.

Chapter 2

About the Exprimental Engine

This chapter describes how turbocharged diesel engines work and in particular the experimental test engine of the trucks of this project.

The two most commonly used engine types today are the petrol engine, also known as the four stroke spark ignited (SI) engine, and the diesel engines, or the compression ignited engine. As petrol engine mostly are common in passenger cars, the diesel engines are mostly used in heavy duty trucks.

In a diesel engine, the fuel is injected directly into the cylinder. Firstly the air is inducted and compressed in the cylinder and thereafter the fuel is inducted. During the compression the temperature is increased to over the self ignition temperature of the fuel. First when the combustion is required to start, the fuel is injected. After a small period of time, when the liquid fuel evaporates and mixes with air, spontaneous ignition occurs. One advantage of this, compares to spark ignited engines, is that negative effects such as knock[1] is limited. The knock occurs because the combustion starts before the whole amount of fuel is injected.

The exprimental prototype test engine is fitted in a Scania 420 truck. A schematic overview of the engine is given in Figure 2.1. As can be seen in this figure the air are first compressed by the compressor and subsequently led through the intercooler. In the intake manifold the air is mixed with burned gases and inducted into the engine. In the engine the fuel is directly injected and mixed with the air and burned gases. Thereafter, the gases are led into the exhaust manifold. In the exhaust manifold, some of the exhaust gases are led back to the intake

[1] Knock occurs in spark ignited engine and can if not handled properly cause severe damage to the engine.

manifold but most of the gases is by the turbin led to the exhaust pipe.

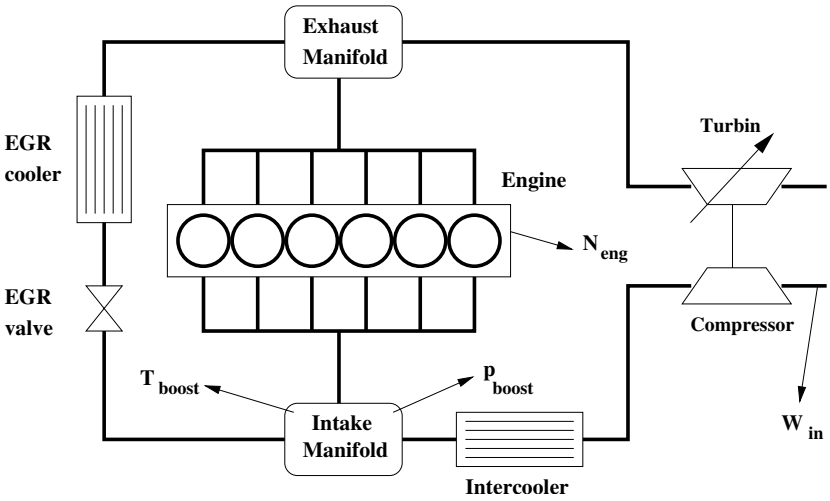

Figure 2.1: Schematic overview of the exprimental test engine.

2.1 Exhaust Gas Recirculation

The prototype engine consists of a Exhaust gas recirculation (EGR) system. EGR is a system which leads some of the exhaust fumes back to the intake of the engine. The concept of EGR has been introduced as a way to reduce nitrogen-oxide (NO_x) production and by this reduce the pollution from the engine. Since NO_x mainly is produced under high pressure and temperature, the way of decreasing the amount of NO_x is by either reducing the temperature or the compression in the combustion chamber.

The EGR system mainly affects the maximum combustion temperature. This is because the EGR mixes cooled exhaust gas and air in the intake manifold and dilute thereafter the normal, unburned gases in the combustion chamber.

There are however two major drawbacks with EGR. The first drawback is that it produces a more complex system of the engine which is more difficult to model and there is not at the moment a sufficiently good model for the system containing the EGR. Because of this, the EGR has to be shut down when running the diagnosis system.

The second drawback is that EGR decreases the power output from the engine and therefore the EGR is only active during low load condition. The use of EGR reduces the formation of NO_x up to 30 % and therefore the EGR can not be shut down for longer periods. If done

it will increase the amount of NO_x above the governmental rules and regulations.

2.2 The Intake System

It is the intake of the system of the engine that is modelled, but there are a number of difficulties in correctly modelling the intake of the system. One of the reason for this is that the intake system contains of several volumes which is the compressor, the inter-cooler and the intake manifold as described in Figure 2.1. All these volumes introduce dynamic into the system resulting in a dynamic and complex system. There are also standing waves in the intake manifold which sometimes result in a negative mass flow between the different volumes. Another problem is that there exist a break turbo which do affect the mass flow sensor in the intake resulting in an increase of mass flow sensed by the sensor but not inducted into the system. It is today not possible to measure when the break turbo is on, and this further complicates the process.

2.3 Sensors and Actuators

The sensors and actuators which are described in this thesis are p_{boost}, W_{in}, W_{bb}, N_{eng} and T_{boost}. The boost pressure sensors, which gives the signal p_{boost}, measures the pressure in the intake manifold. The mass flow sensor measures the mass flow before the compressor and produces the signal W_{in} . The estimated mass flow , W_{bb}, gives the the estimated mass flow from the black box model (see Section 5.1). Finally, the engine speed actuator produces the signal N_{eng}. The boost temperature sensor measures the temperature in the intake manifold and gives T_{boost}. Where in the engine these sensors are located can be viewed in Figure 2.1.

2.4 Faults to be detected

In this thesis the measured mass flow signal W_{in} will be examined for fault detection. Why choosing W_{in} for fault detection is that the mass flow sensor is the sensor which has the highest probability to have a fault. The accuracy for this sensor is not as good as the accuracy for the other sensors which are taken into consideration. Faults in p_{boost} and T_{boost} may also be detected, as well as other types of faults. A discussion about this can be read in Section 5.3.

There are two types of faults that will be examined and these faults are bias and gain faults. If there is a gain fault of θ_g and a bias fault

of θ_B, these faults can be described accordingly:

$$\begin{aligned} W_{gain} &= \theta_g W_{in} \\ W_{bias} &= \theta_b + W_{in} \end{aligned}$$

2.5 Rules and Regulations

In the OBD regulations on heavy duty trucks there is a demand of an on board diagnosis system which are defined in the regulations of EURO 4 and EURO 5. All new engines from 1 October 2005 must be certified with the OBD directives included in EURO 4, for EURO 5 the date is 1 October 2008. One year after these dates all vehicles and engines sold, registered and taken into service must comply with the directives.

The regulation of EURO 4 includes diagnosis. The threshold not to be exceeded and to be monitored by the OBD system is 7 g/kWh nitrogen oxide and 0.1 g/kWh particulates. In EURO 5 more stringent conditions not decided yet is to be monitored.

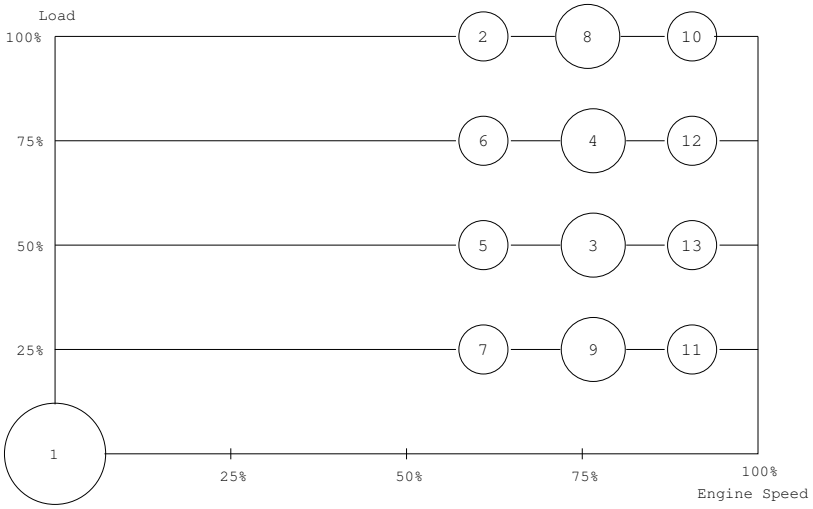

Figure 2.2: The figure describes the ten minutes long on board diagnosis cycle. Each circle describes each stationary point in the cycle. The size of each cycle is proportional to the weighting of that operating point and the number in the circles describes the order of the operating points.

2.5.1 OBD Test Cycle

The OBD test cycle is used to verify if the engines meets the criteria included in EURO 4. During the cycle, the engine speed and the load are changed during approximately ten minutes at a specified pattern which is described in Figure 2.2. The cycle contains of thirteen stationary point. The engine works in all these stationary for 40 seconds each. The transition time for the engine in between the stationary points holds for 20 seconds. The procedure is described as:

1. One fault is simulated or implemented.
2. The engine is preconditioned in three OBD test cycles, with engine startup and shutdown.
3. The engine is operated in one OBD test cycle, with engine startup and shutdown.

This procedure is repeated four times. The OBD system must all four times detect the fault, for the engine to meet the criteria for OBD systems mentioned in EURO 4.

2.6 Measurements

The measurement data was logged from trucks while driving on a test track. There was two separate trucks with the same kind of engine which was driving for 12.5 hours and 5 hours respectively. The driving was quite extreme. Some routes were with a lot of steep slopes while others were driving on a bumpy track. The data was collected at Scanias test track from 27 to 31 may 2002, with a sampling frequency of 20 Hz. A measurement program in windows named Gredi was used to log all the signals.

All signal processing in this thesis is based on all logged data from both of these two individual trucks. There is also 6 startups from the first truck and 34 start up...

Chapter 3

Theory Background

In this chapter, the theory background for the test quantity production chapters (see chapter 4 to 10) will be explained. The theory that has been gathered in this chapter is, as the name of the chapter implies, theory background and some new thoughts can be found in the following chapters.

3.1 Diagnosis

Diagnosis can be explain as for a process, in this case an engine, there are observed variables for which there is a knowledge of what is expected as normal. The task of diagnosis is to, from the observations and the knowledge, generate a diagnosis, i.e. to decide whether there is a fault or not. Including in diagnosis is also isolation of fault, this will however not be dealt with in this thesis.

Model based diagnosis is based on having a process and also a model of the engine. Comparing the model with the actual process then makes the diagnosis. An overview of how a diagnosis system is set up is shown in Figure 3.1.

The diagnosis system is run on the same input (i.e. input of signals) as the engine and the outputs (i.e. output of signals) from the engine are inputs to the diagnosis system. From these inputs the diagnosis system produces a statement,S, that tells if there is a fault or not and ideally which fault it is.

Comparing a test quantity, TQ, with a threshold J produces the statement S. This J can be adaptive or fixed and the TQ is supposed to express the difference between the engine and its model. The TQ should be small (ideally zero) when there is no fault in the system and it should be large when a fault is present i.e. if TQ exceeds the threshold there is a fault detected.

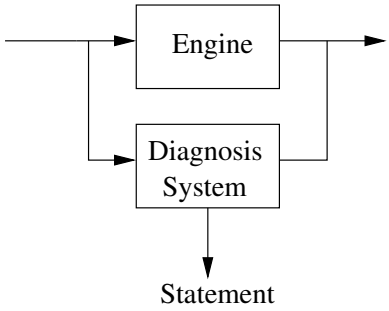

Figure 3.1: Overview of a diagnosis system.

When a test quantity is created it is based on a residual. This residual is the difference between a estimated value and a measured value. When signal processing this residual a test quantity is received. There are a number of ways how to create this residual and how to signal process this residual which will be discussed and compared in this thesis.

3.2 Hypothesis Testing

There is a set of observations $\mathbf{x}=(x_1, \ldots, x_n)$ from a distribution and a certain null hypothesis has to be tested. If there is for example a bias fault in a sensor then the distribution will depend on a parameter θ_b which here is 0 when there is no bias fault and $\neq o$ if there is a bias fault. The null hypothesis H_0 is then the fault free case i.e. $\theta_b = 0$. There are however not only one type of fault but several faults (which fault to be examined is discussed in Section 2.4 and Section 5.3). The test which are described in this thesis is if there is no fault or there is *any* fault out of all the possible faults. Therefore only one binary[1] hypothesis test is to be considered.

Another approach would be to use structured hypothesis testing [1] but the workload of doing such a test is much heavier. Isolation of fault is also a important part in structured hypothesis testing but there is no intention of fault isolation here. This is the reasons why this more direct approach to the problem is taken instead of the method with structured hypothesis testing.

If a test quantity $TQ(\mathbf{x})$ is defined, which is a function from the observations $\mathbf{x}$ to a scalar value, a comparison with a threshold J can be made. A declaration of a *critical region* C is also made which is a

[1] A binary test means that the outcome of the hypothesis test is one, out of two possible decisions

part of the region which TQ varies over. The following significance test may be used

$$\text{if } TQ \in C \text{ reject } H_0$$
$$\text{if } TQ \notin C \text{ do not reject } H_0$$

and a significance level can be defined as

$$\alpha = P(TQ \in C) \text{ if } H_0 \text{ is true}$$

The significance level here is the same as the false alarm rate for the system.

3.3 False Alarm Rate

Assume that there is a probability of $\frac{1}{1000}$ that a particular truck produces at least one false alarm during one year. Call this event A. This assumption is made because one of thousand trucks sold is allowed to produce one false alarm one time in one electric system during one year.

There exist 20 assumed independent electric system of which everyone is able to produce false alarms. That the OBD system alarms one time in one year produces a false alarm is called the event B_0 and that another system alarms one time in one year is called the event B_i for $i = 1 \dots 19$. Every time the engine is started, one test is to be made. This test longs for 600 second, which is the length of the OBD cycle. The shortest time between two tests is therefore $T_T = 600$ seconds. If the assumption that the truck is driven 3500 hours each year is made the maximum number of tests n during one year will be:

$$n = 3500 \cdot \frac{3600}{T_T} = 21000$$

This is of course an exaggeration and the number of tests is much lower.Here there is an assumption that the maximum number of tests for a year never exceeds 4000 tests If these tests are assumed to be independent then the event that the OBD system alarms at the i:th startup is called C_i. The event that the trucks has one failure in one year can be written accordingly:

$$A = B_0 \cup B_1 \cup \cdots \cup B_{19} \tag{3.1}$$

The probability that event A happen is:

$$P(A) = \sum_{i=0}^{19} P(B_i) = 20 \cdot P(B_i) \tag{3.2}$$

Here, the first equal sign holds because of the assumption of independence and the second equal sign holds because of the assumption that the events B_i happen with the same probability. The probability that the OBD system alarms one time in one year is according to (3.2):

$$P(B_0) = \frac{P(A)}{20} \tag{3.3}$$

The probability that the OBD system alarms one time in one year can also be written:

$$P(B_0) = \sum_{i=1}^{4000} P(C_i) = 4000 \cdot P(C_i) \tag{3.4}$$

Here, the first equal sign holds because of the assumption of independence and the second equal sign holds because of the assumption that the events C_i happen with the same probability. The probability that the OBD system will produce one false alarm during one test is can hence be written:

$$P(C_i) = \frac{P(B_0)}{4000} = \frac{P(A)}{20 \cdot 4000} = \frac{1}{1000} \cdot \frac{1}{20} \cdot \frac{1}{4000} \tag{3.5}$$

Here, the first equal sign holds because of (3.4), the second equal sign holds because of (3.3) and the last equal sign holds because that event A happen with a probability of $\frac{1}{1000}$.

The probability for four consecutive tests to produce a false alarm should be equal to (3.5) and if assume independence as before the probability for detect one false alarm will be:

$$(P(A \cap B \cap C))^{\frac{1}{4}} = (1.25 \cdot 10^{-8})^{\frac{1}{4}} = 0.0106 \tag{3.6}$$

This equation then gives the result $\alpha = 0.01$ which is the false alarm rate (or significance level) to be used here.

3.4 Missed Detection Rate

There is demand from the government that a fault has to be detected when running the OBD-cycle. The fault has to be detected four times in a row in the OBD-cycle (see section 2.5.1). The Assumption is made that a 10^{-2} chance of missed detection is the same as that a fault can be detected. When taken the fact that the fault has to be detected four times in a row into account the missed detection rate will be $1 - (1 - 10^{-2})^{\frac{1}{4}} = 0.0025$.

3.5 Threshold

A fixed or an adaptive threshold may be used when deciding the threshold J. When using a fixed threshold one may need to look at the histogram of the test quantity TQ. If instead the adaptive threshold is used, the TQ needs to be normalised with the adaptive threshold before examine the histogram.

Because of the very low false alarm rate, statistical methods are used. With statistical methods, the probability to find values in regions where there is low probability of finding values can be decided with comparatively small amount of data. Two approaches will be examined in this thesis and which are further described section 3.5.1 and 3.5.2.

3.5.1 Gaussian Distribution

One approach is to assume that the test quantity is Gaussian distributed. Why choosing Gaussian distribution is because when examine different test quantities, Gaussian distribution was a distribution that quite well fitted the observed test quantities and the thresholds are then based on that distribution. The definition of quite well is of course ambiguous but one has to choose a distribution and the Gaussian was chosen here. The Gaussian cumulative distribution function is defined as

$$\Phi(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} e^{\frac{-(t-\mu)^2}{2\sigma^2}} dt \tag{3.7}$$

where μ is the mean value and σ is the covariance. The probability that the stochastic variable X will be within the value a and $-a$ should be $1-\alpha$ and it can be written

$$P(-a < X < a) = \Phi(\frac{a}{\sigma}) - \Phi(-\frac{a}{\sigma}) = 1 - \alpha \text{ if } X \in N(0, \sigma) \tag{3.8}$$

From this equation the threshold J is set to a. It is though important to notice that the assumption of Gaussian distribution is made and if the real distribution is *too* distinguished from the Gaussian distribution then the threshold will be wrongly set.

3.5.2 Tail Distribution Estimation

Another approach suggested by [2] is to estimate a exponential distribution to the tail distribution of the test quantity. Why doing this is because, as said before, it is only the tail that is of interest while setting thresholds for very low false alarm rates. Most distributions are also

approximately exponentially distributed in the tail of their distribution and it is therefore a good choice for a estimation.

Adapt the following distribution to the tail (starting at h_0) of the test quantity (or the test quantity normalised with a adaptive threshold):

$$p_{TQ}(x) = \frac{1}{\mu} e^{-\frac{x}{\mu}} \tag{3.9}$$

With the false alarm α and the estimated mean value of the exponential distribution $\hat{\mu}$, the threshold J can be chosen as

$$\int_J^\infty \frac{1}{\hat{\mu}} e^{\frac{-(x-h_0)}{\hat{\mu}}} dx \tag{3.10}$$

or, if solving the integral

$$J = h_0 - \hat{\mu} \ln(\alpha) \tag{3.11}$$

3.6 Sample Kurtosis

When setting the threshold then it is only the tail of the distribution that it is of interest. This is because it is only the regions with low probability that is of interest. If there are values when there is low probability of finding one, assumed having Gaussian distribution, the assumption of Gaussian distribution is not correct and therefore the threshold can not be set to a.

One way of investigate if the distribution of the test quantity is near Gaussian is to look at the histogram of the test quantity and compare it with a Gaussian distribution. Another method is to use the *sample kurtosis* which is defined as

$$\kappa = \frac{E(x-\mu)^4}{\sigma^4}; \tag{3.12}$$

where $E(x)$ is the expected value of x. Kurtosis is a measure of both peakedness and tail weight and the interpretation is not straightforward but one can in most cases look at it as an measurement of the "flatness" or the "peakness" of the distribution. For a more throughout explanation of the concept one may look in [3].

The kurtosis of the Gaussian distribution is 3. Distributions that are more outlier-prone than the Gaussian distribution have kurtosis greater than 3 and distributions that are less outlier-prone have kurtosis lesser than 3. Looking at the distribution of a test quantity then the kurtosis of that test quantity should be no greater than 3.

3.7 Power Function

When using hypothesis testing described in section 3.2, the null hypothesis H_0 should not be rejected when it is true. The mistake to reject H_0 when H_0 is true is called TYPE I error and this is the false alarm rate α.

Similarly, not to reject H_0 when the alternative hypothesis H_1 is true is called TYPE II error and is the chance of missed detection denoted β. The possible faults can be summarised:

TYPE I error - false alarm rate α or $1-\alpha$ chance of accepting a value within the acceptable boundaries.

TYPE II error - missed detection β or $1-\beta$ chance of rejecting a value not within the acceptable boundaries.

and from this the power function $h(\theta)$ can be defined as

$$h(\theta) = P(\text{reject } H_0|\theta) = P(T > J|\theta) \tag{3.13}$$

where θ here is a variable that the distribution depends on. With TYPE I and TYPE II errors in mind, the power function can also be described as

$$\begin{aligned} &\text{if } \theta \in H_1 \text{ then } h(\theta) = 1 - \beta(\theta) \\ &\text{if } \theta \in H_0 \text{ then } h(\theta_0) = \alpha \end{aligned}$$

where $\theta = \theta_0$ when $\theta \in h(0)$ and $\theta \neq \theta_0$ otherwise. The critical region C is chosen as to keep the probabilities of both types of errors small. However both probabilities can not be arbitrarily small because a decrease in α results in an in increase in β. Since there is a demand as for keeping α small, an assignment of the TYPE I error probability α is done. The search is then for a critical region C of the sample space so as to minimise the TYPE II error probability for θ.

3.7.1 Estimating the Power Function

The power function $h(\theta)$ may be estimated by using simulations since it is very hard or even impossible to derive the power function analytically. The method used here is called Monte Carlo simulation and can be described as follows:

1. An assumption of a distribution of noise in the data is made.
2. The parameter θ is fixed for which $h(\theta)$ is calculated.
3. A large amount of data is generated from a Scania truck while driving.

4. For this data series, the test quantity TQ_i is calculated
5. All the n values TQ_i is collected in a histogram.
6. By using the fixed threshold J, $h(\theta)$ can be estimated.
7. Go back to step 2 and fix a new θ.

If not using fixed threshold the same methodology can be used with the exception of normalisation the test quantity TQ with the adaptive threshold $J(x)$.

The power function can be evaluated both for bias faults and for gain faults. The gain faults will in this thesis be evaluated from -50% to +50% and the bias fault from -50% to +50% of the mean of the signal with the exception that the sensor is assumed not to give negative values.

Chapter 4

A Test Quantity Algorithm

A test quantity algorithm is proposed. Given measurement data, a residual is created based on a model of the intake engine. At first, some of the noise in the signal is reduced. Thereafter, some of the values are rejected. After that the signal is normalised and the outliers are disregarded. Finally, it will produce a test quantity TQ which is to be thresholded.

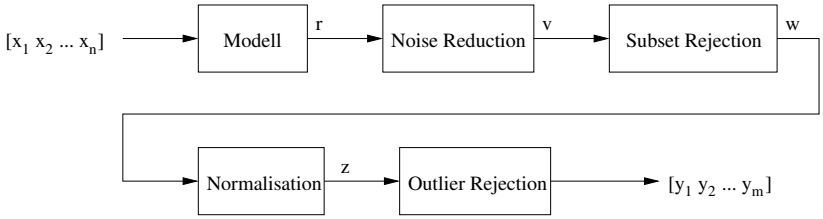

Figure 4.1: The test quantity algorithm schematically described.

The algorithm is schematically described in Figure 4.1 where $[x_1\ x_2\ \ldots x_n]$ is the measured data and $[y_1\ y_2\ \ldots y_m]$ is the test quantity TQ produced. This test quantity shall be thresholded to decide if there is a fault or not. Firstly there exist measuring data $\mathbf{x}$ which applied to a model produces a residual $\mathbf{r}$. How to choose model is explained in Chapter 5. Noise is reduced from the signal $\mathbf{r}$, producing $\mathbf{v}$. How this is done is thoroughly explained in Chapter 6. Some of the values are thereafter rejected and the remaining values are the signal $\mathbf{w}$. Which values to reject and how this is done is discussed in Chapter 7. This signal is furthermore normalised giving $\mathbf{z}$. A discussion about normalisation can be read in Chapter 8. At the end, outlier rejection is

made eventually producing the test quantity $\mathbf{y}$, also denoted TQ. The outlier rejection algorithm is explained in Chapter 9. This test quantity TQ is then thresholded. How to threshold is discussed in Chapter 10. In the following chapters the different signals will be denoted according the nomenclature described in figure 4.1.

Each block in the algorithm will be thoroughly explained in the forthcoming chapters and the goal here is to produce both a pragmatic and a general algorithm that can be easily reused and modified.

Chapter 5

Different Models

There are three different models that is to be examined in this thesis. These models are two static models and one dynamic model. These models are described in the following sections in this chapter. All the models is valid when the EGR is shut off but can be replaced by newer and better model when they are available.

5.1 Scanias black box model

There has been a development of a model at Scania for the intake of the system for some times. This model will in this thesis be viewed as a "black box" which mean that the input and the output of the system only will be examined without worrying about what is happening inside the "black box". This model will be called $ScBB$. This model is static[1] and producing an estimated mass flow in the intake of the system. The residual, i.e. in this case the difference between the measured mass flow and the estimated mass flow, will be denoted r_{ScBB}.

5.2 The Volumetric Efficiency Model

The measure used to measure the effectiveness of an engines induction process is the *volumetric efficiency*, η_{vol} (see e.g. [4]) . The volumetric efficiency is defined as the volume flow rate of air into the intake system, $\dot{V}_a$, divided by the rate at which volume is displaced by the piston, $\dot{V}_d$:

$$\eta_{vol} = \frac{\dot{V}_a}{\dot{V}_d} = \frac{2 \cdot 60 R_{air} T_{boost} W_e}{p_{boost} N_{eng} n_{cyl} V_d} \tag{5.1}$$

[1] That the model is static means that no old information is included in the model, the opposite is a dynamic model.

The estimated mass flow W_e are then compared with the measured mass flow in the intake of the system producing a residual:

$$r_{SEv} = W_{in} - W_e \tag{5.2}$$

The volumetric efficiency of the engine can reach values above unity due to standing waves in the intake manifold but is for the test engine between 0.89-0.92. The volumetric efficiency is usually displayed in a 3-D plot depending on number of revolutions, N_{eng}, and the boost pressure, p_{boost} and this is also the concept that is been using here. The volumetric efficiency map has been produced by driving in a motor test cell at Scania.

5.2.1 The Volumetric Efficiency Map

In the model that is being used here, there is a need for a volumetric efficiency map. The method for producing this map is by driving in a motor test cell. In this motor test, the load and the engine speed are changed while the boost pressure, the mass flow of air and the boost temperature are measured. From these measurement a map is developed, where the values between the measured values are interpolated.

One problem with the mass flow sensor, W_{in}, that is mounted on the engine is that there has to be a certain amount of flow of air for the sensor to work correctly. This amount of air into the intake system is not always enough so when the amount of air is low, the mass flow sensor gives far too low values, i.e. in a certain operating range, the sensor is not working correctly. The operating range was examined in plots from measured data taken from trucks while driving.

There is much individual variation in the mass flow sensor, W_{in}, for different trucks. Therefore, it is hard to determine in which region the mass flow sensor works correctly or the operating range for the mass flow sensor W_{in}. This means that the operating range for the mass flow sensor can not be decided from the amount of mass flow. Instead, it has to be decided from the pressure boost sensor and the engine speed, which do not vary so much between different individuals. It was found out that, when the pressure boost sensor, p_{boost}, is between 100-104 kPa and the engine speed, N_{eng} is between 1000-1500 rpm then the volumetric efficiency map is not correct. In one of the trucks it differentiates as much as 55 % from the correct value. The data received in this operating range is disregarded when further signal processing the signal.

When running in the motor test cell, there are slightly different conditions compares to, when the engine is placed in a truck. This means that there need to be some slight adjustment in the map, to gain the best available model of the engine. There are also unknown individ-

ual variations between engines and how the sensors are implemented in different trucks. To examine these individual variations and how much the motor test cell deviates from the real model, data from more trucks need to be examined.

When examine the signals from the two trucks, the mean of the residual of the signals deviated from zero with 0.0186 kg/s respectively 0.0130 kg/s for the two different trucks. Here, the volumetric efficiency map is compensated for this by multiplying the volumetric efficiency map with 0.94 or the mean is moved 0.0153 kg/s towards zero. This is the same as that the assumption is made that the real volumetric efficiency is 6 % lower than the volumetric efficiency produced in the test cell.

5.3 Fault Modelling

Power function is a measure of how good the system can detect faults (see Section 3.7). As described in Section 2.4, there are bias and gain faults in the signal W_{in}, which will be examined. From (5.1) and (5.2) the residual can be written accordingly:

$$r_{SEv} = W_{in} - \frac{p_{boost} N_{eng} n_{cyl} V_d \eta_{vol}(p_{boost}, N_{eng})}{2 \cdot 60 R_{air} T_{boost}} \tag{5.3}$$

The volumetric efficiency n_{vol} depends on the boost pressure p_{boost} and therefore a fault in the mass flow signal ,W_{in}, affects r_{SEv} different than how a fault in the boost pressure signal affects the residual. This means that the power function for fault in the boost pressure signal will be different to the power function for fault in the mass flow signal. Only power functions for faults in the mass flow sensor is considered when optimising the test quantity algorithm. The reason for this that the mass flow signal is the most unreliable signal of the signals in the engine and it is most probably that there will be a small bias or gain fault in this signal which need to be detected. This does not mean that faults in the boost pressure signal and the boost temperature signal can not be detected, because faults in these signals can be detected. A fault in either the boost temperature signal. T_{boost}, and a fault in the boost pressure signal, p_{boost} will affect r_{SEv} in (5.3). But the algorithm is not optimised with respect to these signals. Other types of faults in the engine may also be detected. If e.g. there is a leakage in the intake of the engine, this fault may affect the pressure boost sensor but not the mass flow sensor and hence the residual will depart from zero. Exactly which faults, except fault in W_{in}, that can be detected needs to be examined further and is out of scope for this thesis.

5.4 A Dynamic Model

If taken the dynamic of the system into account, an dynamic model can be used. A dynamic model is produced and analysed in the following sections.

5.4.1 A Model Based on an Observer

An observer can be used to model the system (For a throughout explanation of the concept observer see e.g. [5]). This observer is derived accordingly:

The ideal gas law is

$$pV = mRT \tag{5.4}$$

and deriving 5.4 applied to the intake system produces the equation

$$\dot{p}_{boost}V_{tot} = (W_{in} - W_e)R_{air}T_{boost} \tag{5.5}$$

where the mass flow, W_e is

$$W_e = \frac{\hat{p}_{boost}N_{eng}n_{cyl}V_d\eta_{vol}(\hat{p}_{boost}, N_{eng})}{2 \cdot 60 \cdot R_{air}T_{boost}} \tag{5.6}$$

A feedback with the estimated pressure denoted $\hat{p}_{boost}$ minus the boost pressure gives the observer

$$\dot{\hat{p}}_{boost} = \frac{R_{air}T_{boost}}{V_{tot}}(W_{in} - W_e) + K(p_{boost} - \hat{p}_{boost}) \tag{5.7}$$

where K is a design variable. A residual is then finally computed as

$$r_{obs} = p_{boost} - \hat{p}_{boost} \tag{5.8}$$

5.4.2 How K affects the Dynamic Model

Assume that there is a fault in the mass flow sensor. Applying a mass flow sensor fault δW_{in} in (5.7) the equation

$$\dot{\hat{p}}_{boost} = \frac{R_{air}T_{boost}}{V_{tot}}(W_{in} + \Delta W_{in} - W_e) + K(p_{boost} - \hat{p}_{boost}) \tag{5.9}$$

will be given. Assume thereafter steady state, or $\dot{\hat{p_{boost}}} = 0$ and using (5.6) in (5.9). This will give the equation

$$0 = \frac{R_{air}T_{boost}}{V_{tot}}(W_{in} + \Delta W_{in}) + Kp_{boost} - (\frac{n_{cyl}V_d}{2 \cdot 60V_{tot}}N_{eng}\eta_{vol}(N_{eng}, \hat{p}_{boost}) + K)\hat{p}_{boost} \quad (5.10)$$

or in another form:

$$\hat{p}_{boost} = \frac{\frac{R_{air}T_{boost}}{V_{tot}}(W_{in} + \Delta W_{in}) + Kp_{boost}}{K + \frac{n_{cyl}V_d}{2 \cdot 60V_{tot}}N_{eng}\eta_{vol}(N_{eng}, \hat{p}_{boost})} \quad (5.11)$$

Now, (5.8) and (5.11) produces the following residual:

$$r_{obs} = p_{boost} - \frac{\frac{R_{air}T_{boost}}{V_{tot}}(W_{in} + \Delta W_{in}) + Kp_{boost}}{K + \frac{n_{cyl}V_d}{2 \cdot 60V_{tot}}N_{eng}\eta_{vol}(N_{eng}, \hat{p}_{boost})} \quad (5.12)$$

If there is no fault, i.e. $W_{in} = W_e$, for $K = 0$, (5.12) will be zero. The absolute value of the difference for the residual when there is a fault and when there is no fault will be

$$abs(r_{fault} - r_{nofault})(K) = \frac{\frac{R_{air}T_{boost}}{V_{tot}}\Delta W_{in}}{K + \frac{n_{cyl}V_d}{2 \cdot 60 \cdot V_{tot}}N_{eng}\eta_{vol}(N_{eng}, \hat{p}_{boost})} \quad (5.13)$$

which has a maximum for $K = 0$, because K is always larger than zero. The difference between the residual when there is no fault and when there is a fault ought to be as large as possible since faults need to be detected. The free parameter $K = 0$ shall therefore be used. Notice that it does not matter if ΔW_{in} is a gain or bias fault in W_{in}, (5.13) holds for both of these faults.

5.4.3 Step Response of the Dynamic System

How the free parameter K affects how fast the system is hard to know because the system is not linear. A linearisation of the system will be given in this section to find out the step response. The step response is a measure of how fast the system is.

Assume that $u = [T_{boost}W_{in}\ \ N_{eng}]^T = [u_1\ \ u_2]^T$ and $y = p_{boost}$ and that η_{vol} is a constant, which is a quite good approximation since $0.89 < \eta_{vol} < 0.92$. The equation (5.5) can then be written as

$$\dot{\hat{p}}_{boost} = k_1u_1 - k_2\hat{p}_{boost}u_2 \quad (5.14)$$

where $k_1 = \frac{R_{air}}{V_{tot}}$, $\eta_{vol} = C$ (here C is a constant) and $k_2 = \frac{n_{cyl}V_dC}{2 \cdot 60V_{tot}}$.

If this equation evaluates for constant number of revolution , i.e. $\dot{N}_{eng} = 0$, the linear equation

$$\dot{\hat{p}}_{boost} = k_1 u_1 - k_3 \hat{p}_{boost} \tag{5.15}$$

where $k_3 = k_2 \cdot N_{constant}$ is given. It is known (see e.g. [5] page 143) that for linear systems, K is a adjustment between how fast the system is and how much disturbances affect the system. The higher K the faster system but also more sensitive to noise. This can be seen in Figure 5.1 which describes the step response for the dynamic system in the upper most plot when $K = 0$ and in the under most plot when $K = 10$.

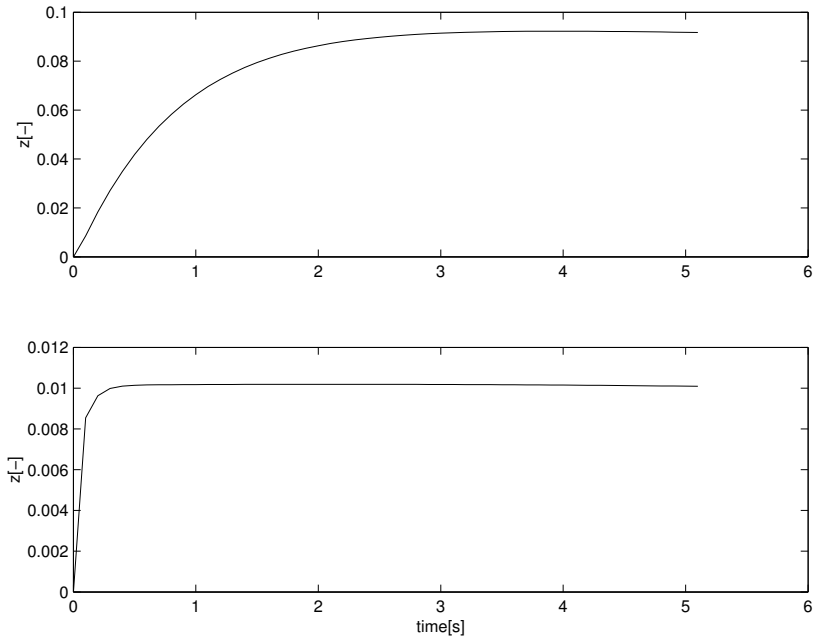

Figure 5.1: The step response of the dynamic model with a fault of 10%. The uppermost plot describes the step response when the feedback is set to zero and the undermost plot describes the step response when the feedback is set to ten.

5.5 Which Model to Use

It has been shown that in steady state, $K = 0$ is the optimum choice for the dynamic model. It is not obvious that this is the optimum choice when not in steady state but this has not been examined in this thesis.

In Figure 5.2 the power functions for the static black box model, the volumetric efficiency model and the dynamic model can be compared.

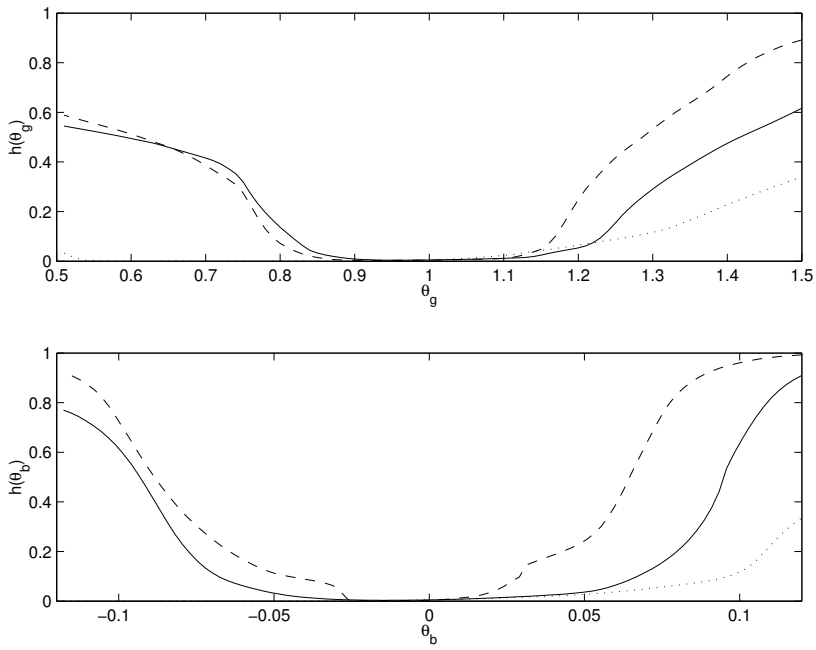

Figure 5.2: The power functions for the volumetric efficiency model (solid), the dynamic model (dashed) and the black box model (dotted).

It can be seen in this figure that of the two static models, r_{SEv} and r_{ScBB}, the volumetric efficiency model is best. When further signal processing the signal, the static volumetric efficiency model will be used. The reason for not using the dynamic model is that it takes a lot of time to simulate the dynamic model. It is worth mentioning that the dynamic model do not have to be better than the static model when running the test quantity algorithm described in Chapter 4 just because the power function for the dynamic model here is better than the power function for the static model. It should be investigated further, which model that produces the best test quantity, but it is out of scope for this thesis.

Chapter 6

Noise Reduction

In the last chapter, the residual r_{SEv} was constructed and in this chapter the signal $\mathbf{r}$ in Figure 4.1 will be examined and the signal $\mathbf{v}$ will be produced.

When examining signals in signal processing, there is often a lot of noise disturbing the signal. This noise needs to be reduced so the information in the signal can be obtained. One common method to reduce the noise in the signal is by low pass filtering the signal and this is also the method used here. Because the system will be implemented in a computer, time discrete filters is to be considered. The most common way of classify time discrete filters is to divide them according to the impulse response:

1. FIR-filter (Finite-duration Impulse Response)
2. IIR-filter (Infinite-duration Impulse Response)

For a more throughout explanation of time discrete filters see e.g. [6]. Before deciding which filter to use, it is important to find out how much data the filter is able to use.

6.1 How Much Old Data to Use

There is a strive to reduce the noise in the signal as much as possible. One intuitive way is to take the mean of the signal of a very long period of time. This is because it is only the low frequency differences in the signal that is of interest. When a fault occur, it is assumed that it does effect the signal over a long period of time. The high frequency differences in the signal are of no interest and shall be reduced as much as possible. Hence, the longer period of time, to take the mean over, the more high frequency differences are reduced, and the better result.

Instead of taken the mean of the signal, it can be low pass filtered with an IIR-filter, with a very low cut off frequency (exactly how to do this is explained in Section 6.3). Also with the low pass filtering, much data shall be used to gain the best performance of the system.

How much data can be taken into consideration? When having a model which only works when the EGR is shut off, the data taken into consideration can only be data from when the EGR is shut off. The EGR is shut off one second at a time and therefore only one second of data can be taken into consideration. When having a model which works with the EGR, all data can be taken into consideration and the result may be improved. Hence, there are two cases, one with EGR and one without.

6.2 FIR-filters

There are certain benefits for choosing FIR-filter compared to IIR-filters. The benefits are:

1. They are always stable.
2. They can be implemented non recursively, which is a chain of delay elements multiplied with constants and then added together.
3. They have linear phase characteristic.
4. They do not oscillate.

There are however one drawback with FIR-filters and that is they sometimes tend to have long impulse response and therefore a long time delay to be effective.

This is a problem only if taken much data into consideration when low pass filtering it. If there is only one second of data to handle, which is the time the EGR is shut off, then this is not a problem. The delay will only be for one second. Consequently, the FIR-filter will be used in this case. But if taken all the data into consideration, FIR-filter will not be the optimum choice. Here an IIR-filter is better because it will have a very short time delay and it will easier be implemented in software.

To use the FIR filter with a low cut off frequency is almost the same as taken the mean of the twenty values within the second of data which can be used. This is also the reason why choosing FIR filter when the EGR is shut off. The reason for using the FIR filter instead of taken the mean of the twenty values is because with the FIR filter, the constants multiplied with the delay elements are not fixed. In this way, the FIR filter may be improved by changing the cut off frequency, and hence the constant before the delay elements. Remez algorithm is used when

designing the FIR-filter. There are however one shortcoming with this method. Faults, which only happen for small amount of mass flows can not be detected. This is important to have in mind while reading this chapter.

6.3 IIR-filters

IIR-filters have infinite impulse response and used here are causal and stable filters. There exists several methods for optimising the filters with respect to different constraints (butterworth filter, chebychev filter etc). Chebychev I filters have fast roll off between pass band and stop band and was chosen here.

There are several design parameters to choose for IIR-filters. The first one to be decided is the order of the filter. The aim is to set the order as low as possible but also have enough element in the filter so the system's dynamic correctly can be handled. Three was the choice here as a good compromise. The cut off frequency has also to be set correctly and how this frequency was chosen can be seen in Section 6.4. The last design parameter is to decide how much peak-to-peak ripple that is allowed in the pass band and 5 dB was chosen here.

The IIR-filter is chosen for the case when the EGR is not shut off. The reason for this is because IIR-filter is faster that FIR-filter and it needs less memory in software when implementing the filter.

6.4 The Cut Off Frequency

The signal is assumed to vary slowly. This assumption is made because the faults is assumed to affect the output during a long time. The cut off frequency need therefore to be set low and frequencies under 0.15 rad/s is to be considered.

In Figure 6.1, a comparison between the power function for chebychev I filters with different cut-off frequencies can be viewed. The cut-off frequency was chosen here to 0.00314 rad/s.

For the FIR-filter there can be seen no difference in the power function for cut-off frequencies of 0.15 rad/s and lower. The cut-off frequency is therefore chosen to be 0.157 rad/s.

6.5 A Comparison Between the Filters

In Figure 6.2 the power function for the different filters applied to the signal can be compared. Here can be seen that the IIR-filter is best, the FIR-filter second best and not applying any filter is worst. It is important to have in mind that the IIR-filter uses more information

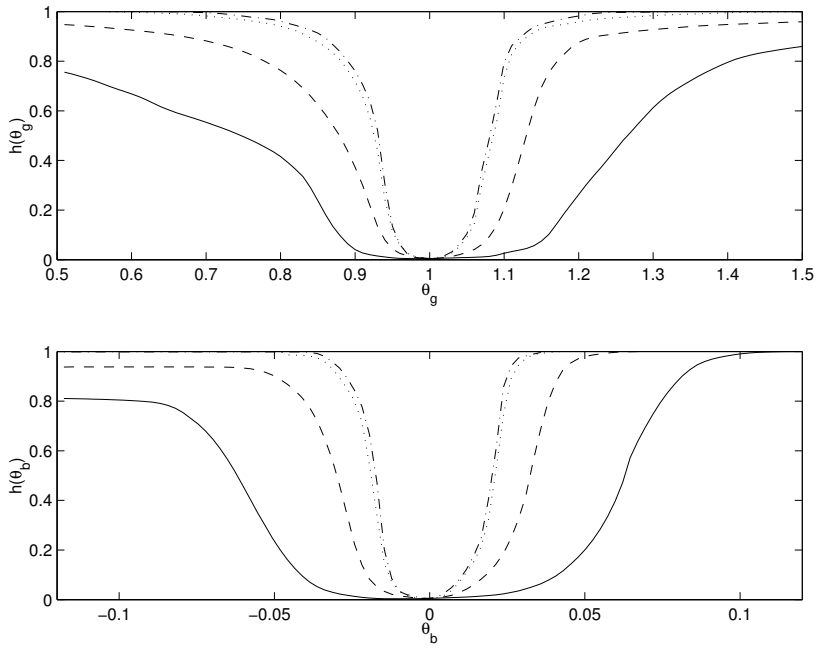

Figure 6.1: Power functions for **r**. It is shown for different cut-off frequencies for the IIR-filter. The different cut-off frequencies are 0.157 rad/s (solid), 0.0157 rad/s (dashed), 0.0314 rad/s (dotted) and 0.00157 rad/s (dash dotted).

that the FIR-filter, and therefore has so much better power functions. The IIR-filter is used when there is a model with EGR and the FIR-filter is used when there is a model without EGR.

In the following chapters the FIR-filtered signal is used when further signal processing is made. It is chosen because at this moment the EGR has to be shut off for the system to work properly.

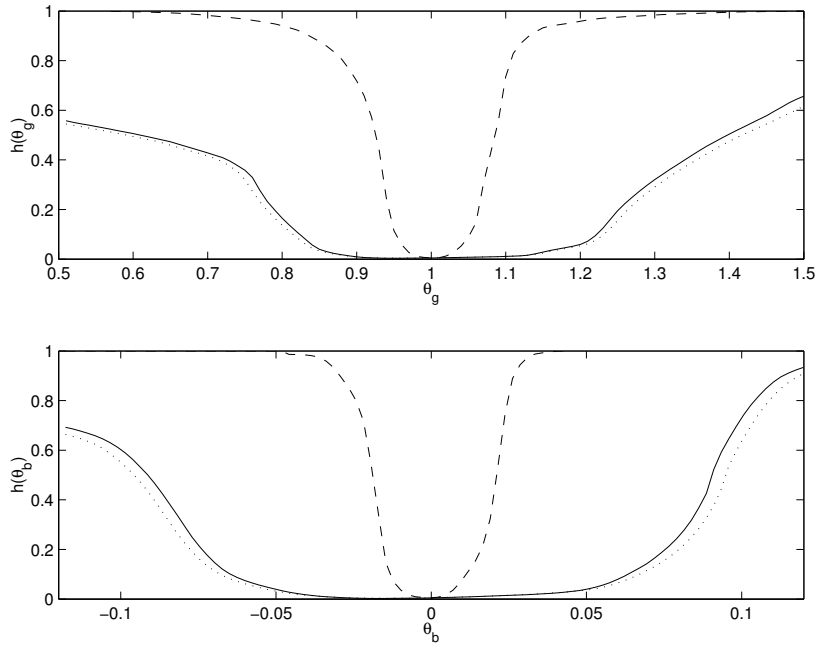

Figure 6.2: Power functions for the IIR-filter applied to the signal (dashed), the FIR-filter applied to the signal (solid) and no filter applied to the signal (dotted).

Chapter 7

Subset Rejection

In the test quantity algorithm, the signal **v** in Figure 4.1will be examined in this chapter and the signal **w** will be produced.

Assume that the model is good when the mass of air flowing into the intake of the engine is high but not so good when the amount of mass flow is low. Then a criterium can be set to decide when the model is valid or not based on the amount of mass flow. This means that there will be a threshold to be set based on the amount of mass flow. If the mass flow exceeds this threshold, the value is taken into consideration but if the mass flow does not exceed this threshold, the value will be disregarded.

If applying this criterium on the signal the result may be improved. How to do this and the improvement of the result when applying this criterium will be discussed in this chapter.

7.1 Validation of the Assumption

If setting a criterium of the signal depending on the amount of mass flow, it is possible to compare the power functions for different criteria on the mass flow. The power function for the residuals with mass flow over 0.3 kg/s can e.g. be compared with the power function for the same residual but with mass flow over 0.4 kg/s. By comparing different power functions with different criteria applied to them an optimum choice of the threshold can be obtained.

In Figure 7.1, power functions with different criteria on the signal can be viewed. It is seen that the power function becomes better when low mass flow values are disregarded, if comparing to the power function when not disregard any values. The assumption that the model is better when the mass of air flowing into the intake of the engine is high, compares to when the amount of mass flow is low, is therefore

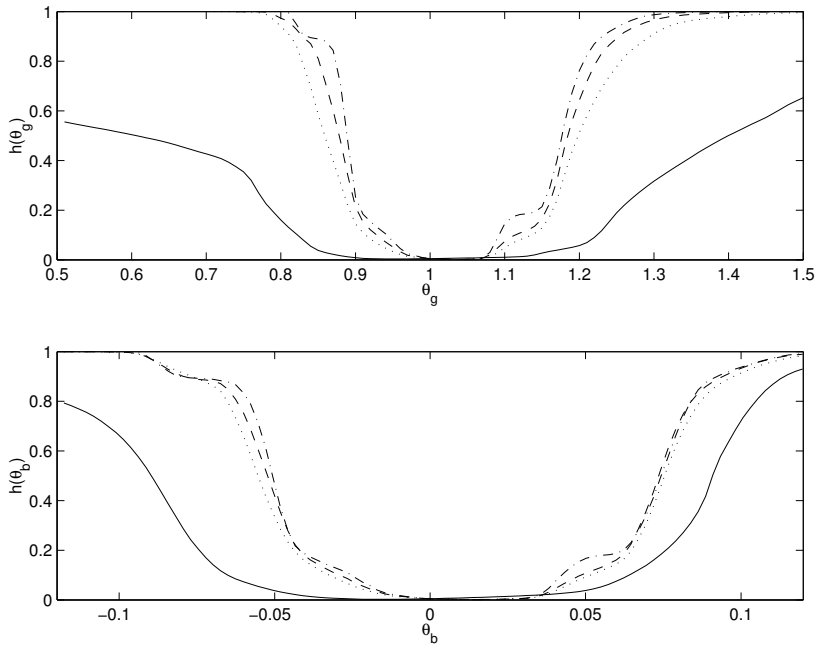

Figure 7.1: For the signal **v**, a comparison between not applying any criteria (solid), the mass flow do not have to exceed 0.35 kg/s (dotted), 0.40 kg/s (dashed) and 0.45 kg/s (dashdotted).

correct.

7.2 Setting the Threshold

The accuracy of the model is as described in Section 7.1 depending on the amount of air flowing into the intake of the engine. In Figure 7.1, power functions for different thresholds can be viewed. Here can be seen that with a higher threshold, the result mostly will be improved.

There is however one limitation on how high the threshold can be set. The higher threshold set the less values left, and with less values it may be hard to find values to base the diagnose on. In Figure 7.2 it is shown how many values in percent that is left for different thresholds. Consequently, the threshold can not be set too high and 0.35 kg/s was chosen as a good compromise. The power function is quite good for this value and there is 28.9% values left.

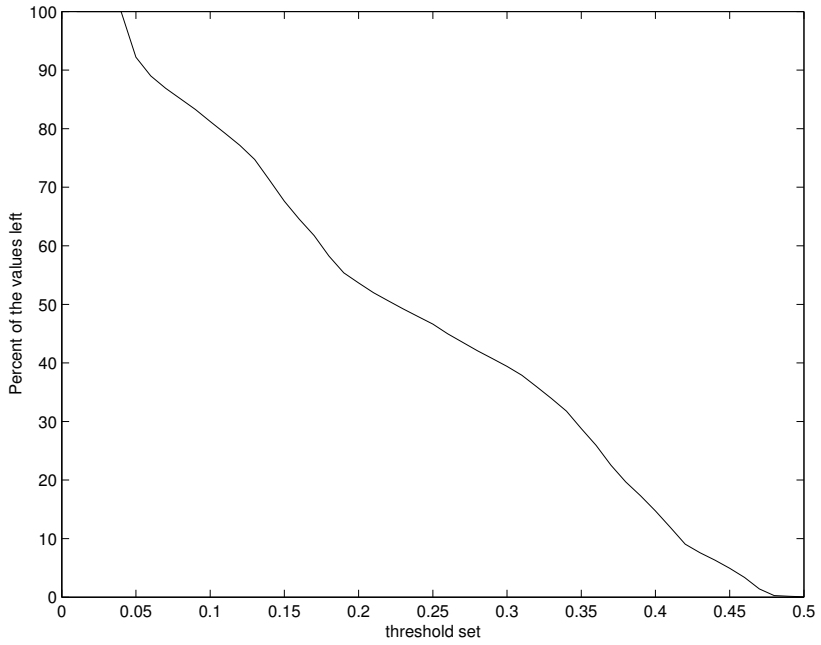

Figure 7.2: The figure describes how much values in percent that are left when applying different threshold values for the mass flow.

7.3 Using Mass Flow as a Test Criterium

When using diagnosis one has to be cautious which signals to use for deciding when a certain criterium has been met. The criterium $W > J$ (here J is the threshold and W the mass flow) can not be used because there is no assurance that the mass flow signal is correct. If this signal is corrupt, the system can be in a state when there will be no tests at all, and consequently no faults will be detected. The problem can be solved accordingly:

If assuming that there will be no more than one fault at a time, $max(W_e, W_{in}) > J$ can be used as the test criterium. If using this test criterium, the system will work even if one sensor is corrupt because W_e does not depend on W_{in} (see Section 5.4.1 for the definition of W_e). If W_e (or W_{in}) has a fault resulting in a too low value for the signal, then the test criterium will hold anyway. The correct signal W_{in} (or W_e) will then be equal to $max(W_e, W_{in})$ and the correct value of the mass flow will be used in the test criterium. But if W_e (or W_{in}) has a fault resulting in too high value for the signal, the result will be different. The corrupt value W_e (or W_{in}) will be equal to $max(W_e, W_{in})$. This value will be higher than the correct mass flow resulting in that less

data will be disregarded and the system will never come into a state where there will be no tests.

Chapter 8

Normalisation

In this chapter, the signal **w** in Figure 4.1 will be examined and the signal **z** will be produced.

When thresholding there are two different criteria that needs to be considered. These criteria is as said in Section 3.7, both to avoid false alarm and missed detection. When examine the histogram of the signal **w** (see Figure 4.1) for different faults, the threshold shall be set to meet the criteria mentioned in Section 3.3. Exactly how to set the threshold is further explained in Chapter 10 but ought to both avoid false alarm and missed detection.

In the uppermost plot in Figure 8.1 the histogram of **w** for the fault free case can be seen. Assume that the threshold is set to 0.058 (exactly how this is done is thoroughly explained in Chapter 10). When there is a fault, as many bars as possible in the histogram shall exceed that threshold (0.058 in this case). In the other two plots, in the same figure, the histogram when there is 20 % fault in the sensor can be viewed. The second plot describes the histogram of the signal when the mass flow is between 0.35 kg/s and 0.40 kg/s. The third plot describes the histogram of the signal when the mass flow exceeds 0.40 kg/s. It can be seen that for higher amount of mass flow the residual depart from zero more than for lower amount of mass flow and hence produce a better result.

A way to take this into consideration is to normalise the residual with the amount of mass flow.In Section 8.1 there will be a suggestion of a method of how to normalise the residual. How the normalisation of the residual will be affected by a gain fault will be examined in Section 8.2 and how it will be affected by a bias fault in Section 8.3.

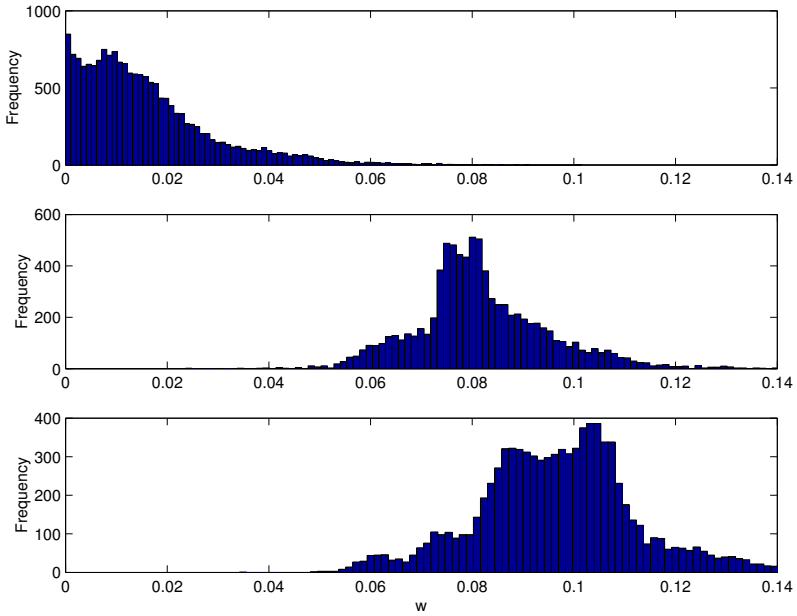

Figure 8.1: Histogram for different faults for signal **w**. The first plot has no fault, the second plot has 20 % fault but only signals where the mass flow is between 0.35 kg/s and 0.40 kg/s is considered. The third plot has also a 20 % fault but here mass flow that exceeds 0.40 kg/s is considered.

8.1 A Method for Normalisation

The idea is that the absolute value of the mass flow of air coming into the system decides the value of the threshold. If J is the threshold and $\mathbf{w}$ is the residual used (see Chapter 4), the following inequity holds for a no fault system:

$$abs(\mathbf{w}) < J(W_{in}) \tag{8.1}$$

In (8.1) there is an adaptive threshold. If dividing the left side of (8.1) with a function which depends on the mass flow, a fixed threshold will be obtained (this function is the normalisation quantity). This is desirable, because then power functions can be used as a performance measure for the signal, which is not possible for adaptive thresholds.

It is however not possible to let the normalisation quantity depend only on the measured mass flow W_{in}, because this signal may be corrupt. If assuming that there will not occur two faults in two different

sensors at the same time, the following normalisation quantity is proposed:

$$W_{norm}(p_{boost}, T_{boost}, N_{eng}, W_{in}) = min(W_e, W_{in}) \tag{8.2}$$

This holds because the estimated mass flow depends on p_{boost}, T_{boost} and N_{eng}. If W_{in} (or W_e) has a fault resulting in a too high value for this signal, the normalisation quantity W_{norm} will be equal to the correct signal W_e (or W_{in}), and the normalisation will work correctly. If W_{in} (or W_e) has a fault resulting in a too low value for this signal, the normalisation quantity W_{norm} will be equal to the corrupt value W_{in} (or W_e). This is however not a problem here. This is because there will be a division with a value which is too low, resulting in too high value for the residual. But one signal was corrupt so there is a fault that needs to be detected. It is desirable to have a large residual when there is a fault. The residual shall be large when there is a fault.

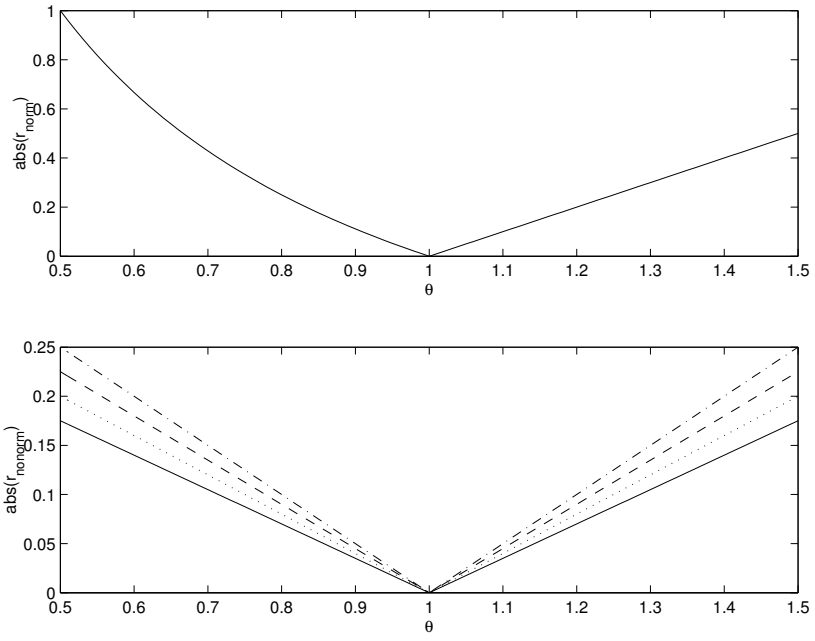

Figure 8.2: In the uppermost plot the $abs(r_{norm})$ with respect to θ can be viewed. In the second plot the $abs(r_{nonorm})$ for different mass flow can be viewed. Mass flow of 0.35 kg/s (solid), 0.40 kg/s (dotted), 0.45 kg/s and 0.50 kg/s can be compared.

8.2 Normalisation Affected by Gain Fault

For gain fault, a comparison between not using the normalisation and using normalisation may be made accordingly:

Assume, there is a gain fault in the signal, the residual will depend on the gain fault θ_g accordingly:

$$r_{norm} = \frac{W_{in} - W_e}{min(W_e, W_{in})} = \frac{\theta_g W - W}{min(\theta_g W, W)} = \begin{cases} \theta_g - 1 & \text{if } \theta_g > 1 \\ \frac{\theta_g - 1}{\theta_g} & \text{if } \theta_g \leq 1 \end{cases} \quad (8.3)$$

If not normalised, the residual will depend on the mass flow according to

$$r_{nonorm} = W_{in} - W_e = (\theta_g - 1)W \quad (8.4)$$

and the plot of the $abs(r)$ for equation (8.3) and (8.4) for different amounts of mass flow can be seen in Figure 8.2 . In this figure, it is important to notice that it is *not* a power function that is plotted and what can be seen on y-axe is not important. What can be seen in the plot is that the residual without the normalisation quantity depart from zero different much depending on the amount of mass flow. This is not so good because the residual shall not depend on the amount of mass flow. If the residual when the amount of mass flow is low, do not depart from zero, when there is a fault, as much as the residual depart from zero when the mass flow is high, it will be harder to detect faults. It can also be seen in the figure that the residual with the normalisation quantity do not depend on the amount of mass flow, which is better.

8.3 Normalisation Affected by Bias Fault

When consider bias fault, the conclusion is different. First the residual with the normalisation factor is to be considered:

$$r_{norm} = \frac{W_{in} - W_e}{min(W_e, W_{in})} = \frac{(W + \theta_b) - W}{min(W + \theta_b, W)} = \begin{cases} \frac{\theta_b}{W} & \text{if } \theta > 0 \\ \frac{\theta_b}{\theta_b + W} & \text{if } \theta_b \leq 0 \end{cases} \quad (8.5)$$

Then the residual without the normalisation factor:

$$r_{nonorm} = (W + \theta_b) - W = \theta_b \quad (8.6)$$

The plot of the $abs(r)$ for equation (8.5) and (8.6) for different amounts of mass flow can be seen in Figure 8.3. In this figure, it is important to notice that it is *not* a power function that is plotted and

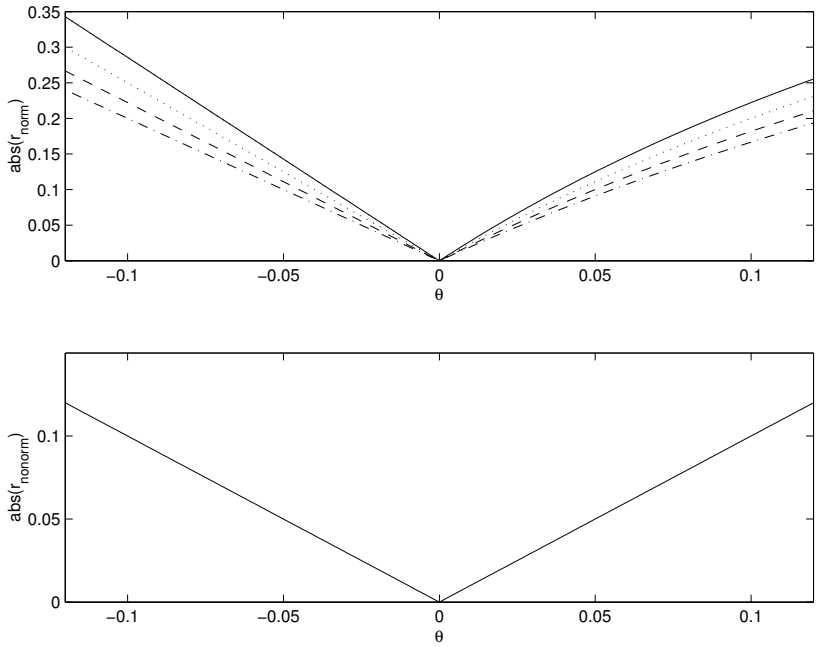

Figure 8.3: In the uppermost plot the $abs(r_{norm})$ for different mass flow can be viewed. In the second plot the $abs(r_{nonorm})$ can be viewed. Mass flow of 0.35 kg/s (solid), 0.40 kg/s (dotted), 0.45 kg/s (dashed) and 0.50 kg/s (dashdotted) is compared in the first plot.

what can be seen on y-axe is not important. What can be seen in the plot is that the residual with the normalisation quantity depart from zero different much depending on the amount of mass flow. This is not so good because the residual shall not depend on the amount of mass flow. If the residual when the amount of mass flow is low, do not depart from zero, when there is a fault, as much as the residual depart from zero when the mass flow is high, it will be harder to detect faults. It can also be seen in the figure that the residual without the normalisation quantity do not depend on the amount of mass flow, which is better.

8.4 Result of Normalisation

It can be seen in Figure 8.4 that the normalisation of the signal **w** do not improve the result, it mostly deteriorates the result. The improvement of the result described in Section 8.2 can not be seen in the upper most plot in Figure 8.4. The normalisation of the residual do not improve the power function for gain fault and the assumption made in the beginning

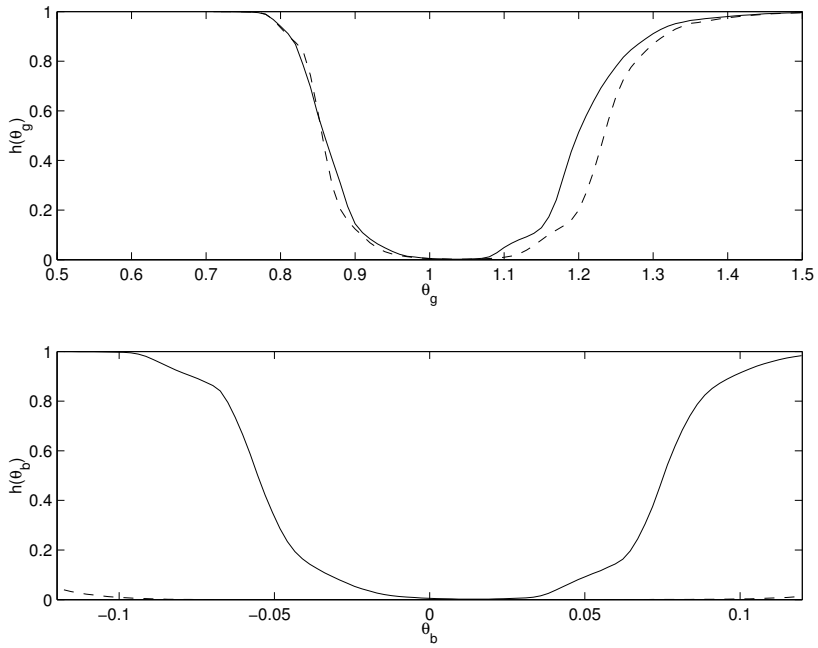

Figure 8.4: A comparison between not using the normalisation quantity (solid) and using normalisation (dashed) can be seen.

of this chapter do not hold. The normalisation do not improve the result and the assumption that the signal **w** depend on the amount of mass flow is not correct.

As can be seen in the under most plot in Figure 8.4 the power function for bias fault will become much worse. The deterioration of the result described in Section 8.3 is much and normalisation with the normalisation quantity (8.2) shall not be used.

It may though exist normalisation factors which do improve the result, but it is out of scope for this thesis to examine any further normalisation quantities.

Chapter 9

Outlier Rejection

In the test quantity algorithm, the signal **z** in Figure 4.1will be examined in this chapter and the signal **y** will be produced.

The histogram for the fault free signal **z** can be viewed in the uppermost plot in Figure 9.2. It can be seen that this signal is quite outlier prone, resulting in a high false alarm rate. In the second plot in the same figure, there is the histogram from the same signal but with a 20 % gain fault. Here, the signal is also outlier prone, resulting in a high missed detection rate. There are a need to keep the false alarm and missed detection rate low (see Section 3.3 and Section 3.4) and by rejecting the outliers in the histograms the result may be improved. An outlier rejection algorithm is proposed which will make the histograms less outlier prone. This algorithm is made accordingly:

1. Because the OBD test holds for ten minutes, there are ten minutes of data to use. Divide the length of this ten minutes in ten equidistant parts of one minute each.

2. Within every period, wait until the system is in a state where the mass flow of air coming into the intake of the air is sufficiently large (see Chapter 7). When the system is in this state, do a subtest. This subtest holds for one second. If the system never will be in this state, neglect this subtest.

3. Reject some of the noise in the signal (see Chapter 6).

4. After ten time periods, take the median of the absolute value of the values within each measurement series, i.e. the result of all the subtests, and reject each value which divert from the median more than 20 %.

5. If there is lesser than four remaining values, then the test is not to be considered reliable and there can be no conclusion of an

alarm.

6. The mean value of the remaining values in each measurement series produces the test quantity.

The aim for this algorithm is to reject the outliers of the signal and produce a better overall performance. Notice that it is assumed that the signal **z** do not depend on the amount of mass flow. That the signal **w** do not depend on the amount of mass flow was the conclusion of Chapter 8 (see Section 8.4) and hence there is no normalisation of the signal **w**. Without normalisation, the signal **w** is equal to the signal **z** and the conclusion of Chapter 8 holds here too. This is important for point four in the algorithm to work correctly. If the signal **z** do not depend on the amount of mass flow, there will not just be e.g. signals with high amount of mass flow that is disregarded and faults in this region can be detected.

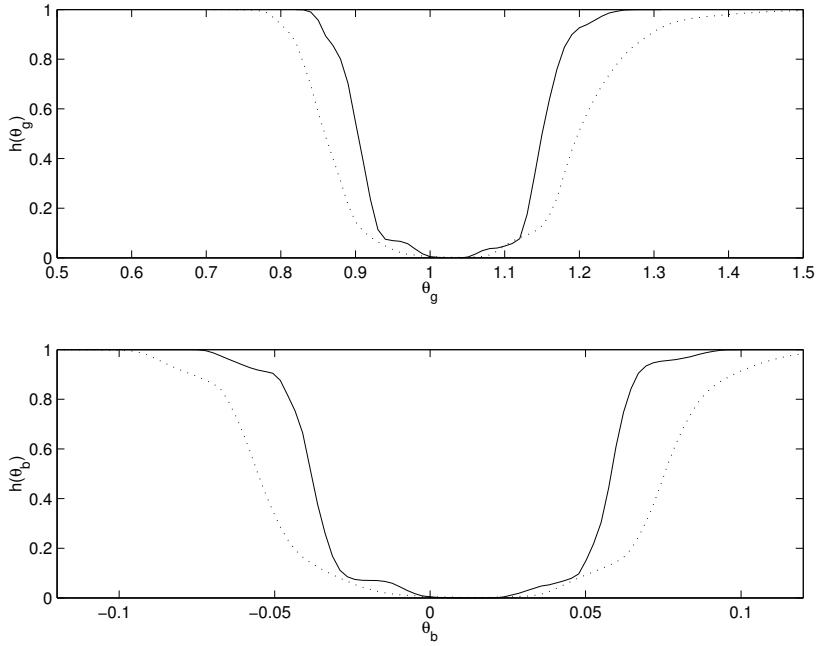

Figure 9.1: Power functions for the signal **y** (solid) and the signal **z** (dotted).

The Figure 9.3 describes the histogram after the outlier rejection algorithm is applied. The first plot describes the fault free case and the second plot describe the histogram when there is a 20% gain fault.

Here, most of the outliers are rejected and the result is improved if compare to Figure 9.3.

That the overall performance is improved can also be seen in Figure 9.1. The power functions for the signal is improved and smaller faults in the sensor can be detected when applying this outlier rejection algorithm.

9.1 Individual Variations

As can be seen in Figure 9.3, there are some data which is situated far from where most of the data is situated. There are a lot of bars around 0.04 in the third plot and around 0.07 in the fourth plot. This deteriorates the result. These data comes from different startup of trucks. If the variation in the different start ups can be handled, the result can be improved significantly. This is though out of scope for this thesis but may be investigated further.

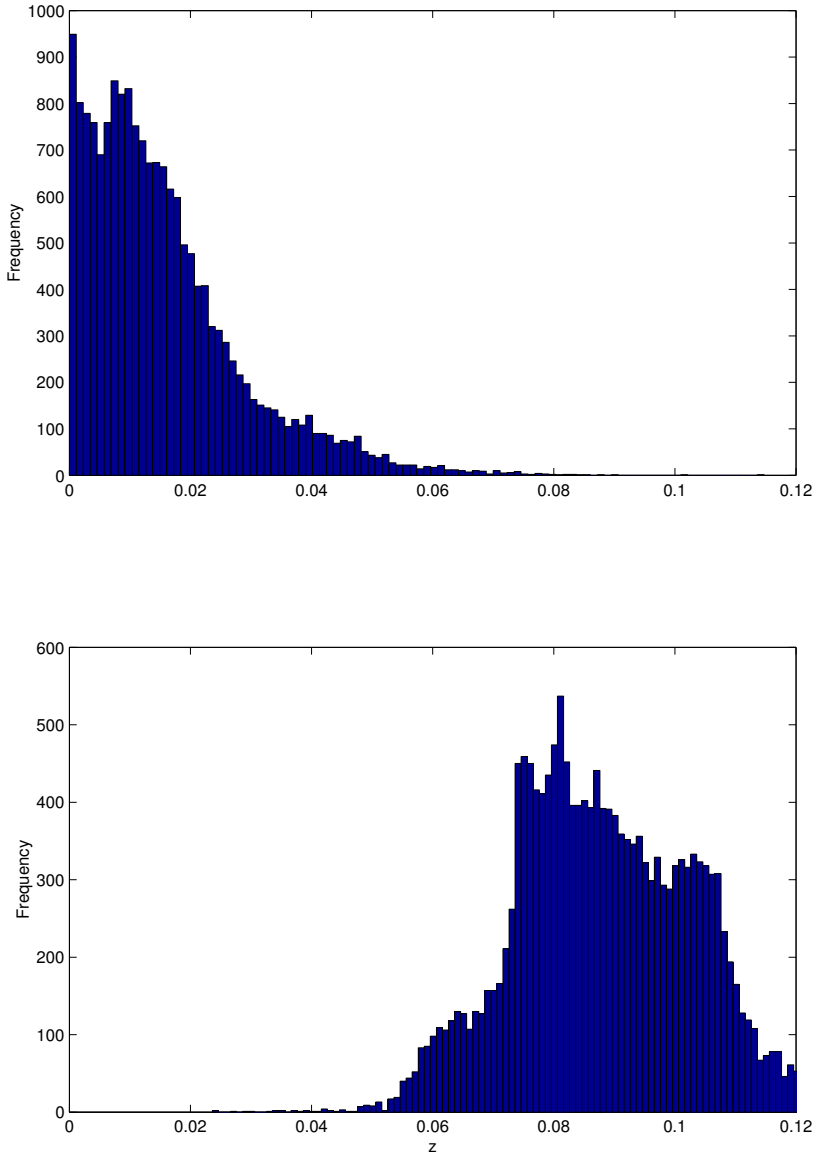

Figure 9.2: The first plot describes the histogram for the fault free signal **z** and in the second plot, there is a 20 % gain fault.

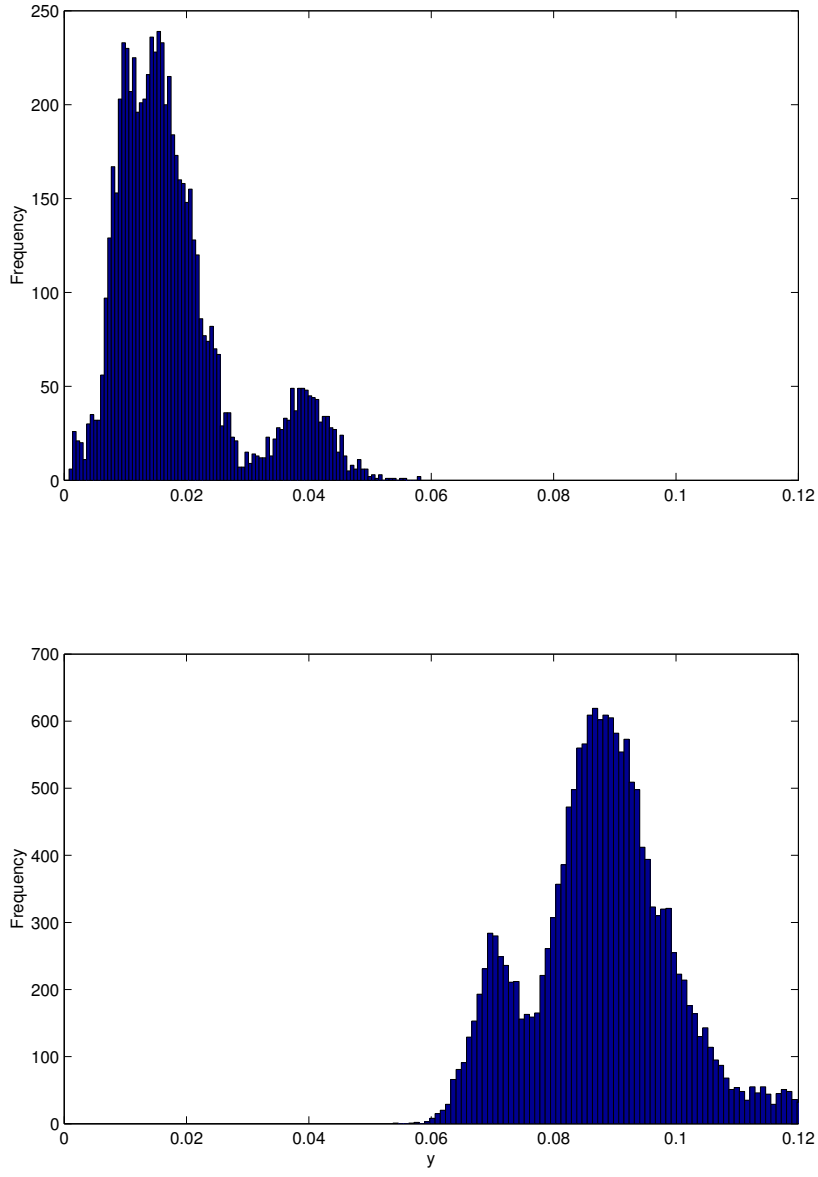

Figure 9.3: The first plot describes the histogram for the fault free signal $\mathbf{y}$. In the second plot, there is a 20 % gain fault.

Chapter 10

Thresholding

When thresholding the data, parametric methods will be used due to low false alarm rates (see Section 3.5). There are two methods that can be used to set the threshold. The first is assuming gaussian distribution and the second is the tail distribution estimation.

10.1 Assume Gaussian Distribution

Assume having gaussian distribution and estimate the data to that distribution gives an estimated mean value and variance of $\hat{\mu}$ and $\hat{\sigma}$ respectively (see Section 3.5.1). If using equation (3.8) and the specified false alarm rate in Section 3.3, the threshold $J_{gauss} = 0.041$ will be given. The estimated gaussian distribution for the signal $\mathbf{y}$ can be seen in Figure 10.1.

When examine the distribution in Figure 10.1 one may notice that the real distribution is not gaussian distributed. In the tail of the distribution there is a lot of bars of data which exceed the estimated gaussian distribution. The signal $\mathbf{y}$ has also a kurtosis of 4.13 which indicate that the distribution is more outlier prone that the gaussian distribution (see Section 3.6). All this indicates that the threshold that will be set based on the assumption of the gaussian distribution probably is not so good.

10.2 Tail Distribution Estimation

The tail distribution theory described in Section 3.5.2 can be used to set the threshold. An exponential distribution is adapted to the tail of the distribution starting at $h_0 = 0.04$. The estimated mean value is $\hat{\mu} = 0.0036$ and the threshold is set, using (3.11) and the specified false alarm rate in Section 3.3, to $J = 0,0566$. Here, there is not so

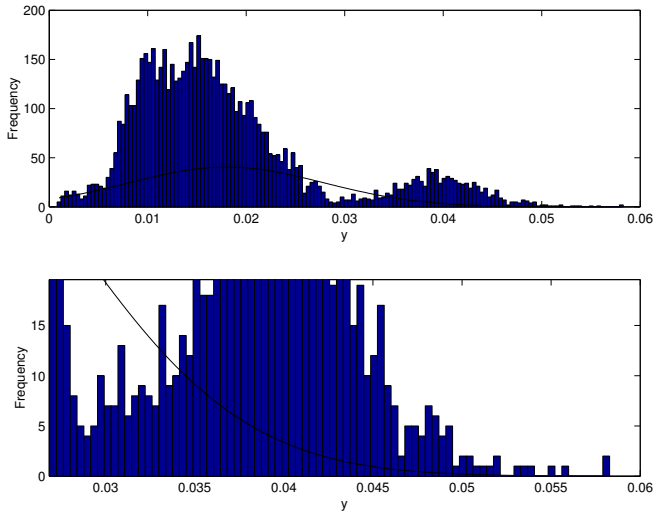

Figure 10.1: Histogram and gaussian distribution for $\mathbf{y}$. The two plots describes the same histogram with different scales.

many bars which exceeds the estimated tail distribution and hence it is a better estimation than the gaussian distribution.

To find out how small faults that can be detected, an exponential distribution is adapted to the tail of the distribution when there is a fault applied to the signal. With the specified missed detection rate in Section 3.4 and the threshold $J = 0.0566$ it can be found out how small faults that can be detected.

In each plot in Figure 10.2, the histogram for the fault free signal and the signal with the smallest fault that can be detected, can be seen. The adapted exponential distribution can also be viewed in these plots. There are four types of fault that can be detected and each fault are plotted in each figure. In the first plot the fault free signal and a signal with a positive gain fault of 20% is plotted. In the second plot, a negative gain fault of 30% is plotted. In the third plot a positive bias fault is plotted. In the fourth plot a negative bias fault is plotted.

In this figure it can be seen that most of the bars in the histogram do not exceed the adapted exponential distribution, indicating that this estimation is quite good. In Figure 10.2, it can also be seen that a positive gain fault of 20%, a negative gain fault of 30%, a positive bias fault of 0.11 kg/s (remember that W_{in} varies between 0.05-0.5 kg/s) and a negative bias fault of 0.09 kg/s can be detected.

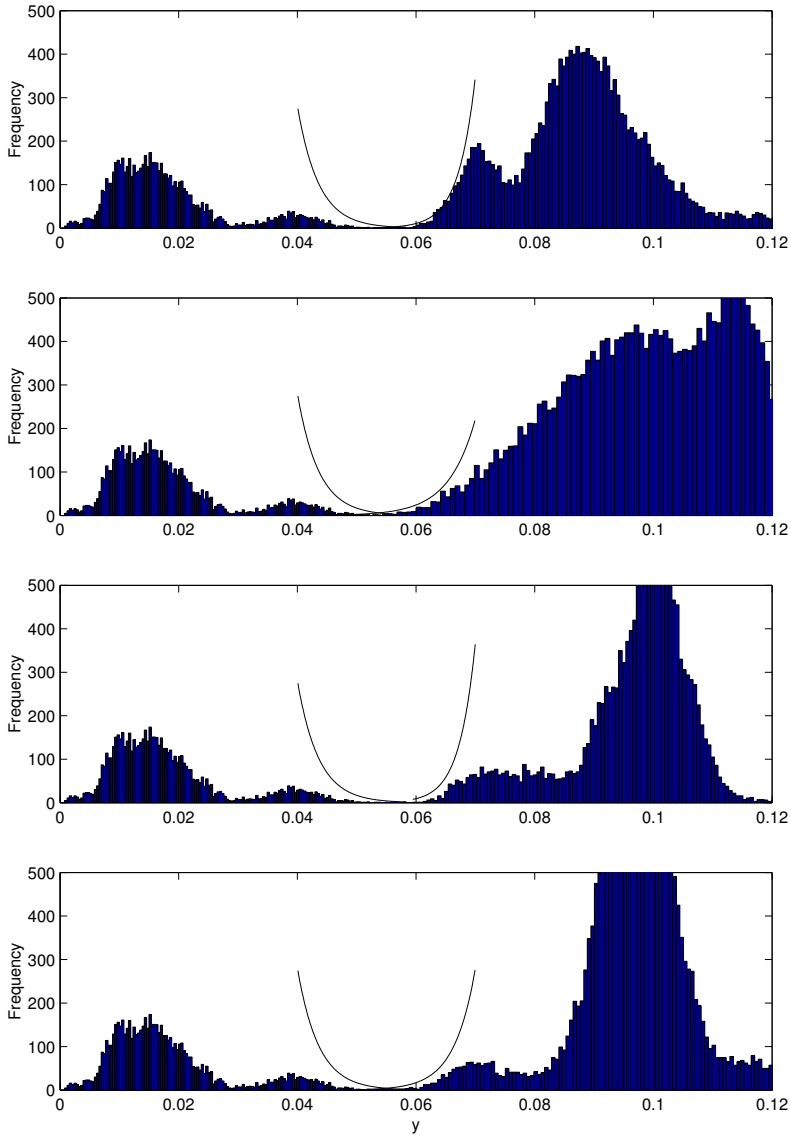

Figure 10.2: Histogram for the fault free signal $\mathbf{y}$ and the signal $\mathbf{y}$ with a fault applied to it.

Chapter 11

Conclusions

The objective with the thesis is to construct a test quantity in a model based diagnostic test which shall be thresholded. There is an algorithm constructed which produces this test quantity. Data has been used for over 17 hours of driving in two separate trucks and a lot of start ups. There is stringent condition on false alarm rate and missed detection rate and statistical methods are used because of this. The models examine in this thesis only work when the EGR is shut off and the EGR is allowed to be shut off only ten seconds each ten minutes. This complicates the process a lot and makes it hard be able to detect small faults. The test quantity algorithm produced in this thesis consists of several "blocks" which each and on (except one block) improve the result.

11.1 Accomplishments

In this thesis, there are especially two of these "blocks" that improves the result much. The first of these blocks is the subset rejection. In this block, values are rejected when the mass flow is larger than a certain threshold. The result is improved significantly and it is shown that the model used, i.e. the volumetric efficiency model, is not good for low mass flows.

The second block which improve the result a lot is the outliers rejection block. This block consists of an algorithm which reject the outliers in the residual and also improve the result significantly. Here can also be seen why the result is not so good. The individual variations in the different start ups of the trucks is large and deteriorates the result a lot. If this individual variations can be solved, the result will be improved a lot.

The result with this test quantity algorithm is that it is with a false

alarm rate of 0.01 and a missed detection rate of 0.0025, it is possible to detect a positive gain fault in the mass flow sensor of 20% a negative gain fault in the same sensor of 30%. It is also possible to detect bias fault in the mass flow sensor. Here the result was 0.09 kg/s and 0.11 kg/s respectively (the mass flow signal varies between 0.05 kg/s to 0.5 kg/s).

11.2 Future Challenges

There are several things needed to investigate to further improve the result. The individual variations between different engines and different start ups of these engines need to be investigated further. The individual variations deteriorates the result a lot and if this could be handled in a good way the result may be improved significantly.

There is a need for a model that works with the EGR, and if having a model with the EGR, it would certainly improve the result. It would also be interested to run an engine in an OBD cycle, with a fault implemented to validate the diagnosis system produced. There are also other things worth looking at. To examine the dynamic model and other kind of normalisation vectors. Faults in other signals such as the boost pressure signal the boost temperature signal may also be examined.

References

[1] Mattias Nyberg and Erik Frisk. Diagnosis and supervision of technical processes. Linköping, Sweden, 2001. Course material, Linköpings Universitet, Sweden.

[2] Fredrik Gustafsson and Jan Palmqvist. Change detection design for low false alarm rates. Technical report, Department of Electrical Engineering, Linköpings Universitet, Linköping, Sweden, 2000.

[3] David Ruppert. What is kurtosis? an influence function approach. *The American Statistician*, 41(1), February 1987.

[4] Lars Nielsen and Lars Eriksson. *Course material Vehical Systems*. Linköping Institute of Technology, Vehical Systems, ISY, Linköping, Sweden, 2001.

[5] Torkel Glad and Lennart Ljung. *Reglerteori (in swedish)*. Studentlitteratur, Lund, Sweden, 1997.

[6] Sune Söderkvist and Lars-Erik Ahnell. *Tidsdiskreta Signaler och System (in swedish)*. Tryckeriet E. Larsson AB, Linköping, Sweden, 1994.

Notation

Symbols used in the report.

Nomenclature

Symbol	Quantity	Unit
n_{cyl}	Number of cylinders	$-$
N_{eng}	Engine speed	RPM
p_{boost}	Boost pressure	Pa
$\hat{p}_{boost}$	Estimated boost pressure	Pa
R_{Air}	Gas constant for air	$J/(kg \cdot K)$
T_{boost}	Boost Temperature	K
T_{eng}	Engine Temperature	K
V_d	Displacement volume	m^3
V_{tot}	Total volume in intake system	m^3
$\hat{V}_a$	Volume flow rate of air into the intake system	m^3/s
$\hat{V}_d$	Volume flow rate of air displaced by the piston	m^3/s
W_{bb}	Estimated massflow of air from the black box model	kg/s
W_e	Estimated massflow of air into the cylinders	kg/s
W_{in}	Massflow of air into the intake system	kg/s
η_{vol}	Volumetric efficiency	$-$

Operators

Operator	Explanation
$\cup$	cup
$\cap$	cap
$\neq$	not equal to
$\in$	in
$\notin$	not in

Abbreviations

Abbreviation	Explanation
EGR	Exhaust Gas Recirculation
NOx	Nitrogen-oxide
OBD	On Board Diagnostics
SI	Spark Ignited