UPTEC X 05 021
APR 2005

ISSN 1401-2138

KARIN HOLMSTRÖM

Nonlinear gene activity-drug sensitivity relationship modelling for improved detection of genes that affect drug sensitivity

Master’s degree project

UPPSALA UNIVERSITET

Molecular Biotechnology Programme

Uppsala University School of Engineering

UPTEC X 05 021	**Date of issue 2005-04**

Author

Karin Holmström

Title (English)

Nonlinear gene activity–drug sensitivity relationship modelling for improved detection of genes that affect drug sensitivity

Abstract

A novel approach for detection of genes whose products affect drug sensitivity was introduced. Two detectors, one based on nonlinear regression and the other on errors in variable modelling, have been tested on simulated data and was, using reciever operating characteristics graphs, compared to a linear state-of-the-art detection strategy, which is based on Pearson's correlation coefficient. On small and nonlinear datasets our modelling detector performed significantly better than the correlation coefficient detector. For example, a dataset size of five points gave under certain circumstances a rate of true positives of the modelling detector that was four times that of the correlation coefficient detector, both at 5% false positives. Our regression detector outperformed the correlation coefficient detector only on very small datasets.

Keywords

Relationship modelling, regression, drug sensitivity, correlation coefficient, detection.

Supervisors

Mats Gustafsson
Dept. of Engineering Sciences, Uppsala University

Scientific reviewer

Anders Isaksson
Dept. of Genetics and Pathology, Uppsala University

Language	Security
English	
ISSN 1401-2138	Classification
Supplementary bibliographical information	Pages **22**

Biology Education Centre Biomedical Center Husargatan 3 Uppsala
Box 592 S-75124 Uppsala Tel +46 (0)18 4710000 Fax +46 (0)18 555217

Nonlinear gene activity–drug sensitivity relationship modelling for improved detection of genes that affect drug sensitivity

Karin Holmström

Sammanfattning

Läkemedelskänslighet, d v s hur stark effekt ett läkemedel har, varierar från patient till patient. Detta är ett problem vid t ex kemoterapi i behandling av cancer. Men denna variation kan utnyttjas för att hitta gener som påverkar just läkemedelskänsligheten. Kännedom om patientens uttrycksnivåer av dessa gener kan användas vid individuell behandling och för att förutsäga behandlingens slagkraft. Dessa gener kan också ge information om hur olika läkemedel fungerar.

Om man känner till patienters läkemedelskänslighet och uttrycksnivåer för olika gener kan man identifiera de intressanta generna på olika sätt. Tidigare har en korrelationskoefficient mellan genuttryck och läkemedelskänslighet använts som mått på hur mycket genen påverkar läkemedelskänsligheten. Denna metod vore smidig och bra om sambandet mellan genuttryck och läkemedelskänslighet vore linjärt. Emellertid är detta inte fallet och i detta projekt har data istället anpassats till en ickelinjär kurva på två olika sätt. Dessa metoder för att hitta gener har med hjälp av simulerade data testats och jämförts med korrelationskoefficientmetoden. Det visade sig att de nya metoderna generellt sett presterar bättre när man har små datamängder.

Examensarbete 20 p civilingenjörsprogrammet i molekylär bioteknik
Uppsala universitet, april 2005

Contents

1 Introduction

To remedy our ill fellow men is a matter of the heart for modern society and probably has been for all societies in all times. Curing is often accomplished with the aid of remedies and medicines. Unfortunately, for many medicines the drug sensitivity varies between patients. The reasons for this variance are complicated and differ with drug and disease, but differences in gene activities between individuals are always bound to play an important role. Knowledge of which genes' expression levels affect drug sensitivity not only provide a prognosis tool but can also be used in individual drug treatment selection and studies of mechanisms of drugs.

Given data on drug sensitivity and gene expression for different cell lines or patients, there is no obvious tool for identifying the genes of interest. Studies have been reported where the cells were divided in the classes drug resistant or drug sensitive, and different methods were used to find the genes whose gene expression profiles differed significantly between the classes [1], [2], [3]. Other studies have correlated the degree of drug sensitivity to gene expression to identify genes of interest. In [4], [5] and [6] Pearson's correlation coefficient (see below) was used to measure the degree of correlation.

Data on gene expression and drug sensitivity can be supplied from e. g. real time RT-PCR or cDNA microarray gene expression analysis and drug sensitivity assays, respectively. In cancer drug sensitivity assays cancer cells are treated with a certain dose of the drug. The fraction of killed cells, called sensitivity index (SI), is determined. The same dose of the drug is used on all samples, and this dose is preferably chosen so that some patients' cancer cells will have a SI of 0 and others of the maximum SI for that specific drug, which is not necessarily 1. Some drugs or doses of drugs cannot kill all the cells, which renders a maximum below 1.

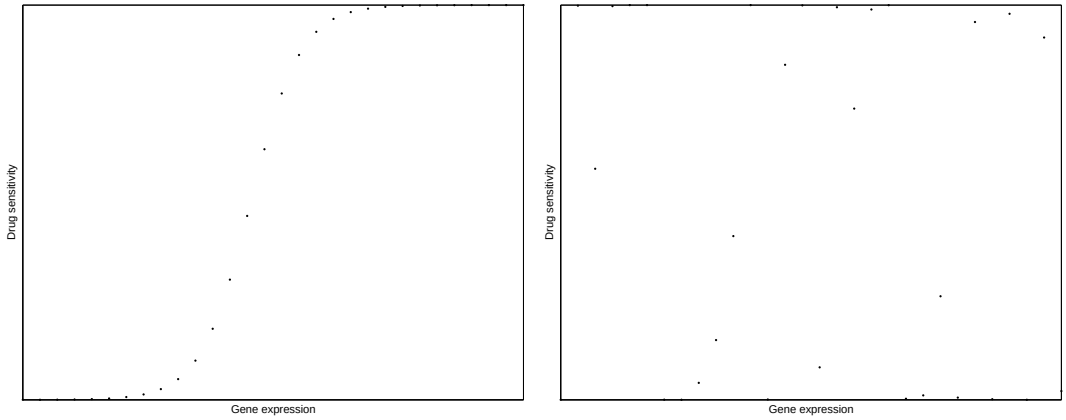
This project's aim was to implement, test and finally use on clinical biological data a novel approach to detect genes whose expression affect drug sensitivity. In this novel approach, nonlinear regression and modelling is used to quantify the correlation between drug sensitivity and gene expression. As mentioned earlier, Pearson's correlation coefficient

$$\rho = \frac{\mathrm{Cov}[X,Y]}{\sqrt{\mathrm{Var}[X]\mathrm{Var}[Y]}}, \quad -1 \leq \rho \leq 1$$

have been used in other studies to measure this correlation. Independence between two variables implies $\rho = 0$, whereas a completely 'linear dependence implies $\rho = -1$ or $\rho = 1$. As SI values always lie in the interval [0 1], a sigmoidal (s-shaped) dependence on gene expression is more likely than a linear one. Sigmoidal dependence is common in nature and, specifically, dose-response curves are known to be sigmoidal. In fact, sigmoidal dependence has been seen in gene expression–drug sensitivity data[1]. Figure 1 shows a sigmoidal gene expression–drug sensitivity correlation and an example of no correlation between drug sensitivity and gene expression. The detectors' task is to distinguish the genes which affect drug sensitivity (Figure 1(a)) from those that do not (Figure 1(b)). Our basic idea[2] is that nonlinear regression and modelling of the data according to a sigmoid curve would lead to detectors more sensitive and accurate than the correlation coefficient detector. In this work, we have compared the performances of our regression detector, our modelling detector and the correlation coefficient detector, which in practice correspond to linear regression. Our regression detector performs regression according to a sigmoid curve and our modelling detector models the data according to a sigmoid curve, which, as opposed to the regression detectors that only account for noise in the observation of drug sensitivity, also take into account noise in the observation of gene expression.

[1] Oral information from Anders Isaksson at Dept of Genetics and Pathology.

[2] Developed by Mats G. Gustafsson at Dept. of Engineering Sciences, Uppsala University, Anders Isaksson at Dept. of Genetics and Pathology, Uppsala University and Claes Andersson at the Linneaus centre for bioinformatics, Uppsala University.



(a) Sigmoidal dependence of drug sensitivity on gene expression.

(b) No dependence of drug sensitivity on gene expression.

Figure 1: Example of a sigmoidal correlation between drug sensitivity and gene expression (a) and an example of absence of correlation between drug sensitivity and gene expression (b).

2 Materials and Methods

2.1 Our detectors

A sigmoidal curve is defined by the equation $y = \frac{a}{1+e^{-w_0-w_1 x}}$. In this work x is gene expression and y is SI. The height of the "step" is equal to a, the slope is determined by w_1 and w_0 determines where on the x-axis the "step" appears, because when $x = -\frac{w_0}{w_1}$ we have $y = \frac{a}{2}$. For the genes of which drug sensitivity is independent, the relation between gene expression and drug sensitivity can be described with

$$y = \mu,$$

where μ is a constant between 0 and 1. Since the data are observations, we model the measurement in our models, so that the independence model is

$$M_0: \quad \tilde{y} = \mu + \epsilon_0, \tag{1}$$

and the sigmoidal model is

$$M_1: \quad \tilde{y} = \frac{a}{1 + e^{-w_0 - w_1 \tilde{x}}} + \epsilon_1, \tag{2}$$

where ϵ_0 and ϵ_1 are normally distributed noise with expectation value 0 and standard deviation σ_0 and σ_1, respectively, i. e. $\epsilon_0 \sim N(0, \sigma_0)$ and $\epsilon_1 \sim N(0, \sigma_1)$. Note that in model M_1 y depends on $\tilde{x}$. In reality y depends on x, but we assume that there is no noise in the observation of x, so that $\tilde{x} = x$. Naturally, there is noise in the observations of gene expression as well. Adding this to the model makes the detector theoretically more

powerful but also computationally more demanding (see section 2.2):

$$M_2 : \begin{array}{l} \tilde{x} = x + \epsilon_x \\ \tilde{y} = \frac{a}{1+e^{-w_0 - w_1 x}} + \epsilon_y \end{array} \tag{3}$$

where $\epsilon_x \sim N(0, \sigma_x)$ and $\epsilon_y \sim N(0, \sigma_y)$. Using Bayesian inference, our detectors compare the probability of one of the two sigmoidal models to the independence model, given data. Specifically, they use the natural logarithm of the ratio $\frac{P(M_i|D)}{P(M_0|D)}, i = 1, 2$ to rank genes and detect those of interest. D is the data, i. e. the observations of gene expression (of one gene) and drug sensitivity. The higher the value of $ln\ \frac{P(M_i|D)}{P(M_0|D)}, i = 1, 2$ for a specific gene, the more probable that this gene affects drug sensitivity. From now on our detector that use

$$ln\ \frac{P(M_1|D)}{P(M_0|D)} \tag{4}$$

will be called the "regression detector", and the one that use

$$ln\ \frac{P(M_2|D)}{P(M_0|D)} \tag{5}$$

will be called the "modelling detector".

2.2 Underlying mathematics

As stated above, our detectors use the value of (4) and (5) to rank and detect genes. The data, D, consists of N observations $(\tilde{x}_i, \tilde{y}_i),\ i = 1, 2, ..., N$, where N is for example the number of patients in a study or the number of cell lines examined. The observations can be put in vectors $(\tilde{\tilde{x}}, \tilde{\tilde{y}})$. Using Bayes' rule on (4), we get

$$\frac{P(M_1|D)}{P(M_0|D)} = \frac{p(D|M_1)P(M_1)}{p(D)} \cdot \frac{p(D)}{p(D|M_0)P(M_0)}$$

(Derivations are shown only for M_1, but are the same for M_2.) Assuming that all models are equally likely given nothing, i. e. $P(M_1) = P(M_0)$, one is left with

$$\frac{P(M_1|D)}{P(M_0|D)} = \frac{P(D|M_1)}{P(D|M_0)}. \tag{6}$$

However, the models give us little information without their parameters. Let us define the sets of parameters for the models as

$$\phi_0 = \{\mu, \sigma_y\}$$
$$\phi_1 = \{w_0, w_1, \sigma_y, a\}$$
$$\phi_2 = \{w_0, w_1, \sigma_y, a, \sigma_x\}$$

and note that $p(D|M)$ is not known, but $p(D|M, \phi)$ is. From here forth, whenever M or ϕ is not indexed, the statement is general for all three models.

BIC
The relation of the known $p(D|, M, \phi)$, and the unknown $p(D|M)$ is

$$p(D|M) = \int p(D|M, \phi)p(\phi|M)d\phi,$$

where $p(\phi|M)$ is the so called prior probability of the parameters. The number of parameters in the parameter set ϕ determines the number of dimensions of this integral. To escape this somewhat horrifying multidimensional integral an approximation called *Bayesian Information Criterion*, BIC, can be used. It states that

$$ln\ p(D|M) \approx ln\ p(D|\tilde{\phi}, M) - \frac{d}{2} ln\ N, \tag{7}$$

where N is the number of observations, d is the number of parameters of the model and $\tilde{\phi}$ is the the MAP (*maximum a posteriori*) estimate of ϕ, i. e. the configuration that maximizes $p(\phi|D, M)$ [7]. Using BIC in (6), we get

$$ln\frac{P(M_1|D)}{P(M_0|D)} = ln\ p(D|M_1) - ln\ p(D|M_0) \approx ln\ p(D|M_1, \tilde{\phi}_1) - \frac{d_1}{2} ln\ N - ln\ p(D|M_0, \tilde{\phi}_0) + \frac{d_0}{2} ln\ N. \tag{8}$$

Let us now examine what $p(D|M, \phi)$ is. We need to embrace the fact that

$$p(D|M, \phi) = \prod_{i=1}^{N} p(\tilde{y}_i, \tilde{x}_i|M, \phi).$$

and recall that in model M_0 and M_1 noise in x is not accounted for, why $\tilde{x}_i = x_i$ for those models. For model M_0, see equation (1), $\tilde{y} \sim N(\mu, \sigma_0)$, as $\epsilon_0 \sim N(0, \sigma_0)$, and hence

$$\begin{aligned} p(x_i, \tilde{y}_i|\phi_0, M_0) &= \{x_i \text{ and } \tilde{y}_i \text{ independent}\} = p(x_i|\phi_0, M_0)p(\tilde{y}_i|\phi_0, M_0) \\ &= \{x_i \text{ independent of } \phi_0 \text{ and } M_0\} = p(x_i)p(\tilde{y}_i|\phi_0, M_0) \\ &= p(x_i)\frac{1}{\sigma_0\sqrt{2\pi}}\ e^{-\frac{1}{2\sigma_0^2}(\tilde{y}_i-\mu)^2}. \end{aligned} \tag{9}$$

Let us leave $p(x_i)$ for now, it will be clear why later on. Accordingly, for model M_1 (equation (2)) we have

$$p(x_i, \tilde{y}_i|\phi_1, M_1) = p(\tilde{y}_i|x_i, \phi_1, M_1)p(x_i|\phi_1, M_1) = p(x_i)\frac{1}{\sigma_1\sqrt{2\pi}}\ e^{-\frac{1}{2\sigma_1^2}\left(\tilde{y}_i-\frac{a}{1+e^{-w_0-w_1 x_i}}\right)^2}. \tag{10}$$

For model M_2 the situation is a bit more complex since we have noise in x (see equation (3)):

$$\begin{aligned} p(\tilde{x}_i, \tilde{y}_i|\phi_2, M_2) &= p(\tilde{x}_i|\phi_2, M_2)p(\tilde{y}_i|\tilde{x}_i, \phi_2, M_2) \\ &= \int_x p(\tilde{x}_i|\phi_2, M_2, x)p(x|\phi_2, M_2)dx \int_x p(\tilde{y}_i|\phi_2, M_2, x)p(x|\tilde{x}_i, \phi_2, M_2)dx \\ &= \int_x \frac{1}{\sigma_x\sqrt{2\pi}} \cdot e^{-\frac{1}{2\sigma_x^2}(\tilde{x}_i-x)^2} p(x)dx \int_x \frac{1}{\sigma_y\sqrt{2\pi}} \cdot e^{-\frac{1}{2\sigma_y^2}\left(\tilde{y}_i-\frac{a}{1+e^{-w_0-w_1 x}}\right)^2} \cdot \frac{1}{\sigma_x\sqrt{2\pi}} \cdot e^{-\frac{1}{2\sigma_x^2}(x-\tilde{x}_i)^2} dx \\ &= \{\text{assume that p(x) is constant}\} \\ &= p(x) \int_x \frac{1}{\sigma_x\sqrt{2\pi}} \cdot e^{-\frac{1}{2\sigma_x^2}(\tilde{x}_i-x)^2} dx \int_x \frac{1}{\sigma_y\sqrt{2\pi}} \cdot e^{-\frac{1}{2\sigma_y^2}\left(\tilde{y}_i-\frac{a}{1+e^{-w_0-w_1 x}}\right)^2} \cdot \frac{1}{\sigma_x\sqrt{2\pi}} \cdot e^{-\frac{1}{2\sigma_x^2}(x-\tilde{x}_i)^2} dx \\ &= \left\{ \int_x \frac{1}{\sigma_x\sqrt{2\pi}} \cdot e^{-\frac{1}{2\sigma_x^2}(\tilde{x}_i-x)^2} dx = 1 \right\} \\ &= p(x) \int_x \frac{1}{\sigma_y\sqrt{2\pi}} \cdot e^{-\frac{1}{2\sigma_y^2}\left(\tilde{y}_i-\frac{a}{1+e^{-w_0-w_1 x}}\right)^2} \cdot \frac{1}{\sigma_x\sqrt{2\pi}} \cdot e^{-\frac{1}{2\sigma_x^2}(x-\tilde{x}_i)^2} dx. \end{aligned} \tag{11}$$

We are now ready to see what equation (8) result in, which is in fact what our detectors base their decisions on. For the regression model equations (4), (8), (10) and (9) give us

$$
\begin{aligned}
ln\ \frac{P(M_1|D)}{P(M_0|D)} &\approx ln\ p(D|M_1,\tilde{\phi}_1) - \frac{d_1}{2} ln\ N - ln\ p(D|M_0,\tilde{\phi}_0) + \frac{d_0}{2} ln\ N \\
&= \sum_{i=1}^{N} ln \left(p(x_i) \frac{1}{\tilde{\sigma}_1\sqrt{2\pi}}\ e^{-\frac{1}{2\tilde{\sigma}_1^2}\left(\tilde{y}_i - \frac{\tilde{a}}{1+e^{-\tilde{w}_0-\tilde{w}_1 x_i}}\right)^2} \right) - \frac{d_1}{2} ln\ N \\
&\quad - \sum_{i=1}^{N} ln \left(p(x_i) \frac{1}{\tilde{\sigma}_0\sqrt{2\pi}}\ e^{-\frac{1}{2\tilde{\sigma}_0^2}(\tilde{y}_i-\tilde{\mu})^2} \right) + \frac{d_0}{2} ln\ N \\
&= N ln\ p(x_i) - N ln\left(\tilde{\sigma}_1\sqrt{2\pi}\right) - \frac{1}{2\tilde{\sigma}_1^2}\sum_{i=1}^{N}\left(\tilde{y}_i - \frac{\tilde{a}}{1+e^{-\tilde{w}_0-\tilde{w}_1 x_i}}\right)^2 - \frac{d_1}{2} ln\ N \\
&\quad - N ln\ p(x_i) + N ln\left(\tilde{\sigma}_0\sqrt{2\pi}\right) + \frac{1}{2\tilde{\sigma}_0^2}\sum_{i=1}^{N}(\tilde{y}_i-\tilde{\mu})^2 + \frac{d_0}{2} ln\ N \\
&= -N ln\left(\tilde{\sigma}_1\sqrt{2\pi}\right) - \frac{1}{2\tilde{\sigma}_1^2}\sum_{i=1}^{N}\left(\tilde{y}_i - \frac{\tilde{a}}{1+e^{-\tilde{w}_0-\tilde{w}_1 x_i}}\right)^2 - \frac{d_1}{2} ln\ N \\
&\quad + N ln\left(\tilde{\sigma}_0\sqrt{2\pi}\right) + \frac{1}{2\tilde{\sigma}_0^2}\sum_{i=1}^{N}(\tilde{y}_i-\tilde{\mu})^2 + \frac{d_0}{2} ln\ N.
\end{aligned}
\tag{12}
$$

For the modelling detector, we use equations (5), (8), (11) and (9):

$$
\begin{aligned}
ln\ \frac{P(M_2|D)}{P(M_0|D)} &\approx ln\ p(D|M_2,\tilde{\phi}_2) - \frac{d_2}{2} ln\ N - ln\ p(D|M_0,\tilde{\phi}_0) + \frac{d_0}{2} ln\ N \\
&= \sum_{i=1}^{N} ln \left(p(x) \int_x \frac{1}{\tilde{\sigma}_y\sqrt{2\pi}} \cdot e^{-\frac{1}{2\tilde{\sigma}_y^2}\left(\tilde{y}_i - \frac{\tilde{a}}{1+e^{-\tilde{w}_0-\tilde{w}_1 x}}\right)^2} \cdot \frac{1}{\tilde{\sigma}_x\sqrt{2\pi}} \cdot e^{-\frac{1}{2\tilde{\sigma}_x^2}(x-\tilde{x}_i)^2} dx \right) \\
&\quad - \frac{d_2}{2} ln\ N - \sum_{i=1}^{N} ln \left(p(x_i) \frac{1}{\tilde{\sigma}_0\sqrt{2\pi}}\ e^{-\frac{1}{2\tilde{\sigma}_0^2}(\tilde{y}_i-\tilde{\mu})^2} \right) + \frac{d_0}{2} ln\ N \\
&= N ln\ p(x) + \sum_{i=1}^{N} ln \int_x \frac{1}{\tilde{\sigma}_y\sqrt{2\pi}} \cdot e^{-\frac{1}{2\tilde{\sigma}_y^2}\left(\tilde{y}_i - \frac{\tilde{a}}{1+e^{-\tilde{w}_0-\tilde{w}_1 x}}\right)^2} \cdot \frac{1}{\tilde{\sigma}_x\sqrt{2\pi}} \cdot e^{-\frac{1}{2\tilde{\sigma}_x^2}(x-\tilde{x}_i)^2} dx \\
&\quad - \frac{d_2}{2} ln\ N - N ln\ p(x_i) + N ln(\tilde{\sigma}_0\sqrt{2\pi}) + \frac{1}{2\tilde{\sigma}_0^2}\sum_{i=1}^{N}(\tilde{y}_i-\tilde{\mu})^2 + \frac{d_0}{2} ln\ N \\
&= \{\text{assume } p(x_i) = p(x)\} \\
&= \sum_{i=1}^{N} ln \int_x \frac{1}{\tilde{\sigma}_y\sqrt{2\pi}} \cdot e^{-\frac{1}{2\tilde{\sigma}_y^2}\left(\tilde{y}_i - \frac{\tilde{a}}{1+e^{-\tilde{w}_0-\tilde{w}_1 x}}\right)^2} \cdot \frac{1}{\tilde{\sigma}_x\sqrt{2\pi}} \cdot e^{-\frac{1}{2\tilde{\sigma}_x^2}(x-\tilde{x}_i)^2} dx \\
&\quad - \frac{d_2}{2} ln\ N + N ln(\tilde{\sigma}_0\sqrt{2\pi}) + \frac{1}{2\tilde{\sigma}_0^2}\sum_{i=1}^{N}(\tilde{y}_i-\tilde{\mu})^2 + \frac{d_0}{2} ln\ N.
\end{aligned}
\tag{13}
$$

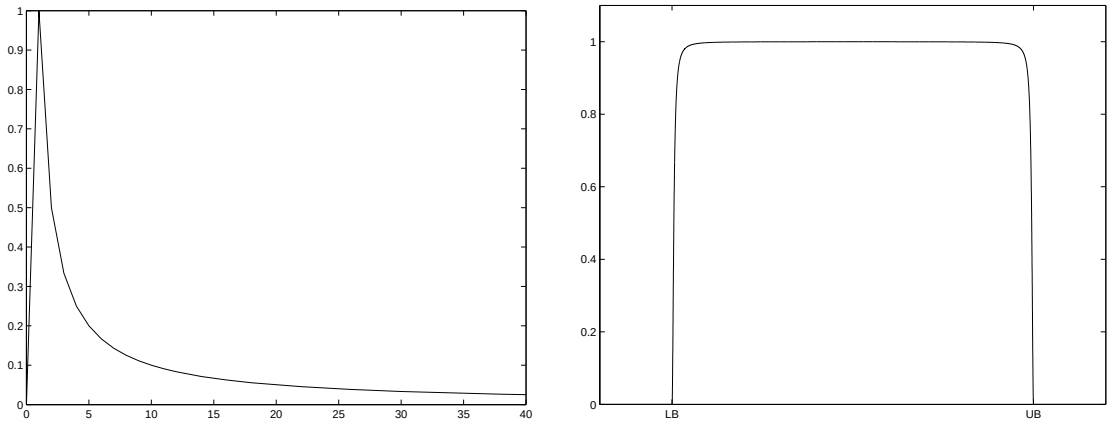
MAP-estimation

To be able to calculate (12) and (13) we need to find the MAP-estimate of the parameters for each model. We search for the values of the parameters that renders the sigmoid or the constant that is best fitted to our data. The derivation of the MAP-estimate is the same for all the models if we now use the general denotion $\phi = \{\varphi_1, ..., \varphi_d\}$ for the parameter sets for the different models.

$$\begin{aligned}
\tilde{\phi} &= \arg\max_{\phi} \; p(\phi|D,M) = \{\text{Bayes' rule}\} = \arg\max_{\phi} \left(\frac{p(D|\phi,M)p(\phi|M)}{p(D|M)} \right) \\
&= \{p(D|M) \text{ independent of } \phi\} = \arg\max_{\phi} \; p(D|\phi,M)p(\phi|M) \\
&= \arg\max_{\phi} \; ln\left(p(D|\phi,M)p(\phi|M)\right) = \arg\max_{\phi} \; \left(ln\; p(D|\phi,M) + ln\; p(\phi|M)\right) \\
&= \arg\max_{\phi} \left(ln \left(\prod_{i=1}^{N} p(\tilde{x}_i, \tilde{y}_i|\phi,M) \right) + ln \left(\prod_{i=1}^{d} p(\varphi_i|M) \right) \right) \\
&= \arg\max_{\phi} \left(\sum_{i=1}^{N} ln\; p(\tilde{x}_i, \tilde{y}_i|\phi,M) + \sum_{i=1}^{d} ln\; p(\varphi_i|M) \right)
\end{aligned} \tag{14}$$

Priors

In equation (9), (10) and (11) the expressions for $p(\tilde{x}_i, \tilde{y}_i|M,\phi)$ are stated. To find the MAP-estimate of the parameters, one also needs the prior probabilities of the parameters, $p(\varphi_i|M)$. The priors are supposed to reflect our *a priori* information about the parameters, e. g. which values we think are more probable than others. The priors of the standard deviation, i. e. σ, a and μ are shown in Figure 2.



(a) Prior probability for σ.

(b) Prior probability for a and μ. The lower and upper boundaries (LB and UB) differ for a and μ.

Figure 2: The prior probabilities of the parameters σ, a and μ.

The prior of the standard deviations should be non-zero only for $\sigma > 0$, and smaller values of σ should be more probable than bigger ones. A discontinuity in $\sigma = 0$, like $f(0) = 0$ but $\lim_{\sigma \to 0^+} f(\sigma) = \infty$, would not be desirable, however, as this would make it harder to find the maximum. Discontinuities in the priors do not

"warn" when you approach forbidden values of a parameter, which may result in one getting stuck at these discontinuities. Therefore we would prefer a steep slope upwards for small $\sigma > 0$. To fit these criteria, the prior of the standard deviation was chosen as $f(\sigma) = \frac{\sigma}{\sigma^2+h^2}$, where h affects (in a not uncomplicated way) at which σ the peak appears, see Figure 2(a). After emprirical testing, $h = \frac{1}{400}$ was decided on.

The priors of a and μ are both almost uniform, but since we want to avoid discontinuities, the smoother function

$$f(a) = \frac{(a-LB)^2}{(a-LB)^2+h^2} + \frac{(a-UB)^2}{(a-UB)^2+h^2} - \frac{(UB-LB)^2}{(UB-LB)^2+h^2}$$

was chosen (equation shown for a, but is the same for μ), see Figure 2(b). LB is the lower boundary and UB is the upper, and the values of these are the only difference between the prior of a and μ. Here h determines the steepness at the boundaries, and was after empirical testing set to $\frac{UB-LB}{200}$. As the SI values can only range from 0 to 1 (although noise can result in values outside this interval), LB and UB for μ were 0 and 1, respectively. The parameter a sets the height of the sigmoid step, as described earlier, and this is not always 1. Therefore, LB and UB for a were 0.5 and 1, respectively.

There is no need to normalize the priors, as constants do not change the position of the global maximum of $p(\phi|D,M)$ which we search for. The parameters w_0 and w_1 have uniform priors ranging from $-\infty$ to ∞ and are thus constant, and need therefore not be taken into the function to be maximized. The grounds for choosing these priors are as follows. The slope, w_1, of the s-shape in the sigmoid could have any value, a negative sigmoid also implies (negative) correlation between gene expression and drug sensitivity. The specific instance of $w_1 = 0$ would result in a constant function, which means that this sigmoid would fit well to data that actually belongs to M_0. However, the BIC approximation (equation (7)), punishes complex models as these have larger values of d. The w_0 parameter determines where on the x-axis the step is sited. Neither for this parameter is there any reason to think that some values should be more probable than others.

Back to the MAP-estimation
We are now ready to finish the derivation of the MAP-estimate. Let us pick up where we left off, at equation (14), but let us examine the models individually from now on. The function that should be minimized to find the MAP-estimate of the parameters of model M_0 is

$$\begin{aligned}
J_0(\phi_0) &= -\sum_{i=1}^{N} ln\ p(\tilde{x}_i,\tilde{y}_i|\phi_0,M_0) - \sum_{i=1}^{d} ln\ p(\varphi_i|M_0) = -\sum_{i=1}^{N} ln\left(p(x_i)\frac{1}{\sigma_0\sqrt{2\pi}}e^{-\frac{1}{2\sigma_0^2}(\tilde{y}_i-\mu)^2}\right) \\
&- ln\left(\frac{(\mu-0)^2}{(\mu-0)^2+0.005^2} + \frac{(\mu-1)^2}{(\mu-1)^2+0.005^2} - \frac{(1-0)^2}{(1-0)^2+0.005^2}\right) - ln\left(\frac{\sigma_0}{\sigma_0^2+0.0025^2}\right) \\
&= -\sum_{i=1}^{N}\left(ln\left(p(x_i)\frac{1}{\sigma_0\sqrt{2\pi}}\right) - \frac{1}{2\sigma_0^2}(\tilde{y}_i-\mu)^2\right) - ln\left(\frac{\mu^2}{\mu^2+0.005^2} + \frac{(\mu-1)^2}{(\mu-1)^2+0.005^2} - \frac{1}{1+0.005^2}\right) \\
&- ln\left(\frac{\sigma_0}{\sigma_0^2+0.0025^2}\right).
\end{aligned}$$

Since $\sqrt{2\pi}$ and $p(x_i)$ is not dependent on ϕ_0, this can be simplified and $J_0(\phi_0)$ redefined as

$$\begin{aligned}
J_0(\phi_0) &= N ln\ \sigma_0 + \frac{1}{2\sigma_0^2}\sum_{i=1}^{N}(\tilde{y}_i-\mu)^2 \\
&- ln\left(\frac{\mu^2}{\mu^2+0.005^2} + \frac{(\mu-1)^2}{(\mu-1)^2+0.005^2} - \frac{1}{1+0.005^2}\right) - ln\left(\frac{\sigma_0}{\sigma_0^2+0.0025^2}\right).
\end{aligned} \tag{15}$$

By similar derivations, we get

$$J_1(\phi_1) = N ln\ \sigma_1 + \frac{1}{2\sigma_1^2}\sum_{i=1}^{N}\left(\tilde{y}_i - \frac{a}{1+e^{-w_0-w_1\tilde{x}_i}}\right)^2 - ln\left(\frac{\sigma_1}{\sigma_1^2+0.0025^2}\right) - ln\left(\frac{(a-0.5)^2}{(a-0.5)^2+0.005^2} + \frac{(a-1)^2}{(a-1)^2+0.005^2} - \frac{0.5^2}{0.5^2+0.005^2}\right) \tag{16}$$

and

$$\begin{aligned} J_2(\phi_2) &= -\sum_{i=1}^{N} ln \int_x \frac{1}{\sigma_y\sqrt{2\pi}} \cdot e^{-\frac{1}{2\sigma_y^2}\left(\tilde{y}_i - \frac{a}{1+e^{-w_0-w_1 x}}\right)^2} \cdot \frac{1}{\sigma_x\sqrt{2\pi}} \cdot e^{-\frac{1}{2\sigma_x^2}(x-\tilde{x})^2} dx \ - ln\left(\frac{\sigma_y}{\sigma_y^2+0.0025^2}\right) \\ &- ln\left(\frac{(a-0.5)^2}{(a-0.5)^2+0.005^2} + \frac{(a-1)^2}{(a-1)^2+0.005^2} - \frac{0.5^2}{0.5^2+0.005^2}\right) - ln\left(\frac{\sigma_x}{\sigma_x^2+0.0025^2}\right) \\ &= -\sum_{i=1}^{N} ln\left(\frac{1}{\sigma_y\sigma_x 2\pi}\int_x e^{-\frac{1}{2\sigma_y^2}\left(\tilde{y} - \frac{a}{e^{-w_0-w_1 x}}\right)^2 - \frac{1}{2\sigma_x^2}(x-\tilde{x})^2} dx\right) - ln\left(\frac{\sigma_y}{\sigma_y^2+0.0025^2}\right) \\ &- ln\left(\frac{(a-0.5)^2}{(a-0.5)^2+0.005^2} + \frac{(a-1)^2}{(a-1)^2+0.005^2} - \frac{0.5^2}{0.5^2+0.005^2}\right) - ln\left(\frac{\sigma_x}{\sigma_x^2+0.0025^2}\right) \end{aligned} \tag{17}$$

To sum up the procedure, given gene expression data of a gene and corresponding SI values the MAP-estimate of the parameters of model M_1, or M_2, depending on which detector is used, and M_0 are found by minimizing $J_1(\phi_1)$ or $J_2(\phi_2)$ and $J_0(\phi_0)$, respectively. Knowing the MAP-estimates, the expression in equation (12) and (13) can be calculated and a ranking value according to the regression detector and the modelling detector, respectively, is determined for this gene.

2.3 ROC graphs

Drawing ROC graphs is a way of visualising the performance of classifiers. The abbreviation reads *Receiver Operating Characteristics*. Every classifier has a false positive rate, FP rate, and a true positive rate, TP rate. The false positive rate is $\frac{FP}{FP+TN}$, where TN is true negatives. The true positive rate is $\frac{TP}{TP+FN}$, where FN is false negatives. Hence, FP and TP rates range from 0 to 1. A perfect classifier that makes no mistakes will have a TP rate of 1 and a FP rate of 0.

A classifier usually makes its decision based on whether a specific value is higher or lower than a certain threshold value. For instance, our regression detector uses the value of $ln\frac{P(M_1|D)}{P(M_0|D)}$ to classify genes as affecting drug sensitivity or not. Different threshold values will result in different FP and TP rates. When constructing the ROC graph for a classifier you successively increase the threshold value from so low that all instances are classified as positives, which yields the point (1,1), to so high that no instance is classified as positive, which yields the point (0,0), or the other way round [8]. A classifier that guess class at random for each instance, will theoretically produce a ROC graph that is a straight line from (0,0) to (1,1). Classifiers that are better than a random guessing one will give a ROC graph where the line moves from (0,0) to (1,1) not straight but bent, closing in on the (0,1) corner. The ROC graph of a perfect classifier is a straight line from (0,0) to (0,1) and on to (1,1).

2.4 Experimental Details

The performances of our detectors have been compared to the absolute value of correlation coefficient detector on simulated data. Our detectors have not been used on biological data, because the time of this project run out.

The gene expression values of the simulated data were uniformly distributed in the interval [0 1000], which corresponds to a 1000-fold difference in gene expression between the two individuals with most extreme expressions. The performances of the detectors (including the correlation coefficient detector) are influenced by the standard deviation of the noise in the data, the size of the dataset (i. e. N, number of observations) and the steepness of the sigmoid slope, which can be seen as a measure of nonlinearity. Therefore these parameters have been varied when producing simulated data. The parameter a does not affect the performance of the correlation coefficient detector significantly if varied between 0.5 and 1 (data not shown). It could however affect the performance of our detectors, but probably not to a great extent. Therefore, a was kept at the constant value 0.8 when producing simulated data.

The values of w_1, which determines the steepness of the sigmoid slope, were 0.01, 0.1 and 0.33. These values implied that the width of the region of the sigmoid with SI values that are not 0 or a, the actual s so to speak, were 1000, 100 and 30 gene expresson units wide, respectively. Only sigmoids with positive slopes were produced, because the sign of the slope should not affect the performance of any of the detectors. Since the data points are uniformly distributed along the gene expression axis, the steeper the slope the fewer datapoints will be in the s-shape region. Specifically, in the case of $w_1 = 0.01$ all datapoints are within the s-shape because the slope is so flat, which gives the data a nearly linear correlation. The w_0 parameter, which determines the site of the s-shape on the x-axis, was picked randomly from the uniform distribution whose boundaries were selected so that the entire s-shape of the sigmoid was within the [0 1000] interval. This uniform distribution was [-1000w_1+5 -5], which needs some explanation. Empirical testing showed that the ends of the s-shape are located at $-\frac{w_0}{w_1} - \frac{5}{w_1}$ and $-\frac{w_0}{w_1} + \frac{5}{w_1}$ when w_1 is positive, which is the case here. Since the ends of the s-shape should be within the interval [0 1000], we have

$$\begin{aligned} 0 &\leq -\frac{w_0}{w_1} - \frac{5}{w_1} \\ \frac{w_0}{w_1} &\leq -\frac{5}{w_1} \\ w_0 &\leq -5 \end{aligned}$$

and

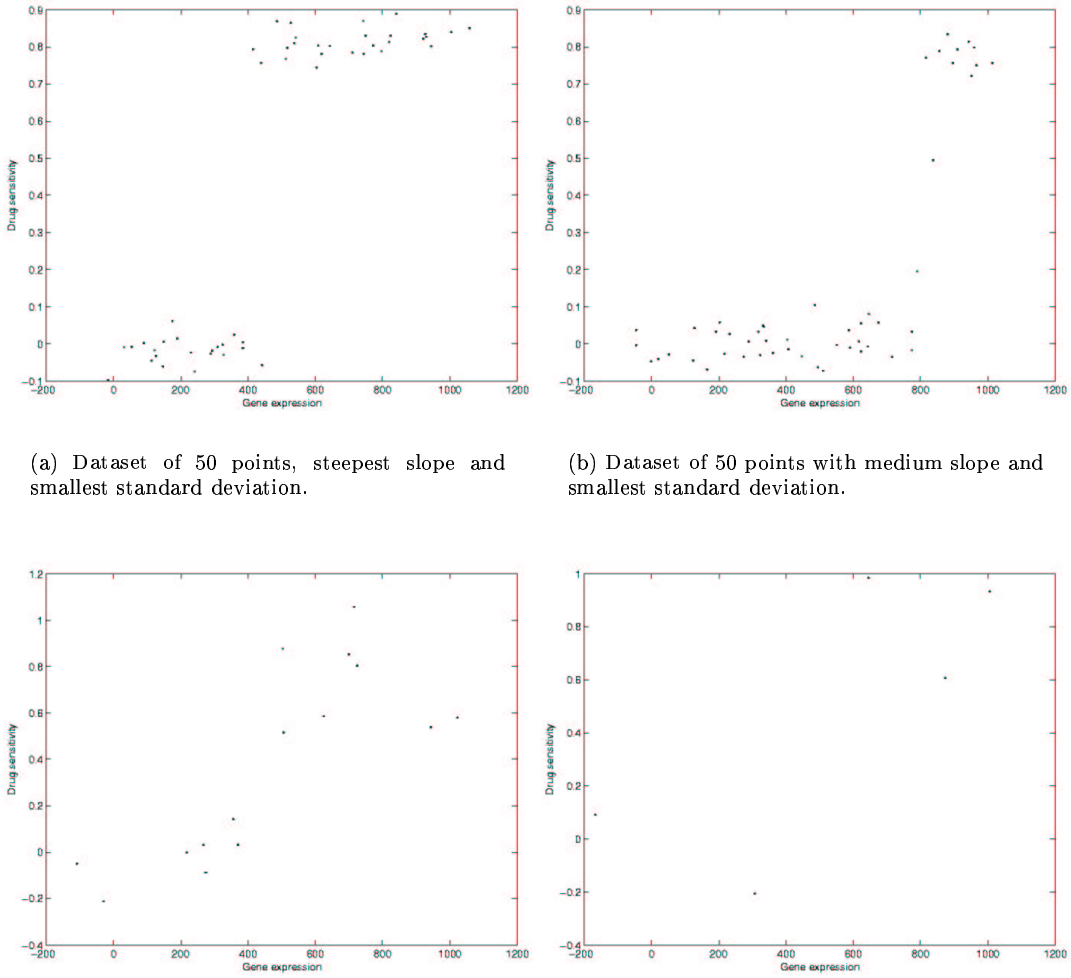
$$\begin{aligned} -\frac{w_0}{w_1} + \frac{5}{w_1} &\leq 1000 \\ \frac{5}{w_1} - 1000 &\leq \frac{w_0}{w_1} \\ -1000w_1 + 5 &\leq w_0 \end{aligned}$$

and hence, $w_0 \in$[-1000w_1+5 -5].

The standard deviation used was adjusted to the range of the data. Expressed in percent the standard deviations were 5%, 15% and 25%. This means that as the range gene expression values is always 1000 units wide, the standard deviation in gene expression was 50, 150 and 250 units, respectively. Since $a = 0.8$ in the simulations, the SI range was 0.8 units, which means that the standard deviations in SI values were 0.04, 0.12 and 0.2 units,

respectively. To get a sense of the meaning of these values, contemplate that 69% of infinitely many samples from a normal distribution will lie within one standard deviation from the expectation value, and 96% of them within two standard deviations. See also Figure 3.

The dataset sizes used were 5, 15, 30, 50 and 75. In Figure 3 four examples of datasets are shown.



(a) Dataset of 50 points, steepest slope and smallest standard deviation.

(b) Dataset of 50 points with medium slope and smallest standard deviation.

(c) Dataset of 15 points with flattest slope and largest standard deviation.

(d) Dataset of 5 points, medium slope and medium standard deviation.

Figure 3: Four examples of simulated datasets.

Consequently, as we had three different slopes, three standard deviations and five dataset sizes, $3 \cdot 3 \cdot 5 = 45$ combinations of parameter settings were used. For each parameter setting 200 sigmoid datasets (the positives) and 200 datasets of model M_0 (the negatives) were produced. The M_0 datasets had the same standard deviation as the corresponding sigmoid datasets, whereas μ was randomly picked from the uniform distribution [0 1]. All

datasets were ranked by the three detectors, and a ROC graph (see section 2.3) was constructed for each of them. The ROC graphs consist of at most 400 points, the event of several datasets having the same rank will lower the number of points. For comparison of the detectors the TP rate corresponding to a FP rate of 5% and 10% were studied. An example of ROC curves of the three detectors is shown in Figure 4.

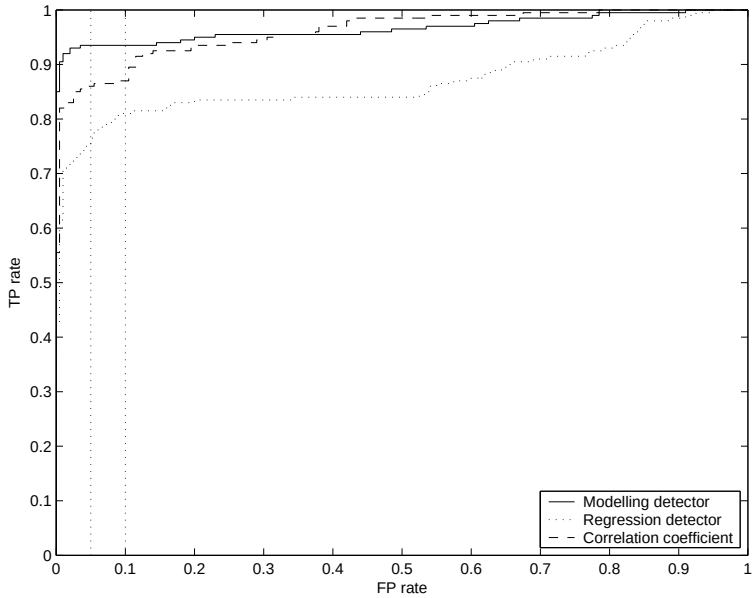


Figure 4: An example of ROC curves of the three detectors. False positive rate 5% and 10% are marked.

As two values were picked from each one of the 45 ROC curves, a total of 90 values were examined. The upper and lower boundary of the 95% *highest posterior density region* [9], from now on denoted LB_{95} and UB_{95}, of the TP rate was calculated for each and every one of these values. These credibility intervals tell us our uncertainty about the true TP rate value. If another 200 positive datsets were produced with the same parameter settings, we might get another TP rate at FP rate 5% for the same detector.

The programming environment MATLAB (Mathworks Inc., USA) version 6.5.0 was imployed for all calculations. MATLAB was run on a server of eight 900 MHz 64-bit processors sharing a primary memory of 16 GB. Running several MATLABs in parallell, approximately three of these processors were at my disposal, why the calculation time to produce these results was about a week and a half.

MATLAB's function fminsearch was used to search for the global minimum of the functions $J_0(\phi_0), J_1(\phi_1)$ and $J_2(\phi_2)$. The minimizer fmincon and fminunc was tried out as well, but was abandoned because fminsearch was most efficient when tested (data not shown) and theoretically most robust. An enhanced steepest descent minimizer was implemented, but this one, too, was abandoned. There were problems with zigzag behaviour and further development was considered to be a waste of time. As fminsearch finds local minima, three different start guesses were given to fminsearch, and the best result (i. e. the local minimum with lowest value of $J(\phi)$) was chosen. The start guesses were randomly picked in the parameter room defined by the uniform distributions $\mu \in$ [0 1]; $\sigma \in$ [0 0.8]; $a \in$ [0.6 1]; $w_1 \in$ [-0.3 0.3] and $w_0 \in$ [-(1000|w_1|-5) -5] if w_1 was positive (see derivation above) or $w_0 \in$ [5 1000|w_1|-5] if w_1 was negative (similar derivation to the case with positive w_1). There was also the constraint that a new start guess was accepted only if the euklidian distance to the previous one was

above a certain threshold value.

The integral in equation (17) and (13) was calculated numerically from $\tilde{x}_i - 5\sigma_x$ to $\tilde{x}_i + 5\sigma_x$. Outside this interval, the $\frac{1}{\sigma_x\sqrt{2\pi}}e^{-\frac{1}{2\sigma_x^2}(x-\tilde{x}_i)^2}$ part of the integrand was assumed to be sufficiently close to 0 to render the entire integrand approximately 0. After empirical testing, division of the integration interval into 800 parts was decided to be a good enough approximation.

The ultimately used code consists of approximately 1300 lines. Much more have been written in total, of course. Additional code has been used for all kinds of testing as well as various detours. All in all, approximately 4000 lines of code has been written.

3 Results

The TP rates including their UB_{95} and LB_{95}, as discussed in section 2.4, are presented in Table 1 and Table 2. One detector is from now on said to be better than another, at a specific parameter setting, if its LB_{95} of the TP rate was higher or equal to the other detector's UB_{95}. Bold values in columns 2 and 5 mean that the detector of that column was better than the correlation coefficient detector. Bold values in column 8 mean that the correlation coefficient detector performed better than the modelling detector, whereas italic values mean that the regression detector was defeated by the correlation coefficient detector. Bold italics values represent cases where the correlation coefficient detector performed better than both of our detectors.

The results of FP rate 5% and 10% differed very little, why only the results of FP rate 5% from here forth are studied. Therefore, all given TP rates are at FP rate 5% unless otherwise stated.

Visualisations of the TP rates can be seen in Figure 5 - Figure 7, where TP rates for the three detectors with 95% highest posterior density boundaries bars are shown for all nine combinations of slopes and standard deviations.

Steepest slope
In Figure 5 we see that our modelling detector performed extremely well for all dataset sizes when the standard deviation was the smallest and very well at the medium standard deviation. Furthermore, it was much better than the correlation coefficient on small dataset sizes. For instance, we see in Figure 5(b) that with a dataset size of five points, the correlation coefficient detector and modelling detector had a TP rate of approximately 0.21 and 0.85, respectively. For the largest standard deviation our modelling detector was better only at the smallest dataset size.

Our regression detector was better than the correlation coefficient for the smallest dataset size at all standard deviations. However, the correlation coefficient was better at the largest dataset sizes, whereas there was no significant difference in performances at the medium dataset sizes.

Medium slope
Again, the modelling detector performed very well for the small and medium standard deviations, see Figure 6, and indeed it was far superior to the correlation coefficient detector for the smallest dataset size and significantly better at the second smallest dataset size. For the largest standard deviation, there was only a significant difference in the performances of these two detectors for the smallest dataset size, but here neither of the detector performed satisfactory.

Our regression detector was better than the correlation coefficient for the smallest dataset size at all standard deviations. However, the correlation coefficient was better at the largest dataset sizes, whereas there was no significant difference in performances at the medium dataset sizes.

Flattest slope
Figure 7 tells us that for the smallest standard deviation there was no significant difference in the performances of the modelling and correlation coefficient detectors, both of them performed extremely well. For the largest standard deviation, the correlation coefficient detector was better than the modelling detector for small dataset sizes. However, for the medium standard deviation, our modelling detector performed better on the smallest dataset.

For the smallest standard deviation both our regression detector and the correlation coefficient detector performed very well, although the correlation coefficient was generally better. For the larger standard deviations, the regression detector was generally better at the smallest dataset size whereas the correlation coefficient was broadly speaking the winner at the other dataset sizes.

Summary
For small dataset sizes and not too flat slopes our modelling detector was much better, i. e. it had a higher true positive rate, than the correlation coefficient detector. On larger dataset sizes these detectors performed, with a few exceptions, equally well. The only circumstances in which the correlation coefficient performed better than the modelling detector was when the slope was the flattest and the standard deviation was large or medium and the dataset sizes were small.

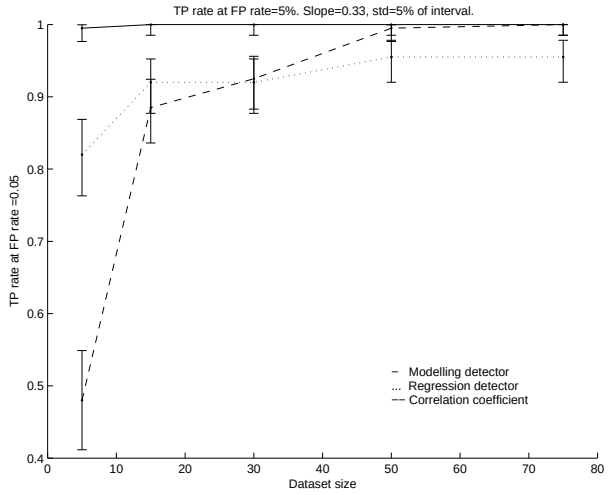
Our regression detector could only defeat the correlation coefficient at the smallest dataset size, while the correlation coefficient was better at the largest dataset sizes. For the medium dataset sizes, there was generally no significant difference in the performances of our regression detector and the correlation coefficient detector.

	TP rates at FP rate = 0.05								
	Modelling det.			Regression det.			Corr. coeff. det.		
Parameter settings	LB_{95}	Value	UB_{95}	LB_{95}	Value	UB_{95}	LB_{95}	Value	UB_{95}
w_1=0.01 ds=5 std=0.05	0.95	0.98	0.99	0.85	0.90	0.94	0.96	*0.99*	1.00
w_1=0.01 ds=5 std=0.15	0.84	**0.89**	0.93	0.71	**0.77**	0.82	0.40	0.47	0.54
w_1=0.01 ds=5 std=0.25	0.11	0.16	0.21	0.37	**0.44**	0.51	0.22	**0.28**	0.34
w_1=0.01 ds=15 std=0.05	0.98	1.00	1.00	0.93	0.97	0.98	0.99	*1.00*	1.00
w_1=0.01 ds=15 std=0.15	0.93	0.96	0.98	0.89	0.93	0.96	0.99	***1.00***	1.00
w_1=0.01 ds=15 std=0.25	0.69	0.75	0.81	0.75	0.81	0.86	0.93	***0.96***	0.98
w_1=0.01 ds=30 std=0.05	0.99	1.00	1.00	0.93	0.97	0.98	0.99	*1.00*	1.00
w_1=0.01 ds=30 std=0.15	0.99	1.00	1.00	0.95	0.98	0.99	0.99	1.00	1.00
w_1=0.01 ds=30 std=0.25	0.84	0.89	0.92	0.81	0.87	0.91	0.99	***1.00***	1.00
w_1=0.01 ds=50 std=0.05	0.98	1.00	1.00	0.97	0.99	1.00	0.99	1.00	1.00
w_1=0.01 ds=50 std=0.15	0.99	1.00	1.00	0.95	0.98	0.99	0.99	1.00	1.00
w_1=0.01 ds=50 std=0.25	0.97	0.99	1.00	0.88	0.92	0.95	0.99	*1.00*	1.00
w_1=0.01 ds=75 std=0.05	0.99	1.00	1.00	0.93	0.97	0.98	0.99	*1.00*	1.00
w_1=0.01 ds=75 std=0.15	0.99	1.00	1.00	0.93	0.97	0.98	0.99	*1.00*	1.00
w_1=0.01 ds=75 std=0.25	0.99	1.00	1.00	0.90	0.94	0.97	0.99	*1.00*	1.00
w_1=0.1 ds=5 std=0.05	0.97	**0.99**	1.00	0.75	**0.81**	0.86	0.24	0.31	0.37
w_1=0.1 ds=5 std=0.15	0.77	**0.83**	0.87	0.70	**0.76**	0.82	0.18	0.24	0.30
w_1=0.1 ds=5 std=0.25	0.31	**0.38**	0.44	0.31	**0.38**	0.45	0.05	0.09	0.13
w_1=0.1 ds=15 std=0.05	0.99	**1.00**	1.00	0.91	0.95	0.97	0.91	0.95	0.97
w_1=0.1 ds=15 std=0.15	0.90	**0.94**	0.97	0.75	0.81	0.86	0.77	0.83	0.88
w_1=0.1 ds=15 std=0.25	0.59	0.66	0.72	0.65	0.72	0.77	0.63	0.70	0.76
w_1=0.1 ds=30 std=0.05	0.99	1.00	1.00	0.93	0.97	0.98	0.99	*1.00*	1.00
w_1=0.1 ds=30 std=0.15	0.95	0.98	0.99	0.91	0.95	0.97	0.93	0.96	0.98
w_1=0.1 ds=30 std=0.25	0.74	0.80	0.85	0.64	0.71	0.77	0.76	0.82	0.87
w_1=0.1 ds=50 std=0.05	0.99	1.00	1.00	0.94	0.97	0.99	0.99	1.00	1.00
w_1=0.1 ds=50 std=0.15	0.99	1.00	1.00	0.88	0.92	0.95	0.95	*0.98*	0.99
w_1=0.1 ds=50 std=0.25	0.91	0.95	0.97	0.75	0.81	0.86	0.85	0.90	0.93
w_1=0.1 ds=75 std=0.05	0.99	1.00	1.00	0.93	0.97	0.98	0.99	*1.00*	1.00
w_1=0.1 ds=75 std=0.15	0.98	1.00	1.00	0.88	0.93	0.96	0.99	*1.00*	1.00
w_1=0.1 ds=75 std=0.25	0.95	0.98	0.99	0.79	0.85	0.89	0.91	*0.95*	0.97
w_1=0.33 ds=5 std=0.05	0.98	**1.00**	1.00	0.76	**0.82**	0.87	0.41	0.48	0.55
w_1=0.33 ds=5 std=0.15	0.79	**0.85**	0.89	0.57	**0.64**	0.70	0.16	0.22	0.28
w_1=0.33 ds=5 std=0.25	0.43	**0.50**	0.57	0.35	**0.42**	0.49	0.08	0.12	0.17
w_1=0.33 ds=15 std=0.05	0.99	**1.00**	1.00	0.88	0.92	0.95	0.84	0.89	0.92
w_1=0.33 ds=15 std=0.15	0.87	**0.92**	0.95	0.75	0.81	0.86	0.69	0.76	0.81
w_1=0.33 ds=15 std=0.25	0.52	0.59	0.65	0.48	0.55	0.62	0.48	0.55	0.61
w_1=0.33 ds=30 std=0.05	0.99	**1.00**	1.00	0.88	0.92	0.95	0.88	0.93	0.96
w_1=0.33 ds=30 std=0.15	0.95	**0.98**	0.99	0.85	0.90	0.94	0.85	0.90	0.93
w_1=0.33 ds=30 std=0.25	0.80	0.86	0.90	0.67	0.74	0.79	0.77	0.83	0.87
w_1=0.33 ds=50 std=0.05	0.99	1.00	1.00	0.92	0.96	0.98	0.98	1.00	1.00
w_1=0.33 ds=50 std=0.15	0.98	1.00	1.00	0.84	0.89	0.92	0.92	0.96	0.98
w_1=0.33 ds=50 std=0.25	0.90	0.94	0.96	0.69	0.76	0.81	0.81	0.86	0.90
w_1=0.33 ds=75 std=0.05	0.99	1.00	1.00	0.92	0.96	0.98	0.99	*1.00*	1.00
w_1=0.33 ds=75 std=0.15	0.98	1.00	1.00	0.91	0.95	0.97	0.95	0.98	0.99
w_1=0.33 ds=75 std=0.25	0.93	0.96	0.98	0.72	0.78	0.83	0.88	*0.93*	0.96

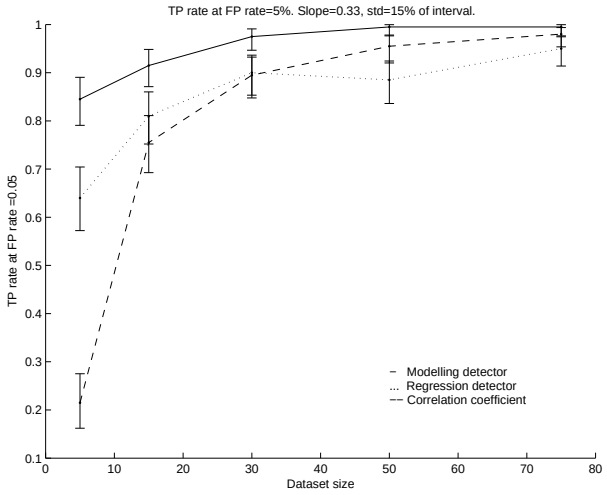
Table 1: TP rates at FP rate = 0.05. Bold values in columns 2 and 5 mean that the detector of that column was better than the correlation coefficient. Bold values in column 8 mean that the correlation coefficient performed better than the modelling detector, whereas italic values mean that the regression detector was defeated by the correlation coefficient. Bold italics values represent cases where the correlation coefficient performed better than both of our detectors.

	TP rates at FP rate = 0.10								
	Modelling det.			Regression det.			Corr. coeff. det.		
Parameter settings	LB_{95}	Value	UB_{95}	LB_{95}	Value	UB_{95}	LB_{95}	Value	UB_{95}
w_1=0.01 ds=5 std=0.05	0.95	0.98	0.99	0.88	0.93	0.96	0.99	*1.00*	1.00
w_1=0.01 ds=5 std=0.15	0.88	**0.93**	0.96	0.76	0.82	0.87	0.74	0.80	0.85
w_1=0.01 ds=5 std=0.25	0.44	0.51	0.57	0.43	0.50	0.57	0.43	0.50	0.56
w_1=0.01 ds=15 std=0.05	0.98	1.00	1.00	0.93	0.97	0.98	0.99	*1.00*	1.00
w_1=0.01 ds=15 std=0.15	0.93	0.97	0.98	0.92	0.96	0.98	0.99	bf*1.00*	1.00
w_1=0.01 ds=15 std=0.25	0.75	0.81	0.86	0.80	0.85	0.89	0.94	bf*0.97*	0.99
w_1=0.01 ds=30 std=0.05	0.99	1.00	1.00	0.93	0.97	0.98	0.99	*1.00*	1.00
w_1=0.01 ds=30 std=0.15	0.99	1.00	1.00	0.95	0.98	0.99	0.99	1.00	1.00
w_1=0.01 ds=30 std=0.25	0.88	0.93	0.96	0.84	0.89	0.92	0.99	bf*1.00*	1.00
w_1=0.01 ds=50 std=0.05	0.98	1.00	1.00	0.97	0.99	1.00	0.99	1.00	1.00
w_1=0.01 ds=50 std=0.15	0.99	1.00	1.00	0.95	0.98	0.99	0.99	1.00	1.00
w_1=0.01 ds=50 std=0.25	0.98	1.00	1.00	0.89	0.93	0.96	0.99	*1.00*	1.00
w_1=0.01 ds=75 std=0.05	0.99	1.00	1.00	0.93	0.97	0.98	0.99	*1.00*	1.00
w_1=0.01 ds=75 std=0.15	0.99	1.00	1.00	0.93	0.97	0.98	0.99	*1.00*	1.00
w_1=0.01 ds=75 std=0.25	0.99	1.00	1.00	0.90	0.94	0.97	0.99	*1.00*	1.00
w_1=0.1 ds=5 std=0.05	0.97	**0.99**	1.00	0.77	**0.83**	0.88	0.58	0.65	0.71
w_1=0.1 ds=5 std=0.15	0.85	**0.90**	0.93	0.73	**0.79**	0.84	0.34	0.41	0.47
w_1=0.1 ds=5 std=0.25	0.55	**0.62**	0.68	0.40	**0.47**	0.54	0.16	0.22	0.28
w_1=0.1 ds=15 std=0.05	0.99	1.00	1.00	0.91	0.95	0.97	0.95	0.98	0.99
w_1=0.1 ds=15 std=0.15	0.90	0.94	0.97	0.79	0.85	0.89	0.86	0.91	0.94
w_1=0.1 ds=15 std=0.25	0.60	0.67	0.73	0.69	0.76	0.81	0.76	**0.82**	0.86
w_1=0.1 ds=30 std=0.05	0.99	1.00	1.00	0.93	0.97	0.98	0.99	*1.00*	1.00
w_1=0.1 ds=30 std=0.15	0.95	0.98	0.99	0.91	0.95	0.97	0.94	0.97	0.99
w_1=0.1 ds=30 std=0.25	0.77	0.83	0.87	0.70	0.77	0.82	0.80	0.86	0.90
w_1=0.1 ds=50 std=0.05	0.99	1.00	1.00	0.94	0.97	0.99	0.99	1.00	1.00
w_1=0.1 ds=50 std=0.15	0.99	1.00	1.00	0.89	0.93	0.96	0.98	*1.00*	1.00
w_1=0.1 ds=50 std=0.25	0.93	0.97	0.98	0.76	0.82	0.87	0.91	*0.95*	0.97
w_1=0.1 ds=75 std=0.05	0.99	1.00	1.00	0.93	0.97	0.98	0.99	*1.00*	1.00
w_1=0.1 ds=75 std=0.15	0.98	1.00	1.00	0.88	0.93	0.96	0.99	*1.00*	1.00
w_1=0.1 ds=75 std=0.25	0.95	0.98	0.99	0.80	0.86	0.90	0.93	*0.97*	0.98
w_1=0.33 ds=5 std=0.05	0.98	**1.00**	1.00	0.82	**0.87**	0.91	0.60	0.67	0.73
w_1=0.33 ds=5 std=0.15	0.82	**0.88**	0.92	0.64	**0.71**	0.77	0.29	0.36	0.42
w_1=0.33 ds=5 std=0.25	0.52	**0.59**	0.65	0.46	**0.53**	0.59	0.18	0.24	0.30
w_1=0.33 ds=15 std=0.05	0.99	**1.00**	1.00	0.88	0.93	0.96	0.90	0.94	0.97
w_1=0.33 ds=15 std=0.15	0.88	**0.92**	0.95	0.77	0.83	0.87	0.75	0.81	0.86
w_1=0.33 ds=15 std=0.25	0.55	0.62	0.69	0.59	0.66	0.72	0.57	0.64	0.70
w_1=0.33 ds=30 std=0.05	0.99	**1.00**	1.00	0.88	0.92	0.95	0.93	0.96	0.98
w_1=0.33 ds=30 std=0.15	0.95	0.98	0.99	0.87	0.91	0.94	0.90	0.94	0.96
w_1=0.33 ds=30 std=0.25	0.82	0.87	0.91	0.71	0.77	0.82	0.85	*0.90*	0.93
w_1=0.33 ds=50 std=0.05	0.99	1.00	1.00	0.92	0.96	0.98	0.99	*1.00*	1.00
w_1=0.33 ds=50 std=0.15	0.98	1.00	1.00	0.85	0.90	0.93	0.93	*0.97*	0.98
w_1=0.33 ds=50 std=0.25	0.90	0.94	0.96	0.75	0.81	0.86	0.82	0.87	0.91
w_1=0.33 ds=75 std=0.05	0.99	1.00	1.00	0.92	0.96	0.98	0.99	*1.00*	1.00
w_1=0.33 ds=75 std=0.15	0.98	1.00	1.00	0.92	0.96	0.98	0.97	0.99	1.00
w_1=0.33 ds=75 std=0.25	0.93	0.96	0.98	0.74	0.80	0.85	0.92	*0.96*	0.98

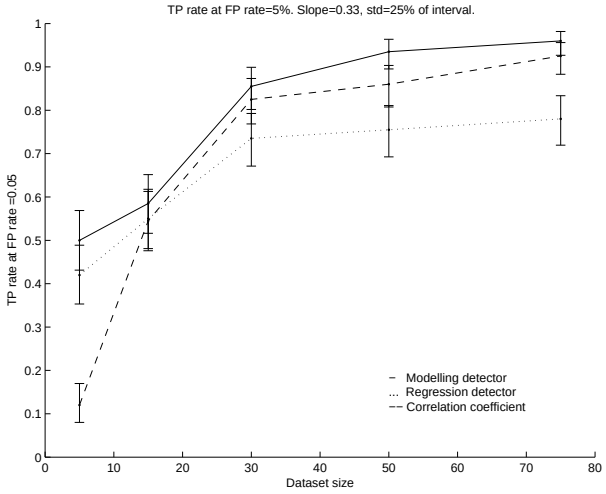
Table 2: TP rates at FP rate = 0.10. Bold values in columns 2 and 5 mean that the detector of that column was better than the correlation coefficient. Bold values in column 8 mean that the correlation coefficient performed better then the modelling detector, whereas italic values mean that the regression detector was defeated by the correlation coefficient. Bold italics values represent cases where the correlation coefficient performed better than both of our detectors.



(a) Smallest standard deviation.

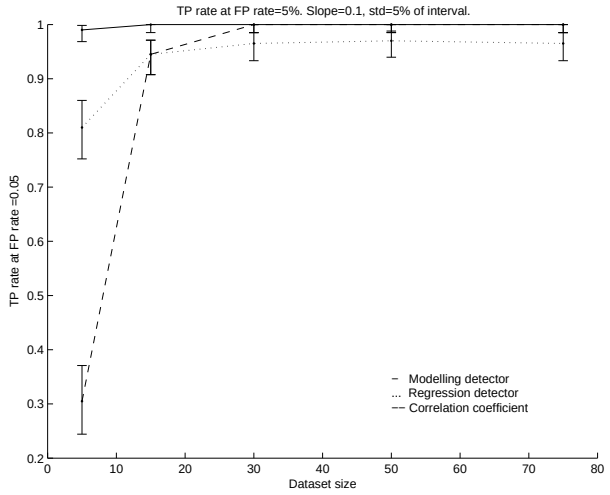


(b) Medium standard deviation.

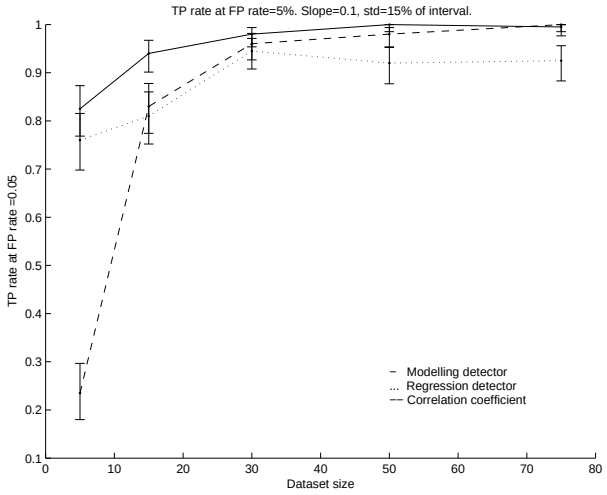


(c) Largest standard deviation.

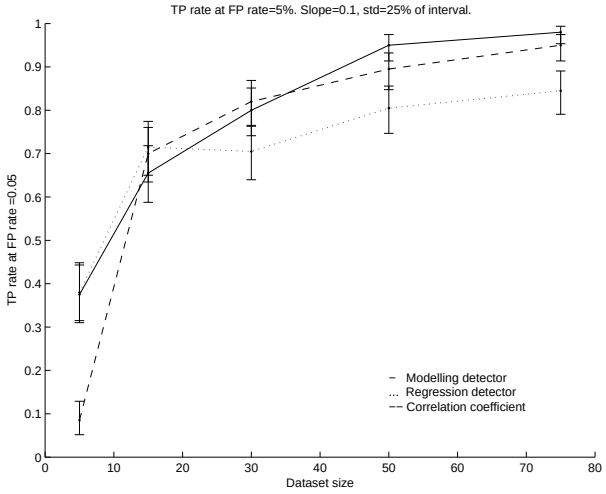
Figure 5: TP rates at FP rate 5% for the three detectors with 95% highest posterior density boundaries bars. The sigmoidal datasets were produced with the **steepest slope**.



(a) Smallest standard deviation.

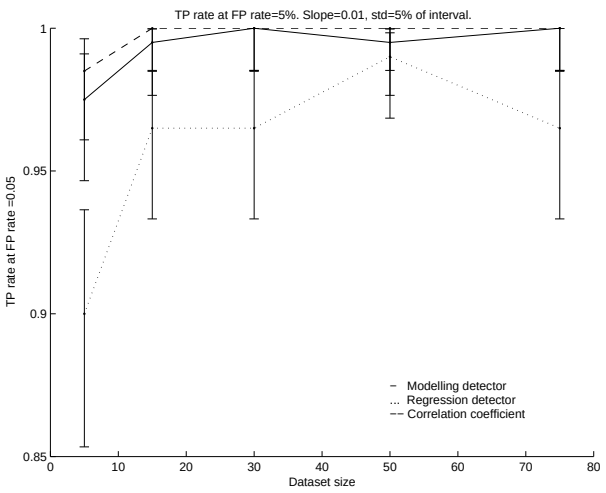


(b) Medium standard deviation.

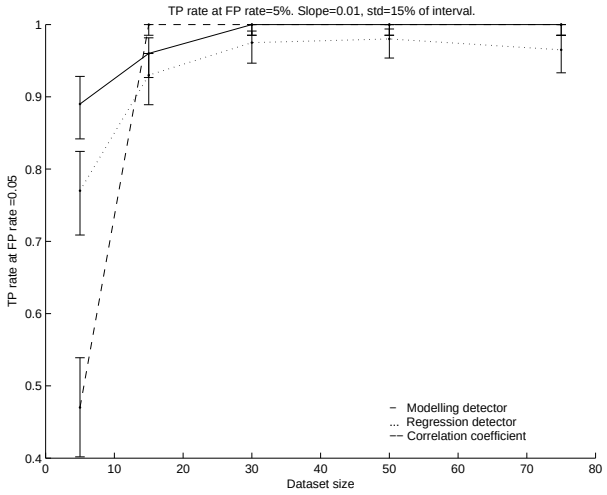


(c) Largest standard deviation.

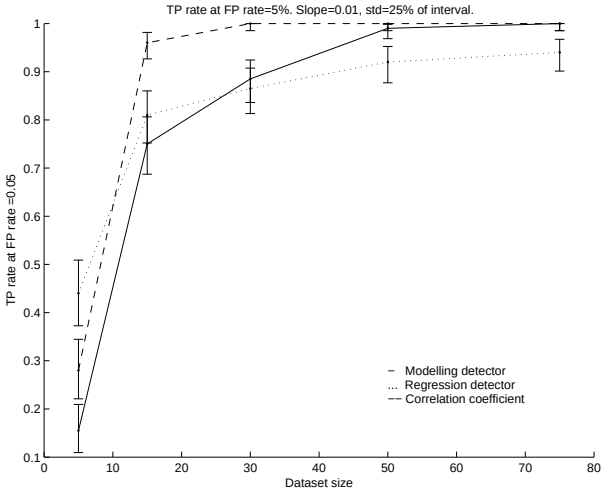
Figure 6: TP rates at FP rate 5% for the three detectors with 95% highest posterior density boundaries bars. The sigmoidal datasets were produced with the **medium slope**.



(a) Smallest standard deviation.



(b) Medium standard deviation.



(c) Largest standard deviation.

Figure 7: TP rates at FP rate 5% for the three detectors with 95% highest posterior density boundaries bars. The sigmoidal datasets were produced with the **flattest slope**.

4 Discussion

4.1 The results

We expected that the correlation coefficient would be outperformed on small datasets and steep slopes, which is evident from the data presented in this thesis. For small datasets and steep slope the modelling detector is substantially better than the correlation coefficient detector, shown here for standard deviations of the noise from 5% to 25% of the range of observed values. For certain circumstances in this work the modelling detector showed a TP rate four times that of the correlation coefficient detector. On larger dataset sizes these detectors performed, with a few exceptions, equally well.

Our other detector, the regression detector, only performed better than the correlation coefficient on a very small dataset size (five points), and was in fact equal to or worse than the correlation coefficient at larger dataset sizes. The results for the larger datasets were somewhat puzzling and we have no explanation for this.

The results clearly show that although the modelling detector is computationally more demanding, it is worth while. Especially on small datasets, less than 20-30 datapoints, where preferably most of the observations have the lowest or highest possible SI values, it outperforms the correlation coefficient detector. Datasets under 20 observations are in no way unreal, sometimes this few cell lines or patients is all that is available. If on the other hand most of the observations have non-extreme SI values, then the standard deviation of the noise affects both detectors to a great extent, and the choice of detector is difficult to make from the results presented in this work.

4.2 Assumptions and approximations

It should be mentioned that the BIC approximation is less accurate for small datasets. This inaccuracy will result in our detectors being randomly better and worse. In this work, however, 400 datasets have been used at a time, which probably evens out the miscalculations and hence our results are reliable.

Both in our detectors and in our simulation of data we have assumed a constant standard deviation of the noise in both SI and gene expression. However, the noise usually varies with the signal, in that a stronger signal gives a larger noise. It is in fact the relative standard deviation that is constant. As the range of SI values is small, this effect can probably be neglected in that case. For gene expression data from cDNA microarray gene expression analysis there is at least one transformation available to stabilize the variance across the full range of expression [10].

In section 2.2 we make the assumption that $p(x)$ is constant on $[-\infty \quad \infty]$. In the simulation of data, we only produce x values in the interval [0 1000], which is not symmetric around 0. This is somewhat inconsistent. Had we produced x values in the range [-500 500] and the code had been adjusted, the same results had probably been reached. Relative gene expressions can indeed be negative.

4.3 Future work

To not be obliged to use transformed microarray data one could develop a detector that model constant relative standard deviation of the noise in gene expression.

The prior of x, i. e. $p(x)$, could be changed into a more realistic one (infinite gene expression is not realistic). This would imply mathematical development of the detectors (see section 2.2).

The simulation of the datasets without correlation between gene expression and drug sensitivity could also be made more realistic. In this work, these datasets were produced according to equation 1. A small standard deviation could therefore result in a dataset with no instances of SI 0.8 or 0. Instead, one could for each sigmoidal dataset produced produce a dataset without correlation between gene expression and drug sensitivity by keeping the SI values from the sigmoid dataset but pick new gene expression values at random. This would be more realistic since each cell line or patient has the same SI no matter which gene's expression one is studying.

5 Acknowledgements

I thank Mats Gustafsson, Anders Isaksson, Claes Andersson and Mikael Wallman, all without whom this work had never been started, proceeded or finished. I also thank Professor Jan Komorowski, head of the Linnaeus Centre for Bioinformatics, for kindly providing a workspace and computer equipment.

References

[1] Holleman A, Cheok M H, den Boer M L, Yang W, Veerman A J P, Kazemier K M, Pie D, Cheng C, Pui C-H, Relling M V, Janka-Schaub G E, Pieters R, Evans W E. *Gene-expression patterns in drug-resistant acute lymphoblastic leukemia cells and response to treatment.* N Engl J Med 2004; 351(6): 533-542.

[2] Ochi K, Daigo Y, Katagiri T, Nagayama S, Tsunoda T, Myoui A, Naka N, Araki N, Kudawara I, Ieguchi M, Toyama Y, Toguchida J, Yoshikawa H, Nakamura Y. *Prediction of response to neoadjuvant chemotherapy for osteosarcoma by gene-expression profiles.* Int J Oncol 2004; 24: 647-655.

[3] Kihara C, Tsunoda T, Tanaka T, Yamana H, Furukawa Y, Ono K, Kitahara O, Zembutsu H, Yanagawa R, Hirata K, Takagi T, Nakamura Y.. *Prediction of sensitivity of esophageal tumors to adjuvant chemotherapy by cDNA microarray analysis of gene-expression profiles.* Cancer Res 2001; 61: 6474-6479.

[4] Zembutsu H, Ohnishi Y, Tsunoda T, Furukawa Y, Katagiri T, Ueyama Y, Tamaoki N, Nomura T, Kitahara O, Yanagawa R, Hirata K, Nakamura Y. *Genome-wide cDNA microarray screening to correlate gene expression profiles with sensitivity of 85 human cancer xenografts to anticancer drugs.* Cancer Res 2002; 62: 518-527.

[5] Szakács G, Annereau J P, Lababidi S, Shankavaram U, Arciello A, Bussey K J, Reinhold W, Guo Y, Kruh G D, Reimers M, Weinstein J N, Gottesman M M. *Predicting drug sensitivity and resistance: Profiling ABC reporter genes in cancer cells.* Cancer cell 2004; 6: 129-137.

[6] Scherf U, Ross D T, Waltham M, Smith L H, Lee J K, Tanabe L, Kohn K W, Reinhold W C, Myers T G, Andrews D T, Scudiero D A, Eisen M B, Sausville E A, Pommier Y, Botstein D, Brown P O, Weinstein J N. *A gene expression database for the molecular pharmacology of cancer.* Nature Genet 2000; 24: 236-244.

[7] Chickering, Heckerman. *Efficient approximations for the marginal likelihood of bayesian networks with hidden variables.* Microsoft research advanced technology division 1996: technical report MSR-TR-96-08.

[8] Fawcett. *ROC graphs: Notes and practical considerations for data mining researchers.* Intelligent enterprise technologies laboratory, HP Laboratories Palo Alto 2003: HPL-2003-4.

[9] Webb. *Statistical pattern recognition (2nd ed).* John Wiley & Sons Ltd 2002: 253-254.

[10] Durbin B P, Hardin J S, Hawkins D M, Rocke D M. *A variance-stabilizing transformation for gene-expression microarray data.* Bioinformatics 2002; 18 (Supplement 1): 105-110.

Read Litterature

- Rang H P, Dale M M, Ritter J M. *Pharmacology (4th ed).* Harcourt Brace and Co Ltd 1999: Chapters 1, 42.
- Jaynes E T. *Probability theory.* Cambridge University Press 2003: Chapters 1, 4, 5, 13, 14, 20.
- MacGregor P F, Squire J A. *Application of Microarrays to the Analysis of Gene Expression in Cancer.* Clin Chem 2002; 48(8): 1170-1177.