Understanding the accuracy of pre-symptomatic diagnosis of sepsis

Problem presented by

Phillippa Spencer

DSTL

ESGI116 was jointly hosted by
Durham University
Smith Institute for Industrial Mathematics and System Engineering

Report author

Jonathan Cumming (Durham University)
Asbjørn Riseth (Oxford University)
Jessica Williams (Oxford University)

Executive Summary

DSTL are interested in the uncertainty associated with the presymptomatic diagnosis of sepsis. This diagnosis results from a complex process involving collection of blood samples, microarray analysis, and statistical prediction via a neural network model.

The study group developed a Monte Carlo simulation method that would allow the rapid generation of an empirical distribution for gene expressions consistent with a given observation and given a specification for the errors involved in the process. Further work extended this method to consider multivariate simulation of multiple genes simultaneously. Sensitivity analyses were performed to identify the most influential sources of error. Additionally, investigations were made into the impact that possible types of uncertainties could have on a classifier such as the one used in this problem.

Version 1.0
May 6, 2016
iv+22 pages

Contributors

Jonathan Cumming (Durham University)
Iliana Peneva (Warwick University)
Asbjørn Riseth (Oxford University)
Michael Tsardakas (Heriot-Watt University)
Jessica Williams (Oxford University)
Liu Zhangdaihong (Warwick University)

Contents

1 Introduction

(1.1) DSTL are interested in the uncertainty associated with the presymptomatic diagnosis of sepsis. This diagnosis results from a complex process involving collection of blood samples, microarray analysis, and statistical prediction via a neural network model. The study group had access to a description of the errors in the process, their natures, sources, and likely distributions, as well as a sample of gene expression data. Fine technical detail of the underlying biological and scientific processes were not available, and so an approach based on detailed mathematical or statistical modelling was not feasible. Additionally, the neural network classifier was not available for study.

(1.2) Therefore, the focus of the study group's activity was in the accumulation of errors involved in the extraction and processing of the blood samples that result in the gene expressions which are input to the neural network. The study group developed a Monte Carlo simulation method that would allow the rapid generation of an empirical distribution for gene expressions consistent with a given observation and given a specification for the errors involved in the process. Further work extended this method to consider multivariate simulation of multiple genes simultaneously. Sensitivity analyses were performed to identify the most influential sources of error. Additionally, investigations were made into the impact that possible types of uncertainties could have on a classifier such as the one used in this problem. The study group also suggest additional work that will further understanding of this phenomenon.

2 Problem description

2.1 Background

(2.1) Research is currently being undertaken to expand the window of efficiency for medical treatment of sepsis through pre-symptomatic diagnosis. This is achieved through an observational clinical study. Blood is taken from consenting elective surgery patients from pre-surgery to treatment end. Some of these patients go on to develop sepsis (3.8%) and the majority recover without developing sepsis. Blood is taken daily. The diagnosis of sepsis has a level of variation between clinicians and hospitals and consensus is reached via a clinical advisory panel where the level of disagreement is analysed. The bloods are stored and then shipped to a laboratory where the RNA or transcriptomic signature is measured by microarray and quantitative methods. The data is retrieved, pre-processed, normalised and undergoes statistical

modelling. This then predicts whether a patient is likely to go on to develop sepsis or not.

(2.2) At every point of this process, from patient to statistical result, there is an associated error or accuracy. There are different data types present and not all of the error points can be considered independent. In order to give the clinician confidence in using this process to assist at point of care, we need to be able to propagate the errors through the complex process to provide an overall uncertainty measurement.

2.2 Framing the problem

(2.3) Fundamentally, the question being asked is: will this patient develop sepsis? Other questions (when will they develop sepsis? what is the cause?) are of interest, but a study of these is predicated on being able to have some confidence in our ability to predict sepsis for a given patient.

(2.4) This assessment can be made by two separate processes:
The clinical diagnosis: made by individual clinicians, but with subsequent consensus opinion made by a clinical advisory panel. This is taken to be the gold-standard for the diagnosis.
The statistical prediction: using a neural network classification model based on the gene expression of a set of key genes obtained through microarray analysis of a blood sample. While statistically accurate at mirroring the clinical diagnosis on the data available, there is no quantification of uncertainty or accuracy associated with the quantities used.

(2.5) The clinical diagnosis is authoritative, though expensive (in terms of time and resource). The statistical prediction relatively cheap in terms of resource, but requires a measure of error in order to quantify the level of confidence that can be placed upon its results. Thus, in order to use the statistical method reliably we require a statement of uncertainty to accompany the statistical prediction. The neural network classifier is treated as a fixed black-box model and modifications or alterations of the classifier were not of interest, therefore the focus of the Study Group fell on the following two areas:

(2.6) **1. Uncertainty propagation** – what is the uncertainty on the gene expression data used as input to the statistical classifier? Given a single gene expression observation, what uncertainty statements could be made about the originating sample given knowledge of the process and its errors? What range of possible gene expressions would be consistent with the observed gene expression given knowledge of the process and its errors? How influential are each of the error sources on the final uncertainty on the gene expression? These questions are addressed in Section 4.

(2.7) **2. Impact of uncertainty on the classifier** – given a generalised version

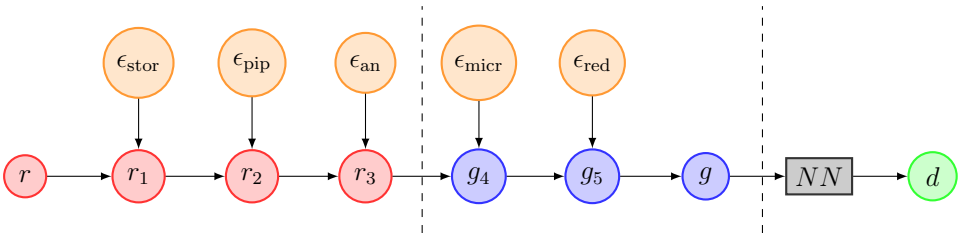


Figure 1: Error model for RNA counts and gene expression.

of the process, what impact could such uncertainties have on predictions made by the classifier? What effects can particular types of error cause? These questions are discussed in Section 6.

3 Sources of uncertainty and error

(3.1) Based on our understanding of the diagnostic process, we considered five main stages at which error may be introduced. These stages are summarised briefly as follows:

1. **Blood Sample** - Blood is first drawn from the patient, before RNA preservatives such as RNAlater are added. The sample is then pipetted so that a sample of the required volume is obtained.
2. **Storage and Shipping** - The blood sample is stored and shipped to a laboratory, while frozen.
3. **RNA Extraction** - The RNA must be isolated from the blood sample, and this is done through pipetting and analyst input.
4. **Microarray Analysis** - This type of analysis uses the RNA counts to determine gene expression data.
5. **Dimension Reduction Analysis** - Statistical modelling is utilised to reduce the data to information most relevant to sepsis diagnosis.

(3.2) These errors are depicted graphically in Figure (1), where each r represents RNA counts and each g represents gene expression data. The ϵ values correspond to the associated uncertainties. The box labelled NN is the neural network classifier, and y is the prediction of sepsis. In order to consider how these sources of uncertainty could impact the diagnosis, we need to quantify each error with regards to its impact on RNA counts and resulting gene expression data.

3.1 Blood Sample ($r \to r_1$)

(3.3) The blood sampling process, as described above, is composed of three main

steps: drawing blood, adding RNA preservatives (such as RNAlater), and pipetting the fluid into a required sample size. Based on provided information about these processes, we were led to the assumption that errors in blood drawing are distributed in a skew-normal manner, errors in pipetting are normally distributed and centred around zero, and errors in RNAlater are distributed bimodally.

(3.4) It seems that error resulting from the blood sample stage would directly result in uncertainty in the sample volume. As long as a sufficiently large sample volume was obtained, we assumed that this ambiguity would not affect the gene expression results, and thus analysed the sample volume uncertainty as a separate problem, the results of which are shown in Section 5.4. The question of how uncertainty in the blood sample volume affects RNA counts remains, however, and may require further consideration.

3.2 Storage and Shipping ($r_1 \to r_2$)

(3.5) It is known that RNA degrades at a fixed rate, d, per day and as the sample is frozen while it is stored and shipped, we make the assumption that errors during this processing stage are due to RNA degradation only. We take the number of days that RNA is in storage to be $t \sim \text{Poi}(\lambda)$, a Poisson distributed random variable. Thus, the RNA counts after shipping and storage will be

$$r_2 = (1-d)^{t+1} r_1, \tag{1}$$

where r_1 is the amount of RNA before shipping and storage, and we assume a minimum storage and shipping time of one day. The decay rate was specified to be $d = 0.0005$. The parameter of the Poisson was not specified, and so was arbitrarily set to be $\lambda = 2$ though it is trivial to adjust this to a more appropriate value.

(3.6) We note here that ϵ_{stor} is a multiplicative error, whereas the other processing errors we consider are additive.

3.3 RNA Extraction ($r_2 \to r_3$)

(3.7) The two errors associated with RNA extraction are due to pipetting error (ϵ_{pip}), and uncertainty incurred by human analysis (ϵ_{an}). If we assume that both pipetting and human analysis are equally likely to result in increased or decreased RNA counts, then these errors can both be represented by normal distributions centred around zero. Initial specification for these errors were that they were within $\pm 1\%$ and $\pm 14\%$ for pipetting and analyst error respectively. These were equated with statements of $\pm 2\sigma$ to derive appropriate parameter values.

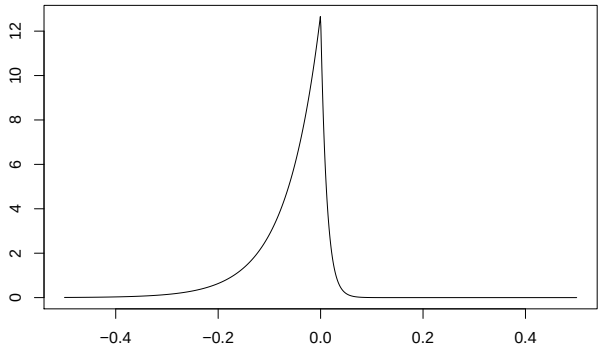


Figure 2: An asymmetric Laplace distribution with parameters $\mu = 0$, $\sigma = 0.01$ and $p = 0.85$.

3.4 Microarray Analysis ($r_3 \to g_4$)

(3.8) Here the RNA is measured by microarray analysis and converted into gene expression data. It has been found [4] that the error distribution that occurs in this process fits an asymmetric Laplace (ALD) distribution [5]. The ALD distribution is formed by two back-to-back exponential distributions of unequal scale. We say that a random variable Y is distributed as an ALD with location parameter μ, scale parameter $\sigma > 0$ and skewness parameter p in $(0, 1)$, if its probability density function (pdf) is given by:

$$f(y|\mu, \sigma, p) = \frac{p(1-p)}{\sigma} \exp\left\{-\rho_p\left(\frac{y-\mu}{\sigma}\right)\right\} \tag{2}$$

where $\rho_p(.)$ is the so called check (or loss) function defined by

$$\rho_p(x) = x(p - \mathbb{I}(x < 0)), \tag{3}$$

with $\mathbb{I}(\cdot)$ denoting the usual indicator function.

(3.9) The asymmetry parameter, p, of the distribution governs the inequality of scale of the distribution, with more extreme values resulting in one side of the distribution having a longer tail than the other. This distribution is illustrated in Figure 2 for the default parameter values used in the analysis.

(3.10) The conversion from RNA count data to gene expression data is given by

$$h(r) = \log_2(r) - m, \tag{4}$$

where m is some median value of the sample[1].

[1]Without the sample information to calculate m, in our analysis we arbitrarily set $m = 10$ to ensure the argument of the logarithm was always positive.

3.5 Dimension Reduction Analysis ($g_4 \to g_5$)

(3.11) In this final stage, the gene expression data is reduced to expression data for the genes most relevant to sepsis diagnosis. Here, we assume that a fixed error (Normal $\pm 5\%$) is accrued due to the amount of information lost through reducing the data set.

4 Uncertainty propagation

(4.1) In general, problems of uncertainty analysis and quantification in complex systems are tackled using Bayesian approaches. This approach has been popularised recently with recent work on the analysis of complex computer models. Therefore, we initially considered approaching the problem from a Bayesian viewpoint – in particular the use of a Bayesian belief network [1]. However, this approach proved unsuccessful due to the lack of sufficiently detailed knowledge about the behaviour of the error processes, and the lack of comprehensive prior information about the quantities involved.

(4.2) Given the error model in Figure 1, the observed gene expression, g, can be expressed simply as a sequence of combinations of additive and multiplicative errors and known transformations to an initial latent RNA count, r. Expressing this formally,

$$y = h(r + \epsilon_{\text{stor}} + \epsilon_{\text{pip}} + \epsilon_{\text{an}}) + \epsilon_{\text{micro}} + \epsilon_{\text{red}} \tag{5}$$

Each of the error terms has a specified distribution and parameters as described in Section 3 and transformation function from RNA count to gene expression, $h(\cdot)$, is as (4).

(4.3) This information is sufficient to apply Monte Carlo [2] sampling to simulate the error and analysis process. Monte Carlo simulation for this problem works by taking an initial value r, sampling an error from each of the error distributions, and then combining as (4), which yields a value g which is consistent with the initial value r and our specification of the error process. In detail, given an initial RNA count, r, we first sample a value of t from its specified Poisson distribution, which we use to produce a value of ϵ_{stor}. Adding this to r yields a single sample from the distribution of r_1, which represents a possible RNA count after storage degradation. We continue this process by sampling a value of ϵ_{pip} to produce r_2, and sampling an ϵ_{an} to produce an r_3. Thus r_3 is now one realisation of an 'observed' RNA count that is consistent with the original 'true' count r and the error process as specified. We can then apply the transformation (4) to map the RNA count into a gene expression, and apply the further errors ϵ_{micro} and ϵ_{red}, to obtain a final realisation of a gene expression, g, that could have originated from the original RNA count, r.

(4.4) While this process only yields a single sample, the strength of Monte Carlo simulation is the ability to repeat the sampling process. Each iteration will yield a different value of g due to the inherent (psuedo-)randomness in sampling from the error distributions. Thus, we can take a single initial r and propagate it through this error process many times to obtain an ensemble of possible realisations of g. From this collection we can then construct an empirical distribution for the gene expression, $g|r$, which directly expresses the uncertainty consistent with this error process and specification, and from which we can infer relevant summary statistics.

(4.5) One complication with this setup is that, in practice, r is an unknown quantity whereas g is observed and it is uncertainty on g that we seek. This presents us with two possible approaches:

(4.6) **Invert the sampling**: If the transformation function, $h(\cdot)$, that maps RNA counts to gene expressions is invertible, we could reverse the error process and use a single observation g to sample from the distribution of r. We can then simply invert (4) for r:

$$r = h^{-1}(g - \epsilon_{\text{micro}} - \epsilon_{\text{red}}) - \epsilon_{\text{pip}} - \epsilon_{\text{an}} - \epsilon_{\text{stor}}, \tag{6}$$

and apply Monte Carlo simulation to this quantity. Thus we now either know or can sample from all the quantities on the right-hand side of the equation and can use the observations directly to produce samples of r.

(4.7) **Adopt a search strategy**: If h is not invertible, then the above approach is not viable. In this case we could adopt a simple search strategy to explore the space of r to seek values which would be consistent with the observations g. This is necessarily more computationally intensive, but could yield regions of plausible values of r consistent with an observed g. (This shares similarity with *Bayesian history matching* in the computer model literature.)

(4.8) In both cases, we arrive at an empirical distribution over values of r which could have been the originating RNA count for the observed gene expression g. To translate these initial RNA counts into uncertainty on the gene expression, we again have two possibilities.

(4.9) **Direct transformation**: Given the empirical distribution of r, we can directly map the RNA counts into a distribution of gene expression by application of h. This quantifies the uncertainty in a latent 'true' gene expression, g^* say. This is not the same as the distribution for g, but its error-free state.

(4.10) **Forward sampling**: Each of the values r in the empirical distribution could have originated the observed g. However, each r itself yields a distribution of possible g values at the end of the error process. If we denote the observed gene expression value as g_0, then basic probability theory shows that

$$f(g|g_0) = \int f(g|r, g_0) f(r|g_0) \mathrm{d}r, \tag{7}$$

and from Monte Carlo simulation we have obtained the empirical distribution $f(r|g_0)$. To fully capture the distribution over possible g, we must further find $f(g|r, y_0)$ and then integrate. The former could be again achieved by the same Monte Carlo sampling of (4), and (given the heavy reliance thus far on Monte Carlo methods) so too could the marginalisation [3]. Note here that this approach yields larger uncertainties than the method above as now we are quantifying the uncertainty in the *observed* gene expression, $g|g_0$, and not the *error-free latent* gene expression, $g^*|g_0$.

(4.11) Due to constraints on time and group size, we focus on the direct transformation method using Monte Carlo on the inverted error process (3) – this is described in Section 5.1. We also performed a similar investigation to the sub-problem of uncertainty in sample volume, which is presented in Section 5.4.

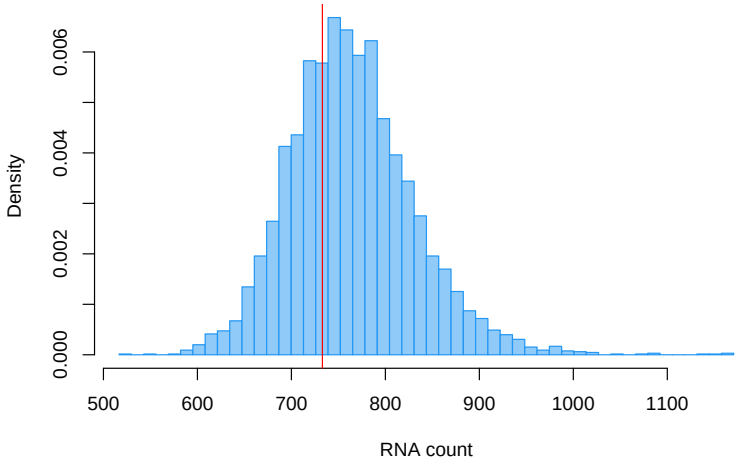
5 Results

5.1 Monte Carlo uncertainty analysis for gene expression

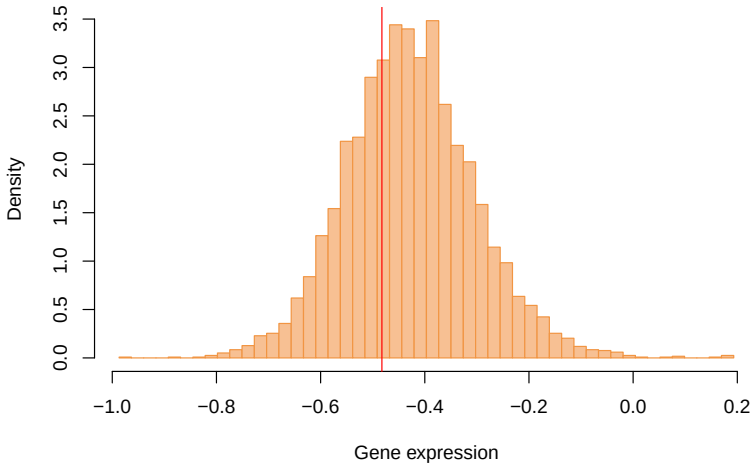
(5.1.1) We now discuss the results of the Monte Carlo uncertainty analysis discussed. Throughout we have used error specifications consistent with those given in Section 3. For the microarray error, the asymmetric Laplace distribution was used. However, no information was available on the skewness parameter for the error associated with this particular microarray. Therefore a representative value has been used to provide moderate skewness; this value can be replaced with a more appropriate value if practical information were available to suggest otherwise.

(5.1.2) Applying the Monte Carlo simulation method to (3) given a single observed gene expression value of $g_0 = -0.4823$ with a Monte Carlo sample size of $N = 5000$ produces a sample of RNA counts with histogram given in Figure 3a. The red vertical line indicates the RNA count value, r_0 that would correspond to g_0 if it were observed without any error or degradation. We observe that as a consequence of the error process, the distribution of the RNA counts has been skewed towards larger values. This is a result of the correction for sample degradation and the strong asymmetry of the Laplace distribution for the microarray error. Consequently, the bulk of the distribution of possible RNA counts is larger than we would expect. Transforming these RNA counts into gene expressions by direct application of $h(\cdot)$ yields the histogram in Figure 3b, which is of similar shape.

(5.1.3) Applying the Monte Carlo simulation process to (4), we can generate an empirical distribution for the gene expressions g given a value of the RNA count r. Since the RNA count is unobservable, for illustration we take a

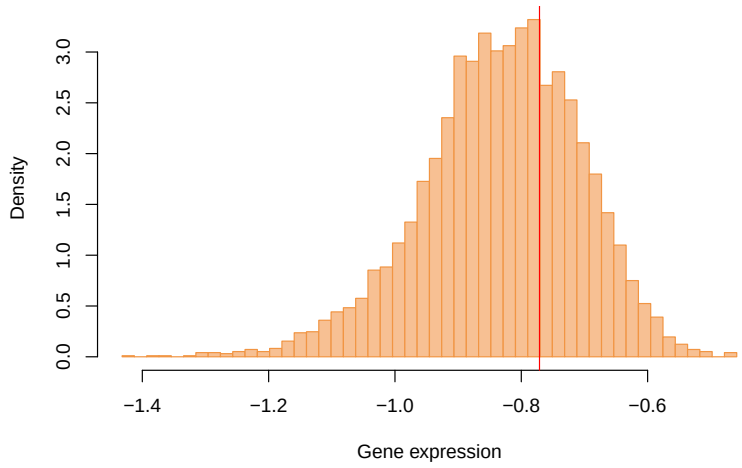


(a) Distribution of RNA counts, r, given g_0. The vertical line indicates the error-free transformation of $r_0 = h^{-1}(g_0)$

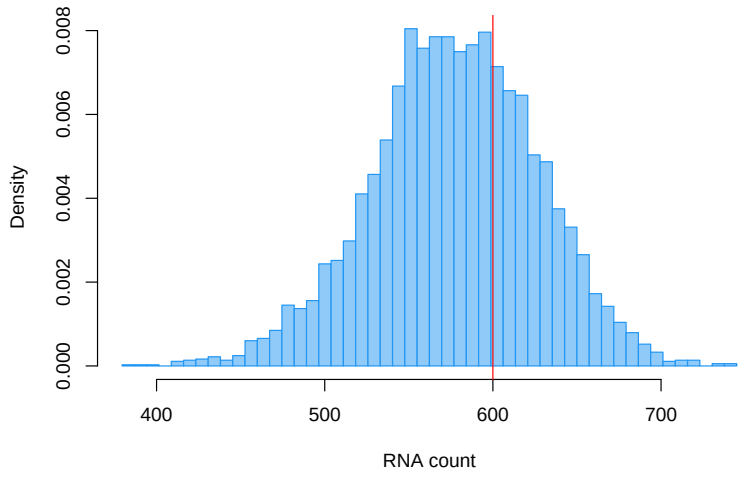


(b) Distribution of latent gene expressions, g^*, corresponding to the above distribution of RNA counts. The vertical line indicates the observed gene expression value.

Figure 3: Monte Carlo simulation results from simulation of given observed gene expression $g_0 = -0.4823$.



(a) Distribution of gene expressions, g, given $r = r_0$. The vertical line indicates the error-free transformation of the RNA count, $g_0 = h(r_0)$.



(b) Distribution of latent RNA counts, r^*, corresponding to the above distribution of gene expressions. The vertical line indicates the initial RNA count value.

Figure 4: Monte Carlo simulation results from simulation of given observed RNA count $r_0 = 600$.

value of $r_0 = 600$, with results given in Figure 4. As we might expect, the distribution now shifts in the opposite direction, with the RNA count degrading over time and the asymmetry of the Laplace error reversed. Thus we observed gene expressions that are more likely to be lower than

we would expect.

5.2 Multivariate simulation

(5.2.1) In the analysis above, we have applied Monte Carlo simulation to propagate the uncertainty associated with a single gene expression value. In practice, this is not a univariate problem and for each blood sample multiple gene expressions are measured. Consequently, a multivariate approach to the uncertainty propagation is required.

(5.2.2) This is (mostly) straightforward, and is achieved by replacing the components of (4) and (3) with vectors rather than scalars. Since the errors are now vector-valued, their associated distributions must also be multivariate. For the Normally distributed errors, a multivariate Normal distribution is simple to sample from given some information on the correlation between the error components. A multivariate version of the asymmetric Laplace distribution exists for higher dimensions [6], though time constraints prevented its implementation here and the microarray error was assumed to be composed of independent univariate ALD errors for each gene. A multivariate approach would also allow for different decay rates for different genes, though this was not considered here.

(5.2.3) By extending the Monte Carlo approach as described above, and using an (arbitrary) correlation between the components of the Normal errors of 0.75 we repeated the analysis of Figure 3 using the vector of all 44 gene expressions as g_0. The result is a 44-dimensional empirical density, which is shown in in the form of 2-dimensional projections between the first four genes.

5.3 Sensitivity analysis

(5.3.1) Given a mechanism for propagating uncertainty through the system, we can investigate how the overall uncertainty in the gene expression or RNA count is affected by each of the sources of error. For this sensitivity analysis, we adopt a one-at-at-time approach: varying each parameter individually over a range of values while keeping others at their specified original values. For each value investigated, we perform the Monte Carlo uncertainty analysis and investigate the effects on the empirical distribution by calculation of summary statistics. Since we are most interested in the spread in the values, we focus on the sample standard deviation. The parameters varied and their ranges are summarised in Table 1.

(5.3.2) The results of the sensitivity analyses are shown in Figure 6 and Figure 7. Each figure shows (leftmost) a plot of the change in the standard deviation of the RNA count due to variation in the named parameter, and then three

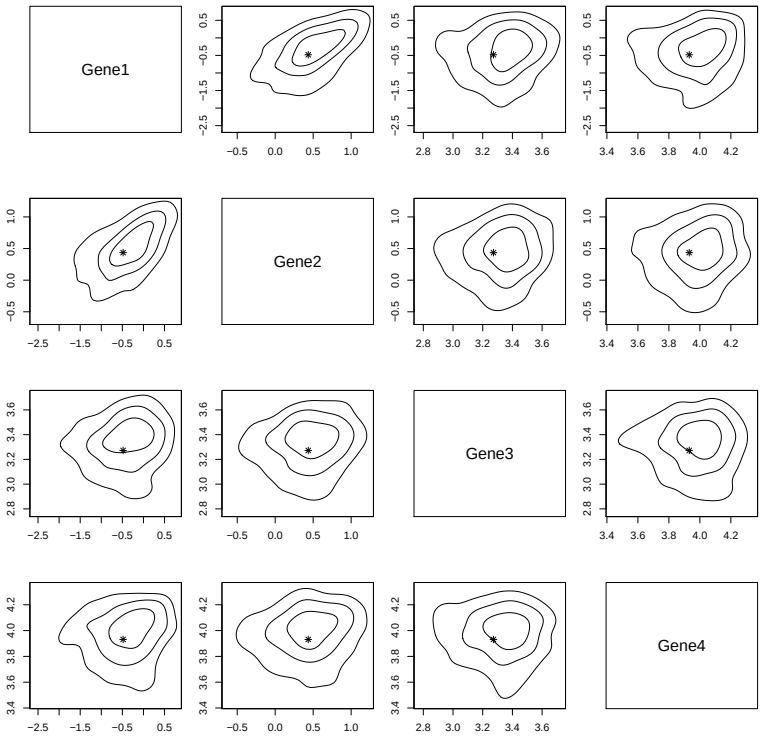


Figure 5: Density plots of empirical multivariate distributions obtained from Monte Carlo sampling for gene expression.

Error source	Min	Default	Max
Analyst error	0%	14%	28%
Pipette error	0%	1%	2%
1− Decay rate	0.9	0.995	0.9999
Storage time Poisson parameter	0	2	10
Microarray ALD σ parameter	0.0001	0.01	0.1
Microarray ALD skewness parameter	0.05	0.85	0.95

Table 1: Summary of model parameters and their ranges used in one-at-a-time sensitivity analysis.

histograms showing the empirical distribution of the RNA count at the lower extreme, mid-point and upper extreme of the range of investigation.

(5.3.3) First, we consider the analyst error which was specified as Normal with range of approximately $\pm 14\%$ of the value. Equating this interval with 2σ of the Normal distribution directly gives a distribution parameter to investigate. The results of the sensitivity analysis clearly indicate that analyst error is a highly influential component with a strong effect on the final standard deviation. At its nominal value of 14% it contributes approximately 40% of the overall variability when all other parameters are held at default values. As the size of the analyst error increases, its Normal distribution begins to dominate the overall distribution making it progressively more symmetric.

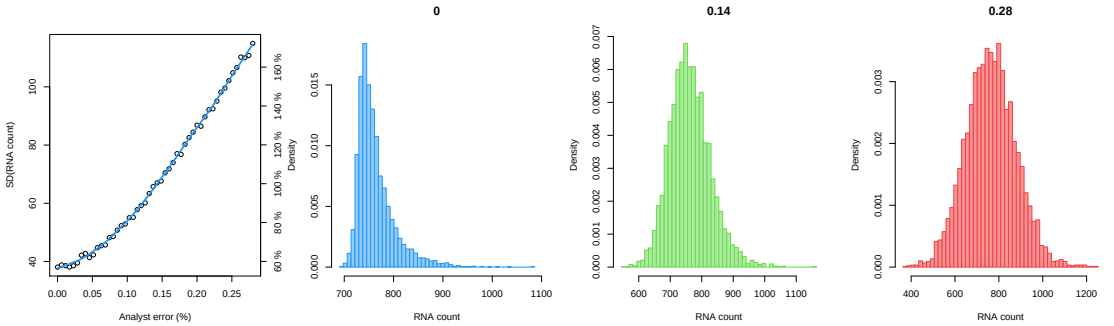
(5.3.4) Conversely, pipetting error is ignorable. Specified as Normal and of the order of $\pm 1\%$ its effects on the final variation are negligible.

(5.3.5) The RNA degradation rate is also a highly influential parameter. While its impact is negligible for a degradation of up to 2% per unit time, beyond this point it has a very strong impact on the spread of the distribution. While this feature is quite striking, since the nominal specified value is well within this range this may in fact be ignorable in practice. Increasing the decay rate induces increased asymmetry as the sampling process attempts to correct for this, induces a shift in the location and an elongation of the upper tail.

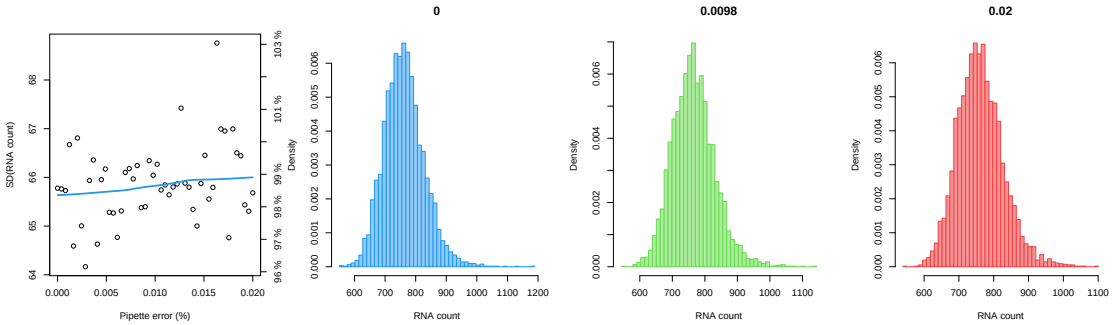
(5.3.6) Storage time was modelled by a Poisson random variable with a rate parameter λ (1), with a default value of $\lambda = 2$. Investigation of the sensitivity to λ displayed no obvious dependence with the overall uncertainty.

(5.3.7) The microarray error was given an asymmetric Laplace (ALD) distribution, with mean 0 and a standard deviation parameter σ and a further skewness parameter p. The standard deviation parameter proved to have exceptionally high sensitivity with relatively small changes in value resulting in huge increases in standard deviation. This is likely attributable to the shape of this particular distribution, with one particularly long tail. Increasing σ will cause this long tail to be reach out even farther making more extreme observations more likely. This is corroborated by the histograms which show the uncertainty distribution becoming dominated by this exceptionally long tail. While this distribution is deemed appropriate in the literature, given its sensitivity in this process some care will be required to ensure that the parameters are properly calibrated or else risk substantial over-statements of uncertainty.

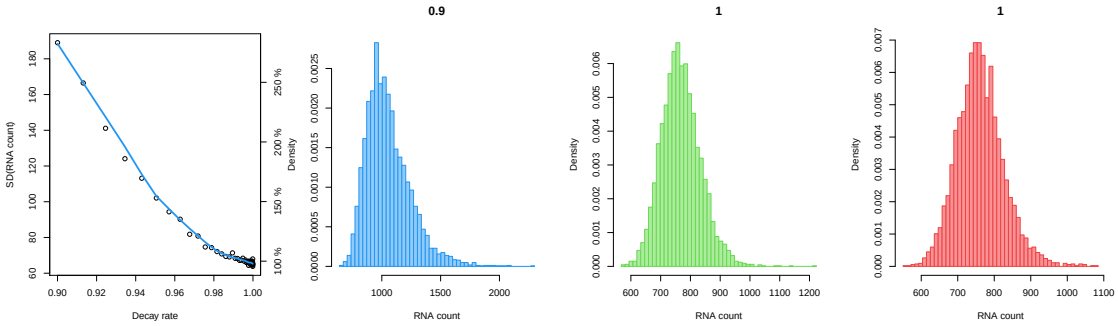
(5.3.8) Finally, the skewness parameter p of the microarray error distribution also proved influential. While generally having only a small effect on the standard deviation for values between 0.2 and 0.8, at the more extreme values of skewness it could increase uncertainty by up to 50%. This is



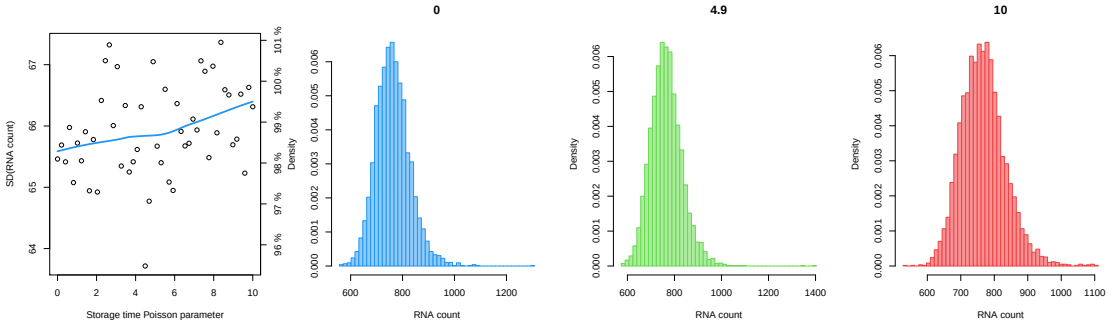
(a) Analyst error



(b) Pipette error

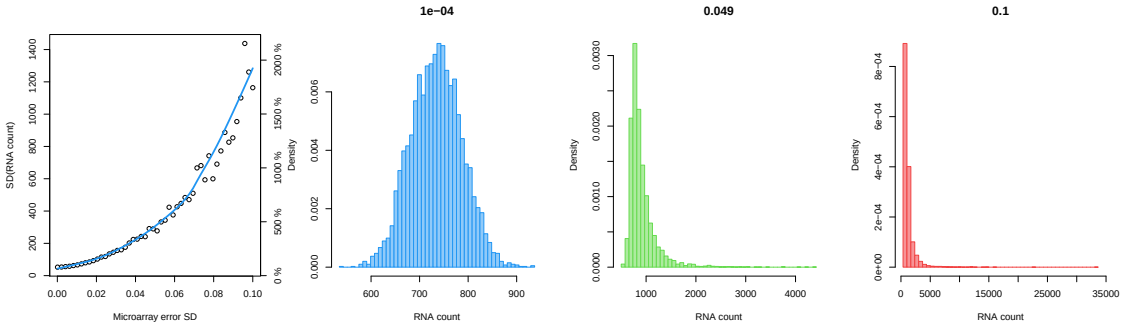


(c) Degradation rate

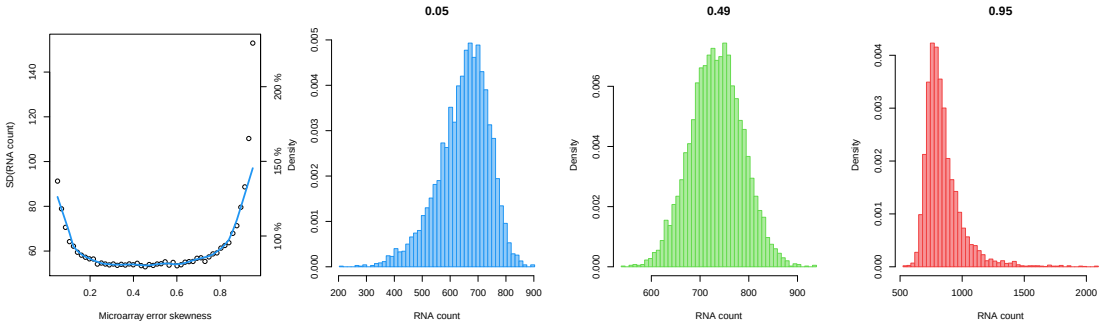


(d) Storage time

Figure 6: Sensitivity analysis of RNA count to changes in the parameters of the data collection and analysis process.



(a) Microarray error σ parameter



(b) Microarray skewness

Figure 7: Sensitivity analysis of RNA count continued.

illustrated by the histograms where the extremes in skewness introduce a long tail in either direction resulting in inflated uncertainty. Smaller values of skewness are either ignorable or dominated by the Normal components resulting in a more symmetric empirical distribution.

5.4 Volume

(5.4.1) During the first stage of the process from patient to sepsis diagnosis, a blood sample is extracted. As explained in Section 3.1, it is most intuitive to describe errors incurred here as affecting the volume of the sample, which may in turn affect the gene expression data, but the nature of this relationship is yet to be understood. However, using the error propagating method, we can perform Monte Carlo simulations to understand the distribution of sample volumes that would be obtained during this stage, assuming that the relevant errors are distributed as described in Section 3.1.

(5.4.2) In Figure 8, we see that the distribution of blood sample volumes is influenced by the bimodal distribution of error from the addition of the RNAlater preservative.

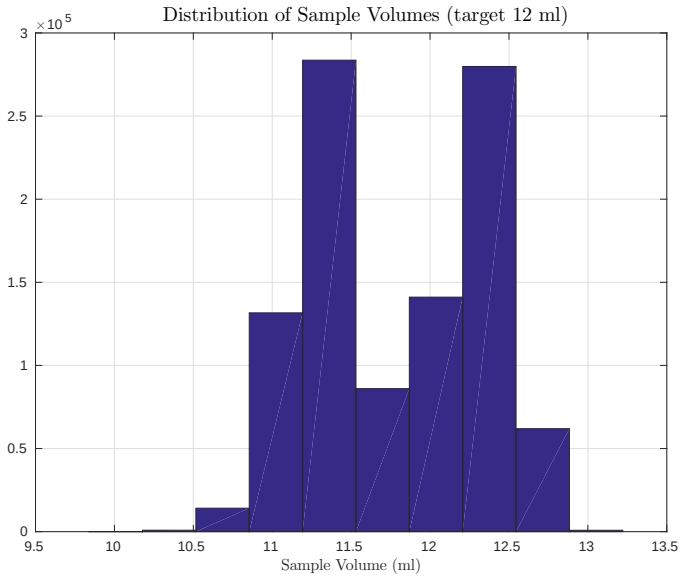


Figure 8: A histogram of the distributed volumes for Monte Carlo simulations with 10^6 trials.

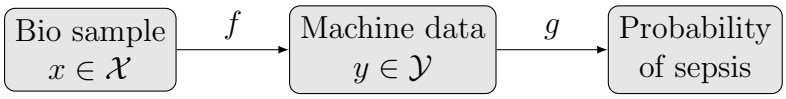


Figure 9: The biological sample is processed into machine readable form, modelled by f. The classifier is trained on data in $\mathcal{Y}$ with certain sepsis outcomes, then used on new patient data to give a diagnosis.

6 Impact of uncertainty on classifier

(6.1) This section considers the impact that errors introduced in the process have on the classification algorithm. We show that different types of errors can have opposite effects. It is therefore important to have an understanding of the nature of the uncertainty in the process to interpret the results from a classifier.

(6.2) Let $\mathcal{X}$ denote the set in which the raw data of the patients is contained. Patient data is processed into machine readable form contained in a set $\mathcal{Y}$. By going from $\mathcal{X}$ to $\mathcal{Y}$, errors occur which we model as events in a sample space Ω. This process is represented by a function $f : (\mathcal{X}, \Omega) \to \mathcal{Y}$, where $f(x, \cdot)$ is a random variable on the sample space Ω that takes values in $\mathcal{Y}$. A diagnosis of sepsis is given by a classification function $g : \mathcal{Y} \to [0, 1]$, representing the probability of the patient developing sepsis. A diagram of the process is shown in Figure 9. We assume that the uncertainty in the process f has a blurring effect on the data: the differences between indicators of sepsis and healthy patients should be less pronounced in $\mathcal{Y}$ than in $\mathcal{X}$.

(6.3) The function g is designed using training data we have collected from pa-

tients. Say we have some data $x_1, \ldots, x_N$ where we know the outcomes $z_i \in \{0, 1\} =: \mathcal{Z}$ of sepsis. The patient corresponding to x_i developed sepsis if $z_i = 1$, otherwise $z_i = 0$. Define a loss function $\ell : [0, 1] \times \mathcal{Z}$ that will compare the prediction from a classifier g with the known outcomes of sepsis. The loss function can be a quadratic penalty $\ell(a, b) = (a - b)^2$, or a weighted penalty that penalises under-prediction of sepsis higher than over-prediction:

$$\ell(a, b) = \begin{cases} 2(a - b)^2 & \text{if } a < b, \\ (a - b)^2 & \text{if } a \geq b. \end{cases} \tag{8}$$

From the data set (x_i, z_i) and the corresponding outcomes $\omega_i \in \Omega$ that represent the process $y_i = f(x_i, \omega_i)$, one can design a classifier function g over some set of functions that map $\mathcal{Y}$ to $[0, 1]$:

$$g = \arg\min_{\hat{g}} \sum_{i=1}^{N} \ell(\hat{g}(y_i), z_i). \tag{9}$$

We wish to understand how the uncertainty that arises from f impacts the classifier g. Say there is an error-free process $f^\dagger : \mathcal{X} \to \mathcal{Y}$ that generates data $y_i^\dagger$ from the data points $x_1, \ldots, x_N$. How will classifiers generated from an error-prone process f differ from the "correct" classifier $g^\dagger$ that arises from $(y_i^\dagger, z_i)$?

6.1 Example

(6.1.1) Let us consider an example of $N = 44$ one-dimensional data points representing a combination of gene expressions. For simplicity, set $\mathcal{X} = \mathcal{Y}$ and let $f^\dagger$ be the identity, i.e. $f^\dagger(x) = x$. The gene expression data $y_i^\dagger = f^\dagger(x_i)$ and their corresponding sepsis outcomes $z_1, \ldots, z_N$ are shown in Figure 10. There is a visible separation between patients with sepsis (negative $y_i^\dagger$) and those without (positive $y_i^\dagger$). This indicates that new patients with a large, negative gene expression y is likely to develop sepsis. Near zero the data is less conclusive, and therefore the true classifier $g^\dagger$ cannot with certainty predict the development of sepsis.

(6.1.2) We will consider two types of errors that blur the separation of data indicating sepsis and healthy patients. The first type moves x_i-values corresponding to sepsis towards the mean of the non-sepsis x_i-values, and vice versa. The second type moves each data point x_i closer to the mean $\overline{x} = \frac{1}{N} \sum_{i=1}^{N} x_i$, which in this example is zero. Their effect on the trained g are opposite of each other, as is shown in Figure 11. The first type errors are modelled by

$$y_i = \begin{cases} x_i + \lambda\beta_i & z_i = 1 \\ x_i - \lambda\beta_i & z_i = 0 \end{cases} \qquad \lambda \approx 9.5, \text{ i.i.d. } \beta_i \sim \text{Beta}(2, 50). \tag{10}$$

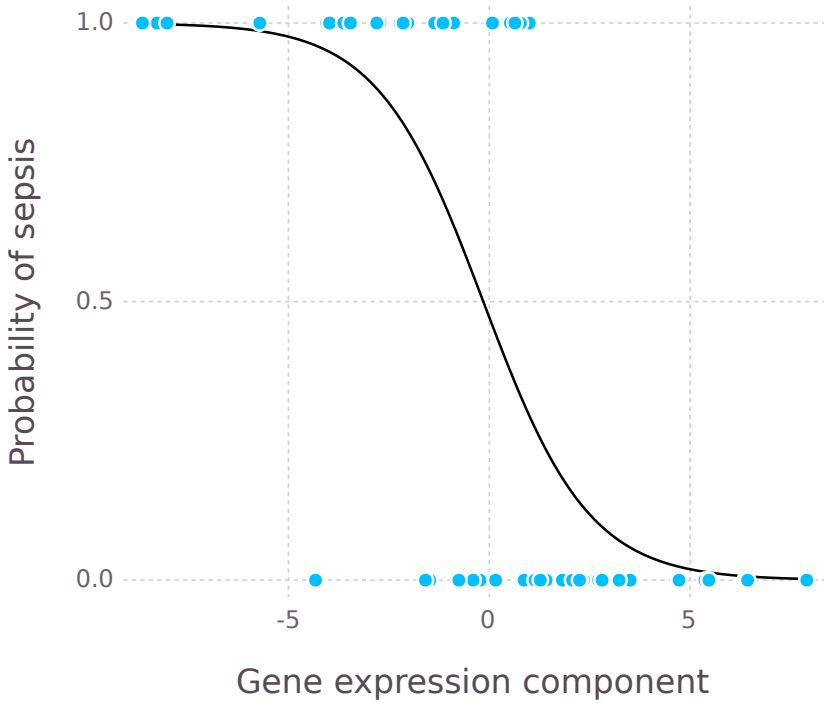


Figure 10: The correct data $(y_i^\dagger, z_i)$ indicates a particular separation between patients who develop sepsis $y^\dagger < 0$ and those who do not. The classifier based on the true data determines a probability of a patient developing sepsis, based on the gene expression y.

The second type errors are given by

$$y_i = x_i(\beta_i + \lambda) \qquad \lambda = 10^{-2},\ \text{i.i.d. } \beta_i \sim \text{Beta}(40, 5). \tag{11}$$

7 Summary and further work

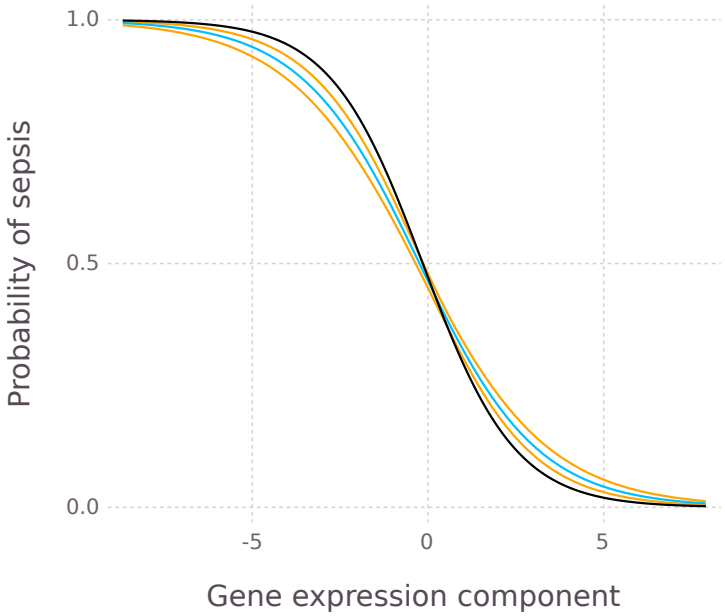
(7.1) In this report we have focussed on the use Monte Carlo simulation methods to propagate uncertainty through a simplified version of the process involved in the production of gene expression data used in the pre-symptomatic diagnosis of sepsis. While a fairly straightforward method, Monte Carlo sampling proved an effective tool for assessing uncertainty on the RNA count or the true error-free gene expression value associated with a give observation. It also presented an opportunity to study the sensitivity of this uncertainty to changes in the parameters of the component errors. This highlighted that the analyst error and the microarray error were most influential on our uncertainty on gene expression. Extension to the methods used enabled a multivariate uncertainty analysis which allowed for the introduction of correlated errors across the expressions of different genes. Finally, the impact that possible types of uncertainties could have on a classifier such as the one used in this problem were also studied.

7.1 Extensions and further work

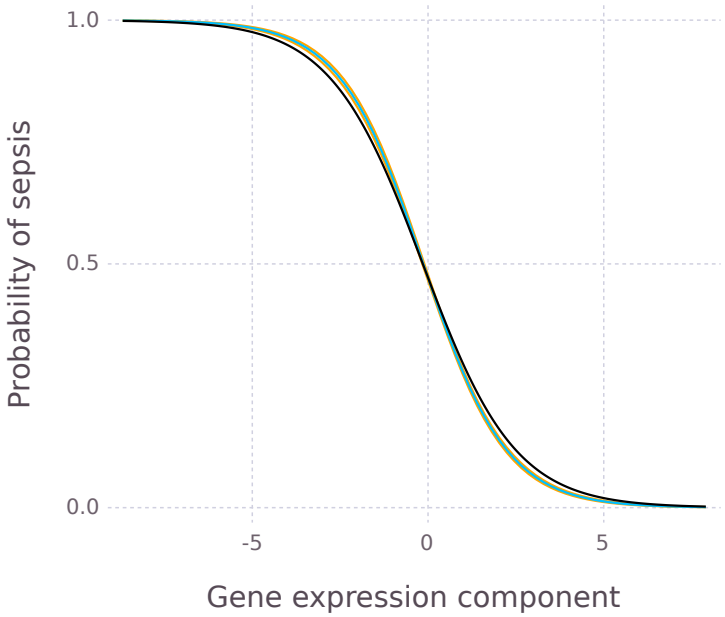
(7.1.1) While the methods developed provide a simple platform upon which to base an uncertainty analysis for this problem, there are a number of possible extensions and developments that could improve the quality and applicability of the methods:

(7.1.2) **1. Incorporation of the classifier**: The ultimate goal is to assess the uncertainty of the prediction of sepsis . Without access to the classifier used to make this prediction, the study group work focussed on the uncertainty on the data that are input to that classifier. Given access to neural network would allow for its incorporation into the Monte Carlo simulation and would result in direct uncertainty assessments on predicted diagnosis. This also would then link directly to the results of Section 6.

(7.1.3) **2. Quality control measurements**: Throughout the process, quality control measurements are made on the integrity of the sample. While these do not contribute to our uncertainty on the gene expressions, they can be considered as observations of the magnitudes of the accumulated errors with poor quality scores associated with samples that have large errors. These quantities could be added to the error model in Figure 1, but this would require adopting a graphical modelling or Bayesian network approach [1].



(a) An additive error results in a classifier that is less pronounced, underestimating the probability of sepsis for $y < 0$ and overestimating it for $y > 0$.



(b) A multiplicative error has less impact on the classifier, but will nevertheless put more weight to a sepsis diagnosis for $y < 0$.

Figure 11: Example of additive and multiplicative errors in f. The black curve shows the error-free classifier $g^\dagger$. A Monte-Carlo approach is used to approximate the distribution of g that arise from the two error processes. The blue is the mean classifier of g, and the orange lines indicate the 2.5% and 97.5% quantiles.

(7.1.4) **3. Calibration of the ALD parameters**: The microarray error's asymmetric Laplace distribution was found to be both influential on the final uncertainty and sensitive to its own parameters. A careful calibration of these parameters to values appropriate to this problem and the type of microarray methods used is recommended.

(7.1.5) **4. Refinement of the multivariate approach**: The multivariate Monte Carlo approach could be improved by incorporating a multivariate ALD for the microarray error, and (similar to the previous point) by ensuring appropriate values for correlations and error variances across genes.

(7.1.6) **5. Connection to the volume process**: In this work, we treated the errors associated in the volume of the sample as an independent problem. However if the sample volume were known to have a direct effect on the RNA counts or gene expression, then the two error processes could be connected.

References

[1] Cowell, R. G., Dawid, A. P., Lauritzen, S. L. and Spiegelhalter, D. J. (1999) *Probabilistic Networks and Expert Systems*. New York: Springer.

[2] Fishman, G.S. (1999) *Monte Carlo: Concepts, Algorithms, and Applications*. New York: Springer.

[3] Robert, C. P., and Casella, G. (2004) *Monte Carlo Statistical Methods*. Springer.

[4] Purdom, E., and Holmes, S. P. (2005) Error Distribution for Gene Expression Data. *Statistical Applications in Genetics and Molecular Biology*, 4:1, article 16.

[5] Yu, K., and Zhang, J. (2005). A three-parameter asymmetric Laplace distribution and its extension. *Communications in Statistics – Theory and Methods*, 34(9-10), 1867-1879.

[6] Kozubowski, T. J., and Podgorski, K. (2000). A Multivariate and Asymmetric Generalization of Laplace Distribution. *Computational Statistics*, 15: 531. Retrieved 2015-12-29.