Incremental Stream Clustering and Anomaly Detection

Jan Ekman, Anders Holst
jan, aho@sics.se

31st January 2008

SICS Technical Report T2008:1
ISSN 1100-3154

Keywords: Incremental Clustering, Anomaly Detection, Bayesian Statistics, Classification

Abstract

This report concerns the "ISC-tool", a tool for classification of patterns and detection of anomalous patterns, where a pattern is a set of values. The tool has a graphical user interface "the anomalo-meter" that shows the degree of anomaly of a pattern and how it is classified. The report describes the user interaction with the tool and the underlying statistical methods used, which basically are Bayesian inference for finding expected or "predictive" distributions for clusters of patterns and using these distributions for classifying and assessing a degree of anomaly to a new pattern. The report also briefly discusses what in general are appropriate methods for clustering and anomaly detection. The project has been supported by SSF via the Butler2 programme.

Contents

Part I

The ISC-tool

1 Introduction

1.1 The "English butler" project

The programme proposal for the "English butler" says that:

> "The long term objective in the BUTLER programme proposal is to provide industrial plants with as much self-surveillance as possible, to make the process autonomous and robust as possible."

A part of the Butler project studies methods and tools for clustering and anomaly detection of a certain kind, incremental stream clustering, ISC for short, which is the subject of this report.

1.2 Anomaly detection

The concept of *anomaly detection* says nothing about the detection approach and it actually says nothing about what to detect. Anomaly detection is sometimes concerned with separating an irregular and hard to define minority of patterns (or data) from a more regular majority. The detection is in this case based on studies and characteristics of the majority and the minority patterns appears as deviant from the found traits of the majority. It is generally the case that this is carried out via some feature extraction, such that the raw input data is not directly analysed but some representative features extracted from the data. In other cases more is known about the anomalies of interest and the features extracted from the raw data is chosen to fit these anomalies.

Anomaly detection often is or has the potential of playing an important part in surveillance in industry. For instance it can be used as a part of:

the maintenance, for example for the detection of mobile parts of the equipment (machinery) on their way to be worn out

the alarm system, for example as an indication of that some fault have arisen

the fault diagnosis system, for example by classifying different anomalies

a surveillance support tool, for example by letting users interactively guide the anomaly detector to discriminate anomalous behaviours from normal ones.

a support tool for operators, for example as an indication that actions may need to be taken.

1.3 Incremental Stream Clustering

In this context incremental stream clustering is regarded as a method, or class of methods, to analyse data. The input data, in general, to such a method is a stream of patterns, where each pattern consists of a set of values. It is assumed that the large majority of the patterns can be divided into naturally occurring and distinct *classes* that are of interest to us for detecting anomalies. A special case is that there is just one such class and all patterns outside that class are anomalous. It may sometimes be the case that some of the classes are considered to consist of abnormal patterns.

The task to be solved by an incremental stream clustering method is to reconstruct or model the pattern classes and use the model to correctly classify a new pattern as either belonging to a certain previously observed class or not belonging to any of these classes. The pattern classes constructed by the method, are called *clusters* and therefore the method is called *clustering*. The clusters are statistical models of the set of the observed cluster patterns. The clusters are typically described by cluster parameters and do not contain any direct information on the observed patterns of the cluster.

The method, in its simplest form, repeatedly takes a new pattern and checks if it is likely to belong to any of the so far constructed clusters. If this is the case the pattern is considered to be a member of the cluster it most likely belongs to and, as a consequence of this, the cluster parameters of this cluster is updated in accordance with the addition of the pattern. If the pattern is not likely to belong to any of the clusters then a new cluster is constructed for this single pattern. The absolutely first pattern observed always give us a new cluster. Depending on the application we may of course modify this simple method, to obtain an improved performance.

2 The ISC-tool

2.1 The general ISC-tool set up

The ISC-tool implementation includes some user interaction and hence it is a slight modification of the method of incremental stream clustering as described above. The user may choose to label clusters and throw away anomalous patterns without any further notice. For the purpose of demonstration the ISC-tool is merged with a pattern generator, from here on just called *the generator*, which the user to some extent can control via a separate user interface. The implemented ISC-tool will be referred to as the *analyser* and the merged analyser and generator will be referred to as the *demonstrator*. The Demonstrator graphical user interface use states to refer to both the classes of the generator and the clusters of the analyser. In this document we use classes and clusters, i.e. a class is a pattern generator state and a cluster is state found by the analyser.

In the example that follows the output from the generator is a stream of patterns, where each pattern consists of 400 values. Each of these values is randomly generated from a normal distribution for some given parameters. Hence, each pattern can be thought of as generated from a class defined by 800 parameters, i.e. mean and standard deviation for each of the 400 values. The user is given the possibility to define several classes and check if the clustering performs well, i.e. the tool is capable of discriminating between patterns generated from different classes. The generator do not give the user the option to choose the the 800 parameters freely for each class. For instance the standard deviation is the same for all values in each class and hence there is in fact only 401 parameter values representing a class in the generator. The user is given the possibility to define several classes and check if the clustering performs well, i.e. the tool is capable of discriminating between patterns generated from different classes.

The analyser assumes no knowledge of the classes and constraints of the generator, other than that each of the 400 values of a pattern is a sample from a normal distribution. Hence, the class parameters is completely unknown to the analyser and the analyser do not know that the standard deviation is the same for all values in each class. In other words the analyser is not tailored to distinguish between the possible classes of the generator

Although the analyser in the demonstrator assumes patterns from normal distributions, the ISC-tool in general has the possibility to analyse patterns from poisson, normal, gamma and chi-square distribution.

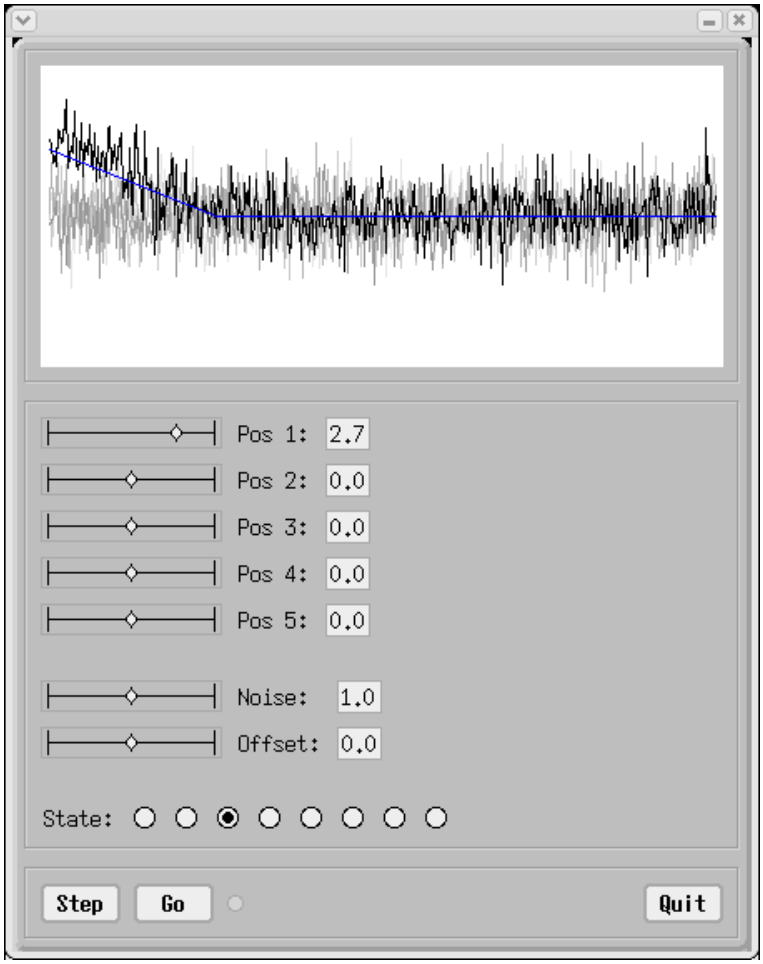

Figure 1: The generator window

2.2 The User Interface of the ISC-tool

The user interface of the ISC-tool consists of three windows, the generator window, the analysis window and the labeling pop-up window.

The generator window, figure 1, lets the user define classes, let her control pattern generation and shows her the generated patterns. The topmost part of the window shows the last pattern in black and let previous patterns fade out in shades of grey. This part of the window also has a blue line showing the mean values of the probability distribution functions of the pattern values. The middle part of the window has seven class parameter slide buttons. The buttons *Pos 1* to *Pos 6* and *Offset*, sets the distributions mean values and the *Noise* button sets their standard deviation. For each class all the 400 distributions will have the same standard deviation. We refer to current settings of these seven slide buttons as the *generator settings* or the *current class settings*. Finally there are eight class selection buttons in this part of the window. The user controls the pattern generation via the buttons Step and Go at the windows bottom line.

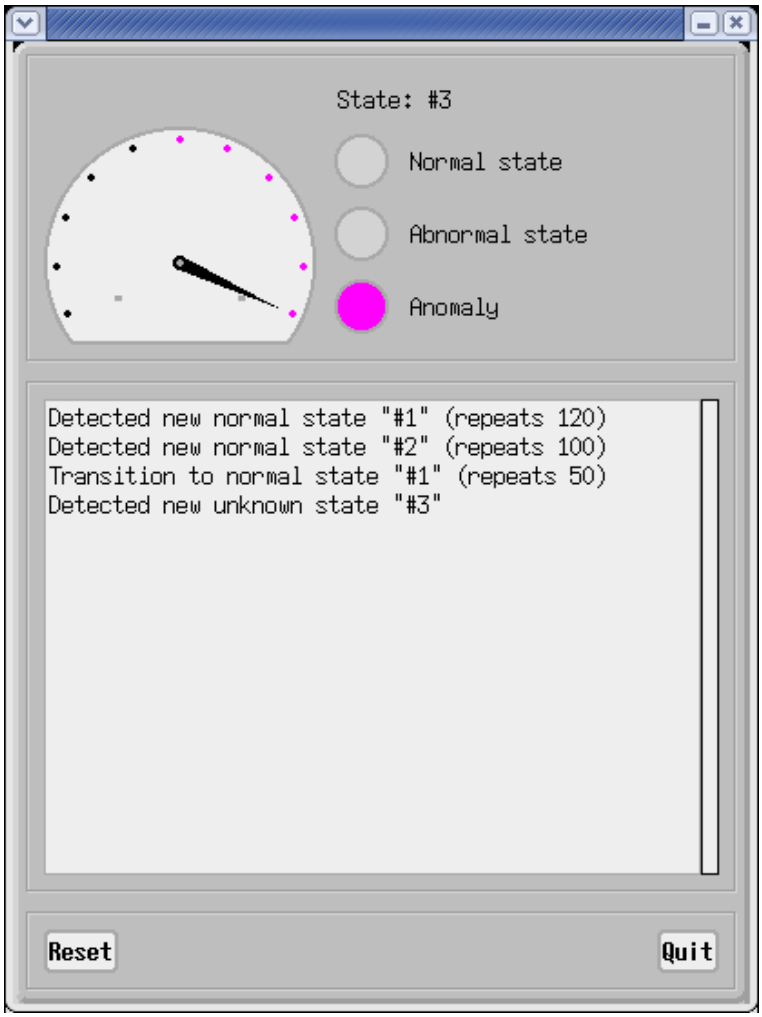

Figure 2: The pattern analysis window

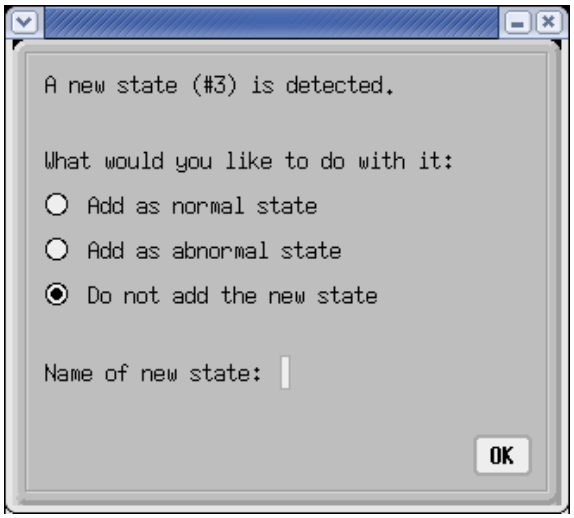

Figure 3: The labeling pop-up window

The analysis window, figure 2, shows analysis results. A top part of the window contains the anomalo-meter, showing us the degree of anomaly of the last generated pattern. Below the anomalo-meter is shown how many consecutive patterns that the analyser found in one and the same cluster and if a pattern is considered anomalous. The user cannot give input to the the tool via the pattern analysis window.

The occurrence of an anomalous pattern makes it possible for a new cluster to arise and this is the only possibility for a new cluster to arise. When an anomalous pattern occurs the labeling pop-up window, figure 3, pops up and asks the user what to do. The default choice "Do not add the new state" is to skip the labeling altogether and throw away the pattern, which has the same effect as if the pattern never occurred. If the user does not want to throw away the anomalous pattern she must chose between adding either a new normal or a new abnormal cluster containing the single anomalous pattern. This is done by pushing the buttons "Add as normal state" and "Add as abnormal state" respectively. The choice of normal or abnormal cluster is only a matter of labeling, i.e. the user chooses a cluster to be abnormal only for the sake that she wants to be noticed of patterns in this cluster.

The user also may want to name (or label) the new cluster via the name field "Name of new state:". We may skip the naming, as we do in the example below. The clusters will in this case be given numbers, e.g. #1, #2, #3 and so on, as default names.

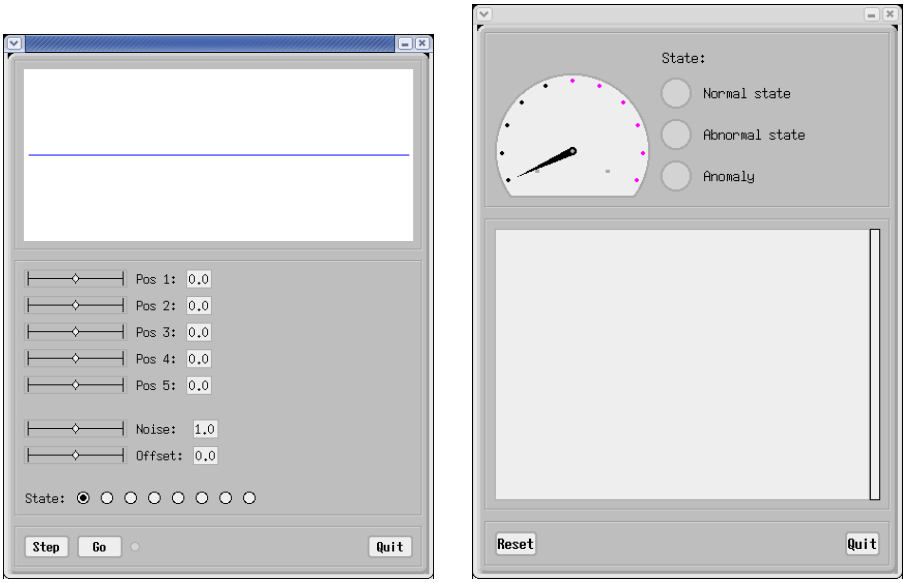

Figure 4: The start up view of the ISC-tool

2.3 Generating and analysing "default class" patterns

Figure 4 show us the start up view of the ISC-tool. The user pushes the Step button which makes the generator generate a first pattern, figure 5. Since the user has not changed the generator settings, the pattern is randomly generated according to the default generator settings. The first pattern is always considered anomalous and hence the labeling pop-up window pops up and asks the user what to do. The user assumes the pattern to be normal and chooses to click at the buttons *Add as normal state* and *OK*. As a result a first normal cluster is introduced, with cluster parameters chosen to fit the pattern and the text in the analysis window changes from *Detected new unknown state* "#1" to *Detected new normal state* "#1", figure 6. Since the user did not name the cluster, the cluster kept its default name #1.

Clicking at the the Step button a second time makes a second pattern be generated and analysed, figure 7. The analyser, given just one pattern as a sample of values from 400 normal distributions, has no knowledge of the standard deviations of these distributions. More over the analysis is based on separate analysis for each of these 400 values. Hence, the analyser cannot conclude anything about the similarity of the very first two patterns. There are many ways to handle this pattern, such as asking for external extra information. In this version of the ISC-tool the two first patterns are, however, assumed to come from the same class. If more information on the classes are present this assumption may be modified, but in cases where the class transitions occurs seldom it is most often the correct choice to let the two first patterns belong to the first class.

The third pattern is generated and analysed as the user once again hits the Step button, figure 8. This time the analyser do check if the received pattern is anomalous. As is shown in the figure the third pattern is found to be nor-

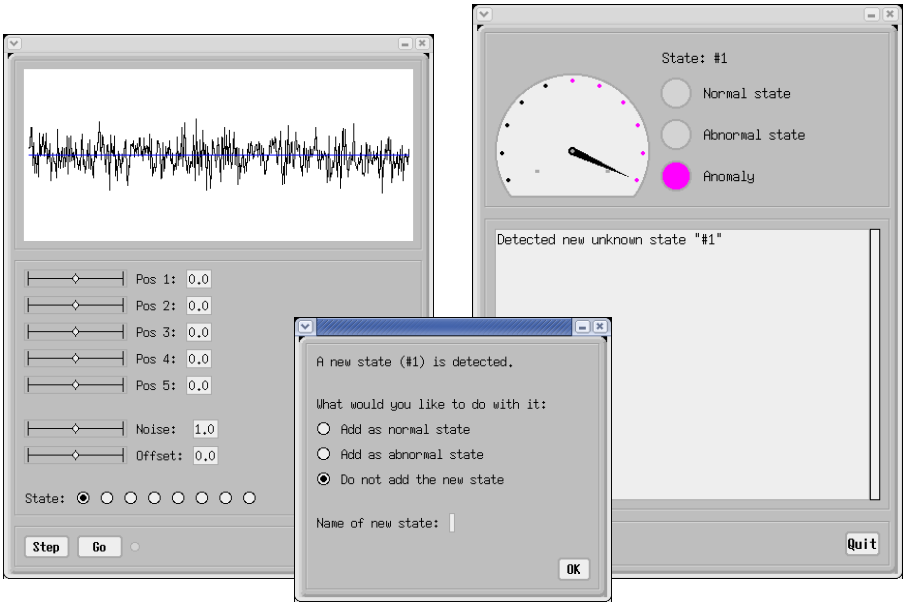

Figure 5: The first pattern received

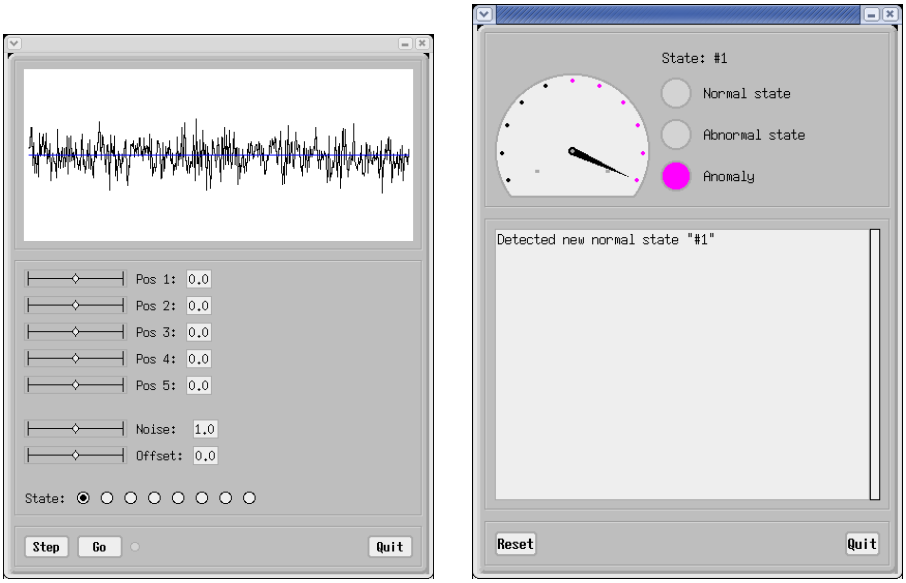

Figure 6: The first cluster is chosen as normal

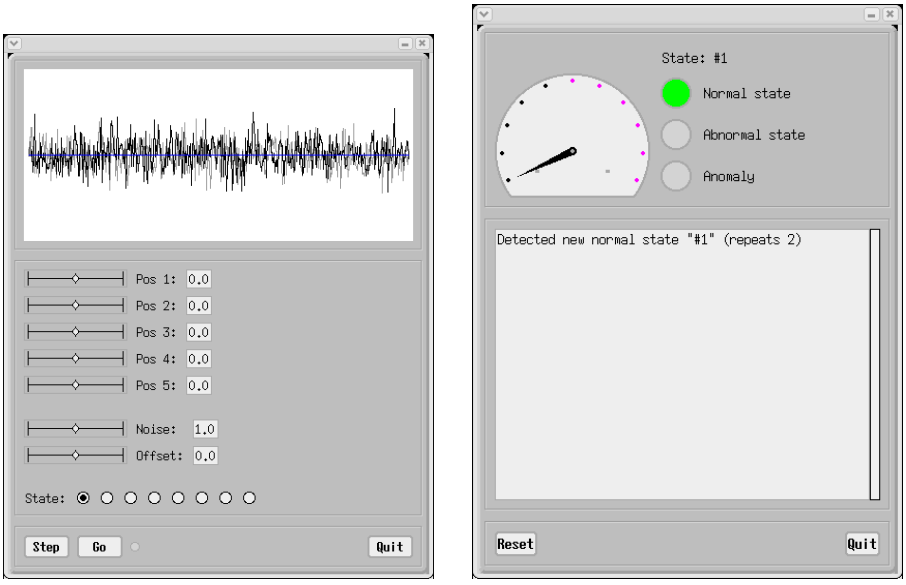

Figure 7: The second pattern is received and classified as belonging to cluster #1

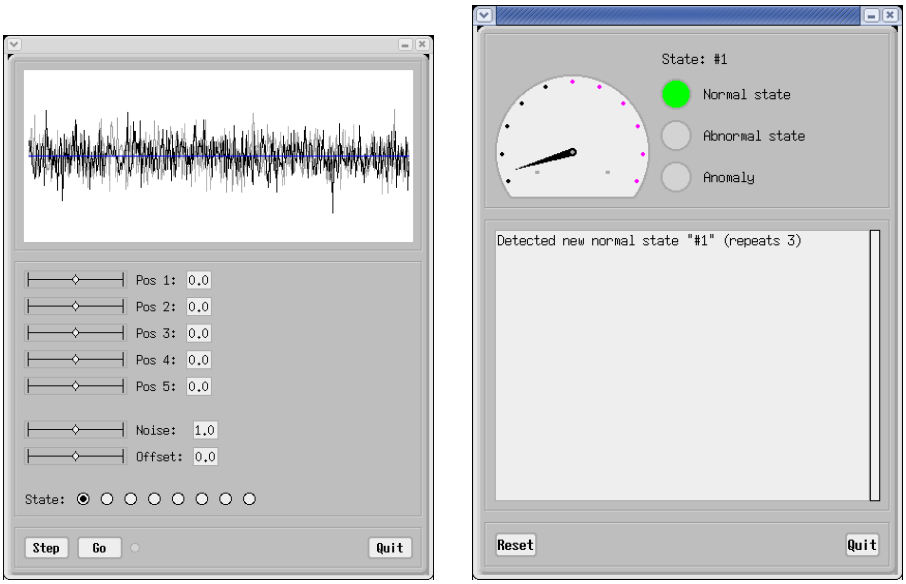

Figure 8: The third pattern is received and classified as belonging to cluster #1

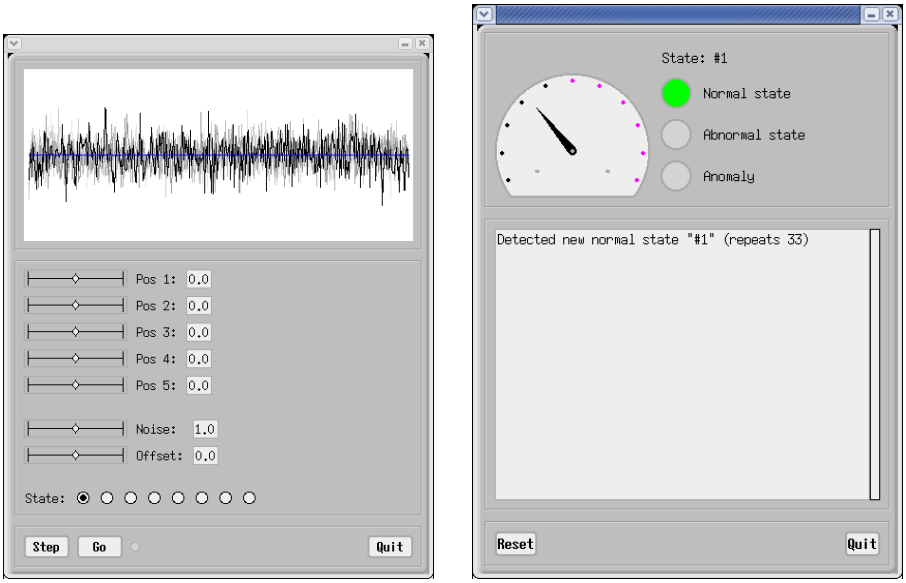

Figure 9: Pattern n:o 33 is received and classified as belonging to cluster #1

mal. The user goes on pushing the Step button without changing the generator settings. The analyser finds all patterns to be normal although some patterns are not so far from anomalous, figure 9. Instead of repeatedly pushing the Step button the user may push the Go button just ones. The ISC-tool will then autonomously generate and classify patterns until the step button is pressed to halt the pattern generation.

2.4 Defining a new class

After having generated 120 patterns from the default generator settings the user now defines a new class by first pushing the second one of the eight class selection buttons, positioned after "State:" in the generator window, and thereafter changing the generator settings, figure 10. The Step button is pressed, a pattern is generated from the new class and the analyser correctly classifies the pattern as anomalous. The labeling pop-up window pops up and asks the user what to do. The user chooses also the new cluster to be normal and again skip to give the cluster a name. The cluster gets the default name #2, figure 11. As a result a second normal cluster is introduced, with cluster parameters chosen to fit the pattern from the new class. The text in the analysis window changes from *Detected new unknown state* "#2" to *Detected new normal state* "#2", figure 12.

The user goes on generating another pattern from the second class. The pattern is assumed to belong to the second cluster, figure 13.

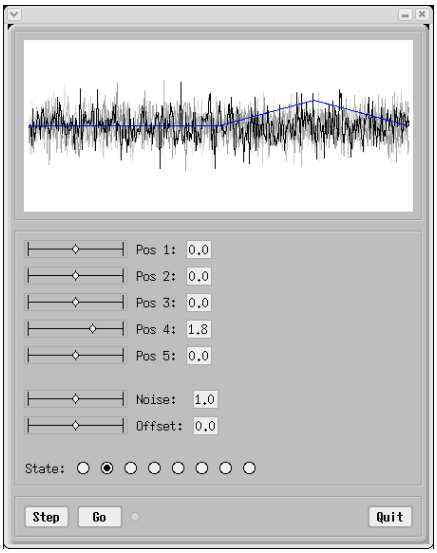

Figure 10: A new class is defined

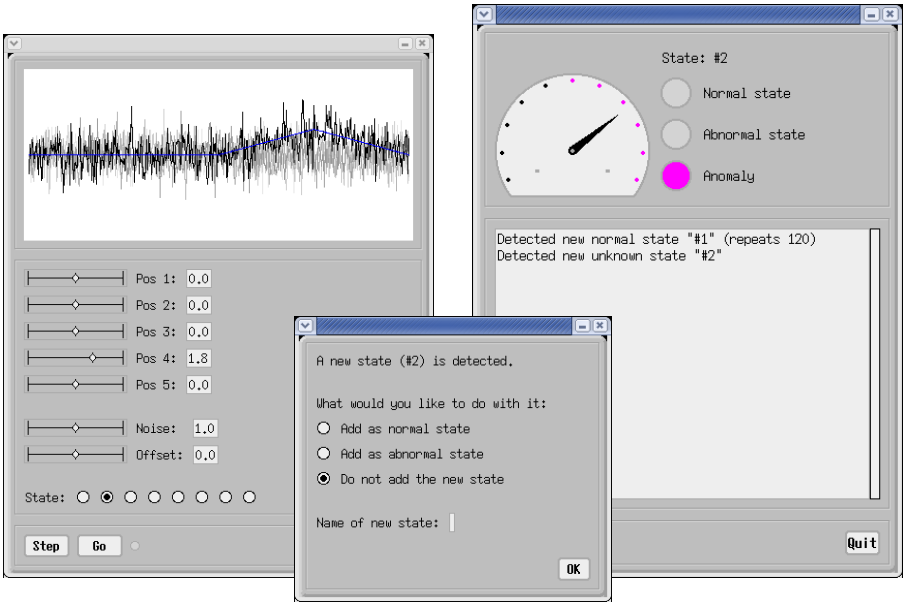

Figure 11: The analyser receives a pattern generated from the new class and classifies it as anomalous

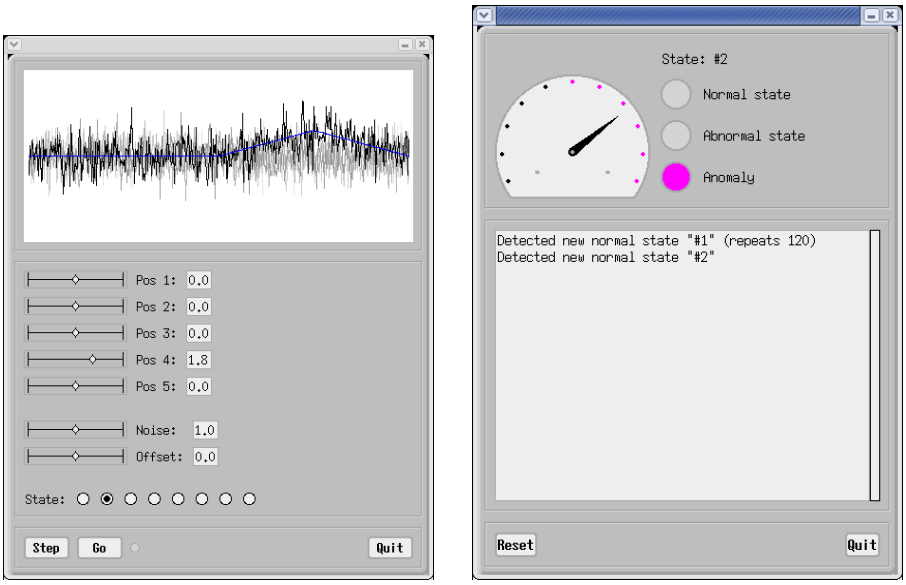

Figure 12: Also the second cluster is chosen to be normal

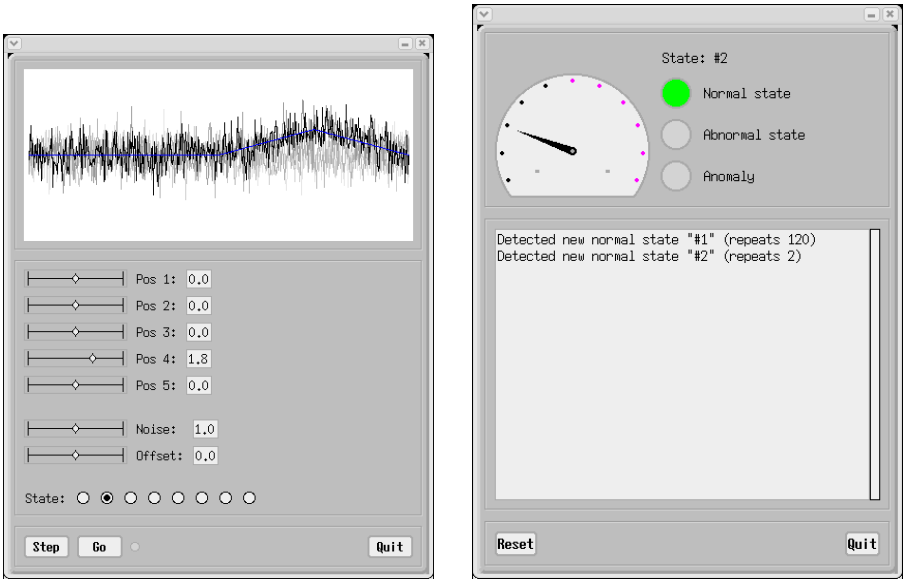

Figure 13: The second pattern from the second class, is classified as belonging to the second cluster

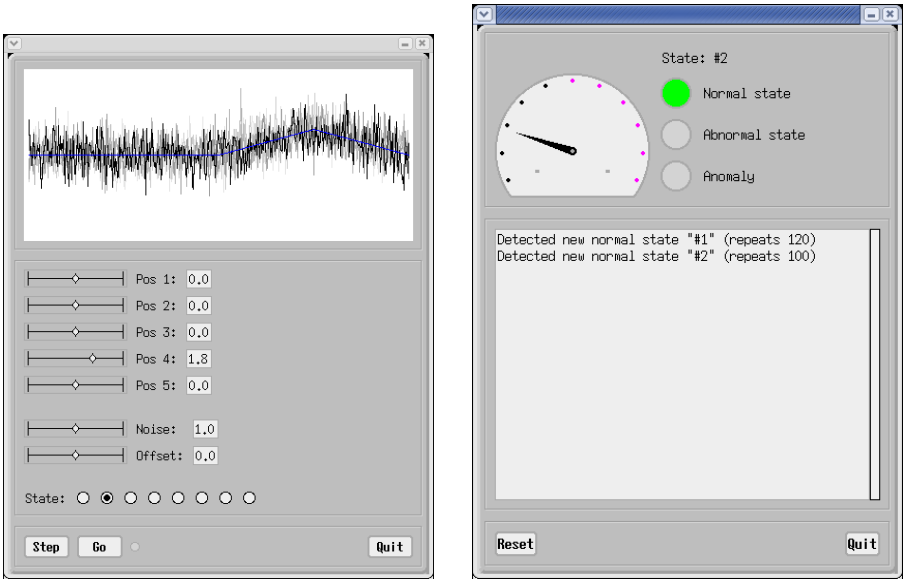

Figure 14: A hundred patterns generated from the second class all belonging to the second cluster

2.5 Switching between classes

After a hundred patterns are generated from the second class, figure 14, all of them classified as belonging to the second cluster, the user decides to generate and analyse some patterns from the first class again. She pushes the first of the eight class selection buttons, positioned after "State:" in the generator window and use the Step- or Go-buttons to generate another fifty patterns from the first class. All of them classified as belonging to the first cluster, figure 15.

Now the user now decides to introduce a third class, pushes the third of the eight class selection buttons, positioned after "State:" in the generator window, change the generator settings and pushes the Step-button. The analyser finds the pattern to be anomalous, figure 16. The user chooses this class to be abnormal and the text in the analysis window changes from *Detected new unknown state* "#3" to *Detected new abnormal state* "#3", figure 17. The user generates a second pattern from the third class and it is classified as belonging to the third cluster, figure 18.

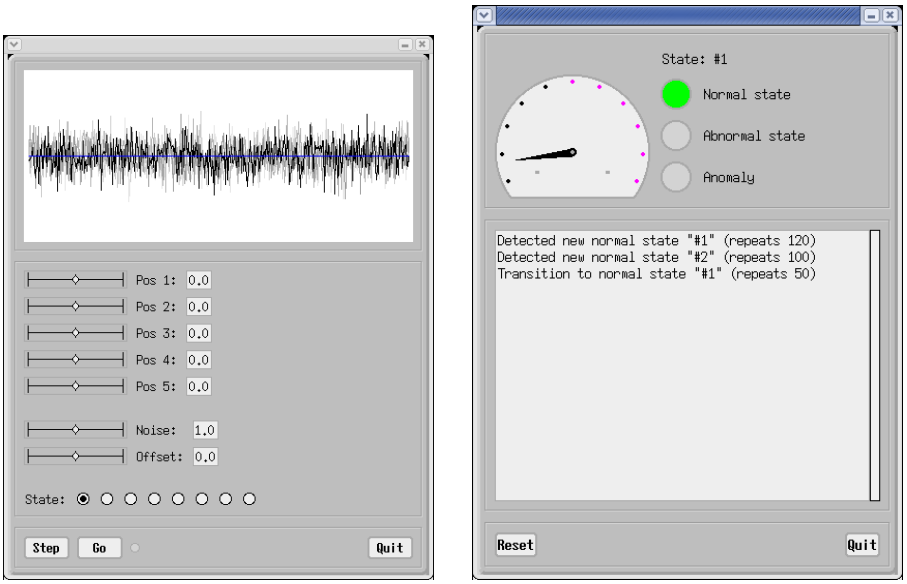

Figure 15: Another fifty patterns generated from the first class all belonging to the first cluster

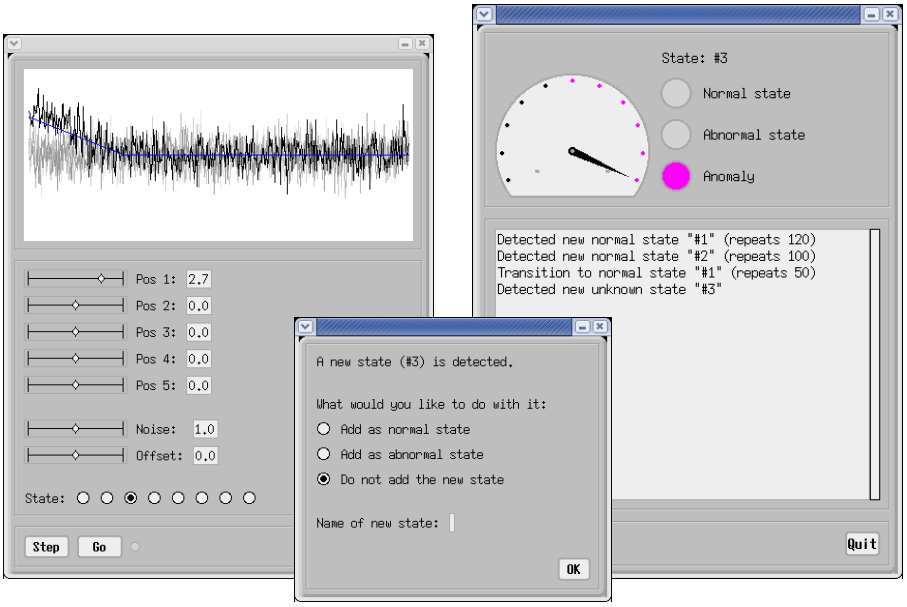

Figure 16: The analyser receives a pattern generated from a third class and classifies it as anomalous

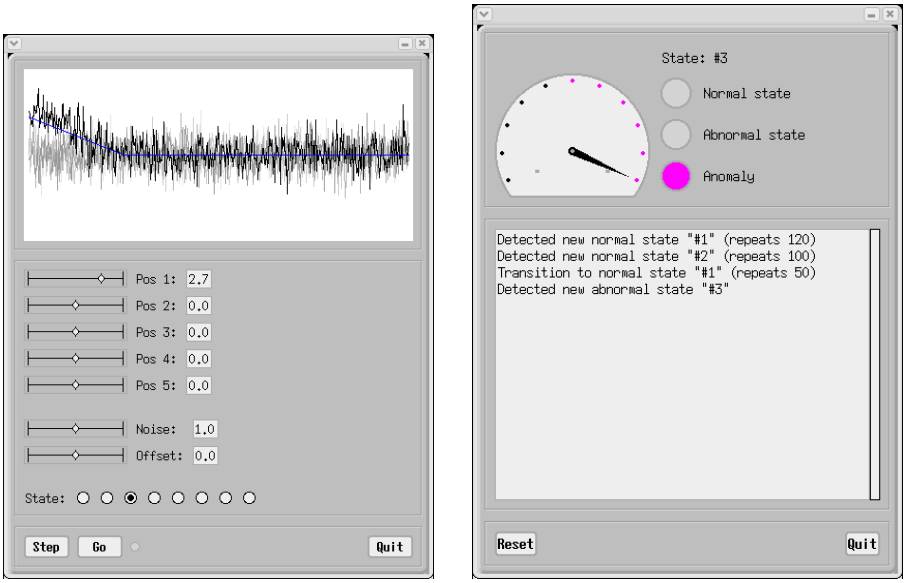

Figure 17: The third cluster is chosen to be abnormal

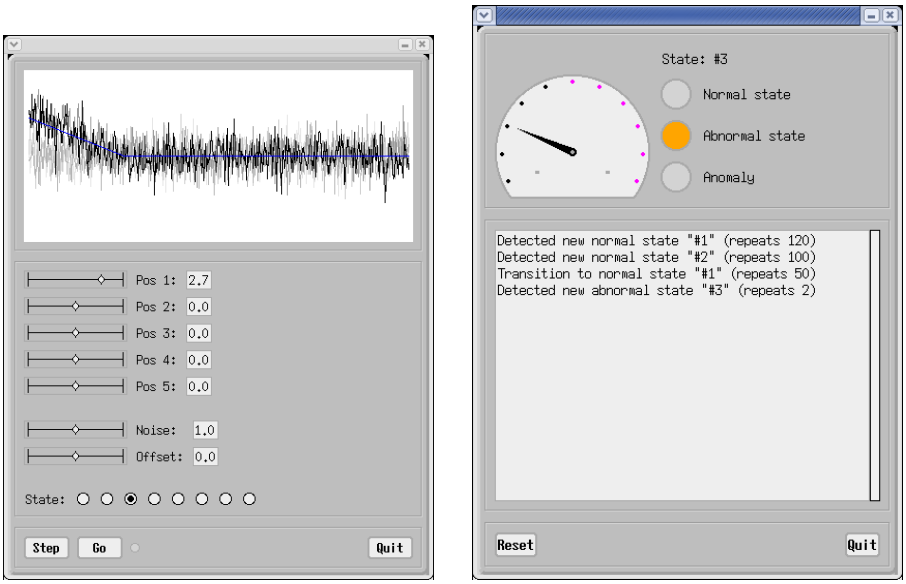

Figure 18: A second pattern, received from the third class, is correctly classified as belonging to the abnormal cluster

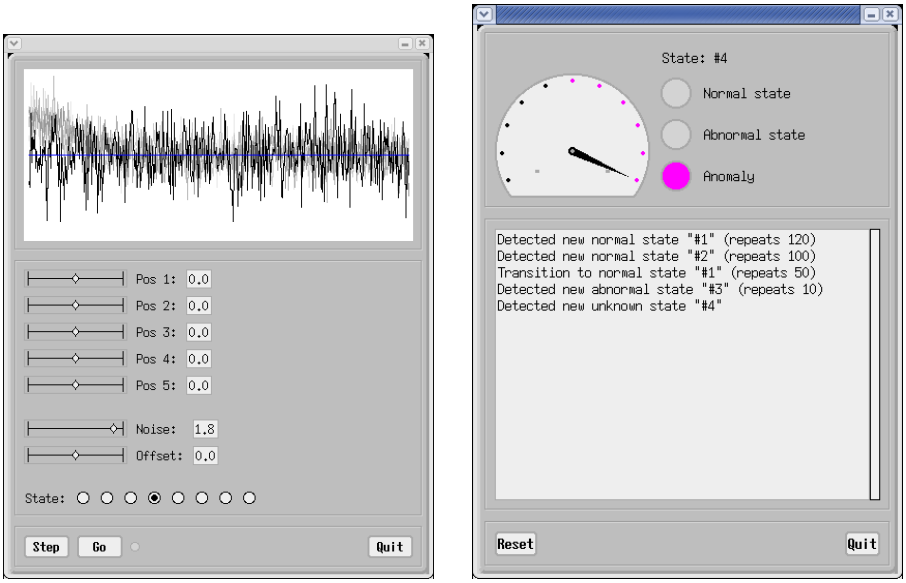

Figure 19: After ten correct classifications of patterns generated from the third class, the analyser receives a pattern from a class with a high noise level and classifies it as anomalous.

2.6 Raising the noise level

The classes defined so far differs in the means of the pattern values distribution but not in the standard deviations. After the user have generated another ten patterns from the third class, all of them classified as belonging to the third cluster, she wishes to increase these standard deviations. She pushes the fourth of the eight class selection buttons, positioned after "State:" in the generator window, drags the Noise slide button up to the value 1.8 and pushes the Step-button. The analyser finds the newly generated pattern to be anomalous, figure 19. The user wishes the analyser to forget that it has ever encountered this anomalous pattern and chooses to push the OK-button in the labeling pop-up window without changing the default choice "Do not add the new state". The next pattern the analyser receives, after the Step-button is pushed once more, is found to be anomalous again, figure 20. This time the user chooses to introduce a new abnormal fourth cluster containing the pattern. After the next push of the Step-button the analyser classifies the generated pattern as belonging to the anomalous fourth cluster, figure 21.

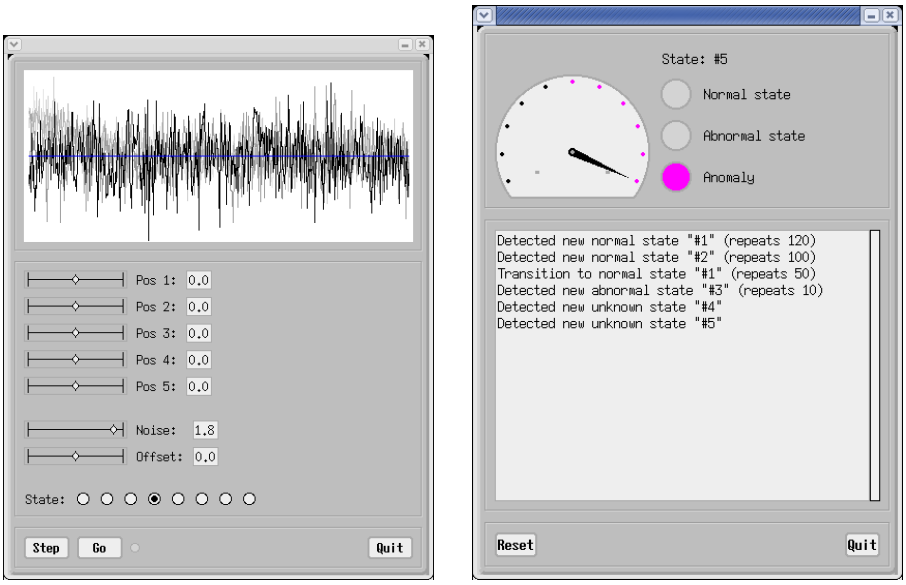

Figure 20: The user has chosen to ignore the anomalous pattern and when a new pattern arrives from the same high noise level class the analyser again classifies it as anomalous.

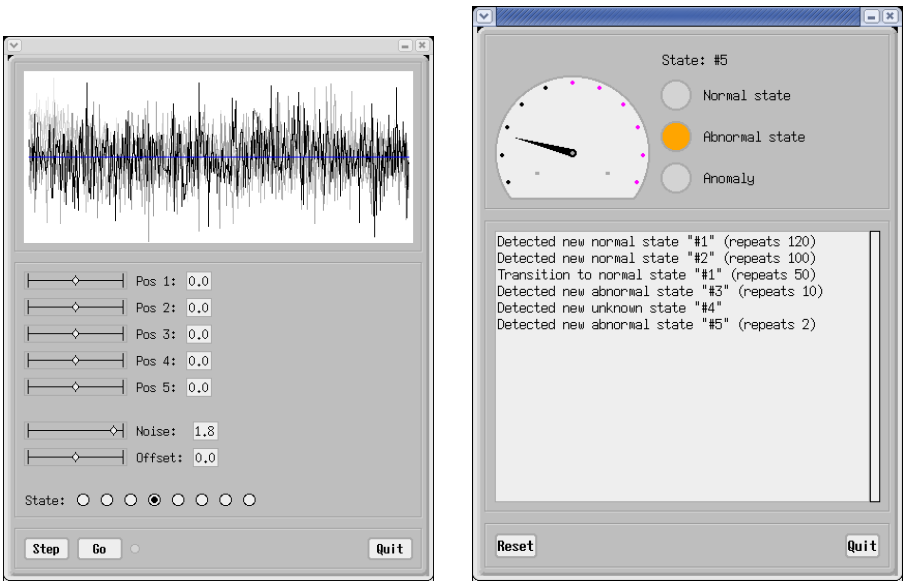

Figure 21: The user chooses to consider the high noise level class to be abnormal and second pattern received from this class is correctly classified

3 Classification and anomaly detection

This section discusses the basic approach of the ISC-tool classification and anomaly detection mechanism. This section also introduces some notation for probability distributions and presents solutions to some definite integrals, that will be referred to in the continuation. The following sections describes the Bayesian inferred formulae implemented in the ISC-tool.

3.1 The classification tasks for the ISC-tool

Concerning the previous section there are basically three tasks of the ISC-tool:

classification deciding which of several clusters a pattern shall belong to, if it is not considered to be anomalous

anomaly detection deciding if a pattern is anomalous or not with respect to a given set of clusters

anomaly assessment assessment of a degree of anomaly to a pattern given a set of clusters (for displaying an anomalo-meter value)

Now, assume that we decide upon how to assess anomaly and detect anomalies, for one cluster, how may we use this to accomplish the three tasks above for a set of clusters ? There are several approaches. For classification, for instance, we may chose between the following. If the pattern is not anomalous then it belongs to the cluster in which it:

(A) has the largest density

(B) is least anomalous

For anomaly detection we have, for instance, the following three approaches. A pattern is anomalous if it is anomalous in:

(i) the most likely cluster.

(ii) each of the clusters.

(iii) the mixture given by taking an equal portion of each cluster.

The approach for classification and anomaly detection in the ISC-tool is (A) and (ii), respectively. In analogy with the approach of anomaly detection, the approach for anomaly assessment is to consider it being the anomaly assessed in the cluster in which a pattern is least anomalous. Moreover, anomaly detection is reduced to choosing an appropriate threshold value for the measured degree of anomaly of a pattern. Hence we can consider the basic tasks in ISC-tool to be just two. Since the approach for classification is (A), the anomaly assessment is not used for the task of classification.

The ISC-tool approach, for a new pattern, can be summarised as follows:

(α) classification by the most likely cluster

(β) anomaly detection by a threshold for assessed anomaly

(γ) assessed anomaly as the minimum assessed anomaly, for all clusters

We observe that there is an incoherence in the approach chosen. Consider for instance the following case. For a given set C of clusters a pattern x is anomalous in all clusters but one, C_1 say. Hence, by approach (β) and (γ), x is not considered anomalous in C and shall therefore belong to one of the C-clusters. Assume that the cluster in which x has the largest density is not C_1 but C_2. By approach (α) we thus shall chose x is to belong to cluster C_2. That is, x is chosen to belong to the cluster C_2 in which it is anomalous and if cluster C_1 is removed from C then x is anomalous and do not belong to C_2 anymore. This incoherence of course a consequence of that the we use non-related measures for pattern classification and pattern anomaly. In the ISC-tool the problem is resolved by modifying approach (α) to (α').

(α') classification of a pattern by the most likely cluster in which the pattern is not anomalous

Hence, in the ISC-tool the pattern x, in the just given example, shall belong to cluster C_1.

3.2 Notation for probability distributions

Table 1 presents the notation for the densities (probability density functions) used in this report. Some of the distribution names in the table are abbreviations: ExpGamma is short for Exponential Gamma, NegBin is short for Negative Binomial and NC-χ^2 is short for Non-Central Chi-square. In the Non-Central Chi-square $I_a(y)$ is a Bessel function of the first kind

$$I_a(y) = \left(\frac{y}{2}\right)^a \sum_{j=0}^{\infty} \frac{\left(y^2/4\right)^j}{j!\Gamma(a+j+1)} \tag{1}$$

In addition to the notation introduced by table above we use $\varphi(x)$ and $\Phi(x)$ to denote the density and cumulative function, respectively, for the standard normal distribution, i.e.$N(0,1)$. That is,

$$\varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} \tag{2}$$

$$\Phi(x) = \int_{-\infty}^{x} \varphi(z)dz \tag{3}$$

Name	Probability density function or Mass function
Beta	$Be(p\|\alpha, r) = \frac{\Gamma(r+\alpha)}{\Gamma(r)\Gamma(\alpha)} p^{r-1}(1-p)^{\alpha-1}$ $p \in [0,1] \quad r, \lambda \in \mathbb{R}_+$
Exponential	$Ex(x\|\lambda) = \lambda e^{-\lambda x} \qquad x, \lambda \in \mathbb{R}_+$
ExpGamma	$Eg(x\|\alpha, s) = \frac{\alpha s^{\alpha}}{(s+x)^{\alpha+1}} \qquad x, \alpha, s \in \mathbb{R}_+$
Gamma	$Ga(x\|r, \lambda) = \frac{\lambda^r}{\Gamma(r)} x^{r-1} e^{-\lambda x} \qquad x, r, \lambda \in \mathbb{R}_+$
NegBin	$Nb(x\|r, p) = p^r \begin{pmatrix} r+x-1 \\ r-1 \end{pmatrix} (1-p)^x$ $x \in \mathbb{N}, \quad r \in \mathbb{Z}_+, \quad p \in [0,1]$
NC-χ^2	$NC^2(x\|\lambda, n) = \frac{1}{2} e^{-(x+\lambda)/2} \left(\frac{x}{\lambda}\right)^{n/4-1/2} I_{n/2-1}\left(\sqrt{\lambda n}\right)$ $x, \lambda, n \in \mathbb{R}_+$
Normal	$N(x\|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(x-\mu)^2} \qquad x, \mu \in \mathbb{R}, \quad \sigma \in \mathbb{R}_+$
Poisson	$Pn(x\|\lambda) = \frac{e^{-\lambda}\lambda^x}{x!} \qquad x \in \mathbb{N}, \quad \lambda \in \mathbb{R}_+$
Student	$St\left(x\|\mu,\, \sigma^2,\, \alpha\right) = \frac{1}{\sqrt{\pi}} \frac{\Gamma\left(\frac{\alpha+1}{2}\right)}{\Gamma\left(\frac{\alpha}{2}\right)} \frac{1}{\sigma\sqrt{\alpha}} \left[1 + \frac{1}{\alpha}\left(\frac{x-\mu}{\sigma}\right)^2\right]^{-(\alpha+1)/2}$ $x, \mu \in \mathbb{R}, \quad \sigma, \alpha \in \mathbb{R}_+$

Table 1: Notation for densities used in this report

In branches of mathematics other than statistics the error function, $erf(x)$, is used rather than $\Phi(x)$, where the relation between $\Phi(x)$ and $erf(x)$ is the following

$$\Phi(x) = \frac{1}{2}\left[1 + erf\left(\frac{x}{\sqrt{2}}\right)\right] \tag{4}$$

For any given density $T(x|\theta)$ we use $X \sim T(\theta)$ to denote that X is a random variable X from a distribution with density $T(x|\theta)$. For instance, $X \sim N(\mu, \sigma^2)$ denotes that X is a random variable X from a distribution with density $N(x|\mu, \sigma^2)$. For a random variable X we often use $p_X(x)$, or just $p(x)$, to denote its known or unknown density and in general we do not distinguish between discrete and continuous random variables. For a discrete random variable, $p(x)$ is just the probability of x. For instance, if $X \sim Pn(\lambda)$ then $p_X = Pn(x|\lambda)$.

3.3 Definite integrals

This section presents some definite used in this report. The following integral, called the gamma integral, is an immediate consequence of the definition of the gamma function. We will often use it together with the theorem: $\Gamma(s+1) = s!$

$$\int_0^\infty x^m e^{-ax} dx = \frac{\Gamma(m+1)}{a^{m+1}} \qquad m > -1, a > 0 \tag{5}$$

The following integral follows from the gamma integral (5), using the substitution u for ax^2.

$$\int_0^\infty x^m e^{-ax^2} dx = \frac{\Gamma\left((m+1)/2\right)}{2a^{(m+1)/2}} \qquad m > -1, a > 0 \tag{6}$$

From this integral in turn we derive the following integral using the substitution u for x^{-1}.

$$\int_0^\infty x^{-m} e^{-ax^{-2}} dx = \frac{\Gamma\left((m-1)/2\right)}{2a^{(m-1)/2}} \qquad m > 1, a > 0 \tag{7}$$

We derive some special cases from the integral (6). Note that the lower limit is $-\infty$ for the following integrals. From $\Gamma(\frac{3}{2}) = \frac{1}{2}\sqrt{\pi}$ and the integral (6) we have

$$\int_{-\infty}^\infty x^2 e^{-\frac{1}{2}x^2} dx = \sqrt{2\pi} \tag{8}$$

From $\Gamma(\frac{5}{2}) = \frac{3}{4}\sqrt{\pi}$ and the integral (6) we have

$$\int_{-\infty}^\infty x^4 e^{-\frac{1}{2}x^2} dx = 3\sqrt{2\pi} \tag{9}$$

From $\Gamma(\frac{1}{2}) = \sqrt{\pi}$ and the integral (6) we have

$$\int_{-\infty}^{\infty} e^{-ax^2} dx = \frac{1}{2}\sqrt{\frac{\pi}{a}} \qquad a > 0 \tag{10}$$

and from this integral, substituting u for $x + b/2a$, we in turn obtain

$$\int_{-\infty}^{\infty} e^{-\left(ax^2+bx+c\right)} dx = e^{b^2/4a-c}\sqrt{\frac{\pi}{a}} \qquad a > 0 \tag{11}$$

3.4 Principal anomaly

The basic idea of how to measure anomaly here is: *the lower density the more anomalous*. Based on this idea this section introduce *principal anomaly*, which we propose is the true measure of anomaly.

Principal anomaly of a pattern z with respect to a random variable X is simply defined as the probability of X having a density greater than or equal to the density in z. Let $A(X, z)$ denote the principal anomaly of a pattern z with respect to X. Then $A(X, z)$ is defined as

$$A(X, z) = P\left(p(X) \geq p(z)\right) \tag{12}$$

where $p(x)$ is the density for the distribution of X. Notice that z is more anomalous the greater $A(X, z)$ is and that $A(X, z) \in [0, 1]$, since principal anomaly is a probability. Notice also that (assuming we accept the idea of lower density as more anomalous) $1 - A(X, z)$ is the probability of a pattern being at least as anomalous as z. The relation of principal anomaly to the likelihood of being anomalous makes it natural to use principal anomaly for defining anomaly, i.e an anomalous pattern is defined as having a principal anomaly above a certain threshold. Let us as an example define anomalous by saying that a pattern z is anomalous if $1 - A(X, z) \leq 0.001$, then the probability for being anomalous, if X is continuous, will be precisely 0.001 and, if X is discrete, that probability is less or equal to 0.001.

Further on in this report we infer conditional or expected anomaly, which is used for the common case of anomaly with respect to a cluster of patterns from a distribution with unknown parameters. This is the case that the ISC-tool encounters.

The principal anomaly is invariant under linear transformations. That is

$$A(aX + b, az + b) = A(X, z) \tag{13}$$

To see that this is true, we use the following theorem on linear transformations of random variables

$$p_{aX+b}(ax + b) = \frac{1}{|a|} p_X(x) \tag{14}$$

For a continuous random variable X with density $p(x)$ the principal anomaly can be written as

$$A(X,z) = \int_{\Omega(z)} p(y)dy, \qquad \Omega(z) = \{y \,|\, p(y) \geq p(z)\} \tag{15}$$

Our concern up till now, in this section, has been both univariate and multivariate distributions. For the special case that X is a random variable of a continuous univariate distribution with a clock shaped symmetric density $p(x)$ with maximum at μ, e.g. the normal distribution, we have

$$\Omega(\mu + z) = \Omega(\mu - z) = \{y \mid \mu - |z| \ \leq \ y \ \leq \ \mu + |z|\} \tag{16}$$

and hence

$$A(X, \mu + z) = \int_{\mu-|z|}^{\mu+|z|} p(y)dy \tag{17}$$

For the special case of a normal distribution, $X \sim N(\mu, \sigma^2)$, we have

$$A(X,z) = \Phi(\frac{|z-\mu|}{\sigma}) - \Phi(-\frac{|z-\mu|}{\sigma}) \tag{18}$$

and equivalently

$$A(X,z) = 2\Phi\left(\frac{|z-\mu|}{\sigma}\right) - 1 \tag{19}$$

Hence, for a random variable X from a normal distribution $A(X,z)$ can be expressed in terms of the error function as follows

$$A(X,z) = erf\left(\frac{|z-\mu|}{\sqrt{2}\sigma}\right) \tag{20}$$

3.5 Deviation

It is not always easy to estimate the principal anomaly. Especially, as in the ISC-tool, where we wish to handle multivariate distributions and where we wish to use as straightforward Bayesian methods as possible to infer the formulae for measuring conditional or expected anomalies given a set of data. Of these reasons another anomaly measure, called deviation, is used in the ISC-tool.

Given a random variable X with density $p(x)$ the deviation $Dev(X,z)$ of a pattern z from X is defined as

$$Dev(X,z) = \frac{E[log\, p(X)] - log\, p(z)}{S[log\, p(X)]} \tag{21}$$

where $S[log\, p(X)] = \sqrt{Var[log\, p(X)]}$ denotes the standard deviation of $log\, p(X)$.

In this section we will derive some properties of deviation that shows that deviation is strongly related to principal anomaly. We will also show that deviation for multivariate distribution with independent variables is easy to obtain from the deviations of the univariate distribution of which the multivariate distribution is composed.

First we observe that $Dev(X, z)$ monotonously increases as $p_X(z)$ decreases. This is a property that $Dev(X, z)$ share with $A(X, z)$ and hence deviation is monotone with respect to the principal anomaly. That is

$$Dev(X, z) \leq Dev(X, w) \Leftrightarrow A(X, z) \leq A(X, w) \tag{22}$$

Another property that deviation share with the principal anomaly is that deviation is invariant under linear transformations. That is

$$Dev(aX + b, az + b) = Dev(X, z) \tag{23}$$

To show this we again use the theorem on linear transformation, equation (14), to obtain

$$log\, p_{aX+b}(ax + b) = log \frac{p_X(x)}{|a|} = log\, p_X(x) - log\, |a| \tag{24}$$

Hence, the random variables $log\, p_{aX+b}(aX + b)$ and $log\, p_X(X)$ only differ by the constant value $-log\, |a|$. From this follows that

$$E\left[log\, p_{aX+b}(aX + b)\right] = E\left[log\, p_X(X)\right] - log\, |a| \tag{25}$$

and

$$S\left[log\, p_{aX+b}(aX + b)\right] = S\left[log\, p_X(X)\right] \tag{26}$$

and from equations (25) and (24) follows that

$$E\left[log\, p_{aX+b}(aX + b)\right] - log\, p_{aX+b}(az + b) = E\left[log\, p(X)\right] - log\, p(z) \tag{27}$$

And from these two last equations it follows that deviation is invariant under linear transformations.

Deviation generalises to multivariate distributions with independent variables in a nice way. Let $X = (X_1, \ldots, X_r)$ be a random variable from a multivariate distribution with independent variables and let X_i have density $p_{X_i}(x)$, then the deviation of a pattern $z = (z_1, \ldots, z_r)$ from X is

$$Dev(X, z) = \frac{\sum_{i=1}^{r} \left(E[log\, p_{X_i}(X_i)] - log\, p_{X_i}(z_i)\right)}{\sqrt{\sum_{i=1}^{r} \left(S[log\, p_{X_i}(X_i)]\right)^2}} \tag{28}$$

By the definition of $Dev(X, z)$, equation (24), we need to show the following in order to show this equation

$$E[log\, p_X(X)] - log\, p_X(z) = \sum_{i=1}^{r} E[log\, p_{X_i}(X_i)] - log\, p_{X_i}(z_i) \tag{29}$$

and

$$S[log\, p(X)] = \sqrt{\sum_{i=1}^{r} \left(S[log\, p_{X_i}(X_i)]\right)^2} \tag{30}$$

We observe that

$$log\, p_X(X) = log\left(\prod_{i=1}^{r} p_{X_i}(X_i)\right) = \sum_{i=1}^{r} log\, p_{X_i}(X_i) \tag{31}$$

hence

$$E[log\, p_X(X)] - log\, p_X(z) = \left(\sum_{i=1}^{r} E[log\, p_{X_i}(X_i)]\right) - \left(\sum_{i=1}^{r} log\, p_{X_i}(z_i)\right) \tag{32}$$

which is equivalent to equation (29). Similarly we have

$$S[log\, p(X)] = \sqrt{Var(log\, p_X(X))} = \sqrt{Var\left(\sum_{i=1}^{r} log\, p_{X_i}(X_i)\right)} \tag{33}$$

and hence

$$S[log\, p(X)] = \sqrt{\sum_{i=1}^{r} Var\left(log\, p_{X_i}(X_i)\right)} \tag{34}$$

which is equivalent to equation (30). Hence, since we have shown equations (29) and (30) it follows that equation (28) holds.

$(z-\mu)/\sigma$	$1-A(X,z)$	$Dev(X,z)$
0	1	-0.70711
1.6449	0.1	1.2060
2.5758	0.01	3.9845
3.2905	0.001	6.9492
3.8905	0.0001	9.9957
1	0.31731	0
2	0.04550	0.70711
∞	0	∞

Table 2: Interpretation of $Dev(X,z)$ in terms of $A(X,z)$ for $X \sim N(\mu,\sigma^2)$

3.6 Deviation for the normal distribution

In this section we show that for a normal random variable X, i.e. $X \sim N(\mu,\sigma^2)$, the following holds

$$Dev(X,z) = \frac{1}{\sqrt{2}}\left[\left(\frac{z-\mu}{\sigma}\right)^2 - 1\right] \tag{35}$$

Observe that $Dev(X,z) = 0$ for $z = \mu \pm \sigma$ as a consequence of this equation. From equation (35) and equation (20) we obtain the following equations for interpreting $A(X,z)$ and $Dev(X,z)$ in terms of each other.

$$A(X,z) = erf\left(\sqrt{\frac{1}{\sqrt{2}}\,Dev(X,z) + \frac{1}{2}}\right) \tag{36}$$

$$Dev(X,z) = \sqrt{2}\left(\left[erf^{-1}\left(A(X,z)\right)\right]^2 - \frac{1}{2}\right) \tag{37}$$

These equations (36) and (37) are important for interpreting the value of $Dev(X,z)$ and for the choice of an anomaly detection threshold.

Table (2) shows z,$1-A(X,z)$ and $Dev(X,z)$ for a sample of values. An example of what can be read off from the table is that if $Dev(X,z) = 7$ then the probability of X being as anomalous as z is just below $1/1000$.

In this section we also show the following generalisation of $Dev(X,z)$ to multivariate normal distributions. Let $X = (X_1,\ldots,X_r)$, where $X_1,\ldots,X_r$ are independent variables from normal distributions, and let $p_{X_i}(x)$ be the density of X_i, then for a pattern $z = (z_1,\ldots,z_r)$ we have

$$Dev(X,z) = \frac{1}{\sqrt{r}} \sum_{i=1}^{r} Dev(X_i, z_i) \tag{38}$$

In order to show equation (35) we shall show that the following holds for the standard normal distribution $Y \sim N(0,1)$

$$Dev(Y,z) = \frac{1}{\sqrt{2}} \left(z^2 - 1\right) \tag{39}$$

Let X be a normal random variable with mean μ and variance σ^2, that is $X \sim N(\mu, \sigma^2)$. Hence,

$$X = \sigma Y + \mu \tag{40}$$

Hence, by (23) and (39) we have

$$Dev(X,z) = Dev\left(Y, \left(\frac{z-\mu}{\sigma}\right)\right) = \frac{1}{\sqrt{2}} \left[\left(\frac{z-\mu}{\sigma}\right)^2 - 1\right] \tag{41}$$

which shows equation (35). Now to show equation (39) observe that

$$Dev(Y,z) = \frac{E[log\,\varphi(X)] - log\,\varphi(z)}{S[log\,\varphi(X)]} \tag{42}$$

From the definitions of $\varphi(x)$ and expected value follows that

$$log\,\varphi(x) = log\,\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} = -\frac{1}{2}\left(log\,(2\pi)\right) - \frac{1}{2}x^2 \tag{43}$$

and

$$E[log\,\varphi(X)] = \int_{-\infty}^{\infty} \left(log\,\varphi(x)\right) \varphi(x) dx \tag{44}$$

hence

$$E[log\,\varphi(X)] = -\frac{1}{2}\left(log\,(2\pi)\right) - \frac{1}{2\sqrt{2\pi}} \int_{-\infty}^{\infty} x^2 e^{-\frac{1}{2}x^2} dx \tag{45}$$

From the integral (8) follows that

$$\frac{1}{2\sqrt{2\pi}} \int_{-\infty}^{\infty} x^2 e^{-\frac{1}{2}x^2} dx = \frac{1}{2} \tag{46}$$

That is

$$E[log\,\varphi(X)] = -\frac{1}{2}\left(log\,(2\pi)\right) - \frac{1}{2} \tag{47}$$

Hence, by (43),

$$E[log\,\varphi(X)] - log\,\varphi(z) = \frac{1}{2}\left(z^2 - 1\right) \tag{48}$$

We have thus simplified the the nominator in equation (42). For the simplification of the denominator in this equation we observe that from integral (8) follows

$$E[X^2] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} x^2 e^{-\frac{1}{2}x^2} dx = 1 \tag{49}$$

and from integral (9) follows

$$E[X^4] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} x^4 e^{-\frac{1}{2}x^2} dx = 3 \tag{50}$$

Hence

$$Var(X^2) = E[X^4] - \left(E[X^2]\right)^2 = 2 \tag{51}$$

That is

$$Var(log\,\varphi(X)) = Var\left(-\frac{1}{2}X^2\right) = \frac{1}{4}Var\left(X^2\right) = \frac{1}{2} \tag{52}$$

and equivalently

$$S[log\,\varphi(X)] = \frac{1}{\sqrt{2}} \tag{53}$$

Hence, from the simplifications of the nominator and denominator in equation (42), equations (48) and (53) respectively, we have

$$Dev(Y, z) = \frac{E[log\,\varphi(X)] - log\,\varphi(z)}{S[log\,\varphi(X)]} = \frac{1}{\sqrt{2}}\left(z^2 - 1\right) \tag{54}$$

and hence we have shown equation (39).

In order to show equation (38) we observe that by equation (26) $S[log\,p_X(X)]$ is the same for all $X \sim N(\mu, \sigma^2)$ and by equation (53) this value is $1/\sqrt{2}$.

$$S[log\,p_X(X)] = 1/\sqrt{2} \qquad \text{for all } X \sim N(\mu, \sigma^2) \tag{55}$$

Hence, from equation (28), follows

$$Dev(X, z) = \frac{\sum_{i=1}^{r} \left(E[log\,p_{X_i}(X_i)] - log\,p_{X_i}(z_i)\right)}{\sqrt{\sum_{i=1}^{r} 1/2}} \tag{56}$$

This equation is equivalent to

$$Dev(X,z) = \frac{1}{\sqrt{r}} \sum_{i=1}^{r} \left(\frac{E[log\, p_{X_i}(X_i)] - log\, p_{X_i}(z_i)}{1/\sqrt{2}} \right) \tag{57}$$

And hence, by equation (55), we have

$$Dev(X,z) = \frac{1}{\sqrt{r}} \sum_{i=1}^{r} \left(\frac{E[log\, p_{X_i}(X_i)] - log\, p_{X_i}(z_i)}{S[log\, p_{X_i}(X_i)]} \right) \tag{58}$$

which is equivalent with equation (38).

4 Bayesian inference

The previous section have shown how to assess anomaly of patterns and classify patterns when the densities are known. This section and the next concerns the more common case that the densities are not known, although their parametric form is assumed to be known, and clusters are given as sets of patterns. For this case we will find counterparts of the concepts introduced in the previous section.

4.1 Expected density

The Bayesian inference presented in this chapter is the same approach as used by for instance: Bernardo (2003) Bayesian Statistics,
"http://www.uv.es/bernardo/BayesStat.pdf".

We assume that $D = \{x_1, \ldots, x_n\}$ are independent samples from a univariate distribution, such that each sample, or pattern, x_i is just a single value. We assume we know of which type the distribution is, i.e. we know the parametric form of the distribution, e.g. a normal distribution, and let θ denote the unknown parameters of the distribution. All our knowledge of the distribution are the parametric form and the sample set D. Our task is to find the density $p(x|D)$, as a function of x. We present the resulting formulae in this section and wait with the explanations till the next. Our approach is to define $p(x|D)$ as the expected density obtained from the samples D.

$$p(x|D) = \int_\theta p(x|\theta)p(\theta|D)d\theta \tag{59}$$

In this equation (59) we will rewrite the posterior density $p(\theta|D)$ using Bayes formula assuming a given prior, possibly improper, density $p(\theta)$

$$p(\theta|D) = \frac{p(D,\theta)}{p(D)} = \frac{p(D|\theta)p(\theta)}{\int_\theta p(D|\theta)p(\theta)d\theta} \tag{60}$$

Inserting equation (60) in equation (59) we get the following formula for the density $p(x|D)$

$$p(x|D) = \frac{\int_\theta p(x|\theta)p(D|\theta)p(\theta)d\theta}{\int_\theta p(D|\theta)p(\theta)d\theta} \tag{61}$$

From the assumption that D consists of independent samples it follows that

$$p(D|\theta) = \prod_{i=1}^{n} p(x_i|\theta) \tag{62}$$

and equation (61) becomes

$$p(x|D) = \frac{\int_\theta p(x|\theta)p(\theta)\left(\prod_{i=1}^n p(x_i|\theta)\right)d\theta}{\int_\theta p(\theta)\left(\prod_{i=1}^n p(x_i|\theta)\right)d\theta} \tag{63}$$

4.2 Observations and explanations

First of all we notice that in equation (59), θ is treated as a random variable. That is, our method is to define the density $p(x|D)$ we search as the expected density, with the parameters θ as random variables. Let's therefore call this method for the *density Bayesian method*. The result of directly and freely taking the uncertainty of θ into account, as the density Bayesian method does, is in general that the obtained density $p(x|D)$ do not have the same parametric form as the distribution from which the samples D were assumed to come. Since the certainty of the parameters θ increases with the number of samples of D, the density $p(x|D)$ will depend on the number of samples of D.

Another Bayesian method is to first estimate the expected values of the unknown parameters θ, using Bayes formula (60), and then put in those estimated values for θ in the parametric form to obtain the density. Let us call this method, the *parameter Bayesian method*. There is a big difference between this method and the density Bayesian in that the density Bayesian method directly take into account our uncertainty about the values of θ, whereas the parameter Bayesian method take into account this uncertainty only through the estimation of the parameters themselves. Hence, in the latter case, the estimated parameters represents both the parameter values and the uncertainty about them. Of course this implies an unsound lack of freedom, i.e. an unsound dependency between of parameter values and the uncertainty about them.

Consider for instance a poisson distribution $Pn(x|\lambda)$, λ being the mean and variance of the distribution. Observing the samples D we wish to chose λ such that $Pn(x|\lambda)$ takes into account the uncertainty about the mean value. Hence, if we at all takes into account this uncertainty, all we can do is to chose a λ that modify the sound choice of mean for our the resulting distribution. This is nothing but a major drawback of the *parameter Bayesian method*.

Let us now consider Bayes formula (60). The formula express that the opinion of θ is changed from the prior $p(\theta)$ to the posterior $p(\theta|D)$, by the observation of the samples of D. The prior represents what is known about θ before observing the samples of D. When we have no such prior knowledge of θ, then a non-informative prior is chosen.

Let us finally consider the notation $p(x|D)$, used in all the equations in the previous section. It is not as direct as it may first look and we must be careful not to missinterpret it. The dependency of D concerns the expected density, as a function of x, via the unknown parameters θ (of a known parametric form). That is, the density at x depends on D through the implicit random variables θ. If θ ceases to be random and becomes a fixed set of values it would follow that $p(x|D) = p(x)$, since we assume that all samples from the distribution are independent from each other.

4.3 Expected density of processes

Let $D = \{(x_1, t_1) \dots (x_n, t_n)\}$ be a given data set of independent samples from a univariate process. This means that the t_i values denotes time intervals and the x_i values are observations during the time interval t_i. That the process is univariate means that each sample, or pattern, x_i is just a single value. That D consists of independent samples means that $D_t = \{t_1, \dots, t_n\}$ are independent samples from some distribution, let's call it the *time distribution*, with parameters τ and that $D_x = \{x_1, \dots x_n\}$ also are independent samples, where each x_i is a sample from a distribution, let us say the *value distribution*, with parameters $\theta(\omega, t_i)$. I.e. the value distribution parameters are functions of both ω and t_i.

The task is, in analogy to the task in the previous section, to find the density $p((x,t)|D,t) = p(x|D,t)$, where the parametric form of the value distributions are known but the *value process parameters* ω are unknown. The time distribution and the *time process parameters* τ are irrelevant for this task. Again our approach is to define $p(x|D,t)$ as the expected density obtained from the samples D, given a prior pseudo distribution $p(\omega)$ of the process parameters.

We observe that x and D_x do not depend on τ. Analogous to the equation (59) and (60) we have, respectively, equations (64) and (65)

$$p(x|D,t) = \int_\omega p(x|t,\omega)p(\omega|D)d\omega \tag{64}$$

$$p(\omega|D) = p(\omega|D_x, D_t) = \frac{p(D_x|D_t,\omega)p(\omega)}{p(D_x|D_t)} = \frac{p(D_x|D_t,\omega)p(\omega)}{\int_\omega p(D_x|D_t,\omega)p(\omega)d\omega} \tag{65}$$

Inserting equation (64) in equation (65) we get the following formula for computing $p(x|D,t)$ analogous to the formula (61) above, for the density function $p(x|D)$.

$$p(x|D,t) = \frac{\int_\omega p(x|t,\omega)p(D_x|D_t,\omega)p(\omega)d\omega}{\int_\omega p(D_x|D_t,\omega)p(\omega)d\omega} \tag{66}$$

From the assumption that D consists of independent samples it follows that

$$p(D_x|D_t,\omega) = \prod_{i=1}^{n} p(x_i|t_i,\omega) \tag{67}$$

Then, in analogy with equation (63), equation (66) becomes

$$p(x|D,t) = \frac{\int_\omega p(x|t,\omega)p(\omega)\left(\prod_{i=1}^{n} p(x_i|t_i,\omega)\right)d\omega}{\int_\omega p(\omega)\left(\prod_{i=1}^{n} p(x_i|t_i,\omega)\right)d\omega} \tag{68}$$

4.4 The statistical modelling in the ISC-tool

In this section we look at how the statistical modelling described in the previous sections are used in the ISC-tool for clustering and anomaly detection.

Recall once again, see section 2.1, that the task of the ISC-tool analyser is to, as correctly as possible, classify patterns. Each new pattern is classified as either belonging to one of the previously observed classes or not belonging to any previously observed class. It is assumed that each of these classes is appropriately described by a probability distribution, that have the same known parametric form for all classes. It may for instance be the case that each pattern consists of 400 values, where we in advance know that each value is a sample from a normal distribution. Although the analyser in the demonstrator assumes patterns from multivariate normal distributions, the ISC-tool in general has the possibility to analyse patterns from any of the distributions: poisson, normal, gamma and chi-square. The ISC-tool assumes no previous knowledge of the patterns other than the parametric form of the class distributions.

The clusters, i.e. the models of the classes, in the ISC-tool are represented by the expected densities, as defined in sections 4.1 and 4.3 and each cluster model is updated whenever a new pattern determined to belong to the cluster.

Given a pattern, the cluster models are used to determine which is the most likely cluster and to assess a degree of anomaly for the pattern with respect to a cluster. The most likely cluster is simply the cluster where the pattern has the largest density, with respect to the given expected densities of the clusters. The degree of anomaly for a pattern with respect to a cluster is the deviation, section 3.5, of the pattern with respect to the cluster. That is, the anomaly degree of a pattern z is $Dev(X, z)$, where X is a random variable that has as density the expected density of the cluster. Classification, anomaly detection and anomaly assessment in the ISC-tool is then done as described in section 3.1.

5 Expected densities for some distributions

In this section we infer the expected densities for clusters of patterns, where these patterns are assumed to be generated from some certain parametric forms, e.g. the normal distribution. The starting point of the derivations in this section are the equations (63) and (68) in sections 4.1 and 4.3 respectively.

5.1 The Expected Poisson density

Assume that $D = \{x_1, \ldots, x_n\}$ are independent samples from a poisson distribution with parameter λ. Using Bayesian inference as described in section 4.1 we will derive the following formula for the density $p(x|D)$, with prior $p(\lambda) = \lambda^{-c}$

$$p(x|D) = \frac{1}{x!} \cdot \frac{\Gamma(x+s-c+1)}{\Gamma(s-c+1)} \cdot \frac{n^{s-c+1}}{(n+1)^{x+s-c+1}} \tag{69}$$

and with s as

$$s = \sum_{i=1}^{n} x_i \tag{70}$$

If c is a natural number then we have, from $\Gamma(s+1) = s!$

$$p(x|D) = \frac{1}{x!} \cdot \frac{(x+s-c)!}{(s-c)!} \cdot \frac{n^{s-c+1}}{(n+1)^{x+s-c+1}} \tag{71}$$

hence

$$p(x|D) = \left(\begin{array}{c} x+s-c \\ s-c \end{array} \right) \left(\frac{n}{n+1} \right)^{s-c+1} \left(1 - \frac{n}{n+1} \right)^{x} \tag{72}$$

That is, $p(x|D)$ is a negative binomial distribution with parameters $s-c+1$ and $n/(n+1)$.

$$p(x|D) = Nb(x|s-c+1, \frac{n}{n+1}) \tag{73}$$

In the ISC-tool we use the prior $1/\lambda$, i.e. $c = 1$. For this case we have

$$p(x|D, c=1) = Nb(x|s, \frac{n}{n+1}) \tag{74}$$

To derive the formula (69) we recall formula (63) which, in the case of a poisson distribution $Pn(x|\lambda)$ with the chosen the prior $p(\lambda) = \lambda^{-c}$, becomes

$$p(x|D) = \frac{\int_0^\infty \lambda^{-c} Pn(x|\lambda) \prod_{i=1}^{n} Pn(x_i|\lambda) d\lambda}{\int_0^\infty \lambda^{-c} \prod_{i=1}^{n} Pn(x_i|\lambda) d\lambda} \tag{75}$$

By the definition of a poisson distribution we have

$$Pn(x|\lambda) = \frac{e^{-\lambda}\lambda^x}{x!} \qquad \prod_{i=1}^{n} Pn(x_i|\lambda) = he^{-n\lambda}\lambda^s \tag{76}$$

where $h = 1/\left(\prod_{i=1}^{n}(x_i!)\right)$. Hence, equation (75) is equivalent to

$$p(x|D) = \frac{\int_0^\infty e^{-\lambda(n+1)}\lambda^{x+s-c}d\lambda}{x!\int_0^\infty e^{-\lambda n}\lambda^{s-c}d\lambda} \tag{77}$$

By the gamma integral, equation (5), we then obtain

$$p(x|D) = \frac{1}{x!}\cdot\frac{\Gamma(x+s-c+1)!}{(n+1)^{x+s-c+1}}\cdot\frac{n^{s-c+1}}{\Gamma(s-c+1)!} \tag{78}$$

which is equivalent to the formula (69) above.

5.2 The Expected Normal density

Assume that $D = \{x_1, \ldots, x_n\}$ are independent samples from a normal distribution with parameters $\left(\mu, \sigma^2\right)$. Given the prior densities $p(\mu) = 1$ and $p(\sigma) = \sigma^{-c}$ we show that we show that the distribution $p(x|D)$ given by Bayesian inference is a student distribution

$$p(x|D) = St\left(x|\tilde{\mu},\, \tilde{\sigma}^2,\, \alpha\right) \tag{79}$$

with parameters $\tilde{\mu}$, $\tilde{\sigma}^2$ and α given by

$$\alpha = n + c - 2 \qquad \tilde{\mu} = \frac{s}{n} \qquad \tilde{\sigma} = \sqrt{\frac{(n+1)\left(r - s^2/n\right)}{n\alpha}} \tag{80}$$

where s and r are the sums

$$s = \sum_{i=1}^{n} x_i \qquad r = \sum_{i=1}^{n} x_i^2 \tag{81}$$

and where the student distribution is defined by

$$St\left(x|\tilde{\mu},\, \tilde{\sigma}^2,\, \alpha\right) = \frac{1}{\sqrt{\pi}}\frac{\Gamma\left(\frac{\alpha+1}{2}\right)}{\Gamma\left(\frac{\alpha}{2}\right)}\frac{1}{\tilde{\sigma}\sqrt{\alpha}}\left[1 + \frac{1}{\alpha}\left(\frac{x-\tilde{\mu}}{\tilde{\sigma}}\right)^2\right]^{-(\alpha+1)/2} \tag{82}$$

In the ISC-tool we we chose $c = 1$, i.e. we use the prior $P(\sigma) = \sigma^{-1}$and hence

$$p(x|D) = St\left(x\,|\,\frac{s}{n},\, \frac{(n+1)\left(r - s^2/n\right)}{n(n-1)},\, n-1\right) \tag{83}$$

As a part of the derivation of the formula (79) we show the following alternative formulation for $p(x|D)$

$$p(x|D) = \sqrt{\frac{n}{\pi(n+1)}} \frac{\Gamma\left(\frac{n+c-1}{2}\right)}{\Gamma\left(\frac{n+c-2}{2}\right)} \frac{\left(r - \frac{s^2}{n}\right)^{\frac{n+c-2}{2}}}{\left(r + x^2 - \frac{(s+x)^2}{(n+1)}\right)^{\frac{n+c-1}{2}}} \tag{84}$$

To show formula (79), recall equation (63) in section 4.1, which in this case of a normal distribution $N(x|\mu, \sigma^2)$ becomes

$$p(x|D) = \frac{\int_{-\infty}^{\infty} \int_0^{\infty} \sigma^{-c} N(x|\mu, \sigma^2) \prod_{i=1}^{n} N(x_i|\mu, \sigma^2) d\mu d\sigma}{\int_{-\infty}^{\infty} \int_0^{\infty} \sigma^{-c} \prod_{i=1}^{n} N(x_i|\mu, \sigma^2) d\mu d\sigma} \tag{85}$$

Let for simplicity $q = 1/\left(2\sigma^2\right)$. Then, by the definition of a normal distribution we have

$$N(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} e^{-q(x-\mu)^2} = \frac{1}{\sqrt{2\pi}\sigma} e^{-q\left(x^2 - 2\mu x + \mu^2\right)} \tag{86}$$

$$\prod_{i=1}^{n} N(x_i|\mu, \sigma^2) = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n e^{-\sum_i q(x_i-\mu)^2} = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n e^{-q\left(r - 2s\mu + n\mu^2\right)} \tag{87}$$

Inserting this in equation (85) we have

$$p(x|D) = \frac{\int_0^{\infty} \sigma^{-(n+c+1)} \left(\int_{-\infty}^{\infty} e^{-q\left((n+1)\mu^2 - 2(s+x)\mu + r + x^2\right)} d\mu\right) d\sigma}{\sqrt{2\pi} \int_0^{\infty} \sigma^{-(n+c)} \left(\int_{-\infty}^{\infty} e^{-q((n\mu^2 - 2s\mu + r)} d\mu\right) d\sigma} \tag{88}$$

The integral in equation (11) is used to solve the inner integrals in this equation. We obtain

$$p(x|D) = \frac{\int_0^{\infty} \sigma^{-(n+c+1)} e^{q\left((s+x)^2/(n+1) - r - x^2\right)} \sqrt{\frac{\pi}{q(n+1)}} d\sigma}{\sqrt{2\pi} \int_0^{\infty} \sigma^{-(n+c)} e^{q(s^2/n - r)} \sqrt{\frac{\pi}{qn}} d\sigma} \tag{89}$$

Let us use g and h defined by

$$g = r + x^2 - \frac{(s+x)^2}{(n+1)} \qquad h = r - \frac{s^2}{n} \tag{90}$$

Using this and $q = 1/\left(2\sigma^2\right)$ in equation (89) and simplifying we have

$$p(x|D) = \frac{\sqrt{\frac{1}{(n+1)}} \int_0^{\infty} \sigma^{-(n+c)} e^{-\frac{1}{2} g \sigma^{-2}} d\sigma}{\sqrt{\frac{2\pi}{n}} \int_0^{\infty} \sigma^{-(n+c-1)} e^{-\frac{1}{2} h \sigma^{-2}} d\sigma} \tag{91}$$

To solve these integrals we use the integral in equation (7). We obtain the following formula for $P(x|D)$

$$p(x|D) = \left(\sqrt{\frac{1}{(n+1)}}\frac{2^{\frac{n+c-1}{2}}\Gamma\left(\frac{n+c-1}{2}\right)}{2g^{\frac{n+c-1}{2}}}\right) / \left(\sqrt{\frac{2\pi}{n}}\frac{2^{\frac{n+c-2}{2}}\Gamma\left(\frac{n+c-2}{2}\right)}{2h^{\frac{n+c-2}{2}}}\right) \tag{92}$$

which simplifies to

$$p(x|D) = \frac{1}{\sqrt{\pi}}\sqrt{\frac{n}{(n+1)}}\frac{\Gamma\left(\frac{n+c-1}{2}\right)}{\Gamma\left(\frac{n+c-2}{2}\right)}\frac{h^{\frac{n+c-2}{2}}}{g^{\frac{n+c-1}{2}}} \tag{93}$$

This is in fact the formula (84). We continue by rewriting g as follows

$$g = r + x^2 - \frac{(s+x)^2}{(n+1)} \tag{94}$$

$$g = r + \frac{x^2(n+1) - (s+x)^2}{(n+1)} \tag{95}$$

$$g = r + \frac{nx^2 - 2sx - s^2}{(n+1)} \tag{96}$$

$$g = r + \frac{1}{(n+1)}\left(n\left(x - \frac{s}{n}\right)^2 - \frac{s^2}{n} - s^2\right) \tag{97}$$

$$g = r - \frac{s^2}{n} + \frac{n}{(n+1)}\left(x - \frac{s}{n}\right)^2 \tag{98}$$

We use this together with the defined parameters

$$\alpha = n + c - 2 \qquad \tilde{\mu} = \frac{s}{n} \qquad \tilde{\sigma} = \sqrt{\frac{(n+1)\left(r - s^2/n\right)}{n\alpha}} \tag{99}$$

to find the following formulae for h and h/g

$$h = r - \frac{s^2}{n} = \frac{\tilde{\sigma}^2\alpha n}{(n+1)} \tag{100}$$

$$\frac{h}{g} = \frac{r - \frac{s^2}{n}}{r - \frac{s^2}{n} + \frac{n}{(n+1)}\left(x - \frac{s}{n}\right)^2} = \frac{1}{1 + \frac{n}{(n+1)}\frac{(x-s/n)^2}{r - s^2/n}} = \frac{1}{1 + \frac{(x-\mu)^2}{\tilde{\sigma}^2\alpha}} \tag{101}$$

that is

$$\frac{h}{g} = \left[1 + \frac{1}{\alpha}\left(\frac{x-\mu}{\tilde{\sigma}}\right)^2\right]^{-1} \tag{102}$$

Let us now insert $\alpha = n + c - 2$ in equation (93) to obtain.

$$p(x|D) = \frac{1}{\sqrt{\pi}}\frac{\Gamma\left(\frac{\alpha+1}{2}\right)}{\Gamma\left(\frac{\alpha}{2}\right)}\sqrt{\frac{n}{(n+1)}}\frac{h^{\frac{\alpha}{2}}}{g^{\frac{\alpha+1}{2}}} \tag{103}$$

that is

$$p(x|D) = \frac{1}{\sqrt{\pi}} \frac{\Gamma\left(\frac{\alpha+1}{2}\right)}{\Gamma\left(\frac{\alpha}{2}\right)} \sqrt{\frac{n}{(n+1)}} \frac{1}{\sqrt{h}} \left(\frac{h}{g}\right)^{\frac{\alpha+1}{2}} \tag{104}$$

Inserting the derived formulae for h and h/g, equations (100) and (102) respectively we obtain

$$p(x|D) = \frac{1}{\sqrt{\pi}} \frac{\Gamma\left(\frac{\alpha+1}{2}\right)}{\Gamma\left(\frac{\alpha}{2}\right)} \frac{1}{\tilde{\sigma}\sqrt{\alpha}} \left[1 + \frac{1}{\alpha}\left(\frac{x-\mu}{\tilde{\sigma}}\right)^2\right]^{-\frac{\alpha+1}{2}} \tag{105}$$

Hence $p(x|D) = St\left(x|\tilde{\mu},\, \tilde{\sigma}^2,\, \alpha\right)$.

5.3 The Expected Gamma density

Assume that $D = \{x_1, \ldots, x_n\}$ are independent samples from a gamma distribution

$$Ga(x|r,\lambda) = \frac{\lambda^r}{\Gamma(r)} x^{r-1} e^{-\lambda x} \qquad x, r, \lambda \in \mathbb{R}_+ \tag{106}$$

where r is a given fixed number. In this section we show that Bayesian inference, assuming the prior density $p(\lambda) = \lambda^{-c}$ gives the following density

$$p(x|D) = \frac{\Gamma(r+\alpha)}{\Gamma(r)\Gamma(\alpha)} \frac{x^{r-1} s^\alpha}{(s+x)^{\alpha+r}} \tag{107}$$

with

$$\alpha = rn - c + 1 \qquad s = \sum_{i=1}^{n} x_i \tag{108}$$

It is also possible to express$P(x|D)$ in terms of the density for the beta distribution

$$p(x|D) = \frac{s}{(s+x)^2} Be\left(1 - \frac{s}{s+x}|\alpha, r\right) \tag{109}$$

To show formula (107), again recall equation (63) in section 4.1, which in this case of a gamma distribution $Ga(x|r,\lambda)$ with prior $p(\lambda) = \lambda^{-c}$ becomes

$$p(x|D) = \frac{\int_0^\infty Ga(x|r,\lambda) \prod_{i=1}^n Ga(x_i|r,\lambda)\lambda^{-c} d\lambda}{\int_0^\infty \prod_{i=1}^n Ga(x_i|r,\lambda)\lambda^{-c} d\lambda} \tag{110}$$

We have

$$\prod_{i=1}^{n} Ga(x_i|r,\lambda) = \gamma\lambda^{rn}e^{-\lambda s} \tag{111}$$

with

$$\gamma = \frac{\prod_{i=1}^{n} x_i^{r-1}}{(\Gamma(r))^n} \tag{112}$$

That is, equation (110) is equivalent with

$$p(x|D) = \frac{\int_0^\infty \frac{\lambda^r}{\Gamma(r)} x^{r-1}e^{-\lambda x}\gamma\lambda^{rn}e^{-\lambda s}\lambda^{-c}d\lambda}{\int_0^\infty \gamma\lambda^{rn}e^{-\lambda s}\lambda^{-c}d\lambda} \tag{113}$$

which simplifies to

$$p(x|D) = \frac{x^{r-1}}{\Gamma(r)} \frac{\int_0^\infty \lambda^{\alpha+r-1}e^{-\lambda(s+x)}d\lambda}{\int_0^\infty \lambda^{\alpha-1}e^{-\lambda s}d\lambda} \tag{114}$$

These integrals are solved using the gamma integral (5) giving us equation (107). The beta distribution can be used to check that the derived density really is a probability distribution, i.e $\int_0^\infty p(x|D)dx = 1$.

5.4 The Expected Exponential density

The density function for the exponential distribution

$$Ex(x|\lambda) = \lambda e^{-\lambda x} \quad x,\lambda \in \mathbb{R}_+ \tag{115}$$

is a special case of the gamma distribution

$$Ex(x|\lambda) = Ga(x|1,\lambda) \tag{116}$$

Hence, assuming that $D = \{x_1,\ldots,x_n\}$ are independent samples from a exponential distribution and assuming the prior density $p(\lambda) = \lambda^{-c}$ we use equation (109) to obtain

$$p(x|D) = \frac{s}{(s+x)^2} Be\left(\frac{s}{s+x}|\alpha,1\right) = \frac{\alpha s^\alpha}{(s+x)^{\alpha+1}} = Eg(x|\alpha,s) \tag{117}$$

with

$$\alpha = n - c + 1 \qquad s = \sum_{i=1}^{n} x_i \tag{118}$$

and where $Eg(x|\alpha,r)$ is a exponential gamma distribution.

5.5 The Expected Non-central Chi-square density

Non-central chi-square distribution is defined by

$$NC^2(x|\lambda,k) = \frac{1}{2}e^{-(x+\lambda)/2}\left(\frac{x}{\lambda}\right)^{k/4-1/2} I_{k/2-1}\left(\sqrt{\lambda k}\right) \qquad x,\lambda,k \in \mathbb{R}_+ \quad (119)$$

where $I_a(y)$ is a Bessel function of the first kind

$$I_a(y) = \left(\frac{y}{2}\right)^a \sum_{j=0}^{\infty} \frac{\left(y^2/4\right)^j}{j!\Gamma(a+j+1)} \quad (120)$$

Assuming that $D = \{x_1,\ldots,x_n\}$ are independent samples from a non-central chi-square distribution with one degree of freedom $NC^2(x|\lambda,1)$, and assuming the prior distribution $p(\lambda) = \lambda^{-c}$ it can be shown that

$$P(x \mid D) = \frac{\left(\frac{s}{n}\right)^a \int_p p^{2b} K_{2b}\left(p\sqrt{(n+1)(s+x)}\right)\cosh\left(p\sqrt{x}\right)h(p)dp}{\left(\frac{s+x}{n+1}\right)^b \sqrt{2\pi x}\int_p p^{2a}K_{2a}\left(p\sqrt{ns}\right)h(p)dp} \quad (121)$$

where

$$s = \sum_{i=1}^{n} x_i \qquad a = \frac{n-2}{4} \qquad b = \frac{n-1}{4} \qquad h(p) = \prod_{i=1}^{n}\cosh(p\sqrt{x_i}) \quad (122)$$

and where the Bessel function $K_\nu(z)$ is defined by

$$K_\nu(z) = \frac{1}{2}\pi\frac{I_{-\nu}(z) - I_\nu(z)}{\sin(\nu\pi)} \quad (123)$$

5.6 The Expected Poisson processes

In this section let $D = \{(x_1,t_1)\ldots(x_n,t_n)\}$ be a given set of samples from a poisson process. In this case the value process parameter, denoted λ, is called *rate* and each x_i is a sample from a poisson distribution with parameter λt_i. That is, the probability of x_i is $Pn(x_i|\lambda t_i)$.

$$Pn(x_i|\lambda t_i) = e^{-\lambda t_i}\frac{(\lambda t_i)^{x_i}}{x_i!} \quad (124)$$

According to the definition of $p(x|D,t)$ in section 4.3 we derive the following formula for the probability $p(x|D,t)$, for the prior $p(\lambda) = \lambda^{-c}$

$$p(x|D,t) = \frac{t^x}{x!}\cdot\frac{\Gamma(x+s-c+1)}{\Gamma(s-c+1)}\cdot\left(\frac{u}{t+u}\right)^{s-c+1}\left(\frac{t}{t+u}\right)^x \quad (125)$$

where

$$s = \sum_{i=1}^{n} x_i \qquad u = \sum_{i=1}^{n} t_i$$

If c is an integer then

$$p(x|D,t) = \begin{pmatrix} x+s-c \\ s-c \end{pmatrix} \left(\frac{u}{t+u}\right)^{s-c+1} \left(1 - \frac{u}{t+u}\right)^{x} \tag{126}$$

Hence, in this case $p(x|D,t)$ is a negative binomial distribution

$$p(x|D,t) = Nb(x|s-c+1, \frac{u}{t+u}) \tag{127}$$

In the ISC-tool we use the prior $p(\lambda) = \lambda^{-1}$ which hence results in

$$p(x|D,t) = Nb(x|s, \frac{u}{t+u}) \tag{128}$$

To derive equation (125) we recall equation (68) in section **??**, which in case of a poisson process with prior $P(\lambda) = \lambda^{-c}$ becomes

$$P(x|D,t) = \frac{\int_0^\infty \lambda^{-c} Pn(x|\lambda t) \prod_{i=1}^{n} Pn(x_i|\lambda t_i) d\lambda}{\int_0^\infty \lambda^{-c} \prod_{i=1}^{n} Pn(x_i|\lambda t_i) d\lambda} \tag{129}$$

Let

$$r = \prod_{i=1}^{n} (t_i^{x_i}/x_i!) \tag{130}$$

and we have

$$P(x|\lambda t) = \frac{e^{-\lambda t}(\lambda t)^x}{x!} \qquad \prod_{i=1}^{n} Pn(x_i|\lambda t_i) = re^{-u\lambda}\lambda^s$$

We insert this in equation (129) and simplify

$$P(x|D,t) = \frac{t^x}{x!} \frac{\int_0^\infty e^{-\lambda(t+u)} \lambda^{x+s-c} d\lambda}{\int_0^\infty e^{-\lambda u} \lambda^{s-c} d\lambda} \tag{131}$$

As in the case of a poisson distribution, section 5.1, we use the gamma integral, equation (5), to solve each of the integrals and obtain

$$p(x|D,t) = \frac{t^x}{x!} \cdot \frac{\Gamma(x+s-c+1)}{\Gamma(s-c+1)} \cdot \frac{u^{s-c+1}}{(t+u)^{x+s-c+1}} \tag{132}$$

which is equivalent with equation (125).

5.7 The Expected Normal process

Let $D = \{(x_1, t_1) \dots (x_n, t_n)\}$ be samples from a normal process, that is each x_i is a sample from a normal distribution $N(x|\mu t, \sigma^2 t)$. Given the prior densities $p(\mu) = 1$ and $p(\sigma) = \sigma^{-c}$ we show that $p(x|D, t)$ is a student distribution

$$p(x|D,t) = St\left(x \,|\, \tilde{\mu},\ \tilde{\sigma}^2,\ \alpha\right) \tag{133}$$

with parameters $\tilde{\mu}$, $\tilde{\sigma}^2$ and α given by

$$\alpha = n + c - 2 \qquad \tilde{\mu} = \frac{ts}{y} \qquad \tilde{\sigma} = \sqrt{\frac{t(t+y)\left(r - s^2/y\right)}{y\alpha}} \tag{134}$$

where s, r and y are the sums

$$s = \sum_{i=1}^{n} x_i \qquad r = \sum_{i=1}^{n} x_i^2 \qquad y = \sum_{i=1}^{n} t_i \tag{135}$$

Recall that the student distribution, with parameters $\tilde{\mu}$, $\tilde{\sigma}^2$ and α is defined by

$$St\left(x \,|\, \tilde{\mu},\ \tilde{\sigma}^2,\ \alpha\right) = \frac{1}{\sqrt{\pi}} \frac{\Gamma\left(\frac{\alpha+1}{2}\right)}{\Gamma\left(\frac{\alpha}{2}\right)} \frac{1}{\tilde{\sigma}\sqrt{\alpha}} \left[1 + \frac{1}{\alpha}\left(\frac{x - \tilde{\mu}}{\tilde{\sigma}}\right)^2\right]^{-(\alpha+1)/2} \tag{136}$$

As a part of the derivation of the formula (133) we show the following alternative formulation for $p(x|D, t)$

$$p(x|D,t) = \frac{1}{\sqrt{\pi}} \frac{\Gamma\left(\frac{n+c-1}{2}\right)}{\Gamma\left(\frac{n+c-2}{2}\right)} \sqrt{\frac{y}{t(y+t)}} \frac{\left(r - \frac{s^2}{y}\right)^{\frac{n+c-2}{2}}}{\left(r + \frac{x^2}{t} - \frac{(s+x)^2}{(y+t)}\right)^{\frac{n+c-1}{2}}} \tag{137}$$

To show formula (133), recall equation (68) in section 4.3, which in this case of a normal process becomes

$$p(x|D,t) = \frac{\int_{-\infty}^{\infty} \int_{0}^{\infty} \sigma^{-c} N(x|t\mu, t\sigma^2) \prod_{i=1}^{n} N(x_i|t_i\mu, t_i\sigma^2) d\mu d\sigma}{\int_{-\infty}^{\infty} \int_{0}^{\infty} \sigma^{-c} \prod_{i=1}^{n} N(x_i|t_i\mu, t_i\sigma^2) d\mu d\sigma} \tag{138}$$

Let for simplicity

$$q = \frac{1}{2\sigma^2} \qquad z = \prod_{i=1}^{n} \sqrt{t_i} \tag{139}$$

Then, by the definition of a normal distribution we have

$$N(x|t\mu, t\sigma^2) = \frac{1}{\sqrt{2\pi t}\sigma} e^{-q(x-\mu t)^2/t} \tag{140}$$

$$\prod_{i=1}^{n} N(x_i|t_i\mu, t_i\sigma^2) = \frac{1}{z}\left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n e^{-\sum_i q(x_i-\mu t_i)^2/t_i} \tag{141}$$

equivalent to (140) and (141) are (142) and (143), respectively

$$N(x|t\mu, t\sigma^2) = \frac{1}{\sqrt{2\pi t}\sigma} e^{-q\left(x^2/t - 2\mu x + \mu^2 t\right)} \tag{142}$$

$$\prod_{i=1}^{n} N(x_i|t_i\mu, t_i\sigma^2) = \frac{1}{z}\left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n e^{-q\left(r - 2s\mu + y\mu^2\right)} \tag{143}$$

Inserting this in equation (138) we have

$$p(x|D,t) = \frac{1}{\sqrt{2\pi t}} \frac{\int_0^\infty \sigma^{-(n+c+1)} \left(\int_{-\infty}^\infty e^{-q\left((y+t)\mu^2 - 2(s+x)\mu + r + x^2/t\right)} d\mu\right) d\sigma}{\int_0^\infty \sigma^{-(n+c)} \left(\int_{-\infty}^\infty e^{-q((y\mu^2 - 2s\mu + r)} d\mu\right) d\sigma} \tag{144}$$

We solve the inner integrals again using the integral of equation (11). This gives us

$$p(x|D,t) = \frac{1}{\sqrt{2\pi t}} \frac{\int_0^\infty \sigma^{-(n+c+1)} e^{q\left((s+x)^2/(y+t) - r - x^2/t\right)} \sqrt{\frac{\pi}{q(y+t)}} d\sigma}{\int_0^\infty \sigma^{-(n+c)} e^{q(s^2/y - r)} \sqrt{\frac{\pi}{qy}} d\sigma} \tag{145}$$

Let

$$g = r + \frac{x^2}{t} - \frac{(s+x)^2}{(y+t)} \qquad h = r - \frac{s^2}{y} \tag{146}$$

Inserting this and $q = 1/\left(2\sigma^2\right)$ and simplifying we have

$$p(x|D,t) = \frac{1}{\sqrt{2\pi}} \sqrt{\frac{y}{t(y+t)}} \frac{\int_0^\infty \sigma^{-(n+c)} e^{-\frac{1}{2}g\sigma^{-2}} d\sigma}{\int_0^\infty \sigma^{-(n+c-1)} e^{-\frac{1}{2}h\sigma^{-2}} d\sigma} \tag{147}$$

The solution of these integrals are given by the integral in equation (7) from which we have

$$p(x|D,t) = \frac{1}{\sqrt{2\pi}} \sqrt{\frac{y}{t(y+t)}} \frac{\left(\frac{2^{\frac{n+c-1}{2}}\Gamma\left(\frac{n+c-1}{2}\right)}{2g^{\frac{n+c-1}{2}}}\right)}{\left(\frac{2^{\frac{n+c-2}{2}}\Gamma\left(\frac{n+c-2}{2}\right)}{2h^{\frac{n+c-2}{2}}}\right)} \tag{148}$$

This is simplified to

$$p(x|D,t) = \frac{1}{\sqrt{\pi}}\sqrt{\frac{y}{t(y+t)}}\frac{\Gamma\left(\frac{n+c-1}{2}\right)}{\Gamma\left(\frac{n+c-2}{2}\right)}\frac{h^{\frac{n+c-2}{2}}}{g^{\frac{n+c-1}{2}}} \tag{149}$$

which is the formula (137). We continue by rewriting g as follows

$$g = r + \frac{x^2}{t} - \frac{(s+x)^2}{(y+t)} \tag{150}$$

$$g = r + \frac{x^2(y+t) - t(s+x)^2}{t(y+t)} \tag{151}$$

$$g = r + \frac{yx^2 - 2tsx - ts^2}{t(y+t)} \tag{152}$$

$$g = r + \frac{1}{t(y+t)}\left(y\left(x - \frac{ts}{y}\right)^2 - \frac{t^2s^2}{y} - ts^2\right) \tag{153}$$

$$g = r - \frac{s^2}{y} + \frac{y}{t(y+t)}\left(x - \frac{ts}{y}\right)^2 \tag{154}$$

Now, since

$$\alpha = n + c - 2 \qquad \tilde{\mu} = \frac{ts}{y} \qquad \tilde{\sigma} = \sqrt{\frac{t(t+y)\left(r - s^2/y\right)}{y\alpha}} \tag{155}$$

we can use equation (154) to find the following formulae for h and h/g

$$h = r - \frac{s^2}{y} = \frac{\tilde{\sigma}^2\alpha y}{t(t+y)} \tag{156}$$

$$\frac{h}{g} = \frac{r - \frac{s^2}{y}}{r - \frac{s^2}{y} + \frac{y}{t(y+t)}\left(x - \frac{ts}{y}\right)^2} = \frac{1}{1 + \frac{y}{t(y+t)}\frac{(x - ts/y)^2}{r - s^2/y}} = \frac{1}{1 + \frac{(x-\mu)^2}{\tilde{\sigma}^2\alpha}} \tag{157}$$

that is

$$\frac{h}{g} = \left[1 + \frac{1}{\alpha}\left(\frac{x-\mu}{\tilde{\sigma}}\right)^2\right]^{-1} \tag{158}$$

Let us now insert $\alpha = n + c - 2$ in equation (149) to obtain.

$$p(x|D,t) = \frac{1}{\sqrt{\pi}}\frac{\Gamma\left(\frac{\alpha+1}{2}\right)}{\Gamma\left(\frac{\alpha}{2}\right)}\sqrt{\frac{y}{t(y+t)}}\frac{h^{\frac{\alpha}{2}}}{g^{\frac{\alpha+1}{2}}} \tag{159}$$

that is

$$p(x|D,t) = \frac{1}{\sqrt{\pi}} \frac{\Gamma\left(\frac{\alpha+1}{2}\right)}{\Gamma\left(\frac{\alpha}{2}\right)} \sqrt{\frac{y}{t(y+t)}} \frac{1}{\sqrt{h}} \left(\frac{h}{g}\right)^{\frac{\alpha+1}{2}} \tag{160}$$

Inserting the derived formulae for h and h/g, equations (156) and (158) respectively we obtain

$$p(x|D,t) = \frac{1}{\sqrt{\pi}} \frac{\Gamma\left(\frac{\alpha+1}{2}\right)}{\Gamma\left(\frac{\alpha}{2}\right)} \frac{1}{\tilde{\sigma}\sqrt{\alpha}} \left[1 + \frac{1}{\alpha}\left(\frac{x-\mu}{\tilde{\sigma}}\right)^2\right]^{-\frac{\alpha+1}{2}} \tag{161}$$

Hence $p(x|D,t) = St\left(x \,|\, \tilde{\mu},\, \tilde{\sigma}^2,\, \alpha\right)$.

Part II

On Bayesian clustering and anomaly assessment

6 Introduction

Part I of this report concerns incremental stream clustering as it is implemented in the ISC-tool. The ISC-tool realises one feasible way to use Bayesian statistics to manage clustering and anomaly detection.

In this second part of the report we are concerned with what ought to be a more accurate statistical approach for clustering and anomaly detection, although we do not know how to or if it is possible to use these approaches in practice.

7 Anomaly assessment

In this section we use Bayesian statistics to infer the expected principal anomaly in a way similar to how the expected density was inferred in section 4.1. We propose that the expected principal anomaly is the true anomaly assessment for a pattern given a clustering of previously observed patterns.

We assume that $D = \{x_1, \ldots, x_n\}$ are independent samples from a univariate distribution. Our task is to find the anomaly $A(X, z|D)$ assuming we know the parametric form of the distribution of the samples. Let θ denote the unknown parameters of the distribution. The approach is to define $A(X, z|D)$ as the expected principal anomaly. We define $A(X, z|D)$ in the same way as the expected density is defined, equation (59). That is

$$A(X, z|D) = \int_\theta A(X, z|\theta)p(\theta|D)d\theta \tag{162}$$

We use equation (60) again to express the posterior $p(\theta|D)$ in terms of the prior $p(\theta)$ to obtain

$$A(X, z|D) = \frac{\int_\theta A(X, z|\theta)p(D|\theta)p(\theta)d\theta}{\int_\theta p(D|\theta)p(\theta)d\theta} \tag{163}$$

For example for the normal distribution it follows from equation (20) that

$$A(X, z|D) = \frac{\int_{\mu,\sigma} erf\left(|z-\mu|/\sqrt{2}\sigma\right) p(D|\mu, \sigma)p(\mu, \sigma)d\mu d\sigma}{\int_{\mu,\sigma} p(D|\mu, \sigma)p(\mu, \sigma)d\mu d\sigma} \tag{164}$$

8 Classification

8.1 The probability of belonging to a class

This section concerns the definition of the probability of belonging to a class under the condition that the class is given by a probability with known density.

Remember that classification is to decide which of several classes a pattern shall belong to, if it is not considered to be anomalous. Remember also that the approach in the ISC-tool for classification is to chose the class in which the pattern has the largest density. We consider this to be the correct way to classify patterns under the condition that the densities are known.

Let $C_1, \ldots, C_m$ be a set of classes represented by distributions with densities $p_1, \ldots, p_m$ respectively. Given a pattern z the ISC-tool classification is to chose the class C_i for which is $p_i(z)$ largest. That is, we chose the class C_i for which

$$p_i(z) = \max_i(p_i(z)) \tag{165}$$

Consider the mixture given by taking an equal portion of each class. Let $\pi_i(z)$ be the probability that a pattern z belongs to class C_i, assuming that z belongs to the mixture. That is,

$$\pi_i(z) = P(z \in C_i \,|\, z \in \bigcup_{i=1}^{m} C_i) \tag{166}$$

Hence, by the definition of conditional probability,

$$\pi_i(z) = \frac{p_i(z)}{\sum_{i=1}^{m} p_i(z)} \tag{167}$$

We observe that the ISC-tool classification is equivalent to choosing the most likely class in the mixture of equal portions of classes. That is

$$p_i(z) = \max_i(p_i(z)) \;\Leftrightarrow\; \pi_i(z) = \max_i(\pi_i(z)) \tag{168}$$

8.2 Bayesian classification

In this section we again consider the common situation that some previously observed patterns are divided into clusters, where each cluster is a set of samples of a class. This class in turn is appropriately described as a probability distribution with known parametric form but with unknown parameters. The classification task is to find the most likely cluster for a given new pattern. We propose that true classification, in this case, is the following classification:

> *A new pattern belongs to the cluster in which it has the maximum expected probability of belonging to.*

Based on this view of true classification we infer some classification formulae in this section. Let thus D be a set of patterns divided into m separate clusters $D_1, \ldots, D_m$ such that

$$D = \{D_1, \ldots, D_m\} \qquad D_j = \left\{x_{(j,1)}, \ldots x_{(j,n_j)}\right\} \tag{169}$$

We let $\pi_j(z|D)$ denote the expected probability for the pattern z to belong to cluster D_j and hence the true classification is

$$z \in D_j \;\Rightarrow\; \pi_j(z|D) = \max_j(\pi_j(z|D)) \tag{170}$$

We assume that the patterns of each cluster D_j are samples from a distribution with density $p(x|\theta_j)$ with known parametric form and with unknown parameters θ_j and let $\pi_j(z|\theta)$ be the probability for the pattern z to belong to cluster D_j, where $\theta = \{\theta_1, \ldots, \theta_m\}$. We also assume that the patterns of D are independent samples from the cluster distributions. In analogy with the definition of expected density, equation (59), we define the expected probability $\pi_j(z|D)$ for the pattern z to belong to cluster D_j as follows

$$\pi_j(z|D) = \int_\theta \pi_j(z|\theta)p(\theta|D)d\theta \tag{171}$$

Hence, using equation (60) to express the posterior $p(\theta|D)$ in terms of the prior $p(\theta)$ we obtain

$$\pi_j(z|D) = \frac{\int_\theta \pi_j(z|\theta)p(D|\theta)p(\theta)d\theta}{\int_\theta p(D|\theta)p(\theta)d\theta} \tag{172}$$

The denominator of the right hand side in this equation is irrelevant for the classification task since it the same for all $\pi_j(z|D)$. We may therefore omit it. That is, let $\alpha_j(z|D)$ be the nominator of the right hand side of the above equation

$$\alpha_j(z|D) = \int_\theta \pi_j(z|\theta)p(D|\theta)p(\theta)d\theta \tag{173}$$

Then the following classification is equivalent to the classification (170)

$$z \in D_j \;\Rightarrow\; \alpha_j(z|D) = \max_j(\alpha_j(z|D)) \tag{174}$$

By equation (167) we have

$$\pi_j(z|\theta) = \frac{p(z|\theta_j)}{\sum_{j=1}^m p(z|\theta_j)} \tag{175}$$

Inserting this in in equation (173) we obtain

$$\alpha_j(z|D) = \int_\theta \frac{p(z|\theta_j)p(\theta)p(D|\theta)}{\sum_{j=1}^{m} p(z|\theta_j)} d\theta \tag{176}$$

From the assumption that D consists of independent samples follows that $p(D|\theta)$ is the product of the sample densities. That is

$$p(D|\theta) = \prod_{j=1}^{m} p(D_j|\theta_j) = \prod_{j=1}^{m} \prod_{k=1}^{n_j} p_j(x_{(j,k)}|\theta_j) \tag{177}$$

For $j = 2$ and for case that the parametric form is the normal distribution we have

$$p(z|\theta_j) = \frac{1}{\sqrt{2\pi}\sigma_j} e^{-\frac{1}{2}\left(\frac{z-\mu_j}{\sigma_j}\right)^2} \tag{178}$$

$$\pi_1(z|\theta) = \frac{p(z|\theta_1)}{p(z|\theta_1) + p(z|\theta_2)} = \frac{1}{1 + \frac{p(z|\theta_2)}{p(z|\theta_1)}} \tag{179}$$

where

$$\theta_1 = (\mu_1, \sigma_1) \qquad \theta_2 = (\mu_2, \sigma_2) \qquad \theta = (\theta_1, \theta_2) \tag{180}$$

we have

$$\frac{p(z|\theta_2)}{p(z|\theta_1)} = \frac{\sigma_2^{-1} e^{-\frac{1}{2}\left(\frac{z-\mu_2}{\sigma_2}\right)^2}}{\sigma_1^{-1} e^{-\frac{1}{2}\left(\frac{z-\mu_1}{\sigma_1}\right)^2}} = \frac{\sigma_1}{\sigma_2} e^{\frac{1}{2}\left(\left(\frac{z-\mu_1}{\sigma_1}\right)^2 - \left(\frac{z-\mu_2}{\sigma_2}\right)^2\right)} \tag{181}$$

such that

$$\pi_1(z|\theta) = \frac{1}{1 + \frac{\sigma_1}{\sigma_2} e^{\frac{1}{2}\left(\left(\frac{z-\mu_1}{\sigma_1}\right)^2 - \left(\frac{z-\mu_2}{\sigma_2}\right)^2\right)}} \tag{182}$$

Hence, by equation (173), we have

$$\alpha_1(z|D) = \int_\theta \frac{p(D|\theta)p(\theta)d\theta}{1 + \frac{\sigma_1}{\sigma_2} e^{\frac{1}{2}\left(\left(\frac{z-\mu_1}{\sigma_1}\right)^2 - \left(\frac{z-\mu_2}{\sigma_2}\right)^2\right)}} d\theta \tag{183}$$

Assuming D consists of independent samples we have

$$p(D|\theta) = \prod_{j=1}^{2} \prod_{k=1}^{n_j} \frac{1}{\sqrt{2\pi}\sigma_j} e^{-\frac{1}{2}\left(\frac{x_{(j,k)}-\mu_j}{\sigma_j}\right)^2} \tag{184}$$

That is

$$p(D|\theta) = \left(\frac{1}{\sqrt{2\pi}}\right)^n \sigma_1^{-n_1}\sigma_2^{-n_2} e^{-\frac{1}{2}\left(\sum_{k=1}^{n_1}\left(\frac{x_{(1,k)}-\mu_1}{\sigma_1}\right)^2+\sum_{k=1}^{n_2}\left(\frac{x_{(2,k)}-\mu_2}{\sigma_2}\right)^2\right)} \quad (185)$$

We insert this in equation (183) and remove the constant term $\left(\frac{1}{\sqrt{2\pi}}\right)^n$ and obtain

$$\alpha_1'(z|D) = \int_\theta \frac{p(\theta)\sigma_1^{-n_1}\sigma_2^{-n_2} e^{-\frac{1}{2}\left(\sum_{k=1}^{n_1}\left(\frac{x_{(1,k)}-\mu_1}{\sigma_1}\right)^2+\sum_{k=1}^{n_2}\left(\frac{x_{(2,k)}-\mu_2}{\sigma_2}\right)^2\right)}}{1+\frac{\sigma_1}{\sigma_2}e^{\frac{1}{2}\left(\left(\frac{z-\mu_1}{\sigma_1}\right)^2-\left(\frac{z-\mu_2}{\sigma_2}\right)^2\right)}} d\theta \quad (186)$$