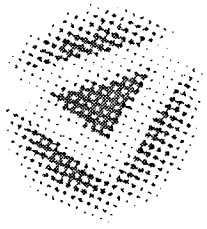
Quality Control of High Throughput Screening

PHILIPPE F. FOUQUART

Master of Science (by Research)
in
Pattern Analysis and Neural Networks

Supervisor: Dr Ian T. Nabney



THE UNIVERSITY OF ASTON IN BIRMINGHAM

September 1997

THE UNIVERSITY OF ASTON IN BIRMINGHAM

Quality Control of High Throughput Screening

PHILIPPE F. FOUQUART

Master of Science (by Research)
in
Pattern Analysis and Neural Networks, 1997

Thesis Summary

High Throughput Screening (HTS) is an effective means to determine the chemical compounds which are efficient on a given biological target. The experiments are carried out using a standard format, the 96-well plate, where six wells are controls whose expected values are known. However the measurement techniques are subject to variation which renders the assessment of an experiment difficult. In the context of quality control of an industrial task, a novelty detection method can be employed to determine abnormal or unusual outputs where the novel points can be defined as the observations which have extreme values compared to other measures observed under the same experimental conditions. The new method proposes to screen an additional set of three plates featuring only control wells which constitue the reference data to compare the plates. This set of plates is used to estimate the distribution of the control values. In the first place, a Gaussian Mixture Model is trained with the EM algorithm. A point is declared 'novel' if its probability is below a novelty threshold. The technique is compared to a traditional approach of outlier detection. The choice of this threshold is investigated together with alternative approaches to the problem.

Keywords: Novelty detection, outliers, Mixture Models, EM algorithm, High Troughput Screening.

Acknowledgements

I am grateful to Pfizer Research for funding the work described in this thesis

I gratefully acknowledge the help of Ian Nabney who supervised the project and provided me with thoughtful criticism and advice. I should like to thank him also for patiently coping with my careless spelling.

The help and suggestions of Andrew Weaver were much appreciated when writing the C code. I would like to thank him for kindly providing the data structure used in the last stage of the project

I must express my thanks to Wilma Keighley for her enlightening answers concerning the application domain.

Finally I am very indebted to Neil Pickles for his help with the practical aspects of the project. I am especially grateful to him for collecting the data and answering many questions about HTS (at least twice each).

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The success of a drug discovery process depends entirely on its capacity either to find new chemical entities or to reveal unknown characteristics of some existing molecules. *High Throughput Screening (HTS)* offers an empirical means to identify novel compounds which act efficiently against a given therapeutic target.

This chapter provides an introduction to HTS to supply the necessary background and understand the quality control of this method. Details can be found in [BKRW97]. The second part focuses on the control wells which are the means of assessing plate quality. We show how the problem of HTS quality control will be tackled using these wells. This final section outlines the practical constraints which should be taken into consideration and gives an overview of this thesis.

1.1 High Throughput Screening

Combinatorial chemistry makes use of automated and miniaturised devices to screen simultaneously a large number of mixtures. Indeed, recent progress in technologies such as bioassays, robotics, computation and data handling now enables large series of experiments, involving tests of thousands or millions of molecules. HTS takes advantage of these advances so that comprehensive collections of compounds can be screened in order to find relevant biological activity. The efficiency of the method relies on the

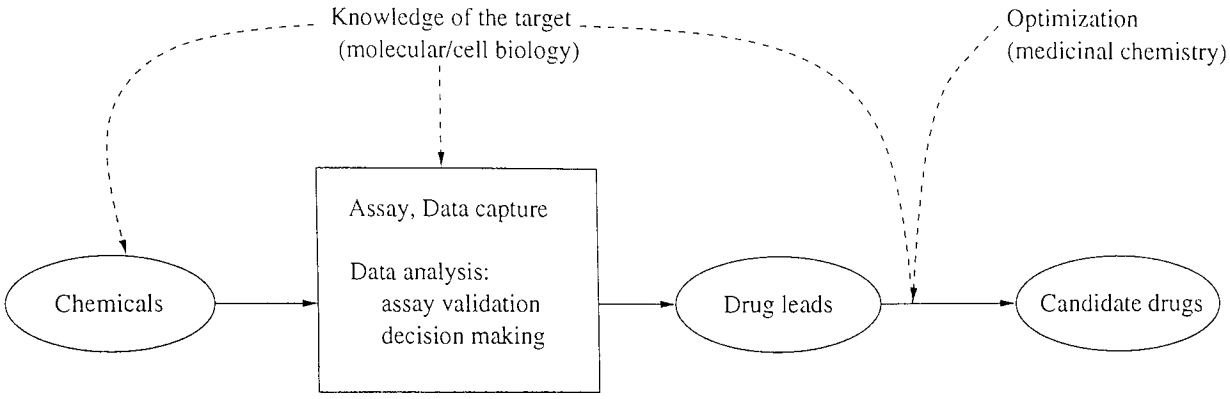


Figure 1.1: **High Throughput Screening Process**

volume of data generated by this technology. Therefore, the more quickly relevant information is found, the more efficient the method. Typically, drug discovery groups examine growing numbers of samples to determine the few compounds, called 'lead compounds', that will progress to the next round of screening, and eventually to the development of a pharmaceutical agent of commercial value.

1.1.1 The HTS process

The HTS process summarised in Figure 1.1 can be divided into five successive steps: compounds supply, assay, data capture, data analysis and sample follow-up.

Compounds supply

Automation is developed as much as possible so as to increase throughput; thus a standard format is required for conveying small liquid samples. The most common is the 96-well plate, each well having a volume capacity of up to 2ml. This type of plate — as well as the other formats : 48, 384 and higher — is also called 'microtitre plate' or 'microplate' because of the small volume of mixture required for each well.

Vast libraries of compounds held by pharmaceutical companies, as the largest source of new potential lead compounds, constitute the basis of the HTS process. Indeed, automated methods enable the relatively rapid synthesis of impressive libraries of compounds. The dry samples are provided in tubes formatted for the microtitre plate format and dissolved afterwards. These are placed for storage into a liquid sample

bank, source of all compounds for HTS.

Assay and data capture

Compared to traditional experiments on a few samples, an assay for HTS has its own requirements. For instance, the handling steps should be limited and the solvent's compatibility ensured. In addition, the difference of incubation time[1] between the microplates of the beginning of a screen and those of the end has to be evaluated by control wells on each plate. The variation which may occur from one plate or one assay to another will be discussed in greater length in Chapter 4.

Various hardware items are involved in the manipulation and the analysis of the 96-well plates for HTS:

- The liquid handling and assay assembly are carried out by manual pipetting devices (fast but restrictive — the same volume is distributed to all the wells — and prone to error) or robotic sample process (accurate and versatile but still slow).
- The separation includes the filtration equipment to harvest the contents of a plate (essentially manual) and the plate washers.
- Signal detection instruments measure the radiometric, fluometric, colorimetric or luminometric activity to estimate the chemical activity of the mixtures. These measures form the data on which this study is based.

The signal measures are saved in databases to be checked and assessed for a reliable interpretation of the data.

[1]The period during which the various compounds of the mixture remain active.

Data analysis

Partly computerised, the data analysis is the key step of the HTS process and will be described thoroughly in Section 1.2.1. It has two goals:

- The *assay validation* ensures the accuracy and the validity of the data with respect to the assay specification,
- The *decision making* determines the 'hits' — mixtures whose activity is considered to be significant — of the assay.

Both operations are manual and conducted by means of a data graphical representation. The assay validation is possible thanks to special wells on each plate dedicated to the control of the assay. The decision making aims at the selection of the wells whose activity is greater than a threshold fixed by the operator. A sample decision interface permits the simultaneous view of a wide range of data in order to assess the bioactivity of a sample to take decisions about its future.

Computer controls are necessary throughout the HTS process including instrument controls (integrated software controls for robots or external computers for liquid handling) and data capture (bespoke programming of the plate reading devices). The contribution of computing systems is more significant in data management, since the capacity to deal with large amounts of data remains the keystone of combinatorial chemistry. Indeed, only powerful databases can manage the massive quantity of information generated by the screening.

Sample follow-up

The samples whose activity is revealed by HTS as active on a given target influence the choice of compounds for further screening in order to maximise the chances of finding a mixture of biological efficiency. On the last stage, they are submitted to lead optimisation which aims at improving the efficiency of the compounds. Contrary

to HTS, this optimisation relies on an existing knowledge base and is performed on small samples which need not be as robust as those of mass screening. If its activity is estimated sufficient, the lead compound is finally validated and added to the company library and database as active towards the target.

1.1.2 Controls and assessment of the assay

To prevent any variation in the analysis of the test sample activity and establish the validity of the assay, *control positions* are always present on each plate. Typically, they consist of:

- a maximum well (100% of activity) ;
- a minimum well (0% of activity);
- a standard well ($\approx$ 50% of activity) .

The controls are generally analysed using softwares which offer different views of the data (graphical and tabular views of the control results). The assistance of a graphical representation of these controls is convenient for the operator and appears to be very efficient to compare controls of the same assay and allow assessment of the experiment. This assessment consists of the de-selection of the controls that reveal anomalous results. These controls can also be effective at detecting rogue plates or handling mistakes. However, it should be noted that this manual data assessment is fairly subjective and may vary from one operator to another. After validation, active samples of the screen can be determined to be submitted to further studies.

1.1.3 Advantages and disadvantages

The main advantage of HTS over traditional chemical schemes is that little information on the structure of the compound is necessary to perform the screening; hence its applicability to any molecular target. In addition, the record of successes as well as failures in databases can be utilised to design further experiments.

The obvious drawback of such an empirical search is that it can be time consuming; hence the need of efficiency. Moreover, the inherent variation of measurement techniques may induce false hits which must be rejected further on. Finally, the range of prospect for a laboratory is limited by the chemical diversity of its own library since the likelihood of finding a new compound by this method depends entirely on the selection of mixture to screen.

The increasing number of screens demands substantial improvements in the HTS process. Besides, as the pressure to find novel therapeutics increases, the cost effectiveness and the speed of HTS become all the more crucial. From this point of view, the introduction of computing systems such as quality control software within an integrated HTS facility may contribute to render the process as efficient as possible.

1.2 Controls and analysis

This section concentrates on data analysis from the viewpoint of the quality control of HTS. To begin with, we present an example of data analysis together with the problems to be solved for this quality control. In the second place, the data which constitute the basis of this study are described.

1.2.1 An example of data analysis

This part introduces a naive example of data analysis for standard plates. Although simplified, it places this work in its biochemical background.

It is generally considered that the data analysis starts with the control well checking even if its object is basically to provide reliable data to the actual data analysis. As stated in Section 1.1.2, the operator inspects a graphical representation of the control data to assess the assay and to detect would-be handling mistakes (such as insertion of an incorrect volume of substrate) or rogue plates. Once spotted, these control values

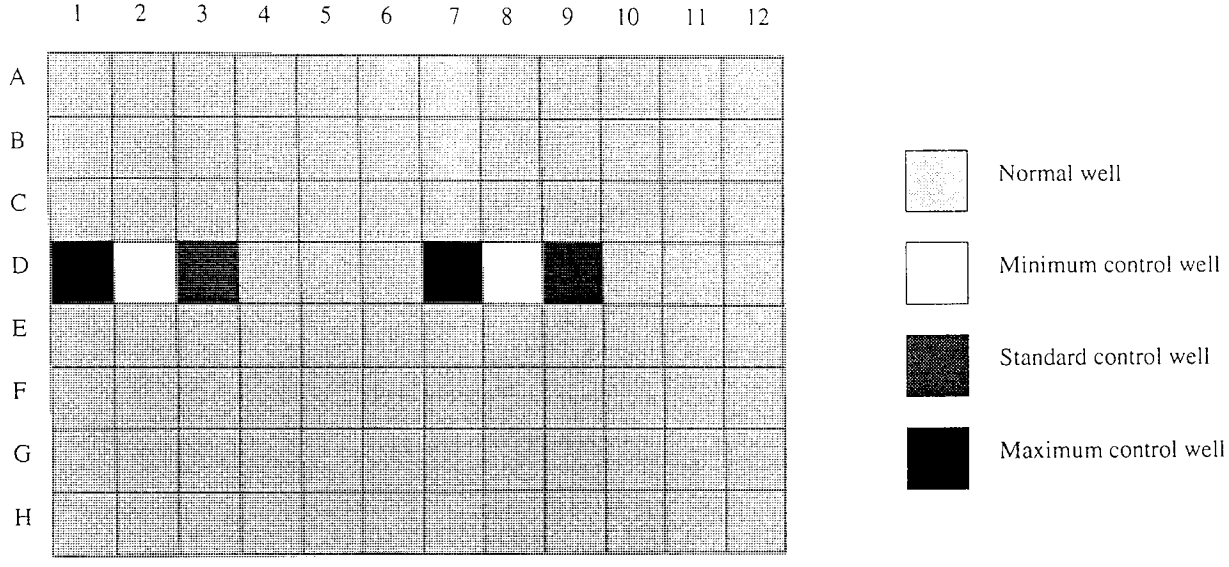


Figure 1.2: **Normal HTS plate**

can be de-selected which ends the assessment of the assay. The rejected control values will thereby not be taken into account for the detection of the 'hits'.

Suppose one wishes to detect a novel enzyme inhibitor, an active compound which prevents the action of an enzyme receptor. Each plate is provided with 90 different substrates (each mixture featuring 20 dry samples) and 6 control substrates (Figure 1.2) where:

- the enzyme causes its normal reaction in a completely uninhibited manner (maximum controls: D1, D7);
- the enzyme is fully inhibited, though present (minimum controls: D2, D8);
- the enzyme is partially inhibited by a compound whose activity on this enzyme is known to be 'average' (standard controls: D3, D9).

For a given screen, all the control wells of the same type (maximum, minimum or standard) of all the plates contain the same mixture.

Because of the variation within an assay, it is important to measure accurately the maximum and minimum activity of this enzyme in the solution: that is to say, to fix

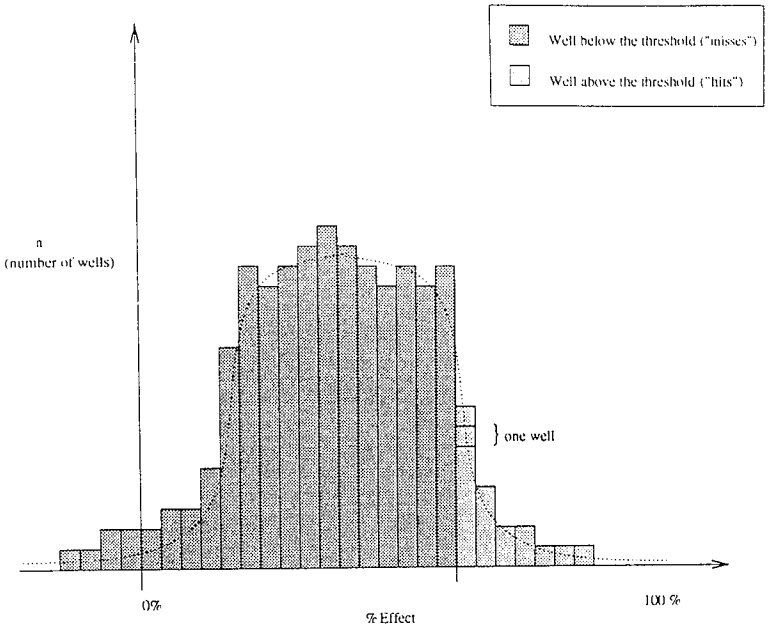


Figure 1.3: **'Hits' determination**

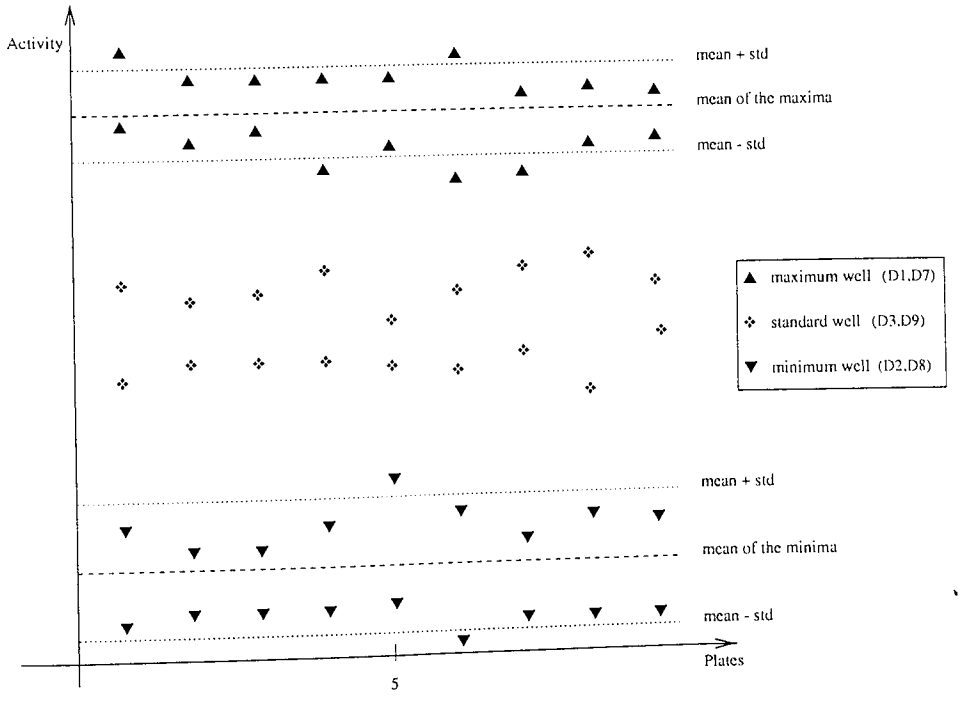


Figure 1.4: **Quality Control**

the boundaries 0% and 100% of Figure 1.3 in order to evaluate the actual inhibition of the enzyme receptor. The minimum and maximum controls are intended for both quality control and calculation of these boundaries. The role of the standard controls is restricted to the quality control.

Ideally, all the maximum (respectively minimum and standard) control values of all the plates for a given screen should be the same since they contain exactly the same solution and measure the same activity. Practically, the activity boundaries (0% and 100%) are estimated as the mean of a selection of controls (respectively minimum and maximum controls) of the assay[2]. This selection is achieved by visual inspection on a graphical representation of the control data similar to Figure 1.4. If the operator considers that a value of a control deviates significantly from the mean of this control, this value can be de-selected. The mean of this control is re-calculated (without the control which was removed). These controls are then checked again regarding this new mean and so on, until all the points are thought to be correct, hence assessment of the assay. The standard deviations (denoted in dash lines) computed for the minimum and maximum controls help the operator to decide whether a point should be kept[3]. This procedure determining whether a given point is an 'outlier' is obviously greatly subjective. Once this task completed for the two controls, the value for the minimum (maximum) activity defining the boundaries 0% (100%) is taken as the average value of those assessed. In other words, the de-selection of a point, say a maximum, affects the analysis in the sense that the value of this control does not influence the computing of the average of the maximum values which gives the estimation for the maximum activity of a compound (*i.e.* the minimum activity for the receptor).

Concerning the first part of this data analysis, we should insist on the fact that if a single point (a control well) is de-selected, there is no consequence whatsoever on the

[2]The screening procedure is generally spread over several different dates for it is too long to be conducted in a row. In this case, the quality control and the analysis is carried out 'assay by assay'.

[3]We shall see in Section 2.1 that the use of these standard deviation error bars which assumes the normality of the data can be put into question.

remaining wells of this plate as far as the data analysis is concerned other than the modification of the activity boundaries. Otherwise stated, if a control is de-selected, no action in practice is taken over other values of the plate even if this may indicate that something has gone wrong with it. On the other hand, if all the six controls are suspicious the corresponding plate can be de-selected.

After assessment, the operator sets a threshold above which the wells are considered as 'hits' (as indicated in Figure 1.3, for a 60% threshold[4]). If the shape of this diagram is consistent with what can be expected for an average screen (according to the operators, a 'Gaussian shaped' plot centred in 50% of activity) the validity of the assay is assessed. The contents of these wells are then stored in the database as being relevant towards the enzyme and will be submitted to further tests.

1.2.2 Other plate formats

Other types of format than the standard 96-well plate can be involved in HTS. This section presents the IC_{50} plates and the *Totals & NSB plates*[5].

The IC_{50} plates are generally employed after a comprehensive screening on normal plates that resulted in the detection of a few lead compounds to determine their optimum concentration. On the IC_{50} plates, the same compounds are disposed on two successive columns from A3 to A12 in different concentrations (Figure 1.5) whereas the first two columns are dedicated to the controls (maximum and minimum). The data analysis is conducted similarly as in Section 1.2.1 to determine the hits, which correspond in this case to the compounds offering the best activity regarding the target together with the optimum concentration.

The *Totals & NSB* plates are generated especially for this study. The 96-well plate

[4]In Figure 1.3, some wells may even have a negative activity with respect to the target (in the previous example, the corresponding compound would activate the receptor instead of inhibiting it).

[5]*Totals* for 'totally inhibited' (the minimum controls) and *NSB* for 'Non Specific Binding' (the maximum controls).

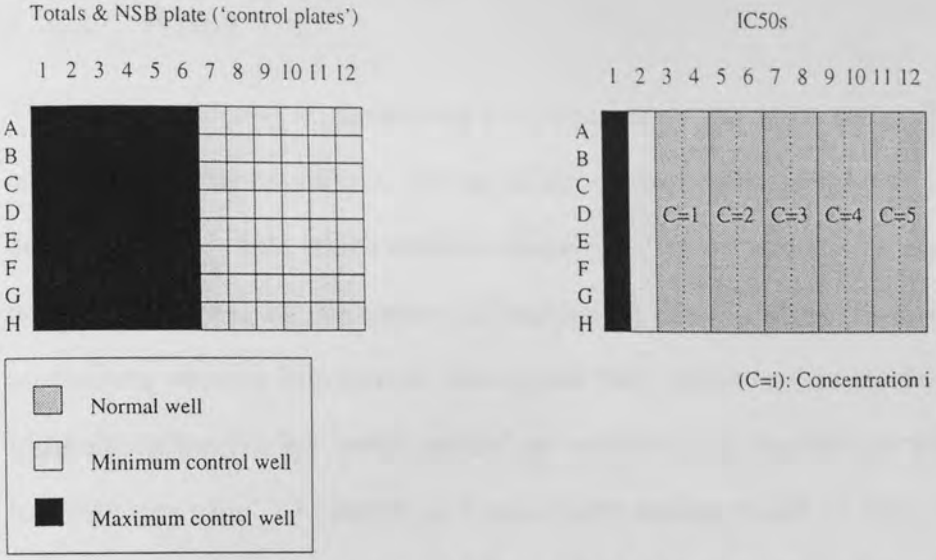


Figure 1.5: ***Totals & NSB's*** **and** IC_{50} **plates**

is separated into two parts, 48 wells for the minimum controls and 48 for the maximum controls (Figure 1.5) whose content was discussed in Section 1.1.2 and 1.2.1.

Note: in the following chapters, we shall try and avoid the terms 'Totals' and 'NSB' and prefer 'mimimum' or 'maximum' control for clarity of the argument. The *Totals & NSB plates* may be referred as the '*control plates*' since they contain only control wells. Similarly, a 96-well plate will be called '*normal plate*'.

1.3 Summary

As frequently emphasised in this introduction, the assessment of HTS data makes use of a great deal of subjectivity. The aim of the study is to determine whether a probability density based method could help the assay assessment exposed in Section 1.1.2.

This section presents the objectives of this project and the methods chosen to achieve them. The last part gives an overview of this thesis.

1.3.1 Aims

This work is aimed at measuring the variance in the HTS data. To do so, we focus on the detection of unusual values of the control wells mentioned in Section 1.1.2 to determine the points which could be de-selected. This task can be regarded as a novelty detection problem in the context of probability density estimation. In other words, the underlying density function of the control well values can be modelled so as to detect unlikely values (called 'novel points' or 'outliers'). It implies the formalisation of this 'novelty criterion' to provide a quantitative measurement of this novelty in numeric (and therefore objective) terms. The software to be provided should point out these outliers and give a numeric evaluation of this novelty.

Requirements

The method for assessing the quality of the HTS data should fulfil the following constraints:

- it should not be computationally expensive even if the time for learning and testing is not crucial. Typically a procedure which takes a few minutes is acceptable;
- as noted above, the control system must pin-point the abnormal plates of the screen; a measure of 'abnormality' should be provided for both the plates and the wells so that the latter can be ordered with respect to this measure;
- the software must produce outputs which can be easily understood and thereby avoid the "black box trauma" of neural computing;
- the method should leave a possibility of automation; it should be designed as a help for the operator who will validate the results of the detection but allows the eventuality of running without any intervention.

We shall refer frequently to these requirements throughout this thesis to justify the decisions that will be taken concerning the novelty detection method.

1.3.2 Approach

Each HTS screen is performed with respect to a given target so the control values of two different screens have *a priori* a different distribution. As a result, from a practical point of view, the approach consisting in 'learning' the distribution of these values requires the screening of three additional plates (the '*Total & NSB plates*' or control plates) for each screen. These very plates are used to train and validate a model for the underlying distribution of the control values. The control values of the normal HTS plates which are 'unlikely' according to this model (the probability density) are declared 'novel' (and have to be pin-pointed to the operator as being 'unusual'). In terms of handling, the additional plates necessitate little extra work: a typical run for HTS features more than 200 plates.

1.3.3 Overview

This thesis consists of four parts. The second chapter presents preliminary works on some HTS data and is divided into three distinct sections. Through popular statistical tests, the first section shows that these data are poorly represented by a single Gaussian. In the second section, standard statistical methods whereby the problem of abnormality detection can be tackled are described. Attention is drawn to the difficulty of using such techniques for the purpose of HTS quality control. Finally, we study the correlation between the three controls: minima, maxima and standards.

The third chapter reviews in detail the model inference framework. First, we define the probability density model involved, Gaussian Mixture Models, and the technique for its training, the *Expectation-Maximisation* algorithm. Second, the data selection is examined and we deal with the problem of choosing a proper novelty threshold for the density. The third part investigates the choice of the model parameters.

The fourth chapter is concerned with the application of this framework. The first part applies a model to the novelty detection of an HTS screen by learning the distribution of the minima and the maxima. An alternative approach to novelty detection, the

'Adaptive Mixture Model', employing the same density model but a dynamic learning procedure is outlined and tested on the same screen. The strengths and weaknesses of the two methods in the context of routine use are discussed. Once a plate is declared novel because of unusual control values, the next step is to spot its abnormal components. The third part treats this problem thanks to the conditional densities of the model previously described. The last part of this chapter is dedicated to the inclusion of the standards in the model.

The results produced in this chapter concern Screen 2 (Appendix A.1). For obvious practical reasons, it was not possible to present in detail the results of the novelty detection on all the screens referenced in Appendix A. This screen was chosen because it features the smallest daily variation (see Section 4.5) and is therefore close to the type of data which could be provided by a automated screening device.

In the final chapter, we present a summary of the study and reflect on additional questions which may constitute an extension of those treated hereafter.

This work is aimed at providing a robust method helping an operator to take a more reliable decision. Bearing in mind the constraints mentioned above, it implies that however appealing or theoretically elegant a method can be, the one and only criterion for choosing or discarding it should remain its practical efficiency.

Chapter 2

Preliminary study

As the training and the validation of the model rely on the control plates, we start with a one-dimensional study of these values.

The quality control of HTS is generally performed comparing the controls (minima and maxima) with one standard deviation from their respective means. This procedure implicitly assumes the normality of the data. The first part of this chapter investigates this hypothesis by testing the normality[1] of the controls. The traditional approach of outlier detection is described and applied to the HTS data. Finally, we test the independence of the three controls and examine the implications of the results regarding the HTS control procedure.

2.1 Goodness-of-fit tests for Gaussian distribution

The two following standard procedures, χ^2 and Kolmogorov-Smirnov, test the goodness-of-fit on a Gaussian distribution (comparing the observed and expected distribution of Figure 2.1). Both sections present briefly the tests and the results obtained when applied on the HTS controls. For the following sections, n denotes the size of the sample

[1] We should be cautious with the term 'normality' since 'normal' and 'Gaussian' are widely considered as synonymous (but some may restrict the former to Gaussian distribution with zero mean and unit variance). To describe such a distribution, we shall prefer 'Gaussian' to 'normal' to prevent confusion with the 'normal' plates. However, in Section 2.1, the denomination 'normality' as in "normality test"refers to 'the condition of being Gaussian'. The word has no implication whatsoever so far as the 'novelty' or 'abnormality' of the plates are concerned.

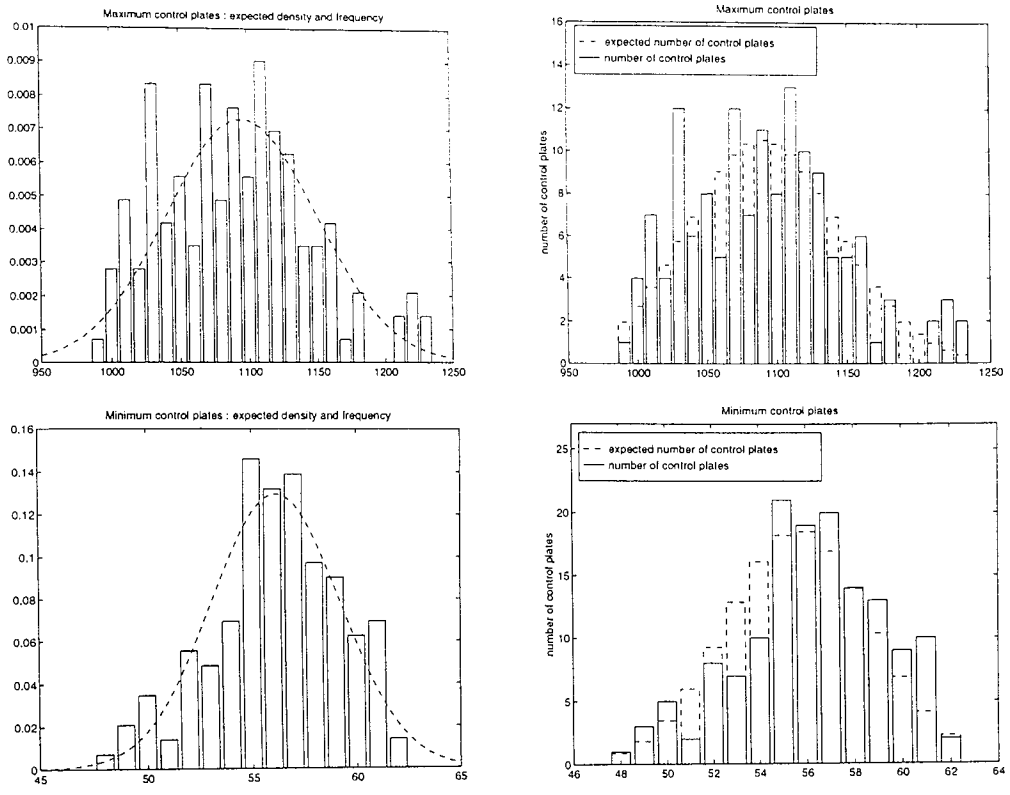


Figure 2.1: Observed and expected distribution (single Gaussian) of the controls (Screen 2)

and k the number of clusters (or bins) of the test.

The two tests-of-fit on a Gaussian distribution are performed using the means $\widehat{\mu}$ and the variances $\widehat{\sigma}^2$ computed in Table 2.1. Both of them test the following hypotheses:

> H_0: *the sample is drawn from the normal distribution with mean* $\widehat{\mu}$ *and variance* $\widehat{\sigma}^2$.
>
> H_1: *the sample is not drawn from the normal distribution with mean* $\widehat{\mu}$ *and variance* $\widehat{\sigma}^2$.

The samples consist of three screens referenced in Appendices A.1.1, A.2.2 and A.3.2. Each of them features three control plates (144 minimum wells, 144 maximum wells).

2.1.1 χ^2 test

The χ^2 test can compare two binned samples to test whether they are drawn from the same distribution. Because it remains one of the most popular goodness-of-fit tests, it is used frequently to compare two densities drawn from continuous variables. This is

Parameters estimation	Minimum plates		Maximum plates	
	$\widehat{\mu}^a$	$\widehat{\sigma}^b$	$\widehat{\mu}$	$\widehat{\sigma}$
Screen 2	56.1528	3.0710	1095.1	54.5483
Screen 1 (1b)	95.98	122.7	3462.9	169.1
Screen 9	27.32	9.04	575.1	74.16

[a]The mean is estimated by $\widehat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i$ where n is the size of the sample

[b]Similarly, for the variance we have: $\widehat{\sigma} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \widehat{\mu})^2}$

Table 2.1: Parameters estimation for statistical tests

the reason why the χ^2 test is applied on the HTS control data to test their normality. Since both controls are continuous variables, the clustering is arbitrary.

Principle

Let O_i be the number of events observed in the i^{th} bin and E_i the expected number according to the known distribution. The test statistic $\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$ has an approximate χ^2 distribution if:

1. no expected frequency is smaller than 1;

2. no more than a fifth of the expected frequencies are smaller than 5.

It may be necessary to combine bins in order to satisfy these conditions (see [MGH89] for detail). The number of degrees of freedom ν is given by the number of bins minus the number of constraints. The number of constraints is the number of estimated parameters plus one.

Application

Since these two parameters are estimates and the sum $\sum_{i=1}^{k} O_i$ is fixed, the number of degrees of freedom ν is $k - 1 - 2$ where k is the number of bins for the test.

These bins have equal width and divide the interval [min(*control*), max(*control*)] (originally) into 25 clusters. Table 2.2 shows the results obtained for the χ^2 statistic.

χ^2 test	Minimum controls		Maximum controls	
Degrees of freedom ν	11 (14 bins)		20 (23 bins)	
Critical values[a]	5%	1%	5%	1%
	19.67	24.72	31.41	37.56
$\chi^2_{Screen\ 2}$	19.7573		35.7496	
Conclusion	*Some evidence for non-normality*		*Some evidence for non-normality*	
Degrees of freedom ν	14 (17 bins)		12 (15 bins)	
Critical values	5%	1%	5%	1%
	23.68	29.14	21.02	26.21
$\chi^2_{Screen\ 9}$	27.9242		24.4020	
Conclusion	*Some evidence for non-normality*		*Some evidence for non-normality*	
Degrees of freedom ν	4 (7 bins)		13 (16 bins)	
Critical values	5%	1%	5%	1%
	9.48	13.27	22.36	27.68
$\chi^2_{Screen\ 1}$	186.4604		22.7099	
Conclusion	*Normality rejected*		*Some evidence for non-normality*	

[a]A significance level of α =1% gives a confidence level, probability of failing to reject H_0 when H_0 is true of 99% (the critical values for 5% are lower than those for 1%). The dual test (a significance level of 99% rejects the normality for the six samples. The Kolmogorov-Smirnov tests will be carried out in the sam way.

Table 2.2: χ^2 tests on maximum and minimum controls

Note: the different numbers of bins are due to the combinations necessary to satisfy the requirements of the χ^2 test on small expected frequencies. The value of the $\chi^2_{Screen\ 1}$ statistic is due to the presence of a great number of outliers in the *control* plates.

2.1.2 Kolmogorov-Smirnov test

The Kolmogorov-Smirnov test can be applied to unbinned distributions that are function of a single independent variable; thus it is particularly suitable for continuous variables such as the HTS control values. It can be more reliable than the χ^2 test in such cases since no arbitrary categories are required.

Principle

The test is based on the cumulative distribution function S_n given by

$$S_n(x) = \frac{1}{n}\#\{y \in E : y \leq x\} \quad .$$

The Kolmogorov-Smirnov statistic D is defined as the maximum value of the absolute difference between the the cumulative distribution function S_n and the expected distribution function F:

$$D = \max_{-\infty<x<\infty} |S_n(x) - F(x)| \quad . \tag{2.1}$$

Application

The value F in x of equation (2.1) is given by the distribution function of the Gaussian of mean $\widehat{\mu}$ and standard deviation $\widehat{\sigma}$: $F(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi\widehat{\sigma}^2}} \exp\{-\frac{(y-\widehat{\mu})^2}{2\widehat{\sigma}^2}\}dy$. Table 2.3 shows the results of the test on HTS control values.

2.1.3 Conclusion

The χ^2 rejects the normality of the data and the Kolmogorov-Smirnov test gives mixed results for the same hypothesis. It suggests that the implicit assumption of normality is not sufficient to carry out the quality control of the data. The following chapters investigate some more complex models for the variation of the control values to improve the representation of the data provided by a single Gaussian model.

Kolmogorov-Smirnov test	Minimum controls		Maximum controls	
Critical values[a]	5%	1%	5%	1%
	0.11132	0.1357	0.11132	0.1357
$D_{Screen\ 2}$	0.0567		0.0580	
Conclusion	*Normality accepted*		*Normality accepted*	
$D_{Screen\ 1}$	0.1154		0.0370	
Conclusion	*Some evidence for non-normality*		*Normality accepted*	
$D_{Screen\ 9}$	0.2615		0.0650	
Conclusion	*Normality rejected*		*Normality accepted*	

[a]For sample size n>100, the critical value D_α can be found to be $D_\alpha = \sqrt{\frac{-ln(\frac{\alpha}{2})}{2n}}$, where $\alpha < 1$ is the significance level of the test.

Table 2.3: Kolmogorov-Smirnov tests on max and min controls

2.2 Outlier detection

Before explaining why the control plates can be used for density inference, it proves interesting to mention in the first place statistical techniques for dealing with the possible presence of outliers in the HTS data. This section shows how the problem would be tackled by standard methods to detect outliers in the univariate case and the limitations of such an approach.

2.2.1 A methodology for univariate problem

A standard approach of outlier detection proposes a two-step procedure:

1. Use points, sequence, box, or normal quantile-quantile plots to spot extreme observations;

2. Apply statistical tests for outliers (also called 'discordancy tests') with an appropriate significance level to determine whether the points selected in 1 differ significantly from the rest of the sample.

There is a plethora of tests available for outlier testing - [BL78] describes 22 tests for the Gaussian case only - depending on various assumptions such as normality.

It should be noted that most of the techniques involved in outlier detection are derived under the assumption of an underlying Gaussian density.

2.2.2 Quantile-quantile plots

We first define the plots mentioned above in the step 1. The quantile-quantile plots can compare two samples suspected to be drawn from the same distribution.

Let $\{y_i\}_{i=1,2,\ldots,n}$ and $\{x_i\}_{i=1,2,\ldots,m}$ two ordered samples with $n \leq m$. For each data fraction $f_i = i/n$ in the smaller sample, $x'(f_i)$ (called 'interpolated quantile') for the largest sample is defined as:

$$x'(f_i) = \begin{cases} x_i & \text{if } n = m \quad , \\ (1-g)x_k + gx_{k+1} & \text{otherwise} \quad , \end{cases}$$

where $h = (m+1)f_i$, k is the integer portion of h and $g = h - k$ (if $m \leq k$, $x'(f_i) = x(m)$).

The quantile-quantile consists in plotting $Q_y(f_i) = y_i$ *versus* $Q_X(f_i) = x'(f_i)$, $i = 1, 2, \ldots, n$. If the two samples are identical, all the plotted points lie on the same line.

The standard normal quantile-quantile plot consists in plotting y_i *versus* $Q_{SN}(f_i)$, where $f_i = (i - \frac{3}{8})/(n + \frac{1}{4})$ and $Q_{SN}(f) = 4.91(f^{0.14} - (1-f)^{0.14})$.

In the general case, the normal quantile-quantile plot for a sample of mean μ and variance σ is derived from Q_{SN} by $Q_N(f) = \sigma Q_{SN} + \mu$.

Unusual trends or clustering on the plot may highlight outliers[2].

[2]The distinction between the two normal quantile-quantile plots is taken into account for consistency with the literature but is not actually necessary; a line remains a line after a linear transformation, an outlier remains an outlier.

2.2.3 Grubbs tests

This part presents a simple method based on the Grubbs's statistics L_k and E_k to test if a subset of a sample $\{x_i\}_{i=1,\dots,n}$ is formed of outliers. The test should proceed as follows:

1. Sort the data $\{x_i\}_{i=1,\dots,n}$ in $\{y_i\}_{i=1,\dots,n}$:

$$y_1 \leq y_2 \leq \dots y_n \quad .$$

2. (a) If the k largest values in the data set are suspected as outliers:

$$L_k = \frac{1}{s_{yy}} \sum_{i=1}^{n-k} (y_i - \bar{y}_L)^2 \tag{2.2}$$

where:

$$\bar{y}_L = \frac{1}{n-k} \sum_{i=1}^{n-k} y_i$$

and

$$s_{yy} = \sum_{i=1}^{n} (y_i - \bar{y})^2 \quad .$$

(b) Conclude that the group of k observations are outliers if the calculated value of L_k is less than the critical value[3] for L_k Grubbs test statistics[4].

3. (a) Similarly, if the k most extreme values are suspected as outliers (some are the largest while others are the smallest ones)

$$E_k = \frac{1}{s_{yy}} \sum_{i=1}^{n-k} (z_i - \bar{z}_E)^2 \tag{2.3}$$

where:

- z_i is the y_i corresponding to the i^{th} smallest $|y_i - \bar{y}_E|$;
- $\bar{z}_E$ average of the y_i corresponding to the $n-k$ smallest deviations.

[3] See the tests referenced N4 and N5 in [BL78] p91 and pp304–306 for the critical values.

[4] Alternative approach suggests the consider the statistic $\frac{t-\bar{x}}{\sqrt{s_{xx}}}$ (or ESD for *Extreme Studentized Deviate*); the strong points and drawbacks of this method being comparable to the use of the second order $\frac{(t-\bar{x})^2}{s_{xx}}$, we chose to present the latter.

(b) Conclude that the group of k observations are outliers if the calculated value of E_k is less than the critical value for E_k Grubbs test statistics.

The limitations and situations where Grubbs tests may fail are discussed in the next section.

2.2.4 Outlier detection in the context of HTS control

This section demonstrates the outlier detection on the HTS data. We discuss the results and underline the weak points of the method in the context of HTS quality control.

Application to HTS controls

The methodology described in Section 2.2.1 is applied on Screen 2. Scatter plots and normal quantile-quantile plots in Figure 2.2 guide the analysis. These plots highlight various suspiciously extreme values. The following controls (marked in Figure 2.2) are *chosen* to test their abnormality[5]:

Minimum controls: 66($D2$), 95($D6$) (Figure 2.2(a));

Maximum controls: 95($D1$), 95($D7$) (Figure 2.2(b));

Standard controls: 128($D3$), 95($D9$) (Figure 2.2(c));

For the standards, one may choose a third point pointed up '???' in Figure 2.2(c) which corresponds to the control D3 of the plate 95 but let us keep it apart for argument's sake.

Table 2.4 shows the results obtained for this outlier detection test.

[5]As in Section 1.2.1, we note '66($D2$)' the first minimum control ($D2$) of the plate 66.

Comments on the discordancy tests

1. The procedure requires judgement on the part of the analyst: it is necessary to choose a *set* of outliers. As a consequence, the test may prove positive even though some of the extreme values of the sample are left aside. Looking back at the standard control 95(D3) in Figure 2.2(c), it would have been sensible to include it in the test set since it differs only slightly from one of the controls tested, the standard control 95(D9). Nevertheless the test succeeded. In a semi-automated method, the former control would not have been detected if such a mistake had been done.

2. The test is sensitive to other outliers: the test may not be conclusive because of the presence of other extreme values in the $n - k$ values considered as 'normal'. This problem is known as '*masking*'. It explains why the test on the maximum controls is not positive. The two controls 95($D1$), 95($D7$) arise naturally in the point plot of Figure 2.2(b), yet these outliers are masked by the relatively high values of the first 40 plates. The difference between these plates and the rest of the screen will be debated more thoroughly in subsequent chapters. The design of the test (a subset tested for abnormality with respect to the rest of the sample) implies that the failure of a test does not necessarily mean that the chosen points are not outliers but rather that these are not the only ones.

More generally, if two points happen to be different from the sample by an order of magnitude, each point will appear both on the numerator and the denominator of equations (2.2) and (2.3). As a result, the corresponding test will not be significant. This is the reason why the procedure making use of graphs is generally reckoned more reliable for outlier detection.

The choice of a proper value for k, the number of outliers to be tested, is obviously crucial. One might not have been worried about choosing *more* outliers than necessary,

Discordancy test	Min 66(D2) 95(D6)		Max 95(D1) 95(D7)		Std 128(D3) 95(D9)	
Critical values	5%	1%	5%	1%	5%	1%
	0.833	0.802	0.833	0.802	0.821	0.794
test statistic	0.5518		0.8184		0.6903	
Conclusion	*Positive*		*Some evidence*		*Positive*	

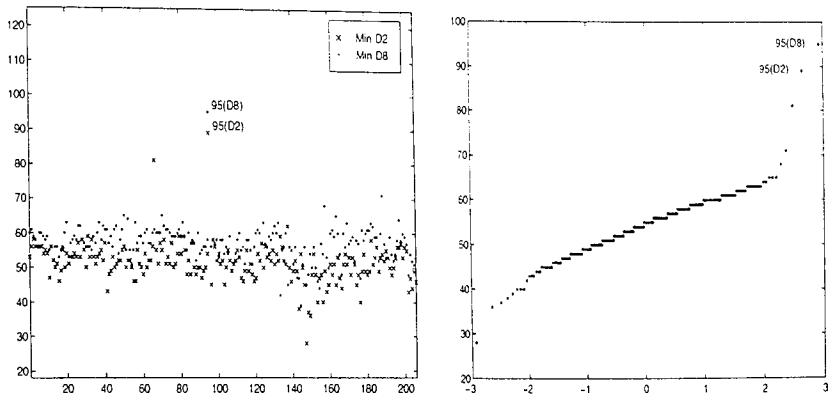
Table 2.4: Discordancy tests

had the validation been carried out manually. The problem is that in the case of the discordancy tests, such a choice would make the test fail. On the other hand, if too *few* outliers are chosen for testing, the test might succeed in the case of large samples despite leaving extreme values undetected as was the case for the standard 95($D3$).

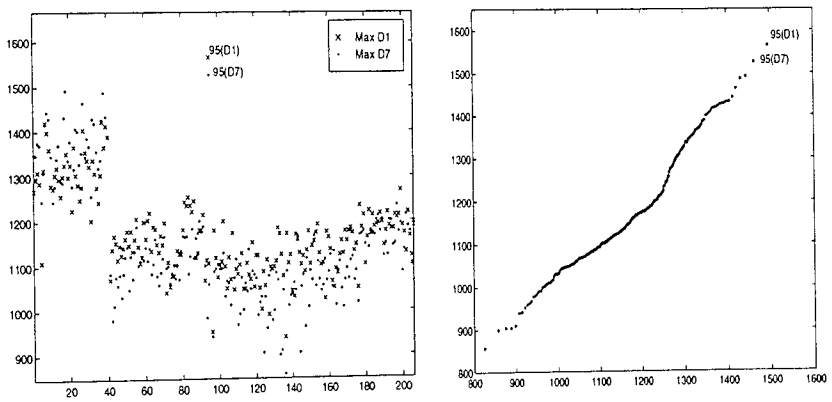
Potential problems in applying the tests

The application of outlier detection in the context of a quality control of HTS data highlighted the following problems:

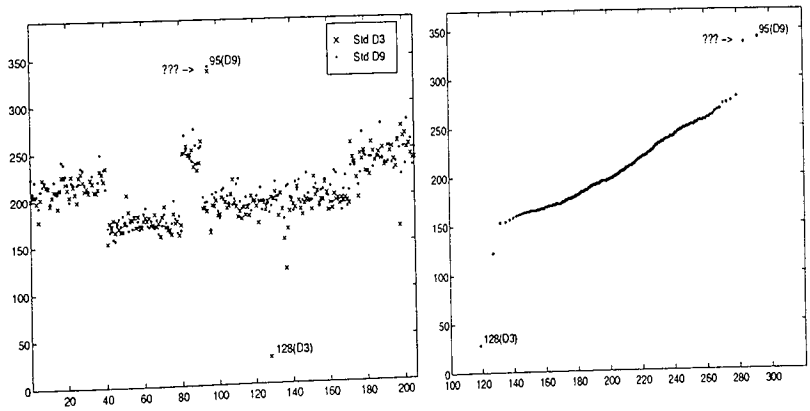
- The method is not based on a density estimation so no description of the data is provided in terms of probability;
- In particular, no measure of novelty is possible, neither is an ordering in the outliers detected;
- This procedure is greatly subjective for the choice of the outliers and the value of k to be tested for; in addition the 'manual' graph inspection leaves no possibility of automation.
- The analysis is more complex in the multivariate case (among other things, the visualisation in a 6-dimensional space is not as easy);



(a) Minimum controls



(b) Maximum controls



(c) Standard controls

Figure 2.2: Point plots (left) and normal quantile-quantile plots (right) for the HTS controls

2.3 Correlation tests

This section presents a popular test for correlation between two random variables X and Y from which two samples $\{x_i\}_{i=1...n}$ and $\{y_i\}_{i=1...n}$ have been drawn. It is based on Pearson's r (or *sample correlation coefficient*) and is applied on the three HTS controls, minima, maxima and standards, of the normal plates.

2.3.1 Principle and application to the controls

The tests of existence of a correlation between the various controls are based on Pearson's sample correlation coefficient given by:

$$r = \frac{s_{xy}}{\sqrt{s_{xx}s_{yy}}}$$

where $s_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})$ and $s_{xx} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$ (if $r = \pm 1$ there exists a linear dependence between X and X). The statistic is given by:

$$t = \frac{r(n-2)^{1/2}}{(1-r^2)^{1/2}}$$

and tests the nullity of the correlation coefficient $\rho = \frac{cov(X,Y)}{\sqrt{var(X)var(Y)}}$. It can be assessed as a Student cumulative statistic[6].

This statistic tests the following alternative:

> H_0: *X and Y are not correlated* ($\rho = 0$);
>
> H_1: *X and Y are correlated* ($\rho \neq 0$).

The results are shown in Table 2.5.

2.3.2 Comments

As expected, the controls of the 96-well plate are all mutually correlated.

[6] *cf* [MGH89] p440 for example.

Independence tests	min/max	min/std	max/std
t-statistic	4.10	2.98	9.69
Critical values t_∞	2.576	2.576	2.576
Conclusion	*Independence rejected*	*Independence rejected*	*Independence rejected*

Table 2.5: Correlation tests (significance level 0.5%)

There is a strong correlation between maximum and standard controls. If we recall that there is no difference between the standards and the normal plates[7] , this strong correlation shows that it makes sense ('statistically speaking') to rely on the controls to assess the quality of the data collected: an unusual variation of the controls would denote an unusual variation of the whole plate.

A significant difference exists between the correlation of maxima and standards and the two other statistics. These two controls may be more sensitive to the experimental conditions than the minima.

[7]The only difference between a standard control and a normal well is that the activity of the standard is known (*cf* Section 1.1.2).

Chapter 3

Novelty detection

Novelty detection aims at determining abnormal or unusual outputs of any industrial task. In the context of the quality control of an experimental scheme, 'novel' points can be defined as the observations which have unusual values compared to other data observed under the same experimental conditions. So far as HTS is concerned, the novel control wells should be studied since they may reveal experimental conditions which may not be those that were intended and therefore should not be taken into account in the activity boundary computation (Section 1.2.1).

From a probabilistic point of view, if the distribution of 'normal values' is known, a novel point is the one which is 'unlikely' for this distribution. Precisely, a point is declared 'novel' if its probability is below a novelty threshold to be determined. As a result, this chapter focuses on:

- modelling the distribution of normal controls;
- defining the 'novelty threshold'.

In this chapter, we motivate the choice of the density model, Gaussian mixture models, and describe the probability density inference technique, the *Expectation-Maximisation* algorithm. In the second place, we show how this model can be trained on the three control plates. The choice of the novelty threshold is then discussed together with its implications regarding the HTS quality control. Finally, we consider

the possibility of data pre-processing and determine the parameters of the model.

3.1 Probability density estimation

This section presents an overview of the techniques used to infer the density of the HTS control values. A density model, the mixture model, is introduced and we describe how the parameters of this model can be estimated by the *Expectation-Maximisation* or EM algorithm in the Gaussian case. Details can be found in [Bis95].

3.1.1 Mixture models

A mixture model represents the underlying density function $p(\mathbf{x})$ of the data as a linear combination of M basis functions:

$$p(\mathbf{x}) = \sum_{j=1}^{M} P(j)p(\mathbf{x}|j) \quad , \tag{3.1}$$

where $P(j)$ and $p(\mathbf{x}|j)$ are respectively the priors (or 'mixing coefficients') and the likelihood that $\mathbf{x}$ is from component j.

The priors should satisfy the constraints:

$$\begin{cases} 0 \leq P(j) \leq 1 \quad , \\ \\ \sum_{j=1}^{M} P(j) = 1 \quad . \end{cases} \tag{3.2}$$

The components $p(\mathbf{x}|j)$ of the mixture are normalised so that: $\int p(\mathbf{x}|j)d\mathbf{x} = 1$, $\forall j$, $j = 1, \ldots, M$. The component densities $p(\mathbf{x}|j)$ are chosen to be Gaussian density functions:

$$p(\mathbf{x}|j) = \frac{1}{(2\pi)^{(d/2)}|\mathbf{\Sigma}_j|^{1/2}} \exp\left\{ -\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_j)\mathbf{\Sigma}_j^{-1}(\mathbf{x} - \boldsymbol{\mu}_j)^T \right\} \quad , \tag{3.3}$$

where $\boldsymbol{\mu}_j$ and $\boldsymbol{\Sigma}_j$ are respectively the mean and the covariance matrix of the component j. The problem of density estimation is therefore to determine the parameters of the model: $\{P(j), \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j,\ j = 1, \ldots, M\}$. This type of model is generally known as 'Gaussian mixture model'.

The elements of the covariance matrix model $\boldsymbol{\Sigma} = [\sigma_{k,l}]_{k,l=1\ldots d}$ are intended to model the covariance[1] $cov(X_k, X_l) = \mathcal{E}[(X_k - \mathcal{E}X_k)(X_l - \mathcal{E}X_l)]$ of the underlying random variables X_k, X_l (it is therefore symmetric). Those commonly used for mixture models can be divided into three different types presented here together with their main properties:

1. Full covariance matrix:

   - no constraint on the model;
   - the number of parameters is $d(d+1)/2$ and the inversion of $\boldsymbol{\Sigma}$ in equation (3.3) is difficult (computationally expensive);
   - the curves of equal density values are ellipses without any constraint on their directions.

2. Diagonal matrix $Diag(\sigma_1^2, \ldots, \sigma_d^2)$ where the σ_i are not necessary equal:

   - ignores the possible correlation between variables (the covariance $cov(X_l, X_k)$ for $k \neq l$ 'is modelled by 0');
   - the number of parameters is d, the inversion of $\boldsymbol{\Sigma}$ is easy;
   - the curves of equal density values are ellipses whose axis directions are given by the standard basis vectors $\begin{bmatrix}1\\0\\\vdots\\0\end{bmatrix}, \ldots, \begin{bmatrix}0\\\vdots\\1\\0\\\vdots\\0\end{bmatrix}, \ldots, \begin{bmatrix}0\\\vdots\\0\\1\end{bmatrix}$ ('the axis of the graph').

3. σI where σ is a real positive number and I the identity matrix:

[1] $\mathcal{E}[X]$ is the expectation of X.

- constrains the elements of the diagonal to be equal and the others to be nought;
- the number of parameters is equal to 1, the inversion of Σ is trivial;
- it is generally used on centred data;
- the curves of equal density values are circles.

The choice of the covariance matrix used for novelty detection will be discussed in detail in Section 3.5.2.

3.1.2 *EM* algorithm

The *Expectation-Maximisation* or EM algorithm provides a effective means to determine the parameters of a mixture model. This section presents briefly this iterative algorithm and the initialisation used.

The updating relations

Suppose we want to find a mixture model (3.1) which describes the distribution of a data set $\mathcal{T} = \{\mathbf{x}_n\}_{n=1,\dots,N}$ where $\mathbf{x}_n = (x_1^{(n)}, \dots, x_d^{(n)})$ is a d-dimensional vector.

Most of the techniques for determining the parameters of a Gaussian mixture model rely on the maximisation the likelihood of the parameters: $\mathcal{L} = \prod_{n=1}^{N} p(\mathbf{x}^n)$ *i.e.* minimising the negative likelihood error given by:

$$E = -\ln \mathcal{L} = -\sum_{n=1}^{N} \ln \left\{ \sum_{j=1}^{M} P(j) p(\mathbf{x}^n | j) \right\} \quad . \tag{3.4}$$

To simplify the notation, let θ be the set of parameters to be determined:

$$\theta = \{P(j), \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j, \ j = 1, \dots, M\} \quad .$$

Let $Q(\theta, \theta') = \mathcal{E}[\ln \mathcal{L} | \theta']$, function of the observed data $\{\mathbf{x}_n\}_{n=1,\dots,N}$. The EM algorithm starts from an initialisation $\theta^{(0)}$ and alternates two steps:

E-step: Find $Q(\theta, \theta^{old}) = \mathcal{E}[\ln \mathcal{L}|\theta^{old}]$.

M-step: Choose θ^{new} to maximise $Q(\theta, \theta^{old})$.

The maximisation of the likelihood $\mathcal{L}$ (**M-step**) is obtained gradually at each iteration (*new*) by conditioning the expectation (**E-step**) on the values of the parameters of the former (*old*). If we recall the expression of E, this is equivalent to choosing θ^{new} for minimising the expectation $\mathcal{E}[E|\theta^{old}]$ leading to a new error E^{new}. In the case of mixture model (3.1), it can be shown[2] that the *new* error admits the upper bound given by:

$$E^{new} \leq E^{old} - \sum_{n=1}^{N}\sum_{j=1}^{M} P^{old}(j|\mathbf{x}^n) \ln\left\{\frac{P^{new}(j)p^{new}(\mathbf{x}^n|j)}{p^{old}(\mathbf{x}^n)P^{old}(j|\mathbf{x}^n)}\right\} \quad . \tag{3.5}$$

As the **E-step** maximise the expectation conditioned on the *old* parameters (E^{old} is fixed), we wish to minimise the second term of (3.5), which will lead to a minimum unless E^{new} is already a minimum.

In the case of Gaussian mixture, the value of $\theta^{new} = \{P^{new}(j)\ , \boldsymbol{\mu}_j^{new}, \boldsymbol{\Sigma}_j^{new},\ j = 1, \ldots, M\}$ can be expressed as a function of $\theta^{old} = \{P^{old}(j), \boldsymbol{\mu}_j^{old}, \boldsymbol{\Sigma}_j^{old},\ j = 1, \ldots, M\}$.

For $\boldsymbol{\Sigma} = Diag(\sigma_1^2, \ldots, \sigma_d^2)$ (the choice of such a covariance matrix is justified in Section 3.2) the minimum of the second term of (3.5) is obtained by differentiation with respect to the parameters $P(j)$, μ_j, $\sigma_t^{(j)}$. This leads to the updating relations:

$$\mu_j^{new} = \frac{\Sigma_n P^{old}(j|\mathbf{x}^n)\mathbf{x}^n}{\Sigma_n P^{old}(j|\mathbf{x}^n)} \quad , \tag{3.6}$$

$$(\sigma_t^{(j)\ new})^2 = \frac{\Sigma_n P^{old}(j|\mathbf{x}^n)(x_t^{(n)} - \mu_t^{(j)})^2}{\Sigma_n P^{old}(j|\mathbf{x}^n)} \quad , \tag{3.7}$$

$$P(j)^{new} = \frac{1}{N}\sum_n P^{old}(j|\mathbf{x}^n) \quad , \tag{3.8}$$

for $t = 1, \ldots, d; j = 1, \ldots M$. Proof and details can be found in [Bis95] for the case $\boldsymbol{\Sigma} = \sigma^2 I$ (change the expression of equation (3.7)).

[2]Using equation (3.4) and Jensen's inequality: $\ln\left(\sum_j \lambda_j x_j\right) \geq \sum_j \lambda_j \ln(x_j)$ if $\lambda_j \geq 0$ and $\sum_j \lambda_j = 1$ (See [Bis95] for example).

Starting values

The performance of the EM algorithm for minimising the error (3.4) can depend on the starting points of the updating equations (3.6), (3.7) and (3.8) . In the training procedure, the parameters are initialised to $\frac{1}{M}$ for $P(j)$ for all j. The centres μ_j are chosen randomly in the interval $[\min(\mathcal{T})\max(\mathcal{T})]$ and the variance as:

$$\min_{i \neq j} || \mu_i - \mu_j || \ .$$

As far as the implementation is concerned, the method checks the values of Σ_j at each step of the EM-algorithm using the Linpack reciprocal condition estimator in order to avoid ill-conditioned matrices.

3.1.3 Why Mixture Models?

Several theoretical arguments can be provided to justify the choice of mixture model as the structure for density inference. First, the models (3.1) have the universal approximation property: they can fit any probability density. Another powerful consequence of using Gaussian mixtures defined by the equations (3.1), (3.2) and (3.3) is that it becomes straightforward to compute the conditional densities because these remain Gaussian (*cf* Section 4.4.1).

From a practical point of view, the mixture models were preferred to other probability density estimators such as Parzen windows estimator because of their speed in evaluating the density at a new data point, which should be regarded as an asset in a routine use.

A drawback is the extra time necessary for training when compared to Parzen windows estimator, for example. In particular, the main critics argue on the slow convergence of the EM if the mixture components are not well separated [XJ95]. This relatively slow convergence should not be considered as a problem since in practice, the control of HTS is not on-line: the extra plates and the standard plates are generally screened on different days. Furthermore, even in an automated procedure, the model

can be trained on the control plates while the normal plates are being screened. This training is a matter of a few minutes in this case which clearly fulfild the requirements settled in Section 1.3.1. As a result, the learning time is not a crucial issue[3] as long as it remains in this order of magnitude. Moreover other non-linear optimisation methods such as gradient based methods are expected to perform poorly on such an ill-separated mixture [XJ95].

The EM algorithm may not be suitable for problems involving several clusters when the starting points do not separate sufficiently the group means. In such a situation, the EM can converge to an inappropriate local minimum for the error (3.4) as reported in ([Rip96], p208) and therefore a poor local maximum for $\mathcal{L}$. In the case of the HTS controls, the nature of the data, a measure of activity for wells containing the *same* mixture, does not suggest separated clusters. The examination of the data provided confirms this intuition (see Figure 3.3 for example). As a result, the complexity M to be determined in Section 3.5.1 does not aim at distinguishing distinct clusters but at determining a more accurate description of the data[4].

The following advantages of the EM overdraw the drawbacks:

- the EM provides a monotonic convergence without the need to set a learning rate;
- the EM gives low computational overhead[5].

[3]The implementation in Matlab of the learning procedure which will be described in Section 3.3 takes 2min 36s on a Sparc (Sun4d) for the data of Screen 2 (206 plates, Appendix A.1) which is quite acceptable in the context of the HTS control since the quality control is not intended to be carried out straight after the screening. Moreover this relatively fast training would permit the integration of the novelty detection scheme into an automated control system.

[4]As mentioned in Section 2.1, the distribution of the control values of HTS can hardly be considered as Gaussian (unimodal). Nevertheless the density function is not well separated so that clusters can not be characterised.

[5]The software is expected to run on micro-computers.

3.2 Number of components

The choice of M, the number of basis functions in the model (3.1) is known to be difficult [LB88]. The available tests for testing the number of components of a mixture model can be divided into two categories:

- the tests based on the likelihood ratio test[6] are the most common (using bootstrapping for example as in [Lac87]),
- the tests based on moment estimators [FL94].

The first problem is that there exists no criterion to determine the optimal choice for the number of components M. Indeed, the tests mentioned above can compare two models in order to choose between 1 and 2 components. Besides, all the tests rely on the hypothesis of homoscedasticity (the basis functions have equal variance) which is not necessarily compatible with the the EM algorithm.

So far, the problem of an adequate choice of number of basis functions in its generality remains unsolved from a theoretical point of view. This is the reason why for this study, the choice of the number of components in the basis is empirical (Section 3.5.1).

3.3 Training and validation procedure

This section explains the procedure for determining a probability density model which describes the distribution of the HTS control values. Cross validation is used to avoid

[6]The problem of deciding the number of components of a model can be formulated in terms of likelihood maximisation: suppose we have an alternative H_0, H_1 for 2 different values (or number) of parameters. H_0 is rejected if :

$$\max_{\theta|H_0} \mathcal{L}(\theta) < \max_{\theta|H_1} \mathcal{L}(\theta).$$

The likelihood ratio test uses the statistic:

$$\lambda^\star = \frac{\max_{\theta|H_0} \mathcal{L}(\theta)}{\max_{\theta|H_1} \mathcal{L}(\theta)}.$$

If the distribution of $\lambda^\star$ is known (which is the case only with serious restrictions on H_i), critical values can be determined and the hypotheses H_0 can be tested *versus* H_1.

over-fitting the data. The criterion for selection among several models is discussed in the second part.

3.3.1 Cross validation

The minimisation of the error (3.4) on a single set $\mathcal{T}$ does not ensure a good performance of the model when presented new data. To find the model which has the best prediction on new data the learning procedure for density inference we proceed by *cross-validation.* The Gaussian mixture model is trained and validated on the sets whose construction is detailed in Figure 3.1. As they constitute the reference regarding the distribution of the controls, the data formed by the control values of the three first plates generate both the training and the validation set. The set C of 2-tuples, random permutation of the control values, is split into two sets C_1 and C_2:

- C_1 is used to generate the training set: a set of 4-tuples $\mathcal{T}$ is created randomly from C_1;
- C_2 is used to generate randomly 4-tuples[7] for the validation set $\mathcal{V}$.

This procedure is repeated ten times. A more complex cross validation procedure for generating training and validation sets can be considered. One might choose to divide C into ten subsets, using nine of them for training and evaluating the error on the last one. Such a procedure can be repeated ten times by changing the validation set for one of the nine training sets (see [Bis95] p374 for example). In practice, this alternative gives similar results on the generalisation error as the one previously mentioned. As a result, we chose the simplest cross validation procedure.

[7]We call 'd-tuple' a vector belonging to a d-dimensional space (d components). The 4-tuples refer to the 2 *Totals* and 2 *NSBs* of the normal HTS plates exposed in Section 1.2.2: $(D2, D1, D8, D7) = (min_1, max_1, min_2, max_2)$. Similarly, 6-tuples will be considered when the standard controls are included to the procedure to create vectors of the form $(min_1, max_1, std_1, min_2, max_2, std_2)$ (Section 4.5).

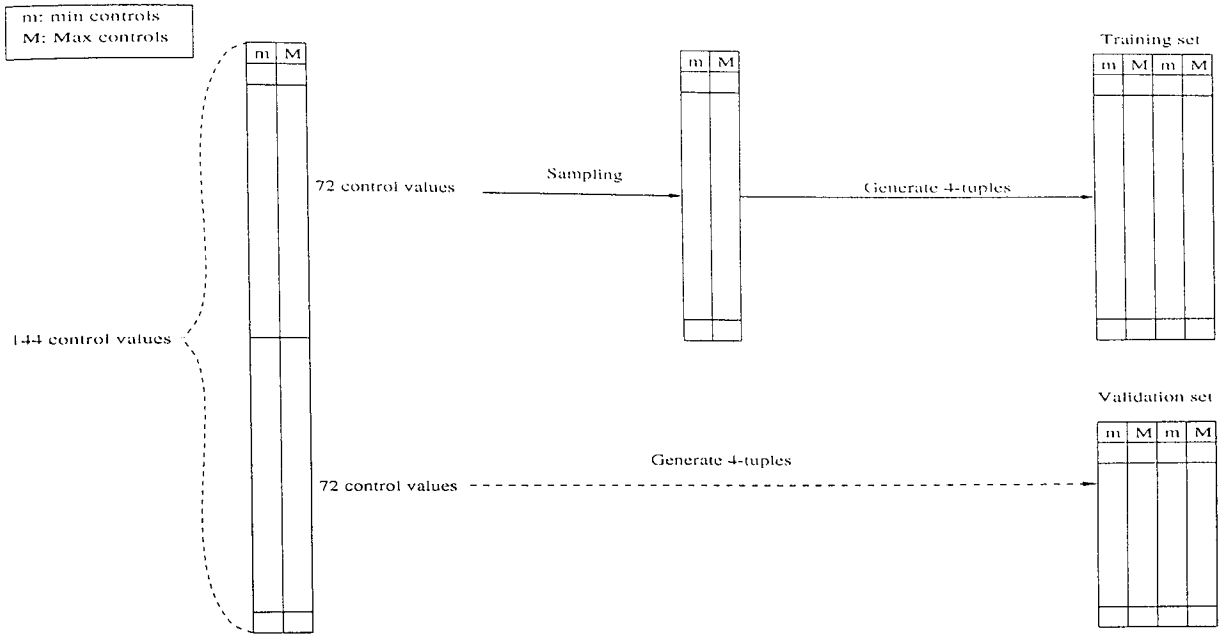


Figure 3.1: Training and Validation set generation

3.3.2 Selection criterion

One of the ten models computed by the procedure indicated in Section 3.3.1 must be chosen according to what is considered to be 'a good model'. The best fit with respect to the generalisation error is kept: our choice will be the probability density which provides the best description of the validation set distribution (the one which gives the smallest error (3.4) on $\mathcal{V}$).

A different criterion of choice among those ten models can be considered. One may select the model inducing the smallest number of novel points in the validation set (using as novelty boundary the smallest value of the density function on the training set). These two strategies differ in principle. The first one prefers the best description of the data. The second, focusing on novelty detection, implies that the training set is 'perfect' so that the fewest elements should be abnormal.. Therefore a good model would be the one that rejects few points. In practice, the two strategies give similar results on detecting novelty on the screen test, HTS Screen Number 2.

3.4 How is the novelty threshold defined?

The advantage of using a novelty threshold is that it ensures the detection to be carried out systematically; regarding the second requirement exposed in Section 1.3.1, it constitutes an objective criterion for deciding whether a point is unusual. We choose to define the novelty threshold as the minimum value of the density function of the validation set. It implies that the controls of the normal plates which have a smaller probability than the smallest probability of the controls of the control plates are declared novel.

Instead of using the training set to fix this novelty threshold, another possibility would be to compute its value as a significance test. Using a set of points sampled from the density function, a value for the threshold corresponding to, say, a 95% novelty rate can be determined (*i.e.* to set the novelty threshold to the 95^{th} percentile of the density function). In this case, we would expect to find ten 'abnormal' plates on a 'normal' screen (normal plates classified as abnormal by chance). This alternative takes advantage of the probabilistic description of the data provided by the probability density. This would have been impossible with the standard statistical techniques described in Section 2.2. Finally, such a choice of threshold is intuitive and easily interpreted. In this respect, it follows the third constraint of Section 1.3.1.

3.5 Model parameters selection

3.5.1 Choice of M: size of the basis

The difficulty concerning the number of basis functions mentioned in Section 3.2 requires an empirical answer to the question "what is the value of M?". This section explains why the value of M is set to 2.

As described in Section 3.3.1, cross validation is used to find a model for each complexity. The data set was made of three control plates (288 wells) that is to say

$3 \times 48 = 144$ minimum controls and 144 maximum controls (Screen 2, Appendix A.1.1).

First, a single Gaussian density is fitted to the data $\mathbf{x} = \{\mathbf{x}^i\}_{i=1..144}$, using the estimators $\boldsymbol{\mu} = \mathcal{E}[\mathbf{x}]$ and $\Sigma = \mathcal{E}[(\mathbf{x} - \boldsymbol{\mu})^2]$. For the complexity $M = 2, \ldots, 6$, the parameters of the Gaussian Mixture Models are obtained by the EM algorithm using the equations (3.6), (3.7) and (3.8) and the initialisation indicated in Section 3.1.2. The first set is used to train the model using re-sampling. The generalisation error of this model is found on the 144 points validation set. This procedure was used ten times for each complexity. Figure 3.2 shows the negative log-likelihood error for the different values of M.

The curve of the error on the validation set clearly shows the improvement between the single Gaussian and the 2 mixture model but does not decrease significantly for higher values. As a result, the model complexity M is set to 2.

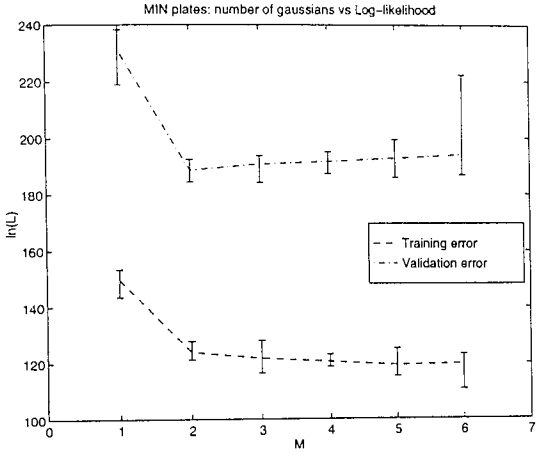
3.5.2 Choice of Σ: $\sigma^2 I$ vs. $\mathbf{Diag}(\sigma_1^2, ..., \sigma_d^2)$

In order to choose the structure of the covariance matrix Σ and a possible data pre-processing, three different procedures are tested:

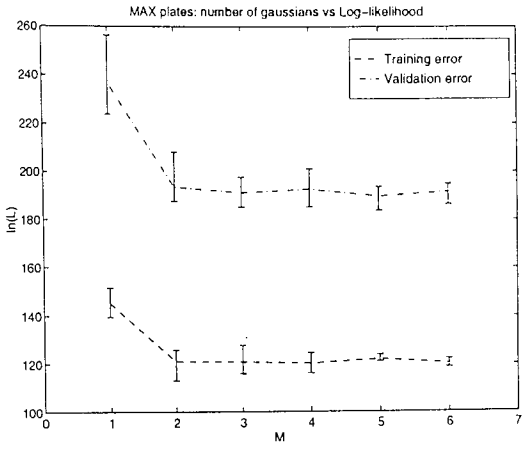
1. Train the model on raw data with $\Sigma = Diag(\sigma_1^2, \ldots, \sigma_d^2)$;
2. Train the model on raw data with $\Sigma = \sigma^2 I$;
3. Train the model on centred data with $\Sigma = \sigma^2 I$.

The generalisation error (negative log-likelihood on the validation set) is then computed for the three cases. The mixture is composed of 2 Gaussians and trained using the same procedure as in Section 3.5.1.

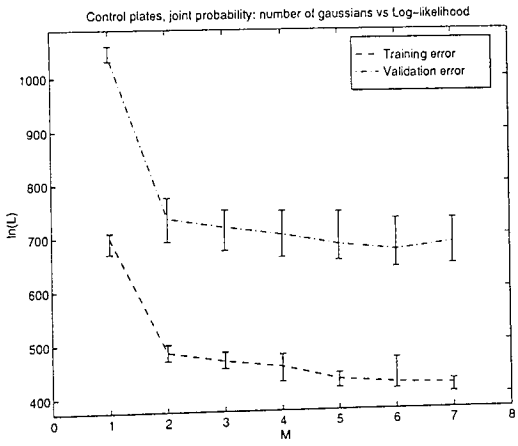
Table 3.1 presents the average of the generalisation error for ten runs. The results for the centred data includes the correction term induced by the normalisation (*cf* Appendix C).



(a) Cross validation 1

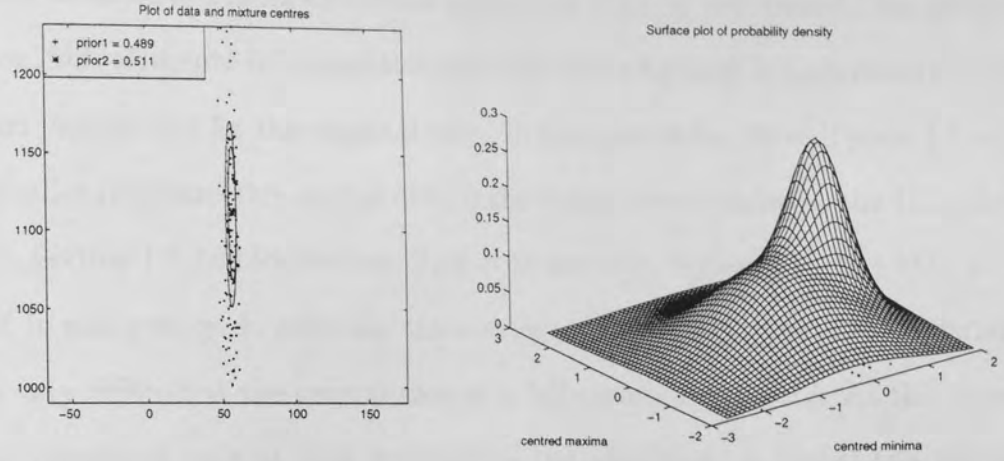


(b) Cross validation 2

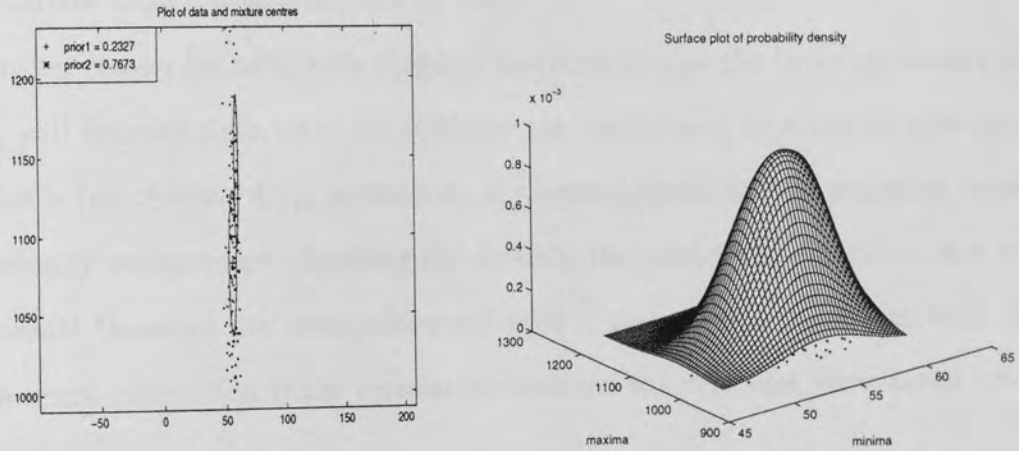


(c) Cross validation 3

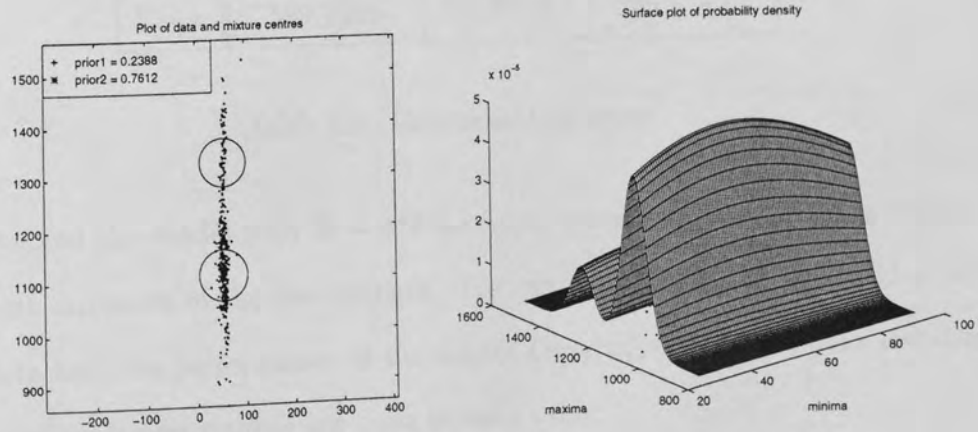
Figure 3.2: **Model order for Gaussian mixture model vs log-likelihood error:** for each number of basis functions, the likelihood of the model is computed ten times. The dash line joints the average of the values which range over the error bars.



(a) centred data



(b) Diag(σ_1^2, σ_2^2), raw data



(c) $\sigma^2 I$, raw data

Figure 3.3: Standard deviation contours and sample data (left) and probability density model (right)

The full covariance matrix described in Section 3.1.1 is not tested. As noted in this section, this structure is computationally expensive because it features $d(d+1)/2$ parameters instead of d for the diagonal case. In the case of the 96-well plate ($d = 4, 5$ or 6), it implies 10 parameters instead of 4. If the method is extended to the IC_{50} plates exposed in Section 1.2.2 which features $2 \times 8 = 16$ controls, it means $16(16+1)/2 = 136$ instead of 16 parameters. In addition, the computation of the inverse of the covariance matrix is very difficult in the general case of a full covariance matrix. As this inverse should be computed once at each iteration of the algorithm, it makes the use of a full covariance matrix all the more expensive. As a result, this matrix structure is inappropriate from a practical point of view.

Another reason for using only diagonal matrices such as the three structures above stated, will become clear when we shall use the conditional densities to spot the unusual wells (see Section 4.5). In the case of diagonal matrices (the marginal variables are mutually independent regarding the model), the conditional densities of a multidimensional Gaussian are straightforward; even if they remain Gaussian, they would require extra calculation if the correlation between the variables were taken into account.

	$Diag(\sigma_1^2, \sigma_2^2)$	$\sigma^2 I$	Centred data
Error	589.9229	734.7716	590.5702

Table 3.1: Generalisation error

As expected the model with $\Sigma = \sigma^2 I$ performs poorly on the raw data because of the different variances of the two controls. The performance of the model $\Sigma = \sigma^2 I$ on centred data and the performance of the model $Diag(\sigma_1^2, \sigma_2^2)$ on raw data are similar. Their respective training times are close to each other.

The model for novelty detection on HTS was trained on raw data with the matrix $Diag(\sigma_1^2, \ldots, \sigma_d^2)$ for four main reasons:

- The small dimension of the problem (4 and 5/6 in the last stage of the project) does not require pre-processing which can be necessary in high dimension problems;

- The complexity of the model can be largely determined by the transformation applied to the data and should be avoided if possible [NCCR+97]. The normalisation involves a loss of information in the data, which may alter the model;

- In particular, the normalisation imposes an extra constraint on the ratio between the two axis of the ellipsis of Figure 3.3 (given by $\frac{\widehat{\sigma_1}}{\widehat{\sigma_2}}$, where the $\widehat{\sigma_j}$ are the standard deviations used for normalisation);

- Because of this small dimension, the use of $Diag(\sigma_1^2, \ldots, \sigma_d^2)$ for Σ is not computationally expensive when compared to $\Sigma = \sigma^2 I$: the gain of the latter in terms of memory allocation and speed is not significant.

Chapter 4

Application

This Chapter presents the results obtained by the novelty detection techniques described in Chapter 3 on HTS data.

In the second place, an alternative method is introduced: the 'Adaptive Mixture Model', whereby the number of components of the mixture can be determined during training. This method is tested on the same screen to compare it to the first approach.

Third, since we are not only interested in spotting rogue plates but also unusual well values, we demonstrate how the conditional densities of the mixture model enable to distinguish within the 4-tuple which component may be abnormal.

Finally, we include the last two controls, the standards, in the novelty detection technique.

4.1 Novelty detection on Screen2

The novelty detection method has been applied on Screen 2 to detect novel plates (the IC_{50} plates were not used for the tests).

The novelty threshold is defined as the minimum of the likelihood function on the validation set. The alternative proposed in Section 3.4 gives similar results for a 1% rejection region.

Figure 4.1-4.4 show the results obtained for this screen. The normal plates are

ordered from 1 to 206 (which correspond to the reference 6-211 in Appendix A.1.1).

The first and the second graphs from top represent respectively the values of the maximum and the minimum wells in a similar way to that used in practice for assessing a screen[1]. On these graphs, the stars '*' denote the first maximum and minimum controls (D1 and D2 in Figure 1.2) of the 96-well plate. The plus '+' denote the maximum D7 and the minimum D8.

The third plot is the negative logarithm of the likelihood[2] of the corresponding 4-tuples and the dotted line represents the novelty threshold above which a point is declared novel. *This graph can be seen as a measure of the novelty of the plate*: the higher this value, the more the corresponding plate differs from those which have been learnt from the control plates (the 'more novel' the plate is)[3].

The novelty detection on the Screen Number 2 declared 73 plates as novel out of 206 plates. The results day by day are summarised in Table 4.1. To compare these results to the manual HTS control, the rejection of a plate does not imply here that all four controls D1, D2, D7 and D8 are 'abnormal' (and would be de-selected in the visual inspection described in Section 1.1.2) but only that at least one component of this 4-tuple is unusual or perhaps that no single value is strictly unusual, but the combination of values is. The question of determining which control(s) is abnormal among the four will be raised in Section 4.4.

[1]The dotted lines of each graph represent $\overline{x}$, $\overline{x}+\sigma_x$ and $\overline{x}-\sigma_x$ used to guide the operator in a real HTS analysis.

[2]The value $p(x)$ taken by a density function p for a point x is generally called the 'likelihood' of this point as mentioned in Chapter 3. Since it can be interpreted as an error (the smaller the likelihood, the greater the error), it is convenient to consider $-log(p(x))$ which has values in the interval $[0, +\infty)$.

[3]For convenience of analysis, the graph has been resized; the circles 'o' on the third graph (plate 7 in Figure 4.1 for example) correspond to values of the negative log-likelihood lying out of the boundaries and therefore denote 'very novel' plates

Date	Total number of plates	Number of rejected plates	Proportion
28/11/96	40	38	95%
06/11/96	40	4	10%
13/11/96	11	2	18%
07/11/96	40	6	15%
12/11/96	40	17	42%
13/11/96	35	6	17%

Table 4.1: Proportion of rejected plates per day (assay)

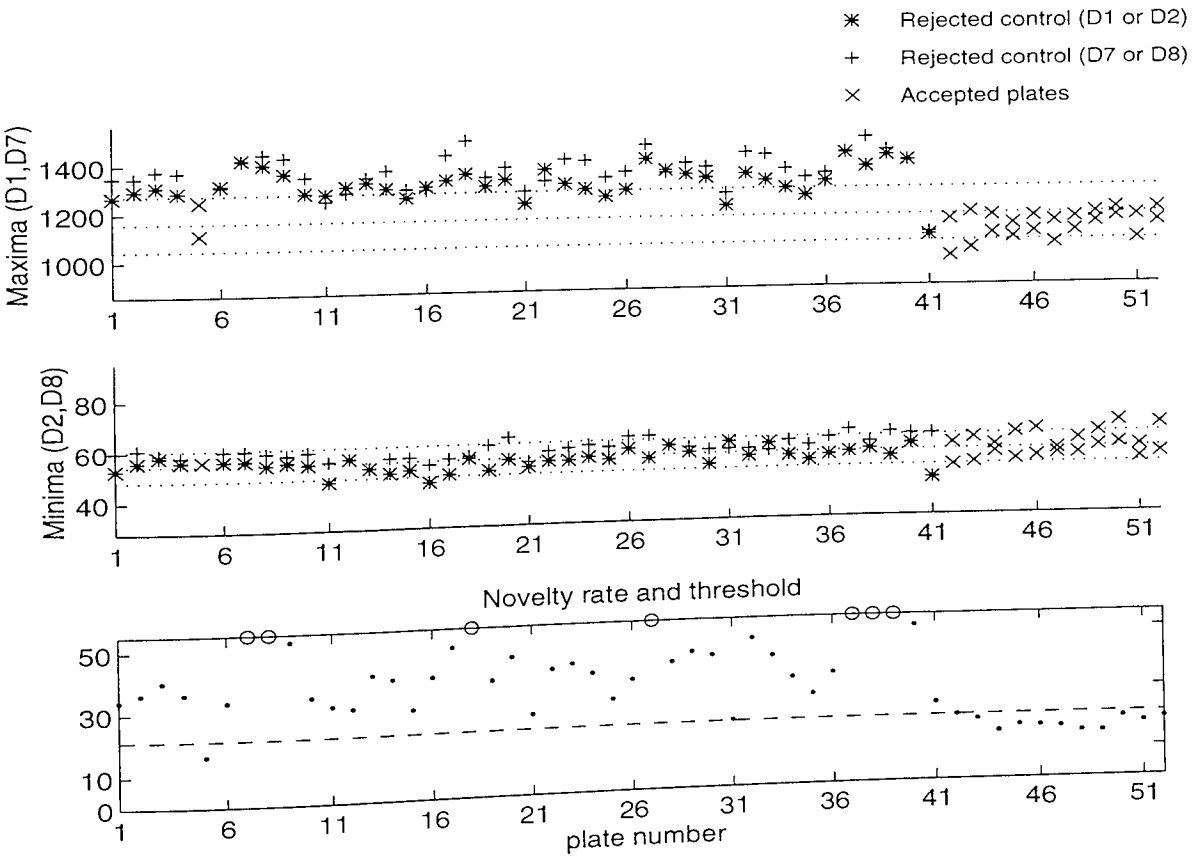


Figure 4.1: **Novelty detection on HTS screen:** plates 1 to 52.

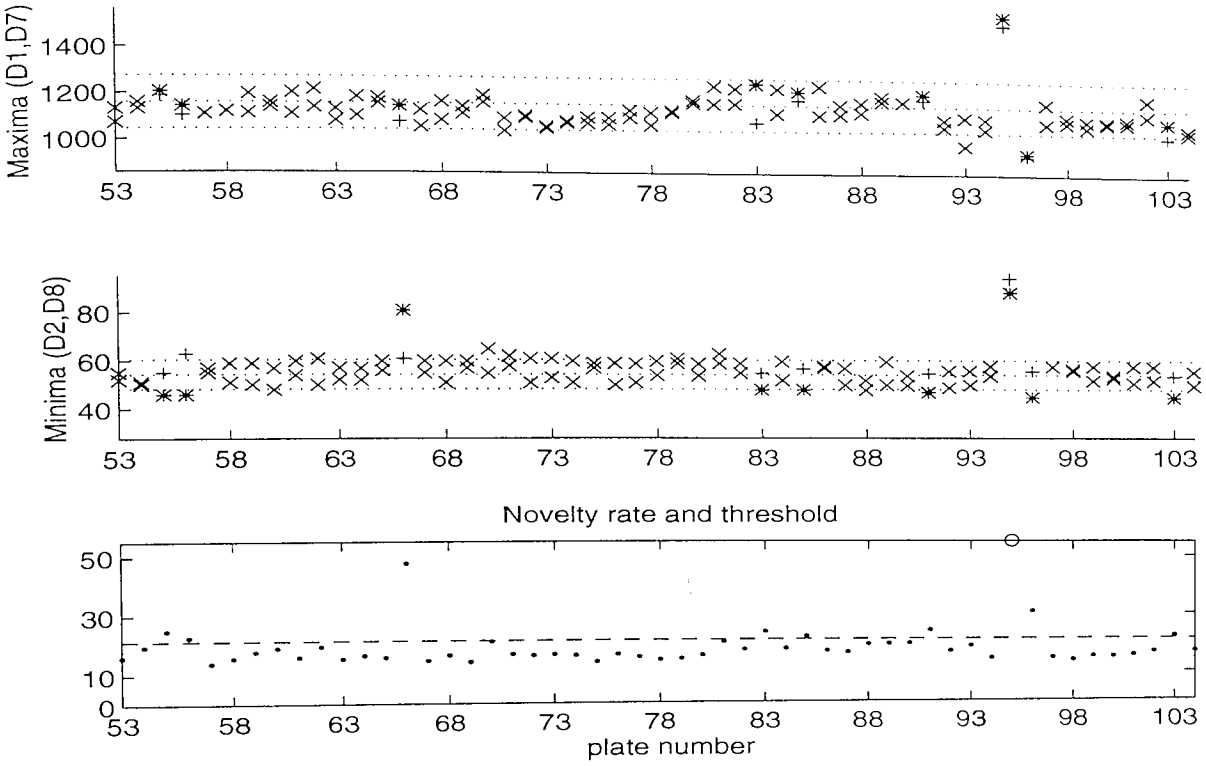


Figure 4.2: **Novelty detection on HTS screen:** plates 53 to 104.

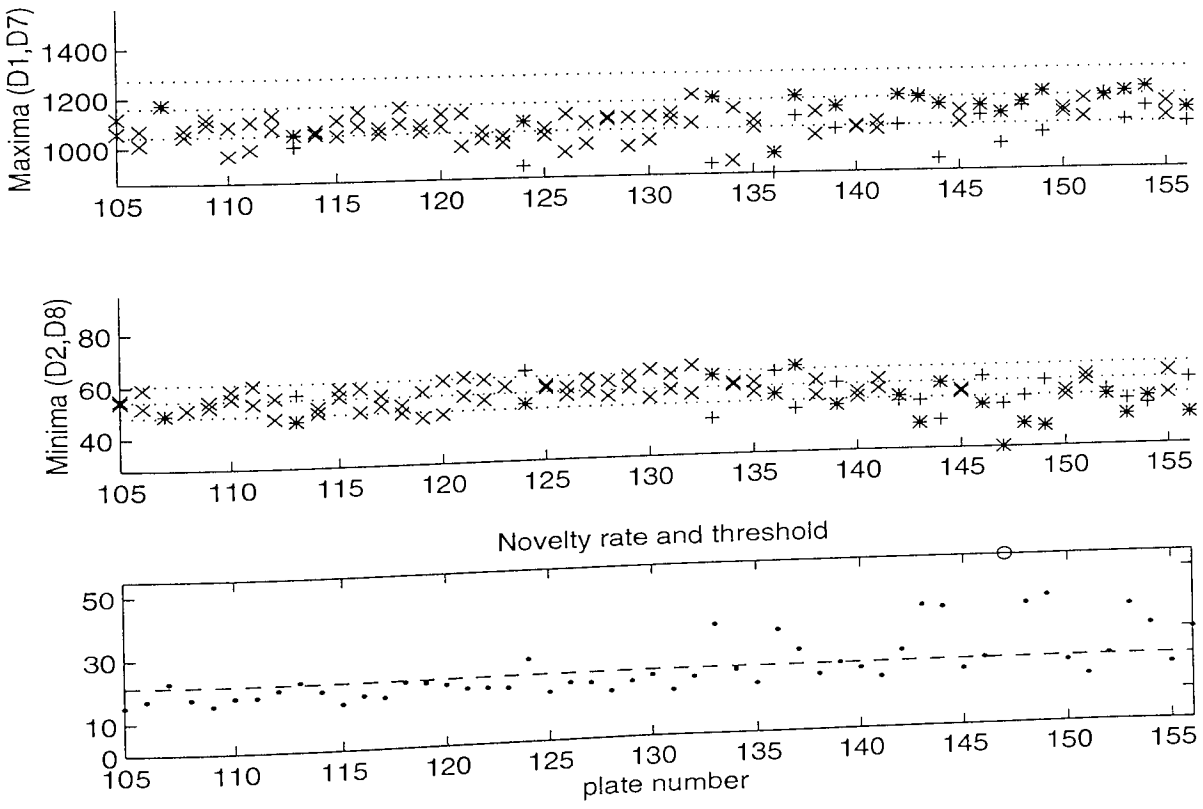


Figure 4.3: **Novelty detection on HTS screen:** plates 105 to 156.

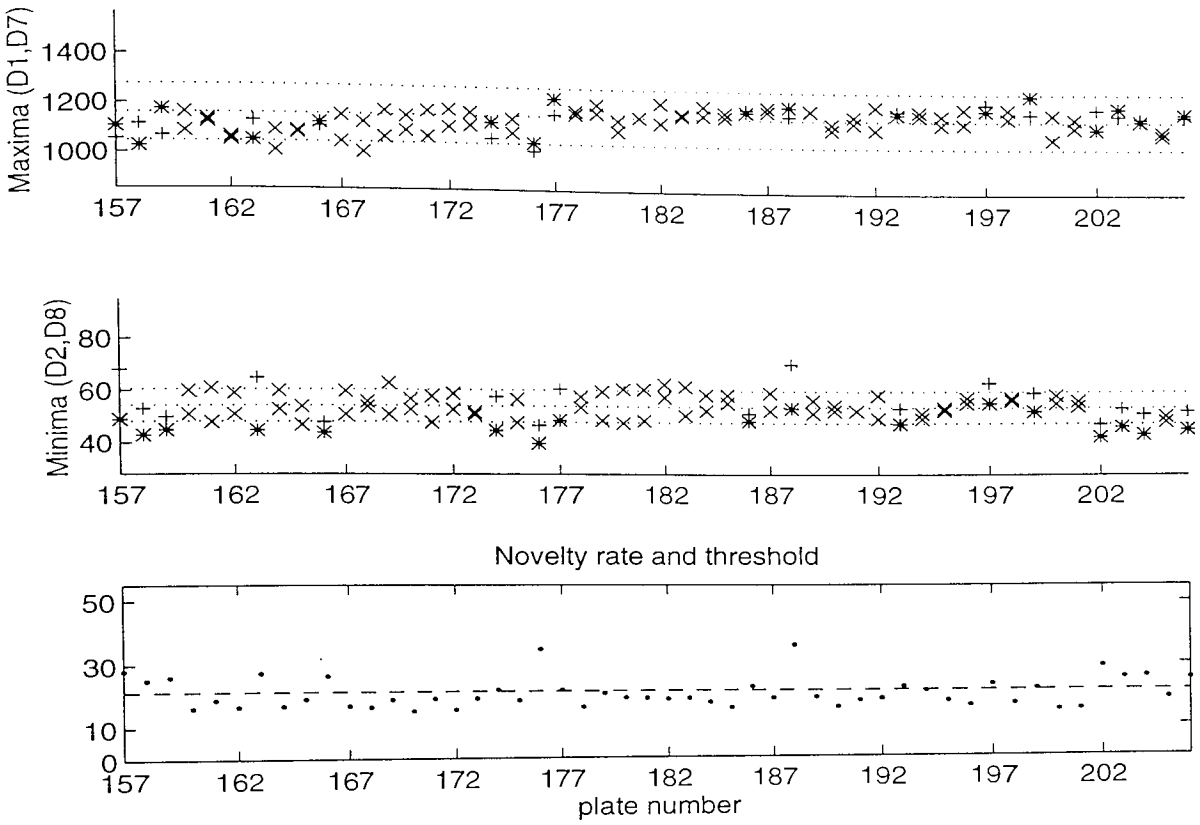


Figure 4.4: **Novelty detection on HTS screen:** plates 157 to 206.

4.2 Discussion

The first remark is that the number of plates declared novel is high (35%). Among the first 41 plates (first assay), all but 2 are declared novel. Indeed, the values of the maximum controls of Figure 4.1 are much greater for the plates 1-40 than those of the other plates while the minimum controls have similar values. According to the Appendix A.1.1, these plates belong to the same assay (were screened on the same day). This variation is due to a systematic difference in the experimental procedure; the period between dilution and screening was longer for this assay than for the others of the same screen so that the reaction is more advanced leading to higher maximum control values for those plates. It explains the dissimilarities between the proportions of rejected plates per assay shown in Table 4.1. If the first assay is omitted, the proportion of rejected plates falls to 19%.

In order to illustrate the fact that a combination of values can be rejected whereas the control values may be acceptable separately, we can compare the plate 91 with the plates 89 and 85 of Figure 4.2. The maximum controls of the plate 91 are similar to those of the plate 89 so are the minimum controls of the plate 91 to those of the plate 85; it is the combination of the four values of the plate 91 which is unusual.

It can be noticed that in Figure 4.1-4.4, some plates may have been rejected although they seem to be similar to accepted ones. In Figure 4.3, for example, the plates 113 and 139 have comparable control values whereas the latter is accepted and the former rejected. The third plot in Figure 4.3 explains this singularity. The two plates have similar novelty values close to the threshold; the first one happens to be above the threshold and the second below. The fact remains that in every method making use of a threshold, such borderline cases occur systematically. Because the software is intended to highlight the control values which are unusual in order to help the operator make a decision, these points around the threshold should not be considered as problematic.

4.3 Adaptive Mixture Model for novelty detection

In view of the difficulty of determining the number of components of a mixture model, an adaptive algorithm for Gaussian mixtures was tested. Detail can be found in [RT94]. The method is based on a stochastic estimation of the parameters of the Gaussian together with a growth criterion for the number of components based on the minimum Mahalanobis distance[4]. Once the training completed, the novelty detection is based on the same criterion.

[4]For the Gaussian density of mean μ and covariance matrix Σ considered in Section 3.1.1, the quantity $\Delta^2 = (\mathbf{x} - \mu)^T \Sigma^{-1} (\mathbf{x} - \mu)$ is called the Mahalanobis distance from $\mathbf{x}$ to μ.

4.3.1 Training procedure

The algorithm uses "reinforcement learning[5]" to maximise the log-likelihood $\sum_{i=1}^{N} \log p(\mathbf{x}_i)$ over all $\mathbf{x}_i$ in the training set. The iterative procedure is defined as follows:

$$\boldsymbol{\mu}_{j,t+1} = \frac{\boldsymbol{\mu}_{j,t} + \alpha_t[P(j|\mathbf{x}_t)\mathbf{x}_t - \boldsymbol{\mu}_{j,t}]}{(1-\alpha_t) + \alpha_t P(j|\mathbf{x}_t)} \quad , \tag{4.1}$$

$$\boldsymbol{\Sigma}_{j,t+1} = \frac{\boldsymbol{\Sigma}_{j,t} + \alpha_t[P(j|\mathbf{x}_t)(\mathbf{x}_t - \boldsymbol{\mu}_{j,t})(\mathbf{x}_t - \boldsymbol{\mu}_{j,t})^T - \boldsymbol{\Sigma}_{j,t}]}{(1-\alpha_t) + \alpha_t P(j|\mathbf{x}_t)}, \; i = 1, \ldots, M \quad , \tag{4.2}$$

where $\mathbf{x}_t$ is a vector randomly chosen in $\mathcal{T}$ and α_t a learning coefficient. The proof by a gradient descent that equations (4.1) and (4.2) converge to a minimum of the error (3.4) can be found in [RT94]. Similarly to the EM algorithm, there is no real restriction on the type of the covariance matrix $\boldsymbol{\Sigma}$. The method was implemented with a full covariance matrix.

The main characteristic of the method relies on its using the *same* threshold noted ϵ_{max} for training and novelty detection, representing the maximum value of a training growth threshold ϵ_t.

In Section 3.3, the number of components of the basis was fixed and determined empirically by cross validation. The basis of the 'Adaptive Mixture Model' to the contrary grows dynamically. At a given time t during the training, if the corresponding $\mathbf{x}_t$ is not properly represented by the model (*i.e.* is 'novel' for the model) a new function is added to the basis (Figure 4.5).

The test value for growth is defined as the greatest activation within the network:

$$\lambda(\mathbf{x}_t) = \max\left\{\Psi(\mathbf{x}_t; \boldsymbol{\mu}_{j,t+1}, \boldsymbol{\Sigma}_{j,t+1}), \; j = 1, \ldots, M\right\} \; , \; t \geq 1 \tag{4.3}$$

where $\Psi(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})^T\right]$. The mixture model grows by one Gaussian according to the criterion:

[5] *Reinforcement learning* is also sometimes called 'learning with a critic'. It describes a learning procedure which gives a feedback from the environment saying whether the result is right or wrong. This definition, which is the common acceptance of reinforcement learning, differs from the present use. In the case of the adaptive mixture model it is characterised according to [RT94] by the learning parameter α_t in equations (4.1) and (4.2) which "cools" this response.

$$\lambda(\mathbf{x}_t) \begin{cases} \leq \epsilon_t & \rightarrow growth \ . \\ > \epsilon_t & \rightarrow no\ growth \ . \end{cases} \tag{4.4}$$

The growth criterion (4.4) can be reformulated using (4.3) as the smallest Mahalanobis distance between $\mathbf{x}$ and the elements of the basis:

$$\min\left\{(\mathbf{x}_t - \boldsymbol{\mu}_j)\boldsymbol{\Sigma}_j^{-1}(\mathbf{x}_t - \boldsymbol{\mu}_j)^T,\ j = 1, \ldots, M\right\} \geq Q_t \tag{4.5}$$

with $Q_t = 2\ln(1/\epsilon_t)$. In other words, the basis grows if the current vector $\mathbf{x}_t$ is too far, in the Mahalanobis distance sense, from the nearest centre.

The growth threshold $0 \leq \epsilon_t \leq \epsilon_{max}$ is initially set as $\epsilon_0 = 0$ and monotonically increases with time[6] according to:

$$\epsilon_t = \min\left\{\epsilon_{max}, \epsilon_{max}\frac{t}{\tau_\epsilon}\right\} \ , \tag{4.6}$$

where τ_ϵ is an integer to be chosen between 1 and the number of iterations of the algorithm (in most cases we used $\tau_\epsilon = N$ where N is the size of the sample). The novelty criterion for a vector $\mathbf{x}$ of the test set becomes:

$$\lambda(\mathbf{x}) \begin{cases} \leq \epsilon_{max} & \rightarrow \mathbf{x}\ is\ novel \ , \\ > \epsilon_{max} & \rightarrow \mathbf{x}\ is\ not\ novel \ . \end{cases} \tag{4.7}$$

4.3.2 Network growth

If $\lambda(\mathbf{x}_t) \leq \epsilon_t$, the new centre and covariance matrix are defined as:

$$\boldsymbol{\mu}_{M+1} = \mathbf{x}_t \ , \tag{4.8}$$

$$(\boldsymbol{\Sigma}_{M+1})_{kl} = \delta_{kl}\frac{1}{d}Tr[C],\ \ l, k = 1, \ldots, d \ , \tag{4.9}$$

where $C = (\boldsymbol{\mu}_{M+1} - \boldsymbol{\mu}_l)\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_{M+1} - \boldsymbol{\mu}_l)^T$ and $\delta_{kl} = 1$ if $k = l$ and 0 otherwise. The priors are all set to $P_t(j) = \frac{1}{M+1}$, $j = 1, \ldots, M+1$.

[6]The initial value ϵ_0 is chosen only for consistency with the increase of ϵ_t. The condition (4.4) does not apply for $t = 0$ where the basis is empty, and must grow at the first step of the algorithm.

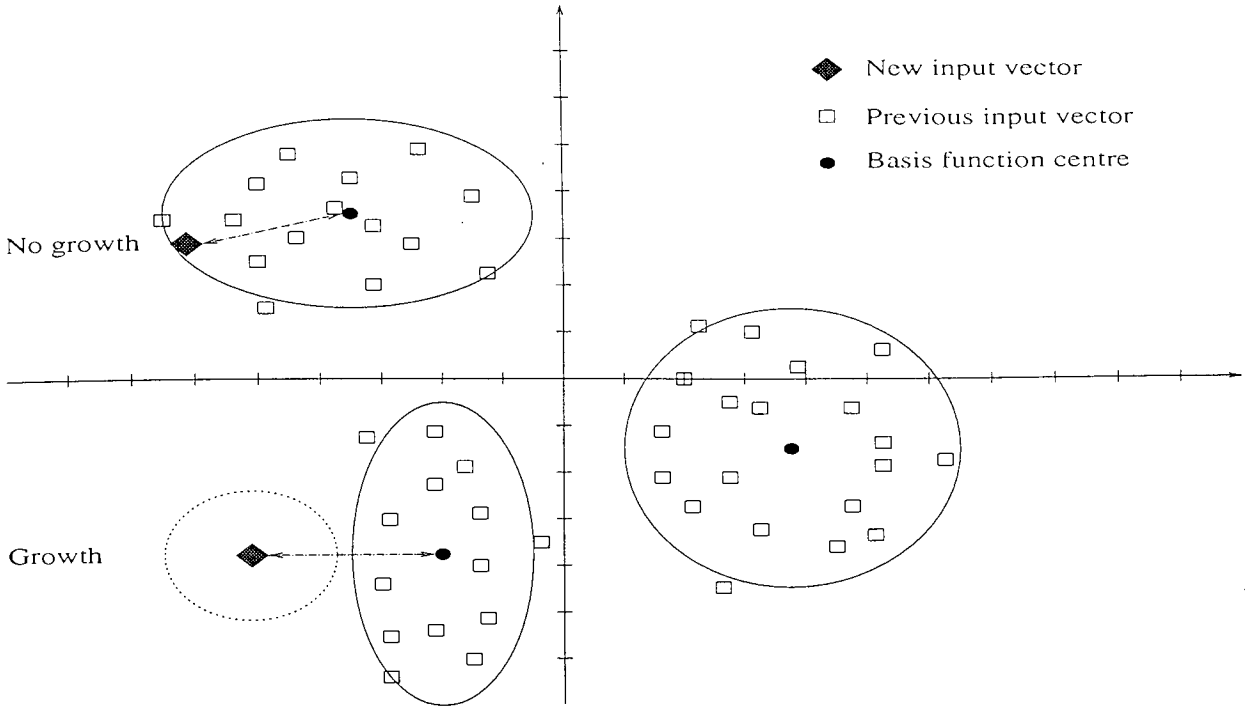


Figure 4.5: Network growth based on Mahalanobis distance for a 2-dimensional data space

It can be noticed that the priors $P_t(j)$ remain the same throughout the training. In that respect, the learning procedure does not provide an 'optimal' choice for the number of Gaussians for describing the probability density (we would expect to update the $P_t(j)$ as well as $\boldsymbol{\mu}_j$ and $\boldsymbol{\Sigma}_j$). However, the 'right' (possibly minimum) number of Gaussians in the basis is not crucial for novelty detection.

4.3.3 Local cooling

The problem of the lack of adaptation for the recently added basis function can be dealt with by allowing both the adaptation gain and the time to be vectors ($\alpha = [\alpha_1, \ldots, \alpha_M]^T$ and $t = [t_1, \ldots, t_M]^T$) and:

$$\alpha_{t_j} = \frac{\alpha_0}{t_j + \tau_\alpha}, j = 1 \ldots M \quad . \tag{4.10}$$

The parameter α_t which appears in equation (4.1) and (4.2) is analogous to a learning rate. It governs the influence of the vector $\mathbf{x}_t$ in the updating of $\boldsymbol{\mu}_t$ $\boldsymbol{\Sigma}_t$. α_0/τ_α gives the

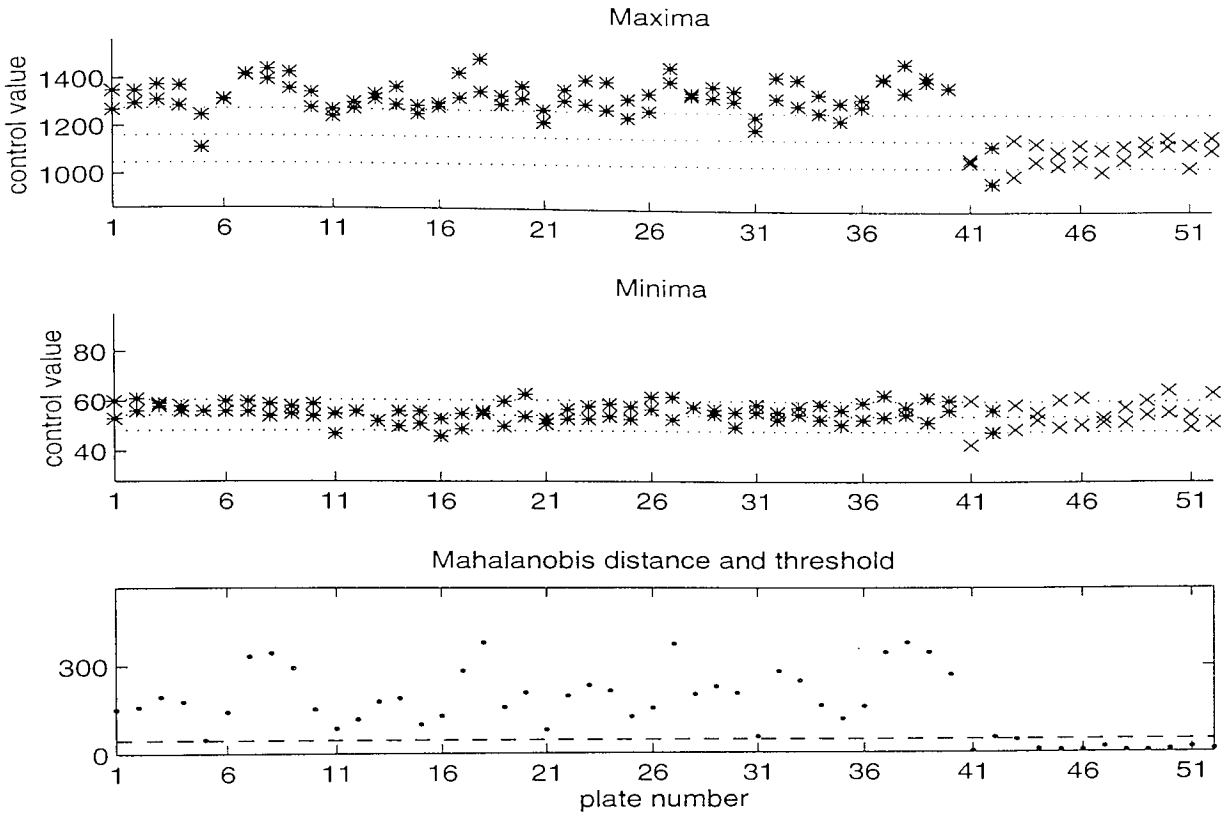


Figure 4.6: **Novelty detection on HTS screen:** plates 1 to 52.

initial value of α_t (for the first updating after addition of a function to the basis). With a M-dimensional time t, the parameter α_t is the same for the updating of two distinct functions of the basis. Thus, the updating equations (4.1) and (4.2) are consistent between the first and the last added basis function.

4.3.4 Application

The Adaptive Mixture Model algorithm is applied on Screen 2 for novelty detection with the parameters: $\epsilon_{max} = 10^{-10}$, $\tau_\epsilon = N$ (so that ϵ_t reaches its maximum value after one iteration through the training data), $\alpha_0 = 0.7$ and $\tau_\alpha = 1$. The results of the detection are shown on Figure 4.6-4.9.

53 plates are declared novel. The training was performed in 4min 20s on a Sparc (Sun4d) with the parameters and leads to a 9 function basis.

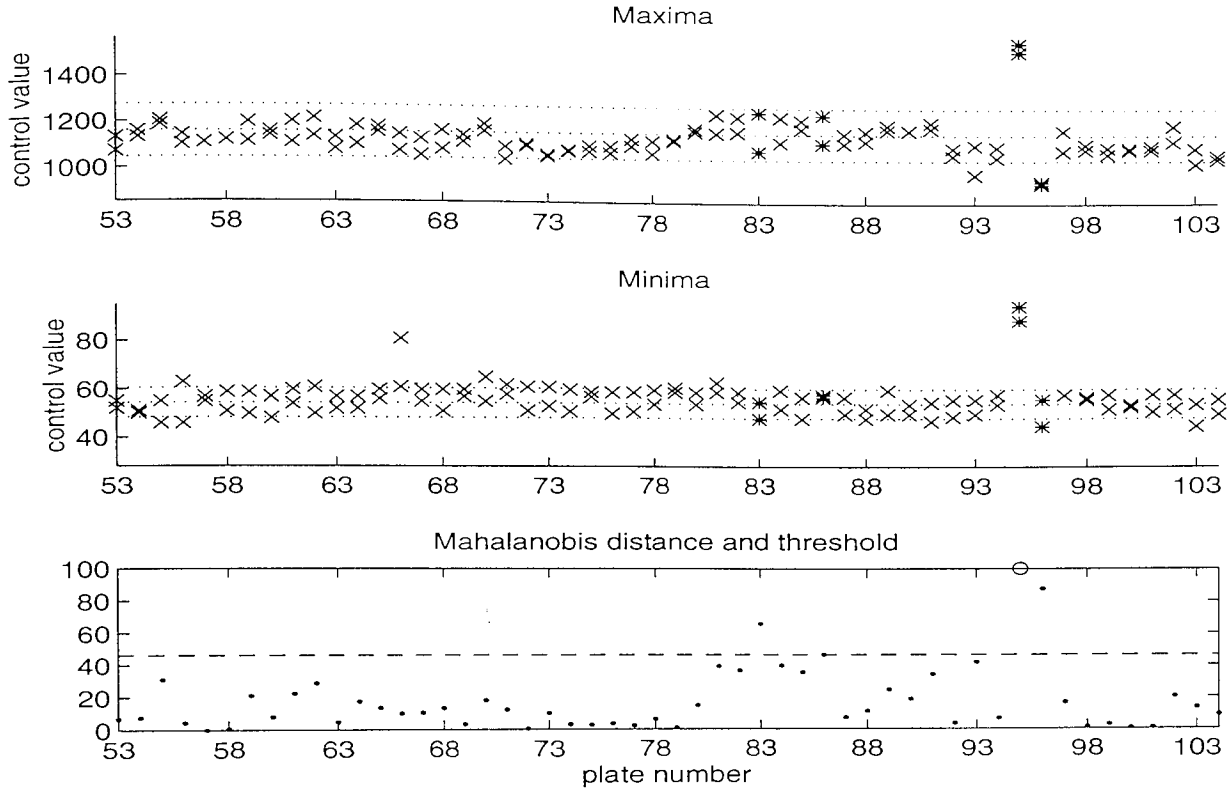


Figure 4.7: **Novelty detection on HTS screen:** plates 53 to 104.

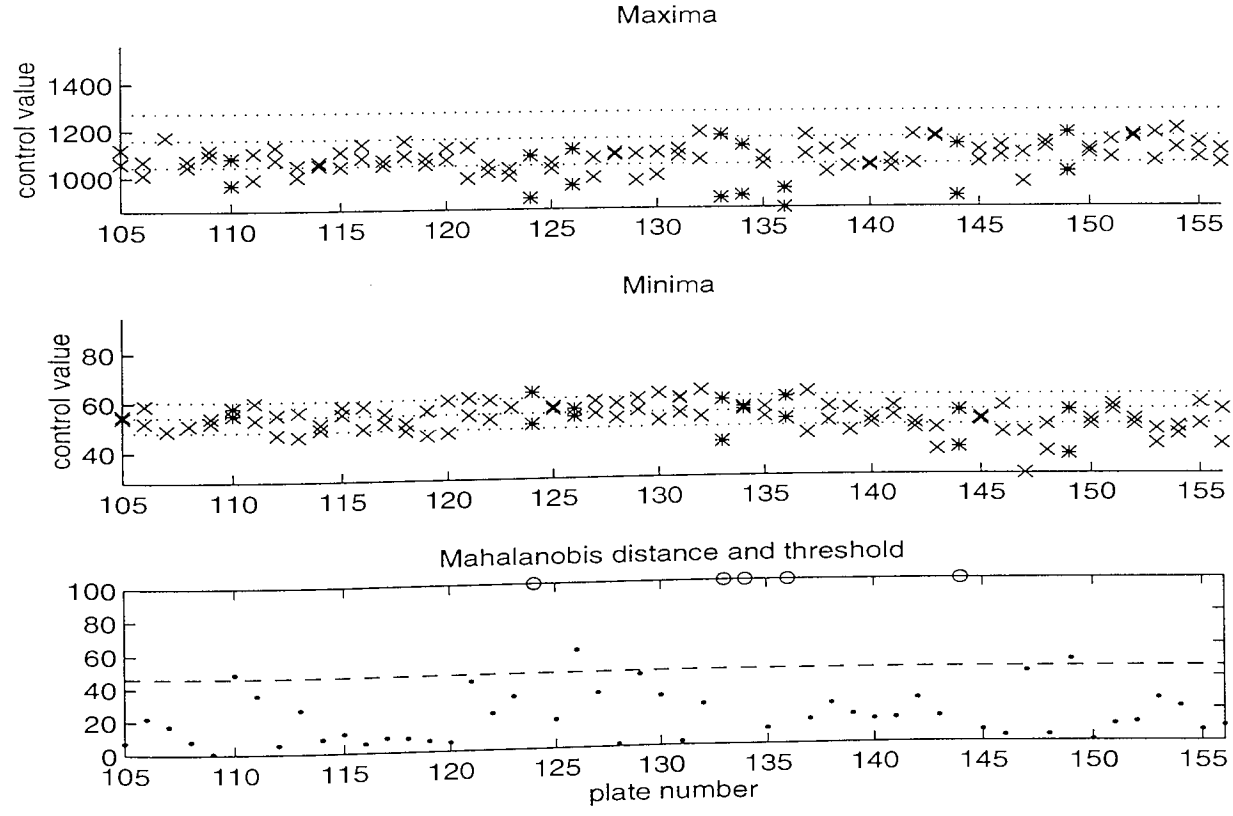


Figure 4.8: **Novelty detection on HTS screen:** plates 105 to 156.

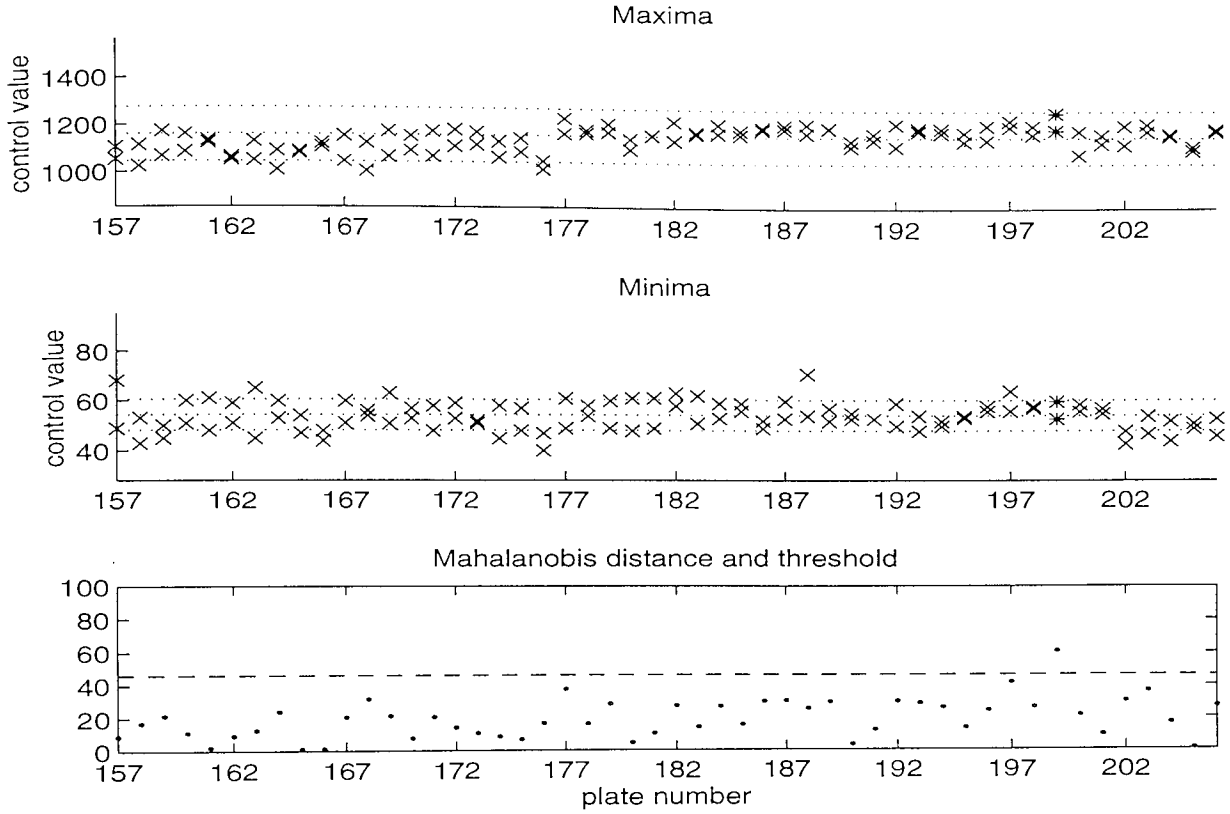
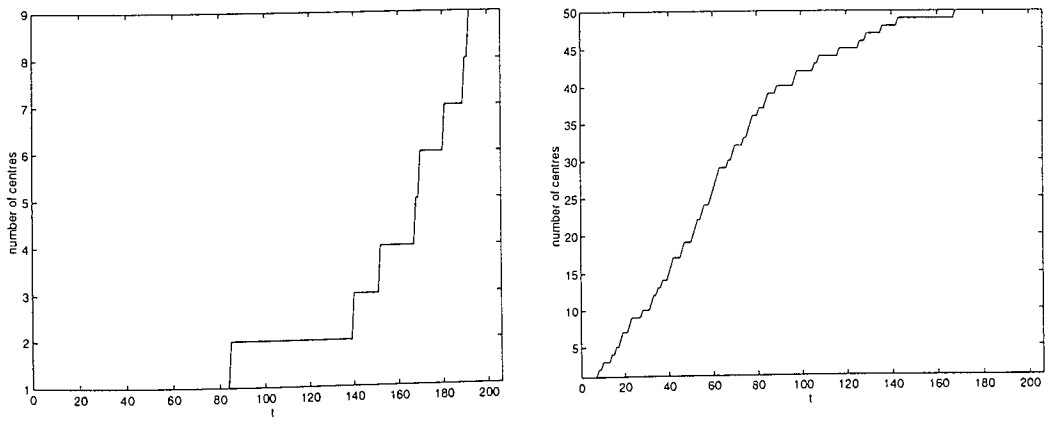


Figure 4.9: **Novelty detection on HTS screen:** plates 157 to 206.



(a) $\epsilon_{max} = 10^{-10}$ (b) $\epsilon_{max} = 10^{-2}$

Figure 4.10: Basis growth during the training

4.3.5 Discussion

First of all, the tests show that the procedure tends to over-fit the data as underlined in Figure 4.6-4.9 (to be compared with the results of Figure 4.1-4.4). The model complexity determination in Section 3.5.1 showed that a 2 Gaussian mixture model gives a good representation of the data whereas a more complex model does not provide significant improvement regarding the negative log-likelihood error. However, it should be emphasised that the Adaptive Mixture Model do not aim strictly at a good representation of the data in terms of probability but to the model which rejects the smallest number of point in the training set.

The method is slow but some reservations have to be brought since the algorithm in our case is implemented using full covariance matrices (in [RT94] the covariance matrices are diagonal). In addition, the fixed value $1/M$ for the priors induces a rapid growth of the basis and the evolution of the basis is highly dependent on the ordering of the training points. As a result the calculation becomes computationally intensive[7]. The cost of the training is certainly more due to the size of the basis rather than the actual complexity of the algorithm: the evaluation of the likelihood of a point with respect to a mixture model is more expensive for a 50 function basis than for a 2 function basis.

Secondly, one may argue that the stochastic approach may be justified in [RT94] since the case study concerns sleep phases whereas so far as the HTS screen is concerned the order of the plates is not important. It should be noticed that in the case study of [RT94] the learning procedure of the adaptive mixture is applied 44 times on the training data; even though the phenomenon is 'naturally' time-dependent, the learning procedure takes actually no account of this property. Moreover, experiments show that for the HTS data, the number of rejected points can vary significantly (they

[7] A version of the adaptive algorithm where the priors are updated as in the EM algorithm by the updating relation $P(j)_{t+1} = P(j)_t + \alpha_t(P(j|\mathbf{x}_t) - P(j)_t)$ (see [RT94]) was also tested. This diminishes significantly the growth (by a third in average); some priors may tend to 0 (and the corresponding Gaussians *de facto* useless in the model). A possibility to overcome this difficulty could be to withdraw the corresponding Gaussians of the model according to a fixed prior threshold but require the choice a such a value which is precisely what we wanted to avoid in this Section.

can double) with the ordering of the training set.

Another drawback of fixed mixing coefficients $P(j)$ is that such an algorithm does not reach for a optimal value for the number of basis functions to properly describe the distribution of the training set and as a consequence can not replace the methods mentioned in Section 3.2. Though not a problem in principle so far as novelty detection is concerned, the growing number of basis functions results in expensive and useless calculations which should be avoided in practice.

Finally, the main drawback is precisely that the novelty threshold is used for training and test since it constitutes the growth criterion. In other words, if the novelty threshold is changed during the test procedure, the detection is not consistent any more (if the threshold is lowered in the test procedure, the novelty detection model is more susceptible to novel points of the test set than to points in training set). In the prospect of an automated scheme such a constraint is not acceptable. In the case of the HTS data, the threshold ϵ_{max} must be very small to prevent the basis from an excessively rapid growth which would induce an intractable computation. Figure 4.10(a) shows the network growth during the training . Figure 4.10(b) underlines the problem of the choice of a suitable parameter for ϵ_{max}. When set to the value chosen in [RT94], the basis grows to 50 functions; the training part takes 13min 21s on the same machine.

With regard to the practical restrictions of Section 1.3.1 we should stress that the choice of the parameter ϵ_{max} as the threshold for growth may prove delicate for the operator. One might find much easier indeed to set a threshold on a probability density as in Section 3.4 than to choose a suitable value for a growth based on a Mahalanobis distance. Besides, the selection of a threshold *before* the training is not necessarily an asset in the context of a quality control (we ignore *a priori* the 'quality' of a screen). In view of the same restrictions, determining the threshold *a posteriori* according to

the probability density leaves a certain control on the novelty detection procedure to the operator and prevents the 'black box trauma' mentioned in the introduction.

4.4 From plates to wells

The material covered in the first part of this chapter has concentrated on the four control wells to determine the unusual plates. The joint distribution and a suitable novelty threshold provide an effective means to determine the novel plates in a HTS screen. In practice the operator concentrates on well values rather than plates to spot abnormal control values. The main focus in the next part of this text is on locating these unusual control wells using conditional densities. The technique is applied on Screen 2.

4.4.1 Conditional densities

The problem can be described more generally. Suppose two continuous random variables by X, Y drawn from two distinct distributions p_X, p_Y. How can values taken by X and Y be compared in terms of probability? As far as HTS is concerned in particular, it is important to determine which component(s) of the vector $(x_1, \ldots, x_d)$ is the most unlikely (*i.e.* which wells of the controls is the most 'abnormal').

In this case, we choose the distribution function to order the components of the d-tuples regarding their 'novelty'. To compare these components we need the distribution of each component X_i conditioned on the other $d-1$ components $p(X_i|X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_d)$. Since the model is a mixture of Gaussians of diagonal covariance matrix (the X_i are independent) these conditional distributions are the marginals whose density is given by $p(y) = \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\{-\frac{(y-\mu_i)^2}{2\sigma_i^2}\}$ which is the projection of $p(x)$ on the axis i.

The measure of the 'contribution' of the component x_i to the rejection of the vector

$\mathbf{x} = (x_1, \ldots x_i \ldots, x_d)$ is given by:

$$\lambda_{\mathbf{x}}(x_i) = \min\{F_{X_i}(x_i), 1 - F_{X_i}(x_i)\} \tag{4.11}$$

where:

$$\begin{aligned} F_{X_i}(x_i) = P(X_i < x_i) &= \int_{-\infty}^{\infty} \ldots \int_{-\infty}^{x_i} \ldots \int_{-\infty}^{\infty} p(y_1, \ldots y_d) dy_1 \ldots dy_d \\ &= \int_{-\infty}^{\infty} p_{X_1}(y) dy \ldots \int_{-\infty}^{x_i} p_{X_i}(y) dy \int_{-\infty}^{\infty} p_{X_d}(y) dy \\ &= \int_{-\infty}^{x_i} p_{X_i}(y) dy \quad . \end{aligned} \tag{4.12}$$

The closer to zero $\lambda_{\mathbf{x}}(x_i)$ is, the 'more novel' the i_{th} well value.

This measure is justified empirically since we are interested in the extreme values of the distribution function, close to 0 and 1. This would not provide a proper means of comparison for values between the centres in the case of a distribution including two distinct clusters for example.

The computation of the conditional densities of a d-dimensional joint density is straightforward: the properties of multivariate normal distributions are easily extended to Gaussian mixture models (the conditional probabilities remain Gaussian). This calculation would not have been as simple with a full covariance matrix (*cf* Section 3.5.2).

4.4.2 Results

The measure (4.11) is applied for on Screen 2 with:

$$F_{X_i}(x_i) = \int_{-\infty}^{x_i} \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\{-\frac{(y-\mu_i)^2}{2\sigma_i^2}\} dy, \; i = 1 \ldots 4 \quad .$$

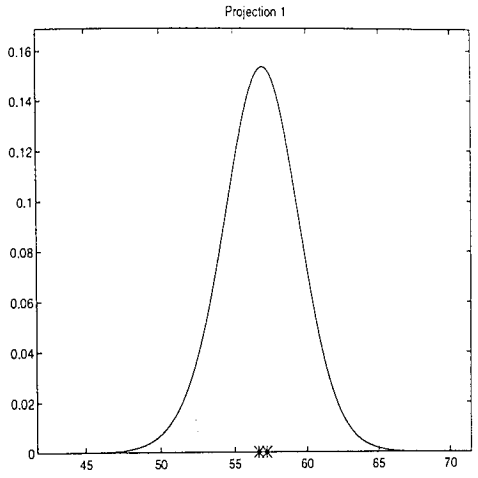
The results are shown in Table 4.2 (The first 40 plates are omitted).

Table 4.3 summarises the results in terms of number of wells rejected per plate for a 5% rejection region (the first 40 plates are omitted). The conditional densities are shown in Figure 4.11.

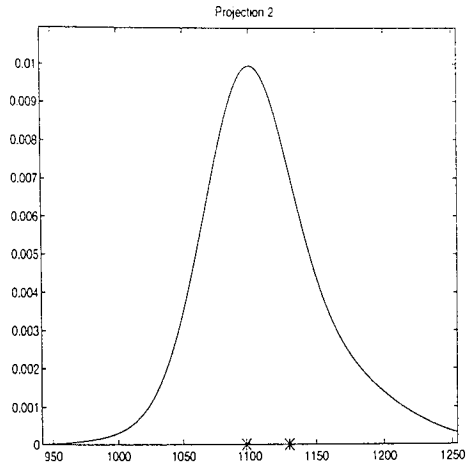
The analysis of such a measure does not provide new information about the variation of the controls but similarly as the log-likelihood for the plates, such a measure

n	loglike	m* min1	M* Max1	m+ min2	M+ Max2
41	2.56e+01	m* 1.3e-06	N M* 1.7e-01	m+ 6.5e-02	M+ 3.5e-01
55	2.92e+01	m* 1.0e-04	N M* 4.1e-02	m+ 2.2e-01	M+ 8.1e-02
66	4.98e+01	m* 2.6e-08	N M* 1.8e-01	m+ 6.5e-02	M+ 3.1e-01
83	2.76e+01	m* 1.1e-03	N M* 7.6e-03	N m+ 2.2e-01	M+ 3.5e-01
85	2.43e+01	m* 1.1e-03	N M* 2.6e-02	m+ 4.7e-01	M+ 9.6e-02
91	2.73e+01	m* 3.5e-04	N M* 3.2e-02	m+ 2.2e-01	M+ 8.1e-02
95	3.51e+02	m* 2.6e-08	N M* 1.5e-04	N m+ 1.1e-13	N M+ 8.6e-12 N
96	3.95e+01	m* 2.6e-05	N M* 4.3e-04	N m+ 3.3e-01	M+ 2.7e-05 N
107	2.63e+01	m* 3.0e-03	N M* 1.0e-01	m+ 4.9e-03	N M+ 1.3e-01
124	3.08e+01	m* 7.7e-03	N M* 3.1e-01	m+ 1.1e-02	N M+ 1.9e-07 N
133	4.07e+01	m* 2.1e-01	M* 1.1e-01	m+ 7.2e-07	N M+ 9.0e-08 N
134	2.55e+01	m* 3.7e-01	M* 3.3e-01	m+ 2.2e-01	M+ 4.0e-07 N
136	4.86e+01	m* 1.8e-02	N M* 1.2e-04	N m+ 1.3e-01	M+ 5.0e-11 N
137	2.47e+01	m* 2.6e-02	M* 1.1e-01	m+ 5.9e-05	N M+ 3.9e-01
142	2.44e+01	m* 3.0e-03	N M* 1.1e-01	m+ 6.5e-04	N M+ 1.1e-01
143	4.30e+01	m* 9.7e-11	N M* 1.5e-01	m+ 6.5e-04	N M+ 1.5e-01
144	4.29e+01	m* 1.4e-01	M* 3.1e-01	m+ 3.5e-09	N M+ 1.9e-07 N
147	7.25e+01	m* 1.3e-22	N M* 3.2e-01	m+ 5.9e-05	N M+ 2.8e-04 N
148	4.03e+01	m* 1.0e-11	N M* 2.7e-01	m+ 1.9e-03	N M+ 4.1e-01
149	4.30e+01	m* 1.0e-12	N M* 9.3e-02	m+ 1.4e-01	M+ 1.3e-02 N
152	2.46e+01	m* 1.1e-03	N M* 1.6e-01	m+ 1.1e-02	N M+ 1.6e-01
153	3.80e+01	m* 6.0e-09	N M* 1.0e-01	m+ 2.1e-04	N M+ 1.5e-01
154	3.17e+01	m* 3.5e-04	N M* 6.3e-02	m+ 1.5e-05	N M+ 4.6e-01
156	2.99e+01	m* 6.0e-09	N M* 4.6e-01	m+ 1.4e-01	M+ 1.0e-01
157	2.44e+01	m* 3.0e-03	N M* 5.0e-01	m+ 3.2e-05	N M+ 1.6e-01
158	2.77e+01	m* 1.3e-06	N M* 2.2e-02	N m+ 8.5e-02	M+ 4.3e-01
159	2.62e+01	m* 2.6e-05	N M* 9.8e-02	m+ 1.1e-02	N M+ 2.3e-01
163	2.76e+01	m* 2.6e-05	N M* 7.7e-02	m+ 1.3e-03	N M+ 3.2e-01
166	2.69e+01	m* 6.1e-06	N M* 3.4e-01	m+ 1.9e-03	N M+ 4.6e-01
176	3.65e+01	m* 6.0e-09	N M* 7.3e-02	m+ 6.5e-04	N M+ 2.4e-02 N
186	2.67e+01	m* 3.0e-03	N M* 6.3e-02	m+ 4.7e-02	M+ 4.9e-02
188	3.63e+01	m* 1.4e-01	M* 3.1e-02	m+ 3.4e-07	N M+ 1.2e-01
193	2.60e+01	m* 1.1e-03	N M* 6.8e-02	m+ 1.4e-01	M+ 4.9e-02
194	2.49e+01	m* 7.7e-03	N M* 8.2e-02	m+ 4.7e-02	M+ 5.3e-02
202	3.59e+01	m* 1.3e-06	N M* 2.9e-01	m+ 1.9e-03	N M+ 2.9e-02
203	2.84e+01	m* 3.5e-04	N M* 2.7e-02	m+ 1.4e-01	M+ 8.4e-02
204	2.97e+01	m* 6.1e-06	N M* 1.2e-01	m+ 4.7e-02	M+ 1.2e-01
206	2.99e+01	m* 1.0e-04	N M* 5.5e-02	m+ 8.5e-02	M+ 8.9e-02

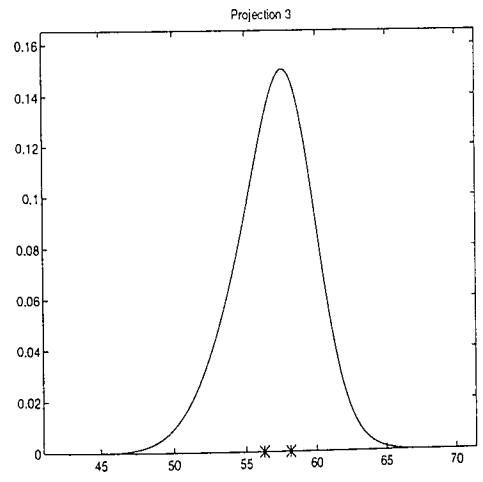
Table 4.2: Results of contribution measure on Screen 2 (N denotes wells whose measure of novelty given by the equation (4.11), is below 5%).



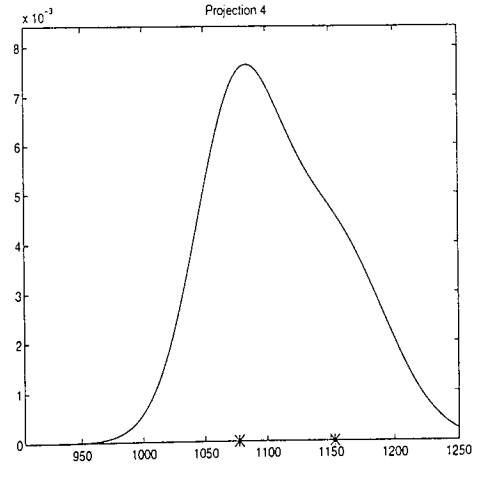
(a) Minimum controls (D2) density



(b) Maximum controls (D1) density



(c) Minimum controls (D8) density



(d) Maximum controls (D7) density

Figure 4.11: Conditional distributions (the stars '*' denote the centres of the basis functions)

number of wells	number of plates	proportion
1	13	17 %
2	53	68 %
3	11	14 %
4	1	1 %

Table 4.3: Repartition of rejected wells for 5 %

enables to quantify the novelty of a well. For instance, a visual inspection of the rejected plate 156 in Figure 4.3 would suggest an analysis such as "the min * (D2) is suspicious whereas the others seem correct"; the conditional densities provide an objective measure with the corresponding controls `m*` flagged '`N`'. In a word, the measure enables to order the wells of the rejected plates in terms of novelty.

If the method was used automatically to reject the unusual values for the standard wells, 156 wells out of 824 would be rejected (18.9 %).

4.5 The standard controls

This section undertakes the final stage of this study: the inclusion of the standard controls in the novelty detection.

As the HTS data come mainly from enzyme assays and research of potential inhibitors or activators, the first part of this section provides a brief summary of the behaviour of such assays. The problems for the HTS controls that arise from this behaviour are described and in particular those of the standard controls. We finally discuss the possibilities to solve them and the results obtained.

4.5.1 Variation of the controls

In Section 4.2, we mentioned the significant difference between the maximum controls of the first assay (first 40 plates) and the rest of Screen 2. This variation was attributed

to biological variance; in other words, the experimental conditions of the two assays were different. For instance, the concentrations of substrate, the time of reaction or the temperature may vary from one assay to the other which induce a systematic difference between assay measures.

The enzyme assays

An enzyme is a biological catalyst. It enhances the rate of a chemical reaction. The activity of an enzyme can be measured by the rate of product formation during the reaction. Figure 4.12 shows schematically a typical curve obtained when the time-course of product formation (the result of the chemical reaction enhanced by the enzyme) is determined. The rate of the reaction at a given time is given by the derivative of this function. Initially, this rate is constant (the time-course is linear) but after a time which depends on the reaction, the rate decreases. This decline may be due to various reasons such as a fall of the substrate concentration, the approach of the reaction equilibrium or more generally some changes in the assay conditions.

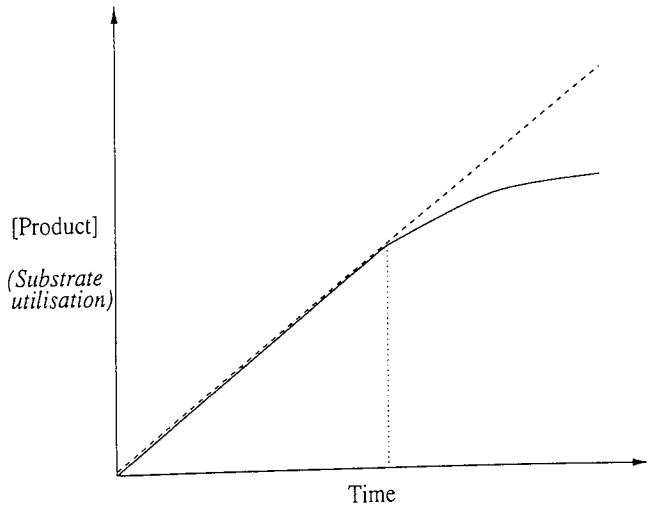
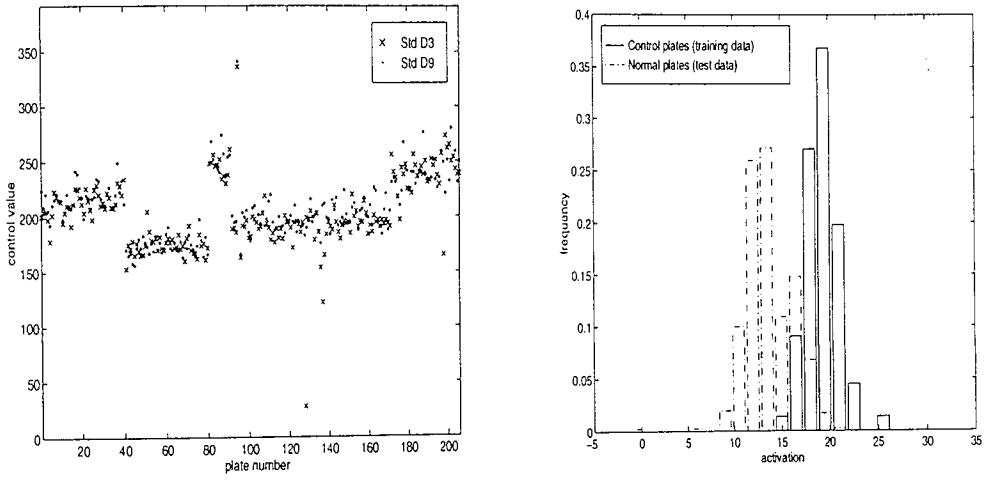


Figure 4.12: Progress curve of an enzyme-catalysed reaction

Since the HTS assay should not be run under these limiting conditions (the non-linear part of the curve on Figure 4.12), the controls (the standards and the maximum controls) which reveal such conditions should be flagged as novel.

The case of the standards

Contrary to the maximum controls the standard controls variation does not depend only on the factors mentioned above. Ideally, the activation (4.13) should remain the same throughout the assay, whatever the time, the temperature or the substrate concentration. However, for some reactions, the incubation time may be too long for the reaction to remain on the proportional area of the reaction curve on Figure 4.12 so that the assertion of a constant activation is no longer true. This variation is clear on the standard controls of Screen 2 (Figure 4.13(a)).



(a) Standard controls

(b) Activations

Figure 4.13: Standard controls and activations (Screen 2)

Five distinct regions emerge from this graph (1-40, 41-80, 81-91, 92-171, 172-206) which correspond to different assays (see Appendix A.1.1) denoting an important variation in the experimental conditions.

How to deal with variation?

It has been suggested that the 'non-stationarity' of the standard controls can be removed by differencing with the maximum controls (respectively the minimum controls) because they share the same evolution during the linear part in Figure 4.12. Thus we form the activation:

$$Activation = \frac{Std - min}{Max - min} \quad . \tag{4.13}$$

Similarly, we define the inhibition of a standard control:

$$Inhibition = \frac{Max - Std}{Max - min} \quad . \tag{4.14}$$

Without day to day assay variation, this ratio should remain 'constant'. Otherwise it denotes a difference in the experimental procedure and the corresponding plate should be detected.

4.5.2 Applications

As previously emphasised, the difficulty in analysing the standard controls comes from their high variability. This section presents three ways of learning the density function of the standards: using directly the standards of the control plates, relying on the normal plates or transforming the data of the control plates for learning.

The strong points and drawbacks are described for these three approaches.

First possibility: standards, minima and maxima alike

The first possibility to take into account the standard controls for novelty detection of HTS is similar to that exposed for the minimum and maximum controls. An additional plate of standard controls can be used as reference to learn the distribution of the raw data. The problem with such an approach is illustrated in Figure 4.14.

The modelled distribution can not be expected to be a good description of the control values because it does not take into consideration the variation of the controls for the whole screen. Indeed, the distribution of the control values on the control plates is significantly different from that of the normal plates as shown in the histograms.

The problem in using the same method is the following. Comparing minima from the control plates to the corresponding controls of the normal plates makes sense: the

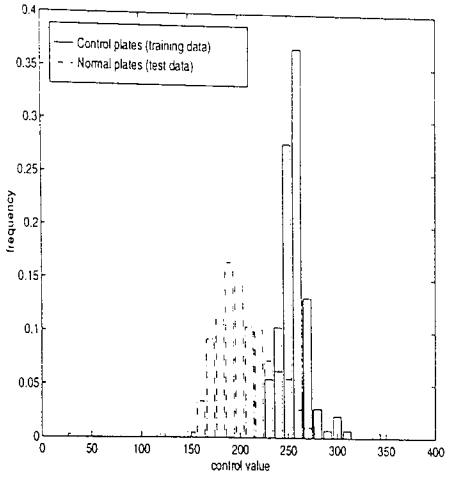


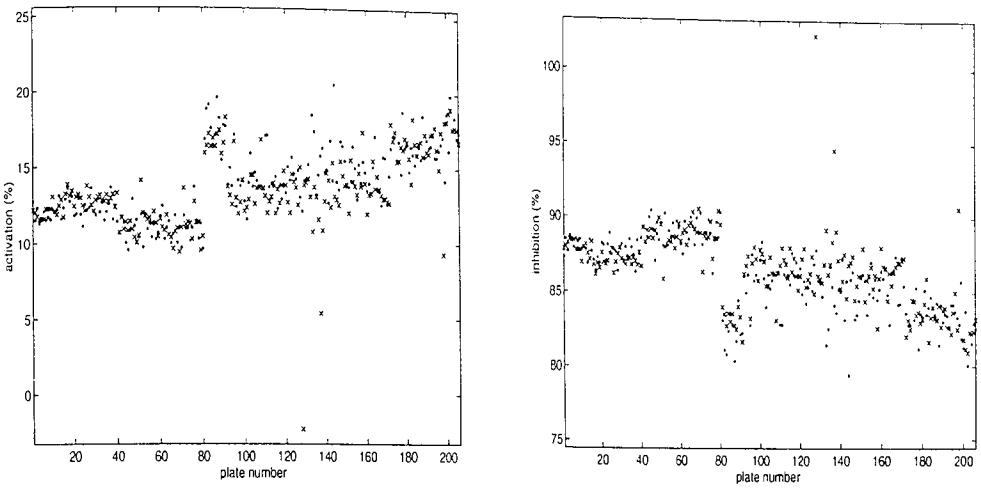
Figure 4.14: Standard controls 1D distribution (Screen 2)

variation is mainly due to handling mistakes (empty wells, double substrate inserted...) and measurement variation. If enough care is taken when screening the control plates, they can provide a 'description of normality' to detect the outliers in the normal plates. The problem with the standards comes partly from the fact that we artificially create 6-tuples from the control plates for the sake of the learning procedure. On the normal plates, the standard and maximum controls are significantly correlated (Section 2.3). As a consequence, if, a standard has an unusually high or low value depending on the experimental conditions, one might suggest that this variation can be removed by differencing with the maxima (inhibition) since maxima and standard controls are very likely to vary in the same way (see Section 2.3). On the control plates though, the maximum and the standard controls can not be expected to be correlated the way they are in the normal plates so that this time variation should be removed by computing the inhibition.

Furthermore, the variation which is due to different days of screening, thus different experimental conditions, is very unlikely to appear on control plates screened on a single day.

If a method similar to the one indicated in Section 3.3 [8] is applied on Screen 2,

[8]Cross validation with 10 iterations; 144 'plates' (6-tuples) constitute the validation set, the training is generated by re-sampling 144 'plates' from a basis of 72 3-tuples $(min, max, standards)$.



(a) Activations (b) Inhibitions

Figure 4.15: Transformed controls (Screen 2)

185 plates are declared novel (198 for the 95^{th} percentile) because of this difference.

The computation of the activation or inhibition does not remove this variation as can be noticed on Figures 4.15(a) and 4.15(b)[9].

An alternative: learning the normal plates directly

The second possibility to deal with the standard controls is to model the distribution of the standards (*via* inhibition) on the *normal* plates.

The main drawback is that potential outliers are present in the training set therefore the model inference may be altered.

If the number of outliers on a given screen is small, the novelty detection does not suffer since their presence will not modify significantly the model. Therefore their probability remains low and they can still be flagged as 'abnormal'. Typically, handling mistakes would be detected if they concerned only a few plates or a few wells. However, if an unusual variation occurs on a larger scale, their probability according to the model

[9]The inhibition (resp. activation) computed for the plate 128 is greater than 100% (resp. smaller than 0%). For this plate the standard control is higher than the maximum control (which most certainly due to a handling mistake).

is higher and this variation in the standard values can not be flagged as abnormal. For example, if a problem occurs with a measurement instrument throughout an assay, it would not be detected.

Finally, from a practical point of view, one of the weaknesses of this procedure is that the quality control of the data cannot start before the whole screen is completed.

If the learning proceeds on the normal plates using two random selections of 103 plates to constitute the training and the validation set (and cross validation with ten runs), *3 plates are declared novel* (20 plates if the novelty threshold is set to the 95^{th} percentile).

Where the standards can make the difference...

A third possibility to deal with the problem of the standards is the following: instead of considering the two standard controls for a given plate, one may consider the absolute value of their difference. This can be justified considering three remarks:

- for a given plate, two standard controls whose values remain on the linear part of the curve in Figure 4.12 do not reveal an 'abnormality'; we can model the tolerated gap between those controls (the 'degree of freedom' on this line of the standard value). This gap can be modelled regardless of the experimental conditions using the control plates;
- the variance of the measures above is already taken into consideration by the maximum and minimum controls therefore the information provided by the standards in this respect is redundant;
- it seems more sensible to consider an unsigned measure of this difference; if a gap D is estimated as normal between the control D3 and D9 ($D3 = D9 + D$), it makes sense to accept the symmetric situation ($D9 = D3 + D$) as acceptable as well.

In terms of quality control and regarding the constraints settled in Section 1.3.1, our replacement of the two standards by their difference in the model is not a problem since these wells are used *only* for the controls contrary to the maximum and minimum controls for which it is important to keep the original values; it is one purpose of the assessment to discard abnormal values of these controls for the activation boundary computation (Section 1.1.2).

One may argue that if a significant variation occurs for the two controls, the difference of the two is not necessarily abnormal. In such a case, that difference would appear on the other controls as well[10], so that the corresponding plate would be flagged as novel all the same. As a consequence, this transformation of the controls should not be considered as a problem but rather as a means of extracting relevant and non redundant information from them.

The results of the novelty detection after pre-processing are shown on Figure 4.16-4.19 (the notations are the same as in Figure 4.1-4.4). The third graph represents the value of the difference between D3 and D9.

62 plates out of 206 are declared novel.

The last column of Table 4.4 shows the contribution of the standards to the novelty detection procedure. Indeed points declared normal by the detection using only four controls (Section 4.1) can be rejected when the standards are taken into account (column 'Diff' in Table 4.4). In particular, the plates 128 and 198 seem to be acceptable so far as the maximum and the minimum controls are concerned (both were accepted in Section 4.1). Yet the examination of the difference between the two standard controls reveals that there is clearly a significant difference between the observed and the expected values which would require further examination of these plates.

The same procedure could have been applied on the maximum and the minimum controls. Remember we want the plate variation due to systematic changes in the

[10]Remember the maximum and the standard controls are highly correlated (Section 2.3).

n	loglike	m*	min1		M*	Max1		m+	min2		M+	Max2		D	Diff	
66	4.08e+01	m*	3.6e-08	N	M*	2.6e-01		m+	7.4e-02		M+	2.8e-01		D	4.7e-01	
95	1.94e+02	m*	3.6e-08	N	M*	2.8e-04	N	m+	1.9e-12	N	M+	3.5e-09	N	D	2.2e-01	
96	3.51e+01	m*	1.8e-03	N	M*	8.8e-04	N	m+	4.1e-01		M+	1.1e-04	N	D	1.5e-01	
124	3.15e+01	m*	4.8e-02		M*	2.9e-01		m+	2.1e-02	N	M+	2.8e-06	N	D	4.0e-01	
128	6.50e+01	m*	1.2e-01		M*	3.6e-01		m+	3.3e-01		M+	3.6e-01		D	1.3e-02	N
133	3.78e+01	m*	1.6e-01		M*	1.4e-01		m+	1.7e-05	N	M+	1.6e-06	N	D	3.3e-01	
134	2.73e+01	m*	4.5e-01		M*	4.3e-01		m+	2.9e-01		M+	4.7e-06	N	D	9.2e-02	
136	4.11e+01	m*	7.7e-02		M*	2.5e-04	N	m+	1.3e-01		M+	8.5e-09	N	D	3.3e-01	
137	3.69e+01	m*	3.9e-02		M*	1.4e-01		m+	4.7e-04	N	M+	3.6e-01		D	1.3e-02	N
143	3.36e+01	m*	1.4e-06	N	M*	2.1e-01		m+	2.9e-03	N	M+	1.1e-01		D	1.8e-01	
144	4.04e+01	m*	2.4e-01		M*	4.1e-01		m+	3.0e-07	N	M+	2.8e-06	N	D	3.3e-01	
147	5.52e+01	m*	1.9e-13	N	M*	2.9e-01		m+	4.7e-04	N	M+	6.6e-04	N	D	2.9e-01	
148	3.29e+01	m*	3.8e-07	N	M*	3.6e-01		m+	6.6e-03	N	M+	4.2e-01		D	2.1e-01	
149	3.51e+01	m*	1.0e-07	N	M*	1.2e-01		m+	2.0e-01		M+	1.6e-02	N	D	3.3e-01	
153	3.23e+01	m*	1.5e-05	N	M*	1.3e-01		m+	1.2e-03	N	M+	1.4e-01		D	4.6e-01	
154	2.86e+01	m*	8.1e-03	N	M*	7.2e-02		m+	1.7e-04	N	M+	4.7e-01		D	9.2e-02	
156	2.81e+01	m*	1.5e-05	N	M*	3.9e-01		m+	2.0e-01		M+	9.3e-02		D	3.7e-01	
158	2.63e+01	m*	3.2e-04	N	M*	4.3e-02		m+	1.3e-01		M+	4.3e-01		D	1.5e-01	
166	2.64e+01	m*	7.9e-04	N	M*	4.4e-01		m+	6.6e-03	N	M+	4.7e-01		D	7.1e-02	
176	3.32e+01	m*	1.5e-05	N	M*	1.0e-01		m+	2.9e-03	N	M+	2.6e-02		D	4.8e-01	
188	3.03e+01	m*	2.4e-01		M*	2.7e-02		m+	9.8e-06	N	M+	8.2e-02		D	8.3e-02	
198	3.10e+01	m*	4.0e-01		M*	3.1e-02		m+	3.3e-01		M+	8.8e-02		D	1.3e-02	N
199	2.71e+01	m*	1.7e-01		M*	1.9e-03	N	m+	1.3e-01		M+	4.2e-02		D	2.1e-02	N
202	2.85e+01	m*	3.2e-04	N	M*	3.9e-01		m+	6.6e-03	N	M+	1.7e-02	N	D	1.8e-01	

Table 4.4: Contribution measure for each of the 5-tuple

experimental conditions to be flagged. If the same method had been applied on the other controls, we would have detected only an unusual variation between the controls of a given plate. The transformation applied only on the standards achieves a balance between the 'acceptable variation' controlled by the minima and maxima and the potential problems on a plate basis controlled by the difference of the standards.

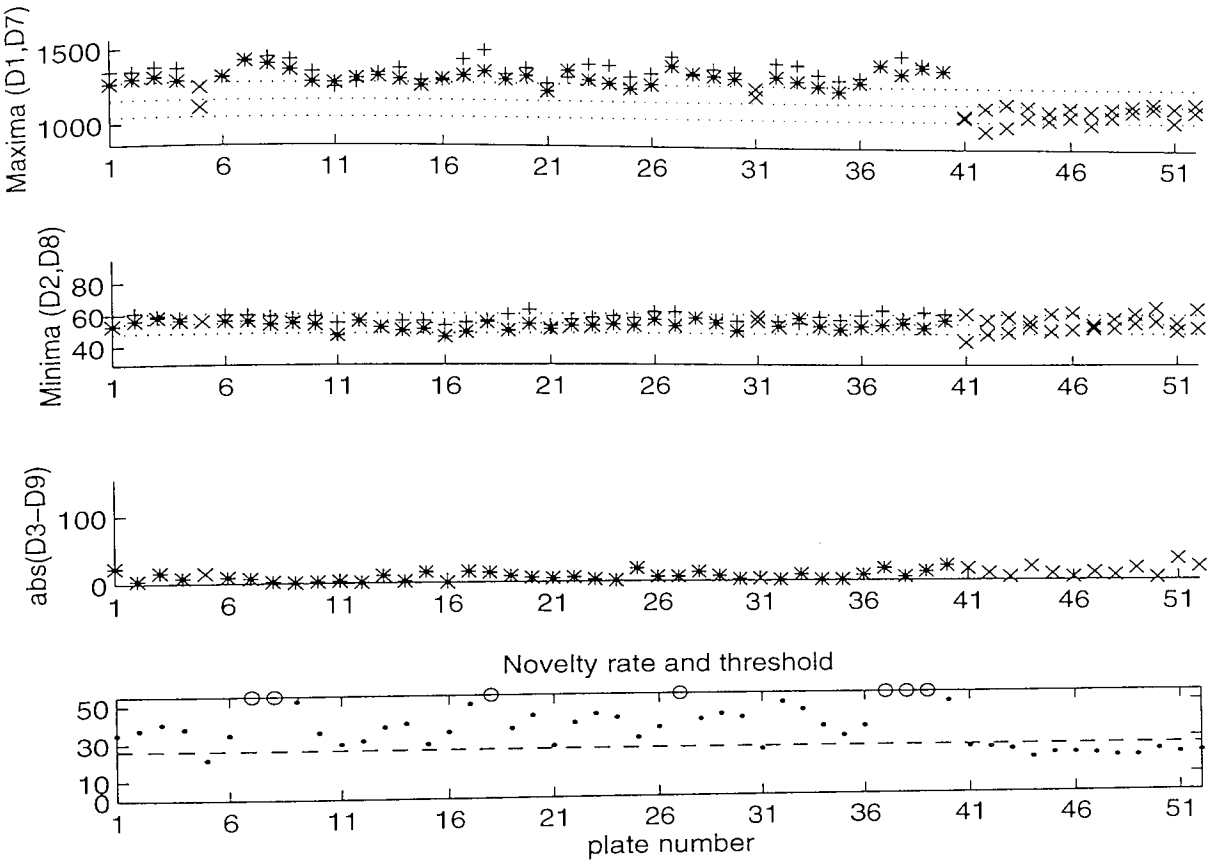


Figure 4.16: **Novelty detection on HTS screen:** plates 1 to 52.

number of wells	number of plates	proportion
1	7	11%
2	47	76%
3	7	11%
4	1	2%

Table 4.5: Repartition of rejected wells for 5 %: *Totals, NSBs & Standards*

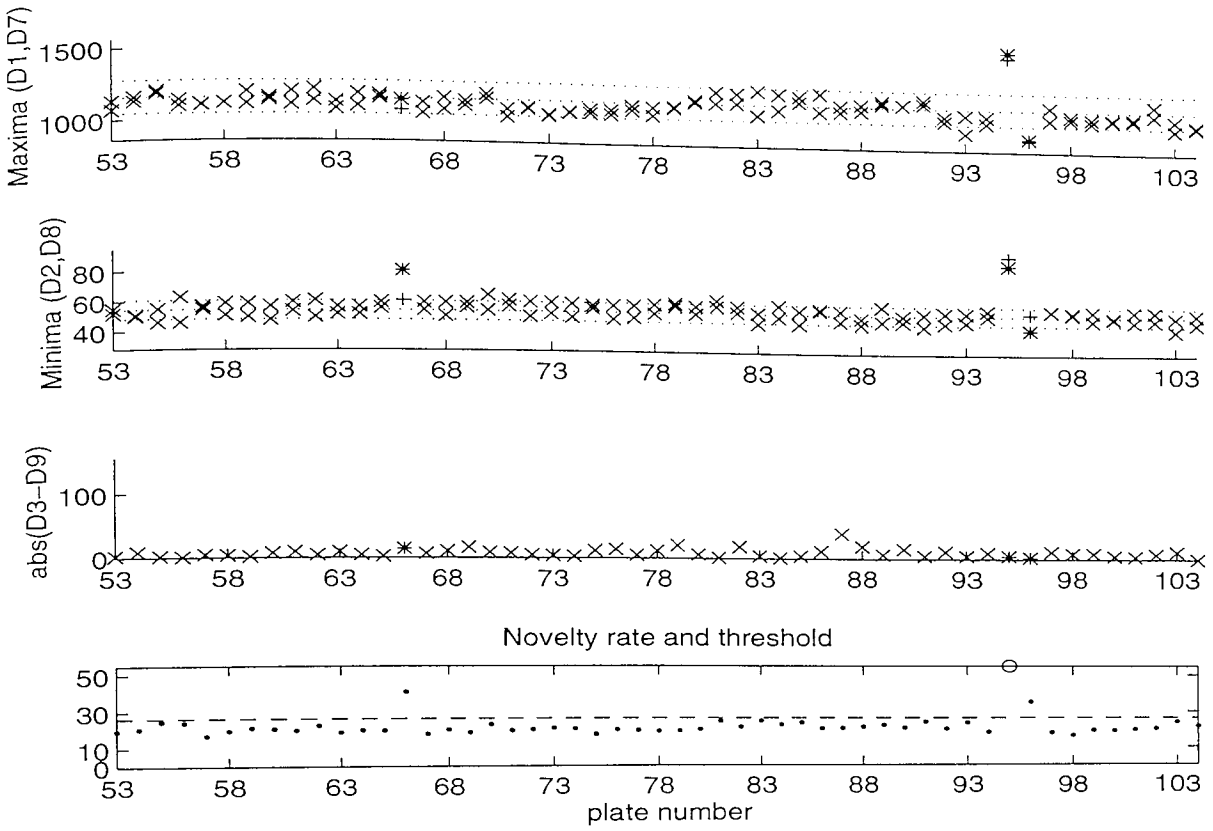


Figure 4.17: **Novelty detection on HTS screen:** plates 53 to 104.

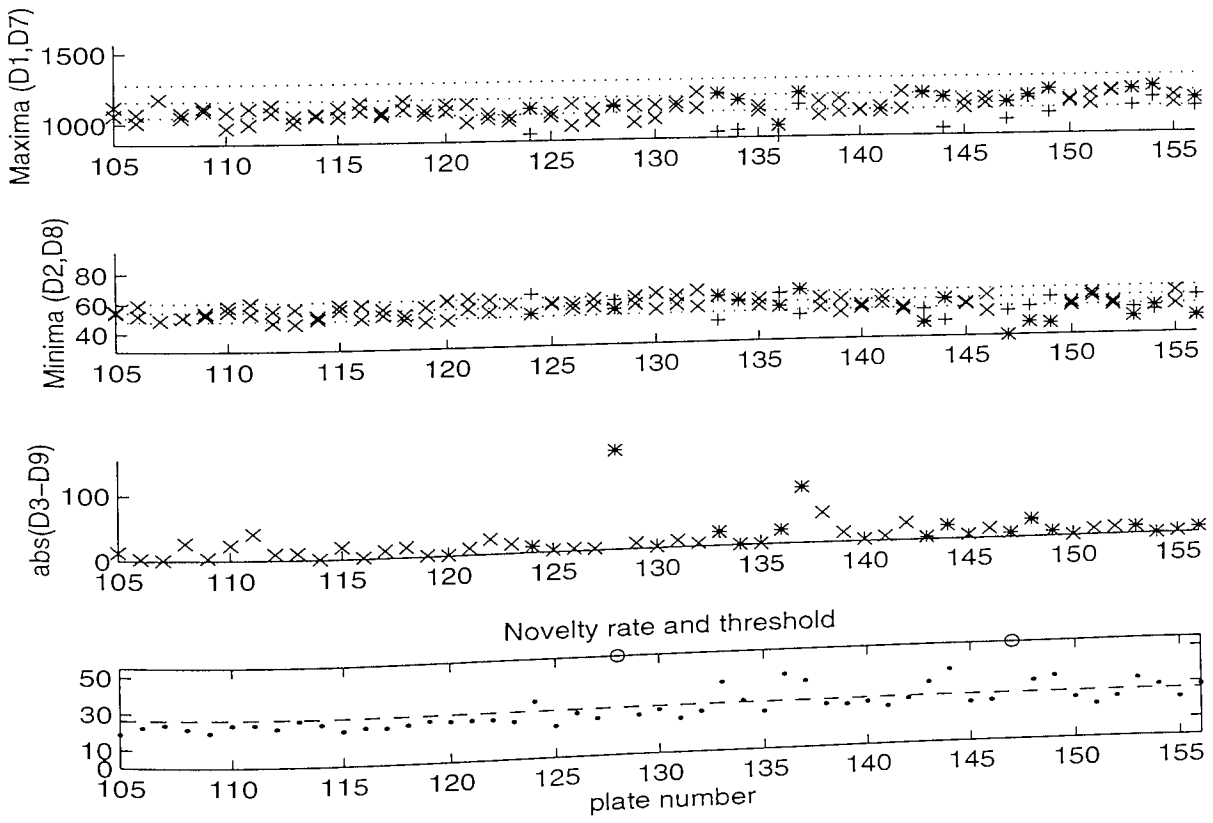


Figure 4.18: **Novelty detection on HTS screen:** plates 105 to 156.

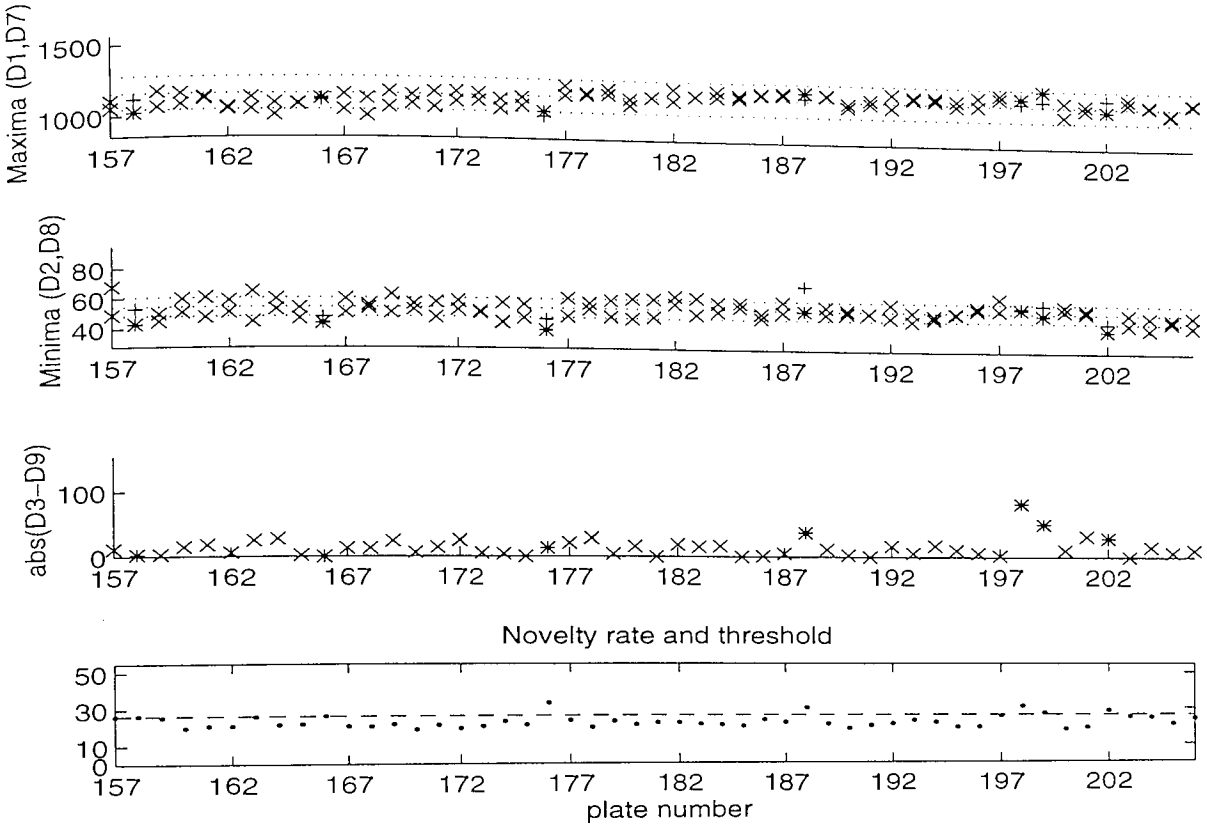


Figure 4.19: **Novelty detection on HTS screen:** plates 157 to 206.

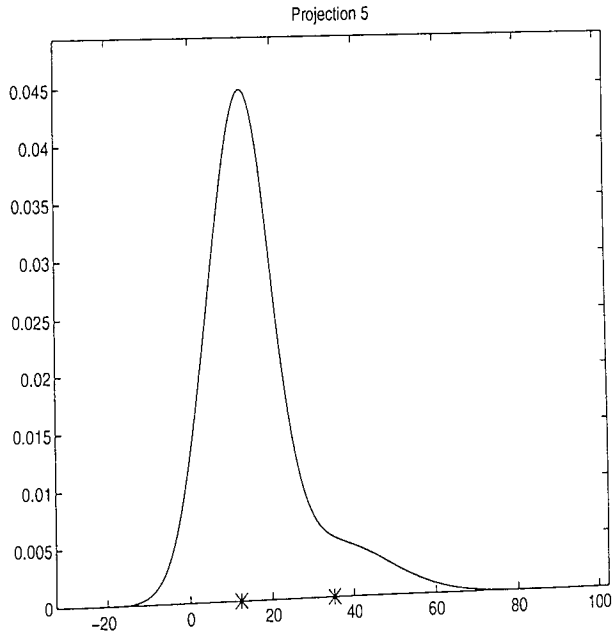


Figure 4.20: Probability density of the difference $|D9 - D3|$ (the stars '$*$' denote the centres of the basis functions)

Chapter 5

Conclusion

This chapter presents the conclusion of this thesis. To start with, we sum up the results of the preliminary study and emphasise the weak points of the traditional approach of outlier detection in the context of an industrial control. The new method to tackle the problem of novelty detection is also recalled. Its advantages are discussed together with cases where it might give poorer results. The last section mentions related issues which can be investigated on the same grounds.

5.1 Results of the preliminary study

In the first place, the correlation tests showed that the control variation due to experimental conditions reflects properly the variation of the standard wells. In other words, it validates the procedure based solely upon the controls for assessment of the data quality.

Second, the severe limitations of the traditional approach of dealing with outliers were put forward:

- Regarding the quality control of HTS data, the method whereby the outlier detection is treated as a problem of hypothesis testing reveals a lack of robustness:
  - The systematic use of various plots as a preliminary stage to testing outliers

is tedious and does not bring any improvement to the current control of HTS data that is to say: it can be subjective and unreliable, in particular when choosing the number of outliers to be tested;

- It does not provide any representation of the data in terms of probability nor any measure of abnormality neither for the plates nor for the wells. Therefore it can not be used for ordering the abnormal plates or controls with respect to this novelty.

- It does not allow the possibility of automation, since the manual graph analysis stage is *highly* recommended to set the number of points to be tested.

These limitations suggest that a method based on probability density inference should be preferred because it would provide a description of the data which can be used as an objective criterion for the determination of abnormal plates and wells.

5.2 New approach

This section describes the approach of the *Quality control of High Throughput Screening* based on density inference. We summarise the method investigated and more particularly what it implies from the user's point of view. We then discuss strong points and potential limitations of the method.

5.2.1 The method

An additional set of three plates called '*control plates*' is added to the beginning of the screen. These plates include the minimum, maximum and the standard controls. The mixture contained by these controls are exactly the same as the mixture of the corresponding controls of the '*normal plates*'. As a result, the measure of the activity of the control plates can be used to assess the variability of the normal plate controls through probability density inference.

Using the principle of maximum likelihood, we have formulated the problem of dealing with abnormal values in terms of density estimation and novelty threshold. A Gaussian mixture model framework was chosen to perform this estimation. It was preferred to other methods because of both practical constraints (computational efficiency, EM algorithm) and theoretical assets (universal estimation property, straightforward conditional densities). Cross validation was used to determine empirically the model complexity and the structure of the the covariance matrix.

The novelty detection proceeds as follows: if the control values of a given plate reveals a low probability, this plate is declared 'novel' or 'abnormal'. This probability constitutes a measure of the abnormality of the plate. In the second place, the conditional densities of the control plates with respect to the model are computed in order to highlight the value(s) which differ significantly from what is expected. The threshold can be modified according to the degree of novelty required for the quality control.

5.2.2 Achievement

Section 4.1 showed satisfactory results in regard to detection based on minimum and maximum controls. The number of rejected plates was higher than expected due to a variation which has been discussed in great length in Section 4.5. The method was improved by inclusion of the standards in the procedure as a fifth component taking into account their difference (Section 4.5.2). Some plates featuring suspicious standard controls were detected although the corresponding plates were accepted by the first scheme.

The negative log-likelihood together with the conditional densities provide a measure of the novelty of the rejected plates or wells. Subsequently it constitutes a quantitative criterion whereby a objective decision can be taken for the quality control of HTS.

The second part of the study was concerned with the choice of a proper threshold.

Two possibilities have been suggested:

- the minimum of the likelihood function on the validation set;
- the critical value for the density probability corresponding to the level of novelty desired.

The adequate choice should be determined by the user. The first possibility has the obvious advantage of not requiring any intervention. It is also intuitive: the controls of the control plates being considered as the reference of normality, all of them should be accepted by the model they infer. On the other hand, the use of a critical value in similar way to a statistical test permits the user to tune the detection as desired.

The high number of rejected plates pointed out in Section 4.1 should not raise concern. First of all, they do reveal a significant difference between normal and control plates so this difference should appear in the detection. Second, if the control is conducted manually the novelty threshold can be adapted in order to fulfil the constraints which may arise from experimental restrictions such as the price of the compounds or the cost of a second run. The contribution of the method in such a case is that it enables the user to validate the plates from the most to the least likely.

5.2.3 Limitations

The limitations of the method were emphasised in Section 4.5. These are mainly due to the fact that the entire variation on the whole screen can not be captured by the two control plates if the screening is conducted under unsteady experimental conditions. The various days of screening involve significant differences between the distributions of the data of each assay. The computation of inhibition (or activation) does not remove this variation (Section 4.5.1). The system will detect *any* variation between the control and normal plates. It is up to the user to determine whether this inconsistency is acceptable.

A potential problem is also the versatility of such a scheme if the screening of the first three plates is not properly controlled. As the density inference proceeds on the control plates, the 'representation of normality' can not be expected to be valid if the screening of these plates is poorly controlled. The outlier detection can cope with rare and significant mistakes on the control plates but if a serious difference occur between control and normal plates, the detection would most certainly detect a great number of novel plates since its reference is inaccurate.

Nonetheless, it can be noticed that if the HTS is eventually automated as it can be Pfizer aim for the future, such a variation will be removed. Both of these problems due to different days of screening would therefore disappear.

5.3 Further studies

This section reviews a few extensions of the argument above related to novelty detection. These concern on the one hand a more thorough study of the normal wells of the normal plates and on the other hand different approaches of the quality control of the HTS data. Since some of these options were investigated, we give a 'flavour' of what can be expected of such studies.

5.3.1 Normal wells and plates

It would certainly prove useful to model the distribution of the normal well values. Goodness-of-fit tests showed that the modelling of the the distribution by a normal distribution gives mitigated results as can be noted in Table 5.1.

If the 'normality' of the distribution for the normal wells of an HTS screen could be modelled such a work would have two advantages:

- a plate which would not have this distribution could be flagged as abnormal and

Screen	5%	1%
Screen 1	20.00	35.65
Screen 1b	81.74	88.59
Screen 2	96.66	98.85
Screen 9	70.81	80.86

Table 5.1: Kolmogorov-Smirnoff normality tests on normal wells

therefore could be examined more thoroughly to determine whether this is either due to an unusual number of 'hits' for the plate or induced by a problem which occurred during the screening.

- a probabilistic definition of a 'hits' could be designed such as a compound which appear in the top n% of a screen.

Besides, it may be interesting to compare two plates from two different assays. This would require only a 'rescaling' of the data which can be obtained by multiplying the values of the wells of a given plate by the activation of this plate.

5.3.2 Novelty detection on the control plates

For the time being, a simple method was implemented to suppress outliers in the first three plates: a point further than 2 standard deviations from the mean is removed from the training set. The tests for outlier detection exposed in Section 2.2 could be used on this purpose. As was emphasised, such tests should be carried out with great care, in particular in an automated procedure. This was the reason why they were not used for detection on the control plates. Nevertheless they might prove useful if the quality control of HTS was intended to be done automatically: a visual inspection of the *control plates* guided by the value of the statistical tests may help the operator to remove those outliers. As a result, the operator would have to inspect three plates instead of 206.

5.3.3 The IC_{50}s

The data for the IC_{50} (Figure 1.5) did not allow to perform substantial experiments to tests the method for this format properly. The method used for the 96-well plate can be easily extended to the IC_{50}s; only the dimension of the data would differ. If the screen is studied plate by plate, the 4/6 dimensional space of our study would be replaced by a 16-dimensional space. If experimental characteristics make the study on one single line of the IC_{50} more sensible, then the network would be trained on a 2-dimensional space.

5.3.4 Detection on a day-to-day basis

As frequently indicated in the previous chapters the controls are subject to an important variation from one day to another. To treat this problem, it is possible to train a model for each assay (whose date that may appear in the header of the data file). The results of a first attempt is shown in Table 5.2[1].

Nonetheless the problems of learning directly on the normal plates exposed in Section 4.5.2 remain. In such an approach, if a substantial set of plates was abnormal it would not be detected and would result in the detection of false hits. Furthermore, if the control for those those plates happen to be correctly modelled by a single Gaussian, such a procedure might result in over-fitting the data because of the small size of the training set. Note that in results on Screen 2 in Table 5.2 the plate 95 which clearly differs from the rest of the screen (which had the highest negative log-likelihood) is not declared novel.

The preliminary studies and a prototype of the method presented in this thesis were initially implemented Matlab. The code detecting novel plates and unusual well values was re-written in C. The final software will be incorporated in the HTS procedure for real-world tests.

[1]The learning procedure is the one used in Section 4.5. It proceeds by cross-validation and includes minimum, maximum and standard controls.

Dates	28/11/96	06/11/96	13/11/96	07/11/96	12/11/96	13/11/96
Plates	1-40	41-80	81-91	92-131	132-171	172-206
Number of rejected plates	1	2	2	0	1	1
Plates rejected	37	51,76	88,89	-	137	176

Table 5.2: Daily detection on Screen 2

Appendix A

Screen references

A.1 Screen 2

A.1.1 HTA and Totals & NSBs plates

Screen Number	2
Number of plates	211
Counter used	Anthos HTII
Control plate	Totals & NSBs: 1-3
	IC50: 4-5
Normal plate	cHTA: 6-211
Date/s of Assay/s	1: 1-5 = 19/11/96
	2: 6-45 = 28/10/96
	3: 46-85 = 06/11/96
	4: 86-96 = 13/11/96
	5: 97-136 = 07/11/96
	6: 137-176 = 12/11/96
	7: 177-211 = 13/11/96

A.1.2 Standard control plates

Screen Number	2
Number of plates	3
Counter used	Anthos HTII
Control plate	Standards Only: 2+3
	Max/Min/Standards: 1
Date/s of Assay/s	1: 1-3 = 14/05/97

A.2 Screen 1

A.2.1 HTA plates

Screen Number	1
Number of plates	115
Counter used	Packard 9912V Microplate Topcount
Control plate	HTA: 1-115
Invalid plates	35 & 57 : Double ligand
	77 & 78 : No assay window
Date/s of Assay/s	1: 1-16 = 01/07/96
	2: 17-46 = 10/07/96
	3: 47-86 = 17/07/96
	4: 87-106 = 18/07/96
	5: 107-114 = 17/09/96
	6: 115 = 03/10/96

A.2.2 Totals & NSB plates

Screen Number	1
Number of plates	6
Counter used	Packard 9912V Microplate Topcount
Control plate	Totals & NSBs: 1-3
Date/s of Assay/s	1: 1-6 = 13/02/97

A.3 Screen 9

A.3.1 HTA plates

Screen Number	9
Number of plates	206
Counter used	Wallac LKB 1205-001 Beta Plate LSC
Control plate	HTA: 1-206
Invalid plates	173-299 : A6 ALL ACTIVE (not relevant)
Date/s of Assay/s	1: 9/1/97 = 1-20
	2: 15/1/97 = 21-50
	3: 16/1/97 = 51-74
	4: 22/1/97 = 75-108
	5: 23/1/97 = 109-134
	6: 24/1/97 = 135-142
	7: 05/2/97 = 143-172
	8: 06/2/97 = 173-206

A.3.2 Totals & NSB plates

Screen Number	9
Number of plates	3
Counter used	Wallac LKB 1205-001 Beta Plate LSC
Control plate	Totals & NSBs: 3
Date/s of Assay/s	1: not provided = 1-3

A.4 Screen 1b (same controls as Screen 1)

A.4.1 HTA plates

Screen Number	10
Number of plates	231 cHTA + 32 HTA
Counter used	Packard 9912V Microplate Topcount
Date/s of Assay/s	1: 1-10 = 30/07/96
	2: 11-16 = 14/08/96
	3: 17-26 = 15/08/96
	4: 27-56 = 21/08/96
	5: 57-93 = 28/08/96
	6: 94-95 = 1/11/96
	7: 96-155 = 5/11/96
	8: 156-209 = 6/11/96
	9: 210-224 = 11/11/96
	10: 240-263 = 21/11/96

Appendix B

Results

B.1 Screen 1

112 plates are declared novel out of 112.

448 wells out of 448.

number of wells	number of plates	proportion
1	0	0%
2	14	12%
3	15	13%
4	83	74%

B.2 Screen 1b

212 plates are declared novel out of 263. 663 wells out of 1052.

number of wells	number of plates	proportion
1	5	2%
2	45	21%
3	80	38%
4	82	39%

B.3 Screen 9

150 plates are declared novel out of 206.

389 wells out of 824.

number of wells	number of plates	proportion
1	28	18%
2	61	38%
3	43	27%
4	27	17%

Appendix C

Computation of the error after normalisation

The problem can be formulated as follows: Suppose we have a sample $\mathcal{T}_1 = \{x_n\}_{n=1}^N$. A mixture model $\mathcal{M}_1$ is trained on this sample using the EM algorithm. In the second place, the linear transformation $\phi: X \rightarrow \frac{X-a}{b}$ is applied on $\mathcal{T}_1$ to find $\mathcal{T}_2$ and a model $\mathcal{M}_2$ is trained on $\mathcal{T}_2$. How can we compare the two models in terms of performance? The problem is schematically shown in Figure C.1. To simplify the notations, we consider the case σI in one dimensional case.

This Section shows that if two models $\mathcal{T}_1$ and $\mathcal{T}_2$ are trained with such a procedure, a linear term must be added to the error $\mathcal{E}_2$ to compare it to $\mathcal{E}_1$. To do so, we compute $\mathcal{E}_1$ on $\mathcal{M}_1 = \{P(j), \sigma_j, \mu_j, j = 1 \ldots d\}$, the mixture model computed by the EM algorithm using the initialisation $\{P(j)^{(0)}, \sigma_j^{(0)}, \mu_j^{(0)}\}$ and $\mathcal{E}_2$ on $\mathcal{M}_2$ the mixture model computed using the transformed initialisation $\{P(j)^{(0)}, \sigma_j^{(0)} b^{-1}, \frac{\mu_j^{(0)}-a}{b}\}$ to show that

$$\mathcal{E}_1 = \mathcal{E}_2 + N \ln(b) \quad . \tag{C.1}$$

First, we show by induction that the parameters of $\mathcal{M}_2$ found by the EM algorithm

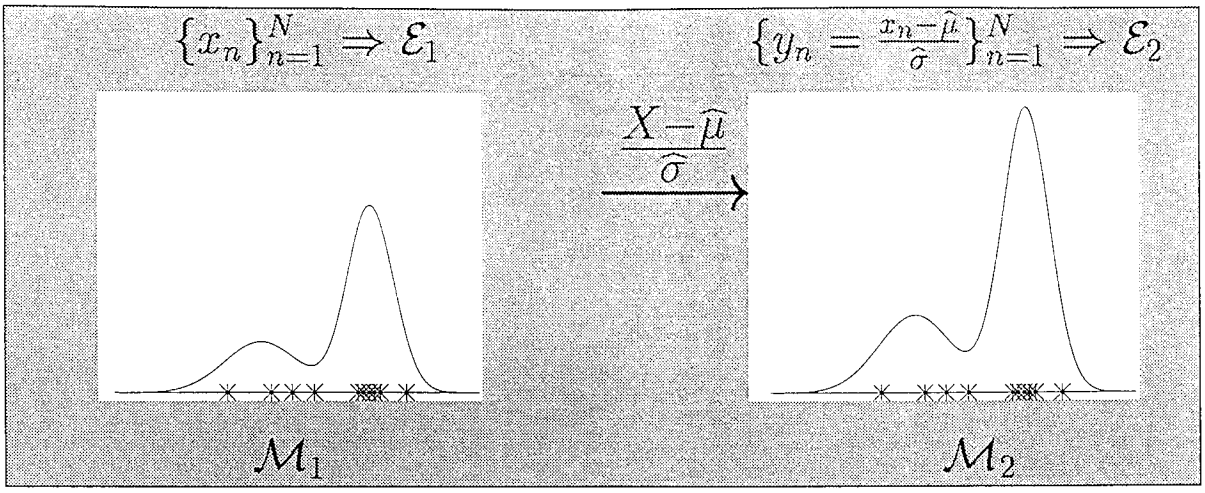


Figure C.1: Normalisation and log-likelihood error on a 2 Gaussian Mixture Model

are given by $\{P(j), \sigma_j b^{-1}, \frac{\mu_j - a}{b}, j = 1 \ldots d\}$. In Section 3.1.2, we had:

$$\mu_j^{new} = \frac{\Sigma_n P^{old}(j|x^n) x^n}{\Sigma_n P^{old}(j|x^n)} \quad , \tag{C.2}$$

$$(\sigma_j^{new})^2 = \frac{\Sigma_n P^{old}(j|x^n) ||x^n - \mu_j^{new}||^2}{\Sigma_n P^{old}(j|x^n)} \quad , \tag{C.3}$$

$$P(j)^{new} = \frac{1}{N} \sum_n P^{old}(j|x^n) \quad . \tag{C.4}$$

Since the distribution of the data is unchanged by linear transformation the priors (C.4) and the posterior probabilities $P(j|x^n)$ remain the same from the raw data to the normalised ($P^{(t)}(j|x^n) = P^{(t)}(j|y^n), \forall x, y, t$).

If the assertion is true until the *old* step, for the *new* step of the training of $\mathcal{M}_2$ we have:

$$\mu_j'^{new} = \frac{\Sigma_n P^{old}(j|y^n) y^n}{\Sigma_n P^{old}(j|y^n)} \quad , \tag{C.5}$$

$$(\sigma_j'^{new})^2 = \frac{\Sigma_n P^{old}(j|y^n) ||y^n - \mu_j^{new}||^2}{\Sigma_n P^{old}(j|y^n)} \quad . \tag{C.6}$$

The equation (C.5) gives:

$$\begin{aligned}
\mu_j^{'new} &= \frac{\Sigma_n P^{old}(j|x^n)(\frac{x^n-a}{b})}{\Sigma_n P^{old}(j|x^n)} \quad , \\
b\mu_j^{'new} &= \frac{\Sigma_n P^{old}(j|x^n)x^n}{\Sigma_n P^{old}(j|x^n)} - a \quad , \\
\mu_j^{'new} &= \frac{\mu_j^{new} - a}{b} \quad . \qquad \text{(C.7)}
\end{aligned}$$

Similarly, from equation (C.6) we find:

$$\begin{aligned}
(\sigma_j^{'new})^2 &= \frac{\Sigma_n P^{old}(j|x^n)||\frac{x^n-a}{b} - \frac{\mu_j^{new}-a}{b}||^2}{\Sigma_n P^{old}(j|x^n)} \quad , \\
b^2(\sigma_j^{'new})^2 &= \frac{\Sigma_n P^{old}(j|x^n)||x^n - \mu_j^{new}||^2}{\Sigma_n P^{old}(j|x^n)} \quad ,
\end{aligned}$$

which gives

$$\sigma_j^{'new} = \sigma_j^{new} b^{-1} \quad . \qquad \text{(C.8)}$$

The proof above is easily reproduced for the first step of the EM algorithm. So finally, the parameters of $\mathcal{M}_2$ found by the EM algorithm are given by $\{P(j), \sigma_j b^{-1}, \frac{\mu_j - a}{b}, j = 1 \ldots d\}$.

The second part proves the relation (C.1). The error for $\mathcal{M}_1$ is given by:

$$\begin{aligned}
\mathcal{E}_1 &= -\ln\Big\{\prod_{n=1}^{N} p(x_n)\Big\} \\
&= -\ln\Big\{\prod_{n=1}^{N}\sum_{j=1}^{M} P(j)\ \frac{1}{\sqrt{2\pi\sigma_j^2}}\exp(-\frac{(x_n-\mu_j)^2}{2\sigma_j^2})\Big\} \quad . \qquad \text{(C.9)}
\end{aligned}$$

Similarly, the error for $\mathcal{M}_2$ is given by:

$$\begin{aligned}
\mathcal{E}_2 &= -\ln\Big\{\prod_{n=1}^{N} p(y_n)\Big\} \\
&= -\ln\Big\{\prod_{n=1}^{N}\sum_{j=1}^{M} P'(j)\ \frac{1}{\sqrt{2\pi\sigma_j^{'2}}}\exp(-\frac{(y_n-\mu_j')^2}{2\sigma_j^{'2}})\Big\} \quad . \qquad \text{(C.10)}
\end{aligned}$$

Substituting (C.7) and (C.8) in (C.10) we deduce

$$\begin{aligned}
\mathcal{E}_2 &= -\ln\left\{\prod_{n=1}^{N}\sum_{j=1}^{M} P(j)\ \frac{1}{\sqrt{2\pi(\sigma_j b^{-1})^2}}\exp\left(-\frac{(\frac{x^n-a}{b}-\frac{\mu_j-a}{b})^2}{2(\sigma_j b^{-1})^2}\right)\right\} \\
&= -\ln\left\{\prod_{n=1}^{N} b\sum_{j=1}^{M} P(j)\ \frac{1}{\sqrt{2\pi\sigma_j^2}}\exp\left(-\frac{(x^n-\mu_j)^2}{2\sigma_j^2}\right)\right\} \\
&= -\ln\left\{\prod_{n=1}^{N}\sum_{j=1}^{M} P(j)\ \frac{1}{\sqrt{2\pi\sigma_j^2}}\exp\left(-\frac{(x^n-\mu_j)^2}{2\sigma_j^2}\right)\right\} - \ln(b^N) \qquad \text{(C.11)}
\end{aligned}$$

and finally, from (C.9) we conclude

$$\mathcal{E}_2 = \mathcal{E}_1 - N\ln(b) \qquad \square$$

The proof is easily extended to the case of the application $\Phi :\ \mathbf{x} \rightarrow \mathbf{B}\mathbf{x} + \mathbf{a}$ where $\mathbf{x}$ and $\mathbf{a}$ are d-dimensional vectors and $\mathbf{B}$ a $d \times d$ invertible matrix. In this case, we find:

$$\mathcal{E}_1 = \mathcal{E}_2 + N\ \ln(\det(\mathbf{B}))\ \ .$$

To summarise, the method used in Section 3.5.2 to compute an error comparable to the one computed on raw data in order to compare two mixture models proceeds as follows:

1. the application $X \rightarrow \frac{X-\widehat{\mu}}{\widehat{\sigma}}$ is used to centre the data;
2. a mixture model is trained on the centred data with $\mathbf{\Sigma} = \sigma I$;
3. the negative log-likelihood error $\mathcal{E}_1$ is computed on the centred model;
4. the error $\mathcal{E}_2$ is computed with respect to:

$$\mathcal{E}_1 = \mathcal{E}_2 + N\ \ln(\prod_{i=1}^{d}\widehat{\sigma}_i)$$

where N is the size of the sample.

Bibliography

[Bis95] C. M. Bishop. *Neural Networks for Pattern Recognition.* Oxford Univ. Press, 1995.

[BKRW97] D. Bojanic, W. W. Keighley, M. J. Russel, and T. P. Wood. Factors for the successfull integration of assays, equipment, robotics and software for high throughput screening. In Devlin, editor, *High Throughput Screening - The Discovery of Bioactive Substances*, pages 493–508, New York, 1997. Marcel Dekker.

[BL78] V. Barnett and T. Lewis. *Outliers in Statistical data.* Wiley, 1978.

[FL94] W. D. Furman and B. G. Lindsay. Testing for the number of components in a mixture of normal distributions using moments estimators. *Computational Statistics & Data Analysis*, 17:473–492, 1994.

[Lac87] G. J. Mc Lachlan. On bootstrapping the likelihood ratio test statistic for the number of components in a normal mixture. *Appl. Statist.*, 36(3):318–324, 1987.

[LB88] G. J. Mc Lachlan and K. E. Basford. *Mixture Models: Inference and Applications to Clustering.* Marcel Dekker, New York, 1988.

[MGH89] R. L. Mason, R. F. Gunst, and J. L. Hess. *Statistical Design & Analysis of Experiments.* Wiley, 1989.

[NCCR+97] A. Nairac, T. A. Corbett-Clarck, R. Ripley, N. W. Townsend, and L. Tarassenko. Choosing an appropriate model for novelty detection. In *5th International Conference on Artificial Neural Networks*, pages 117–122. IEE, July 1997.

[Rip96] B. D. Ripley. *Pattern Recognition and Neural Networks*. Cambridge University Press, 1996.

[RT94] S. Roberts and L. Tarassenko. A probabilistic resource allocating network for novelty detection. *Neural Computation*, 6:270–284, 1994.

[XJ95] L. Xu and M. I. Jordan. On convergence properties of the em algorithm for gaussian mixtures. AI Memo 1520, MIT, January 1995.