DIAGNOSTIC TESTING: MODEL ESTIMATION AND DECISION SUPPORT USING GRAPHICAL MODELS.

Erik Jørgensen, Erik.Jorgensen@agrsci.dk,
Michael Höhle & Søren Højsgaard
Danish Institute of Agricultural Sciences, PO - Box 50 DK-8830 Tjele, Denmark

Many agricultural decision problems can be treated within the framework of diagnostic testing. Test observations are made sequentially in to order to classify a production unit as normal/unnormal, diseased/healthy etc. The central design problem is to find the optimal classification based on the outcome of a test. As in other decision support systems a major aspect is how to specify the relevant parameter values to be used for finding the optimal classification Using pregnancy testing on sows as an illustration, the paper demonstrates how graphical modeling and Influence Diagrams offers a coherent framework for solving this specification problem. The same structured stochastic model is used for estimation of model parameters and for optimisation of the decisions concerning the diagnostic testing. The uncertainty in the knowledge of the model parameters is readily incorporated.

1. Introduction

Diagnostic testing is often carried out without adequate consideration concerning the cost involved and the potential benefit. Such tests are usually performed in order to make some treatment decision, where the outcome of each test serves as decision criteria. The test indicates the most likely state of the animal and thus the optimal treatment of the animal. However, the tests are seldom perfect indicators of the underlying state, and precision and test setup regarding calibration of device, dependency between repeated testing etc. are important, when optimal decisions are considered.

Usually, the test is based on a measurement of some characteristic on a continuous scale, and the result is transformed to a binary test outcome based on some threshold or cut-off value. Ideally, the threshold value is chosen such that the resulting test outcome obtains the optimal combination of false negatives and true positive. Note, that visual assessment of traits in categorical classes, e.g., behaviour category or disease severity, may be considered within the same framework. Typically, the final diagnostic decision will be based on a sequence of such tests, e.g., a visual inspection by the farmer, a more precise clinical evaluation by a veterinarian, and in some cases a precise confirmatory laboratory test. Thus, we need to specify a model containing all these observations, as well as their dependency to the underlying state, and the dependency between the different observations. Parameter estimation in models of such structured stochastic systems have recently become much easier, due to methodological developments in graphical modelling and Markov Chain Monte Carlo estimation (Gilks et al., 1996)

The test sequence as well as the choice of threshold influence the risk of selecting the wrong treatment. Depending on the cost of this, the expected income of the decisions will vary. The optimal threshold value is the value with maximal expected income. Thus, to optimise the decision we need to take into account both the uncertainty related to the outcome of the decision and the uncertainty related to our lack of knowledge of the test in question. The optimal characteristics of each of these tests can not be considered individually, but every test need to be considered simultaneously. Smith & Slenning (2000) presents a general introduction to decision analysis within veterinary diagnostic testing but using decision trees as the framework to describe the problem. We will pursue a more recent representation of such decision problems, namely Influence Diagrams, see e.g., Jensen (2001). Apart from being a much more compact representation of the decision problem, the advantage is that it is closely related to graphical modelling.

The focus of the present paper is to illustrate how the use of Influence Diagrams and related graphical models establishes a coherent link between data analysis and subsequent decision analysis. We have chosen to illustrate the methodology using the pregnancy testing problem, because it contains several of the important elements that needs to be considered.

2. Pregnancy testing

Successful matings that results in pregnancy is a key figure for successful pig production, and the pregnancy state of the animal is an important parameter in production management. We will not discuss the problem in detail but will refer to Meredith (1989), who presents further aspects related to the pregnancy testing.

The use of pregnancy testing in modern pig production may be summarised as follows. Three weeks after mating sows are inspected with respect to recurrence of oestrus. If they show oestrus they will be mated, and if they do not show oestrus, they will be inspected using a pregnancy test device, approximately a week later. Depending on the outcome of this test, the sow is either culled, kept or selected for later re-testing with the test device. This repeated test (e.g. a week later) leads to decisions concerning keeping or culling.

The test devices measure certain physical characteristics of the sow. These characteristics are also influenced by other factors than pregnancy state, introducing dependencies between repeated tests. In addition, it is well-known that the test outcome depends on the skill of the operator.

3. The estimation problem

From the description above it is obvious that the precision of the pregnancy test devices is an important parameter when selecting decision strategies. Fortunately, this precision may be studied experimentally.

3.1. Data material

Data used in this study originates from a Danish investigation by the National Pig Committee performed in order to compare different devices for pregnancy testing of sows. The experimental design allowed for estimation of the precision of different test devices as well as the repeatability of the test with respect to sow and operator. A total of 44 sows were tested, by 6 persons (operators) each using 6 different test devices, i.e., a total of 1584 tests. Due to a coding error the outcome is missing for one sow for device 2 used by operator 2. Thus, the data set consists of 1583 observations with sow number $\{1, \cdots, 44\}$, operator id $\{1, \cdots, 6\}$, device id $\{1, \cdots, 6\}$, true pregnancy state {pregnant, not pregnant}, test outcome {negative, positive}, and days from mating to testing $[30, 38]$. A total of 28 sows were pregnant. The sows were selected for the study according to pregnancy state, that is the pregnancy rate is not representative.

3.2. Model

We will assume that the testing device measures a latent continuous response, e.g. hormonal level, or ultrasonic reading. The model of the latent continuous response is:

$$V_{ij} = \mu_D D_i + S_i + O_j$$

where μ_D is the effect of pregnancy state i, $D_i \in \{0, 1\}$ indicates the pregnancy state of sow i, S_i is effect of the ith sow, and O_j is the effect of the jth operator. The effect of sow and operator were treated as random, $S_i \sim \mathcal{N}(0, \sigma_S^2)$ and $O_j \sim \mathcal{N}(0, \sigma_O^2)$. The binary outcome of the test T_{ijk} is 1, if $V_{ij} + \nu_{ijk} > \lambda_k$ and 0 otherwise, where ν_{ijk} follows the logistic distribution with mean 0 and device dependent variance σ_{Mk}^2, and λ_k is the device dependent threshold value.

$$\Pr(V_{ij} + \nu_{ijk} > \lambda_k) = \Pr(\nu_{ijk} > \lambda_k - V_{ij}) = \frac{1}{1 + \exp(y_{ijk})},$$

where $y_{ijk} = \frac{(V_{ij} - \lambda_k)\pi}{\sigma_{Mk}\sqrt{3}}$ follows the standardised logistic distribution.[1]

$\mu_D = 5$ was assumed and non-informative prior distributions were selected for the parameters. The prior distribution of the variance parameters were specified using the distribution of the corresponding precision parameters ($\tau_x = 1/\sigma_x^2$). These parameters $\{\tau_S, \tau_O, \tau_{M1}, \ldots, \tau_{M6}\}$ were assumed gamma distributed with parameters $(0.001, 0.001)$. The prior distribution of λ_k where assumed uniform in the interval from -2 to 10, i.e., in the relevant interval. The estimation of the posterior distribution of model parameters were performed using the Markov Chain

[1] Alternatively the normal distribution might have been chosen for ν_{ijk}. However, this causes numeric instability in the estimation procedure

Monte Carlo (MCMC) approach via the program WinBUGS (Gilks et al., 1993). The estimations were based on 10000 iterations of the MCMC chain, following a burn-in period of 5000 iterations. In Fig. 1 the model is shown as a graphical model, a so called doodle. The doodle corresponds to a WinBUGS program.

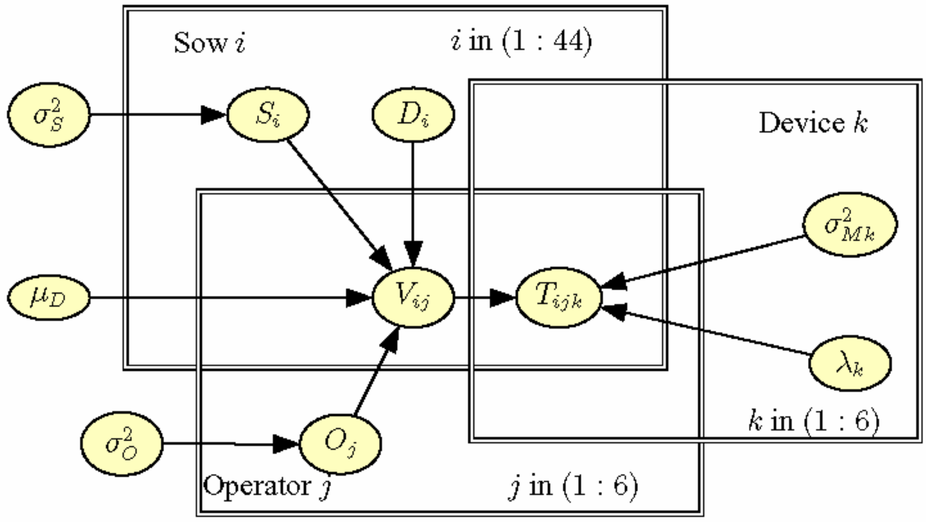

Figure 1: BUGS doodle representing the threshold model

3.3. Results

The results showed marked difference between sows and test devices and only a slight variation in the effect of the operator of the device.

The estimated value of the standard deviation between sows $\sigma_S = 0.85$ is more than three times as large as the standard deviation between operators, $\sigma_O = 0.26$. However, the credibility intervals overlaps so σ_S is not significantly larger than σ_O. The precision of the different devices varies markedly. Expressed as standard deviations, σ_{Mi}, varies from 1.2 (device 4) to 2.9 (device 3). Similarly, the threshold values of the 6 devices differed. Kernel density estimates of the distribution is shown in the left part of Fig. 2 in left part of the figure. Especially, device 6 differed from the rest with a markedly higher threshold value. This resulted in very different values for the sensitivity (Se– conditional probability of positive test results given pregnancy) and specificity (Sp– conditional probability of negative test results given no pregnancy) of the test device. While the precision or the device specific standard deviation σ^2_{Mk} is usually specific to the device components and physical characteristic used in the measurement, the threshold value is a simple matter of adjusting the alarm threshold to a desired value. The estimated Se and Sp values can be used to calculate so-called ROC curves (Hanley, 1998) that shows the dependency between Se and Sp when the threshold value varies (Fig. 2 right).

Even though device 1 and 6 differ markedly in Se and Sp, ((0.96,0.68) vs. (0.87,0.86)) the difference may in fact be due to different adjustment of the threshold for alarm, as the standard

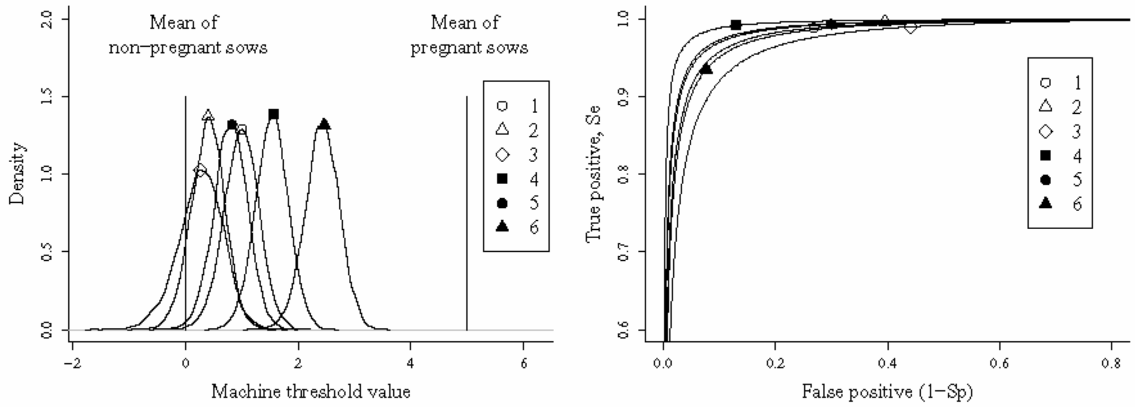

Figure 2: Posterior distribution the threshold parameters τ_i of the 6 devices(left), and ROC- curves based on model estimates (right). The markers indicate estimated Se and Sp

deviation were comparable ($\sigma_{M1} = 2.1$ and $\sigma_{M6} = 2.2$). Device 4 is obviously the best and device 3 the poorest of the 6 devices tested. Note that the Se of machine 3 (0.94) is actually higher than the Se for device 6 (0.87). Another observation is that the producers of the test devices (except device no. 6) put a higher priority on Se rather than Sp.

This naturally leads to the questions such as: is this adjustment of the test devices sensible, how much more can I pay for the different types of test devices?

These questions are discussed in the next section.

4. The decision problem

In this section, the evaluation of different strategies for use of pregnancy testing devices will be discussed from decision theoretic point of view.

The decision framework is the usual one in modern pig production as described in section 2. The decision problem can be modeled using a so-called Influence Diagram, as illustrated in Fig. 3.

The central part of the problem replicates the statistical model in Fig. 1. The D node is the (unobservable) pregnancy state of the sow, V is the latent pregnancy signs of the sow (and the person/operator performing the test) and T_1 and T_2 are test outcomes (oval nodes indicates stochastic variables). In addition, the outcome of the test depends on the device specific standard deviation σ_M. The figure represents the use of the same device for the two tests.

In the statistical model the threshold values λ were treated as a stochastic variable to be estimated. In the decision context, the λ nodes represent the decision concerning the threshold value to use in each of the two tests. (Decisions are indicated using rectangular node). The arrow between H and λ_1 indicates that the outcome of the heat detection is known when the decision is made.

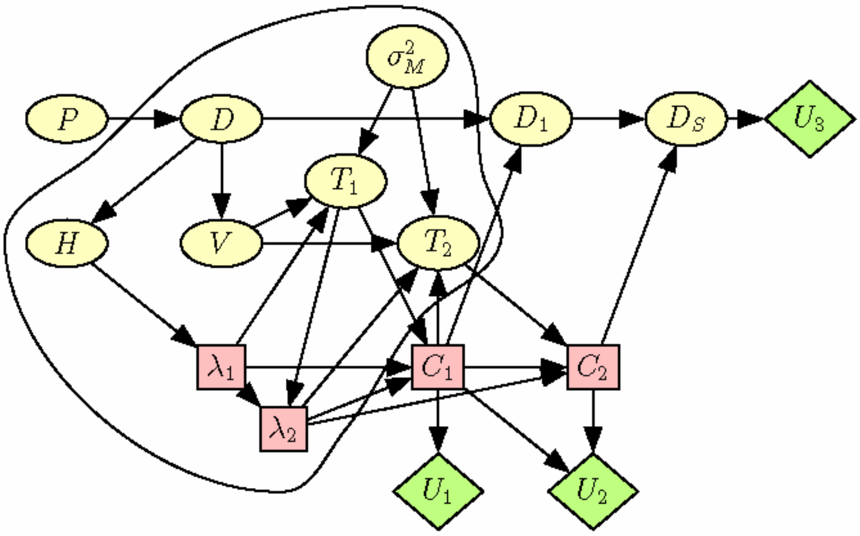

Figure 3: Influence Diagram representing the decisions related to pregnancy testing. The enclosed area corresponds to the statistical model

The remaining nodes are additions compared to the statistical model. The P node indicates the herd pregnancy rate, and the H node the heat observation. C_1 and C_2 are the decisions related to the outcome of the pregnancy tests. The alternative decisions in C_1 is to keep the sow, to cull the sow or to perform the second pregnancy test. C_2 is based on the second test and the alternatives are keep or cull. For each test a cost is implied (U_1 and U_2) (Rhombic nodes indicates cost or utility). The nodes D_1 and D_s models the pregnancy state of the sow after it is affected by the decisions C_1 and C_2, i.e., the decision cull results in the nonpregnant state. Finally, a reward U_3 is obtained depending on the state of D_s. If the animal is pregnant the reward corresponds the marginal return of a litter of pigs, if not, the reward is zero. We will restrict the analysis to the case of a negative outcome of heat detection.

The quantification of the network is relatively standard based on the parameter estimates from the statistical analysis. In addition, the utility nodes needs to be specified. In the following calculations reasonable costs in Danish Kroner (DKK) have been chosen, of course these cost should be adapted to the conditions in each herd.

The program Hugin (http://www.hugin.com) was used for specification of the Influence Diagram (ActiveX library ver. 5.3). A user interface programmed in Borland Delphi 5.0 was used for calculation of optimal strategies for different scenarios.

4.1. Results of decision analysis

The maximum yield that may be obtained using the testing devices can be calculated rather easily. The perfect information in this case is certain knowledge of the animal's pregnancy state. If we obtain perfect information, only the pregnant animals need to bed fed until farrowing, if we obtain no information we have to feed all the sows, which has not shown oestrus. Of course this figure depends on the pregnancy rate as well as the precision of the heat detection.

The higher the pregnancy rate, the lower the potential value of the test device. (Note, that replacement costs is not included in the calculations, so it is a rather optimistic figure). For example, a pregnancy rate of 0.90 results in a maximum value of 13 DKK, whereas a pregnancy rate of 0.80 results in approx. 30 DKK. Similarly, with an increase to 0.80 of heat detection rate 17.1 DKK, and a decrease to 0.70, 26 DKK. The average value is 21.5 DKK. Thus, the value of the pregnancy testing is in that interval.

In Fig. 4 the expected marginal return from using the 6 different devices is shown with a single testing. Furthermore the optimal Se and Sp settings. Note that the values are based on a free test, i.e., no labour, machine cost etc. is included. In the right part of the figure, similar values are shown in relation to the precision of each device. In addition, the figure show the value using two repeated tests with each device.

Device	Marginal return, DKK	Se	Sp
No Test	1405.5		
1	1411.6	0.97	0.65
2	1415.0	0.97	0.75
3	1406.9	0.97	0.38
4	1420.0	0.97	0.88
5	1414.4	0.97	0.74
6	1410.5	0.97	0.60
Perfect	1427.0	1.00	1.00

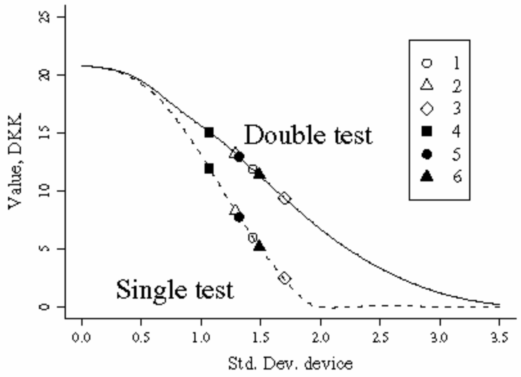

Figure 4: Expected return from optimal strategy, free test, machine precision known

As shown, the optimal value for Se is identical (97 %) for every device, and is a little higher but rather closely to the observed values shown in Fig. 2, except for device 6. The optimal values for Sp are surprisingly low in some cases. These values refer to single testing for the average heat detection rate and pregnancy rate. Using the Influence Diagram it is easy to calculate the optimal use of the pregnancy test, under different conditions. Furthermore, the optimal combination of two test devices may be considered, e.g., a cheap first test, followed by a more elaborate and expensive test. From an economic point of view the main problem is that very few sows that has not already shown heat are non-pregnant.

Of course, it is of interest to explore several other aspects, but hopefully, the few examples mentioned illustrates the potential of the approach.

5. Discussion

As indicated in the previous pages, graphical modeling offers a comprehensive framework for both data and decision analysis. Because the stochastic nature of the problem is inherent in the modeling, the inclusion of parameters and background knowledge with different precision introduces no further complications. The graphical structure gives a natural representation of issues such as repeated measurements and other kinds of dependent measurements, as illustrated with the latent V variable which models the underlying physical characteristic measured by the test devices (Fig. 3). Even though the decision tree approach are still suggested as the method for solving these problems e.g., Smith & Slenning (2000), it should be realised that Influence Diagrams is a far better choice because of the more compact representation. Actually, the original motivation for the Influence Diagram was to make a more compact representation of decision trees (Howard & Matheson, 1984). The differences in the algorithms for finding optimal solutions are a later development.

We have used the pregnancy test problem to illustrate the method. However, a very wide range of problems within agricultural production may be handled within the same framework. Problems like estimation of disease prevalence, early warning using automatic registration equipment, BSE-testing schemes etc.

Thus, the necessary framework exists for handling these common decision problems, both concerning quantification and optimisation. However, surprisingly few applications are available. The reason may be that sequential decision problems is inherently complex to solve, if they become much larger than the example used. But because of current research within in example artificial intelligence, it is reasonable to expect that at least finding approximate solutions to far more complicated problems will be feasible in the future, see e.g., Lauritzen & Nilsson (2001).

LITTERATURE

Gilks, W., A. Thomas, & D.J. Spiegelhalter (1993). A language and program for complex Bayesian modelling. *The Statistician,* 43 pp. 169–178.

Gilks, W.R., S. Richardson, & D.J. Spiegelhalter (1996). *Markov Chain Monte Carlo in Practice.* Chapman & Hall, London.

Hanley, J.A. (1998). Receiver Operating Characteristics ROC Curves. In P. Armitage & T. Colton, editors, *Encyclopedia of Biostatistics,* pp. 1134–1139. J. Wiley & Sons, Chichester.

Howard, R.A. & J.E. Matheson (1984). Influence Diagrams. In R.A. Howard & J.E. Matheson, editors, *Readings on the Principles and Applications of Decision Analysis,* pp. 719–762. Strategic Decisions Group, Menlo Park, California.

Jensen, F.V. (2001). *Bayesian Networks and Decision Graphs.* Springer-Verlag, New York.

Lauritzen, S.L. & D. Nilsson (2001). Representing and Solving Decision Problems with Limited Information. *Management Science,* 47 pp. 1235–1251.

Meredith, M.J. (1989). Pregnancy diagnosis in higher performance pig herds – is it a waste of resources? *Pig News and Information,* 10 pp. 477–480.

Smith, R.D. & B.D. Slenning (2000). Decision analysis: dealing with uncertainty in diagnostic testing. *Preventive Veterinary Medicine,* 45 pp. 139–162.