Bayesian Networks and the Problem of Unreliable Instruments

Luc Bovens, University of Colorado at Boulder
e-mail: bovens@spot.colorado.edu
Stephan Hartmann, University of Konstanz
e-mail: stephan.hartmann@uni-konstanz.de

Abstract. We appeal to the theory of Bayesian Networks to model different strategies for obtaining confirmation for a hypothesis from experimental test results provided by less than fully reliable instruments. In particular, we consider (i) repeated measurements of a single testable consequence of the hypothesis, (ii) measurements of multiple testable consequences of the hypothesis, (iii) theoretical support for the reliability of the instrument, and (iv) calibration procedures. We evaluate these strategies on their relative merits under idealized conditions and show some surprising repercussions on the variety-of-evidence thesis and the Duhem-Quine thesis.

How can experimental test results from less than fully reliable instruments (LTFR instruments) provide confirmation for a scientific hypothesis? A range of strategies has been discussed in the literature (e.g. in Franklin 1986, 165-191), but no attempt has been made to give a Bayesian analysis of these strategies. This is unfortunate, since such an analysis proves to be rewarding in many respects. First, it enables us to construct a taxonomy of strategies. In scientific practice, these strategies occur in mixed forms. The models permit us to isolate certain general strategies and to draw some perspicuous analytical distinctions within each *genus*. Second, it shows that under certain constraints these strategies are indeed legitimate strategies: it is possible for a hypothesis to receive strong confirmation, even when scientific instruments to test them are less than fully reliable. Third, it yields rather surprising claims about the conditions under which specific strategies for dealing with LTFR instruments are more and less successful.

Why has there been so little interest in Bayesian circles in the status of experimental reports from LTFR instruments? The task of modeling even the simplest strategies is daunting. We need more powerful tools to do the job: here is where Bayesian Networks come in handy. Over the last two decades, the theory of Bayesian Networks has been developed in artificial intelligence on the dual pillars of graph theory and the theory of conditional independence structures. Although the theory certainly has some philosophical roots, philosophers of science have done little to harvest its fruits. This is what we intend to do in addressing the question at hand.

The literature contains examples galore of how test results from LTFR instruments can yield a respectable degree of confirmation. We will investigate the following types of strategies by modeling them under some plausible idealizations:

Strategy 1. Repeated measurements with a single LTFR instrument or single measurements with multiple independent LTFR instruments of a single testable consequence of a hypothesis yield *the same* test results.

Strategy 2. Repeated measurements with a single instrument or single measurements with multiple independent LTFR instruments of multiple testable consequences of a hypothesis yield *coherent* test results.

Strategy 3. We find support for the LTFR instrument in an *auxiliary theory* which may or may not be dependent on the hypothesis under investigation.

Strategy 4. The LTFR instrument is *calibrated* against the test results of a single or of multiple independent instruments that are more reliable than the LTFR instrument.

1. Modeling Confirmation with a LTFR Instrument

Consider a very simple scenario. Let there be a hypothesis, a testable consequence of the hypothesis, a LTFR instrument and a report from the LTFR instrument to the effect that the testable consequence holds or not. To model this scenario, we need four propositional variables (written in italic script) and their values (written in roman script):

(1) *HYP* can take on two values: HYP, i.e. the hypothesis is true and $\overline{\text{HYP}}$, i.e. the hypothesis is false;

(2) *CON* can take on two values: CON, i.e. the testable consequence holds and $\overline{\text{CON}}$, i.e. the testable consequence does not hold;

(3) *REL* can take on two values: REL, i.e. the instrument is reliable and $\overline{\text{REL}}$, i.e. the instrument is not reliable;

(4) *REP* can take on two values: REP, i.e. there is a positive report, or, in other words, a report to the effect that the testable consequence holds and $\overline{\text{REP}}$, i.e. there is a negative report, or, in other words, a report to the effect that the testable consequence does not hold.

A probability distribution over these variables contains 2^4 entries. The number of entries will grow exponentially with the number of propositional variables. To represent the information in a more parsimonious format, we construct a Bayesian Network.

A Bayesian Network organizes the variables into a *Directed Acyclical Graph* (DAG), which encodes a range of (conditional) independences. A DAG is a set of *nodes* and a set of *arrows* between the nodes under the constraint that one does not run into a cycle by following the direction of the arrows. Each node represents a propositional variable. Consider a node at the tail of an arrow and a node at the head of an arrow. We say that the node at the tail is the *parent node* of the node at the head and that the node at the head is the *child node* of the node at the tail. There is a certain heuristic that governs the construction of the graph: there is an arrow between two nodes iff the variable in the parent node has *a direct influence* on the variable in the child node.

In the case at hand, whether the testable consequence holds is directly influenced by and only by whether the hypothesis is true or not; whether there is a report to the effect that the testable consequence holds is directly influenced by and only by whether the testable consequence holds or not and by whether the instrument is reliable or not. Hence, we construct the basic graph in figure 1.1 in which the node with the variable *HYP* is a *parent node* to the node with the variable *CON* and the nodes with the variables *CON* and *REL* are in turn parent nodes to the node with the variable *REP*.

Furthermore, *root nodes* are unparented nodes and *descendant nodes* are child nodes, or child nodes of child nodes etc. E.g., *HYP* and *REL* are root nodes and *CON* and *REP* are descendant nodes of *HYP* in our graph.

From DAG to Bayesian Network, one more step is required. We need to stipulate a probability distribution for the variables in the root nodes of the graph and a conditional probability distribution for the variables in the other nodes given any combination of values of the variables in their respective parent nodes.

Let us turn to our example. First, we take care of the root nodes, i.e. we assign a prior probability to the hypothesis and to the reliability of the instrument:

(1.1) $P(HYP) = h$ with $0 < h < 1$

(1.2) $P(REL) = r$ with $0 < r < 1$.

Second, consider the node with the variable *CON* which is a child node to the node with the variable *HYP*. We take a broad view of what constitutes a testable consequence, that is, we do not require that the truth of the hypothesis is either a necessary or a sufficient condition for the truth of the testable consequence. Rather, a testable consequence is to be understood as follows: the probability of the consequence given that the hypothesis is true is greater than the probability of the consequence given that the hypothesis is false:

(1.3) $P(CON \mid HYP) = p > q = P(CON \mid \overline{HYP})$.

Third, consider the node with the variable *REP*, which is a child node to the nodes with the variables *CON* and *REL*. How can we model the workings of an unreliable instrument? Let us make an idealization: We suppose that we do not know whether the instrument is reliable or not, but if it is reliable, then it is fully reliable and if it is not reliable, then it is fully unreliable. Let a fully reliable instrument be an instrument that provides maximal information: it is an instrument that says of what is that it is, and of what is not that it is not:

(1.4) $P(REP \mid REL, CON) = 1$

(1.5) $P(REP \mid REL, \overline{CON}) = 0.$

Let a fully unreliable instrument be an instrument that provides minimal information: it is an instrument that is no better than a randomizer:

(1.6) $P(REP \mid \overline{REL}, CON) = P(REP \mid \overline{REL}, \overline{CON}) = a$ with $0 < a < 1.$

Let us call a the randomization parameter. We can now construct the Bayesian Network by adding the probability values to the graph in figure 1.1.

Although the direct-influence story behind the construction of a Bayesian Network is only of heuristic value, the arrows have a precise probabilistic meaning: they carry information about the independence relations between the variables in the Bayesian Network. The following is the fundamental Graph Construction Rule:

(*GCR*) A variable represented by a node in the Bayesian Network is independent of all variables represented by its non-descendant nodes in the Bayesian Network, conditional on all variables represented by its parent nodes.

Hence, our Bayesian Network is constructed on grounds of the following (conditional) independences:

(1.7) $HYP \perp REL$

(1.8) $CON \perp REL \mid HYP$

(1.9) $REP \perp HYP \mid REL, CON$.

(1.7) says that if one does not know any values of the variables, then coming to learn that the instrument is reliable or that the instrument is unreliable does not alter the prior probability that the hypothesis is true. This is a plausible assumption as long as one's reasons for believing that the instrument is reliable are independent of the truth of the hypothesis. In section 4, we will investigate what happens when this assumption is violated. (1.8) says that if one knows no more than that the hypothesis is true or that the hypothesis is false, then coming to learn in addition that the instrument is reliable or that it is unreliable does not alter the probability that the testable consequence holds: as long as one does not know what report the instrument provides, coming to learn about its reliability teaches nothing about the testable consequence. (1.9) says that if one knows no more than that some definite values of *REL* and *CON* are instantiated, then coming to learn in addition that some definite value of *HYP* is instantiated does not alter the probability of REP: the chance that the instrument will yield a positive or a negative report is fully determined by whether the instrument is reliable and whether the testable consequence holds or not; once this information is known, the truth or falsity of the hypothesis itself becomes irrelevant. The latter two assumptions seem beyond reproach.

The Bayesian Network also represents a series of other conditional independences, e.g. $REP \perp HYP|CON$. The theory of Bayesian Networks presents an axiomatic structure, viz. the semi-graphoid structure, which permits us to derive these independences from the conditional independences that can be read off by (*GCR*). It also contains a convenient criterion, viz. the d-separation criterion, which permits us to read these same conditional independences directly off of the graph. For the details, we refer to the relevant literature.[1]

What's so great about Bayesian Networks? A Bayesian Network contains information about the independence relations between the variables, probability assignments for each root node and conditional probability assignments for each child node. A central theorem in the theory of Bayesian Networks states that a joint probability distribution over any combination of values of the variables in the Network is equal to the product of the probabilities and conditional probabilities for these values as expressed in the Network (Neapolitan 1990, 162-164). For example, suppose we are interested in the joint probability of HYP, $\overline{\text{CON}}$, REP and $\overline{\text{REL}}$. We can read the joint probability directly off of figure 1.1:

(1.10) $P(\text{HYP}, \overline{\text{CON}}, \text{REP}, \overline{\text{REL}}) =$

$$P(\text{HYP})P(\overline{\text{REL}})P(\overline{\text{CON}}\,|\,\text{HYP})P(\text{REP}\,|\,\overline{\text{CON}},\overline{\text{REL}}) = h(1-r)(1-p)a.$$

Standard probability calculus teaches us how to construct marginal distributions out of joint distributions and subsequently conditional distributions out of marginal

[1] The axioms for semi-graphoids were developed by Dawid (1979) and Spohn (1980) and are presented in Pearl (1988, 82-90) and Neapolitan (1990, 193-195). For details on the d-separation criterion, see Pearl (1988, 117-118), Neapolitan (1990, 202-207) and Jensen (1996, 12-14).

distributions. When implemented on a computer, Bayesian Networks provide a direct answer to such queries.

We are interested in the probability of the hypothesis given that there is a report from a LTFR instrument that the testable consequence holds. This probability is P*(HYP) = P(HYP | REP) = P(HYP, REP) / P(REP). For ease of representation, we will abbreviate (1-x) as $\overline{x}$.

$$(1.11)\quad P^*(HYP) = \frac{\sum_{CON,\,REL} P(HYP)P(REL)P(CON \mid HYP)P(REP \mid CON, REL)}{\sum_{HYP,\,CON,\,REL} P(HYP)P(REL)P(CON \mid HYP)P(REP \mid CON, REL)}$$

$$= \frac{h(pr + a\overline{r})}{hr(p-q) + qr + a\overline{r}}.$$

We measure the degree of confirmation that the hypothesis receives from a positive report by the difference:

$$(1.12)\quad P^*(HYP) - P(HYP) = \frac{h\overline{h}(p-q)r}{hr(p-q) + qr + a\overline{r}}.$$

Note that P*(HYP) - P(HYP) > 0 iff p>q. To have some numerical data, let h = r = a = 1/2 and let p = 3/4 and q = 1/4. Then P*(HYP) = 5/8 and P*(HYP) - P(HYP) = 1/8.

We know now how to model the degree of confirmation that a hypothesis receives from a single positive report concerning a single testable consequence of the hypothesis by means of a single LTFR instrument. This basic model will be the paradigm to model complex strategies to improve the degree of confirmation that can be obtained from LTFR instruments.

2. Same Test Results

Suppose that we have tested a single testable consequence of the hypothesis by means of a single LTFR instrument. We have received a positive report, but we want to have additional confirmation for our hypothesis. We might want to run more tests of the very same testable consequence. Now there are two possibilities. Either we can take our old LTFR instrument and run the test a couple more times. Or we can choose new and independent LTFR instruments and test the very same testable consequence with these new instruments. First, we will show that both of these substrategies can be successful: If we receive more reports to the effect that the testable consequence holds, either from our old instrument or from new and independent instruments, then the hypothesis does indeed receive additional confirmation. Second, we are curious to know which substrategy is the better strategy assuming that we do indeed receive more reports to the effect that the testable consequence hold. In other words, which substrategy yields a higher degree of confirmation? Is there an univocal answer to this question, or is one substrategy more successful under certain conditions, while the other strategy is more successful under other conditions? To keep things simple, we will present our analysis for *one* additional test report, either from the same or from different LTFR instruments.

Let us first model the degree of confirmation that the hypothesis receives from an additional positive report from the same LTFR instrument. In figure 2.1, we add a node to our basic graph to represent the binary variable *REP2* and substitute *REP1* for *REP*. Just like *REP1*, *REP2* is directly influenced by *REL* and *CON* and so two more arrows are drawn in. We impose a condition of *symmetry* on the probability distribution P for

this graph: also for this second report the instrument is either fully reliable or it is fully unreliable with the same randomization parameter a.

Secondly, we model the degree of confirmation that the hypothesis receives from an additional confirming report from a second independent LTFR instrument. In figure 2.2, we add a node to our basic graph for the variable *REL2* which expresses whether the second instrument is reliable or not and add a node for the variable *REP2* which expresses whether the second instrument provides a report to the effect that the testable consequence holds or not. *REP2* is directly influenced by *REL2* and *CON*: we draw in two more arrows. To keep matters simple, we impose a condition of *symmetry* on the probability distribution P' for this graph: there is an equal chance r that both instruments are reliable and if the instruments are unreliable then they randomize at the same level a. To compare the scenario with one instrument to the scenario with two instruments we need to impose a *ceteris paribus* condition: for this reason we postulate the same values h, p, q, r and a for the probability distributions P and P'.

The instruments are independent of one another. What this means is that

(2.1) $REPi \perp REPj \mid CON \quad \forall i, j = 1,2$ and $i \neq j$.

Suppose that we know that the consequence holds or we know that the consequence does not hold. Then there is a certain chance that we will receive a report to the effect that the consequence holds. Now whether we receive another report to this effect or not, does not affect this chance. An independent instrument may not always provide us with an accurate report, but it is not influenced by what other instruments report. It can be shown by standard techniques in the theory of Bayesian Networks that (2.1) is a conditional independence that can be read off from the graph in figure 2.2.

Are these strategies successful? The strategy of searching out an additional report from the same LTFR instrument about the same testable consequence provides additional confirmation to the hypothesis iff

(2.2) $\Delta P = P(HYP \mid REP1, REP2) - P(HYP \mid REP1) > 0$.

The strategy of searching out an additional report from a different LTFR instrument about the same testable consequence provides additional confirmation to the hypothesis iff

(2.3) $\Delta P = P'(HYP \mid REP1, REP2) - P'(HYP \mid REP1) > 0$.

In the appendix we have spelled out a standard procedure to investigate this question and have shown that the answer is affirmative in both cases. We will contrast this result later with our results for the second strategy.

We turn to the question whether, *ceteris paribus*, the hypothesis receives more confirmation from a second positive report from one and the same LTFR instrument or from independent LTFR instruments. We calculate the following difference:

(2.4) $\Delta P = P'(HYP \mid REP1, REP2) - P(HYP \mid REP1, REP2)$.

Following our standard procedure,

$$(2.5)\ \Delta P = \frac{a^2 h \bar{h} r \bar{r} (1 - 2\bar{a}\bar{r})(p - q)}{(a^2 \bar{r} + hr(p - q) + qr)(a^2 \bar{r}^2 + 2ar\bar{r}(q + h(p - q)) + r^2(\bar{h}q + hp))}.$$

Since $0 < a, h, r < 1$ and $p>q$, $\Delta P > 0$ iff $1 - 2\bar{a}\bar{r} > 0$. The graph in figure 2.3 represents this inequality. For values of (r,a) above the phase curve, $\Delta P > 0$, i.e. it is better to receive reports from two instruments; for values of (r,a) on the phase curve, $\Delta P = 0$, i.e. it does not make any difference whether we receive reports from one or two instruments; for values of (r,a) below the phase curve, $\Delta P < 0$, i.e. it is better to receive reports from one instrument than from two instruments.

Do these results seem plausible at some intuitive level? There are two conflicting intuitions at work here. On the one hand, we are tempted to say that confirming results from two instruments is the better way to go, since independence is a good thing. On the other hand, if we receive consistent positive reports from a single instrument, then we feel more confident that the instrument is not a randomizer and this increase in confidence in the reliability of the instrument benefits the confirmation of the hypothesis. For higher values of r, the former consideration becomes more weighty than the latter: there is not much gain to be made anymore in our confidence in the reliability of the instrument(s) and we might as well enjoy the benefits of independence. For lower values of a, the latter consideration becomes more weighty: if we are working with an instrument which, if unreliable, has a low chance of providing positive reports, then consistent positive reports constitute a substantial gain in our confidence in its reliability, which in turn benefits the confirmation of the hypothesis.

3. Coherent Test Results

The second strategy to raise the degree of confirmation for a hypothesis is to identify a range of testable consequences which all can be assessed by a single or by multiple independent LTFR instruments. Let us draw the graphs for two testable consequences.

Following our heuristic, the hypothesis (*HYP*) directly influences the testable consequences (*CONi* for i=1,2). Figure 3.1 represents the scenario in which there is a single instrument: each testable consequence (*CONi*) conjoint with the reliability of the single instrument (*REL*) directly influences the report about the consequence in question (*REPi*). Figure 3.2 represents the scenario in which there are two independent instruments: each testable consequence (*CONi*) conjoint with the reliability of the instrument that tests this consequence (*RELi*) directly influences the report about the consequence in question (*REPi*). We define a probability distribution P for the DAG in figure 3.1 and a probability distribution P' for the DAG in figure 3.2. We impose the *symmetry* condition within each distribution and the *ceteris paribus* condition between distributions for all the relevant parameters.

We can now check whether our strategies are successful. We show in the appendix that the strategy is always successful with multiple instruments:

$$(3.1)\ \Delta P = P'(HYP \mid REP1, REP2) - P'(HYP \mid REP1) > 0.$$

But with a single instrument, the strategy is not always successful. We show in the appendix that

$$(3.2)\ \Delta P = P(HYP \mid REP1, REP2) - P(HYP \mid REP1) > 0 \text{ iff } a\bar{r}(p + q - a) + pqr > 0.$$

In figure 3.3, we fix a=.5 and construct phase curves for high, medium and low range values of the reliability parameter r. In figure 3.4, we fix r=.5 and construct phase curves for high, medium and low range value of the randomization parameter a. Since we have stipulated that p>q, we are only interested in the areas below the straight line for p=q in both figures.

The areas in these graphs in which $\Delta P<0$ are certainly curious: for certain values of p, q, a and r, we test a first consequence of a hypothesis, receive a positive report and are more confident that the hypothesis is true; then we test a second consequence of the hypothesis with the very same instrument, receive once again a positive report...but this time around our degree of confidence in the hypothesis drops! How can we interpret these results? Notice that the effect is most widespread for (i) lower values of r, (ii) higher values of a and (iii) lower values of p. To get a feeling for the magic of the numbers, let us look at this range of values, where the effect occurs *par excellence*. Hence, let us consider instruments, which are not likely to be reliable, and, if unreliable, have a high chance of providing a positive report, and testable consequences which are unlikely to occur when the hypothesis is true (though of course are still more likely to occur than when the hypothesis is false). Considering (i), we do not have much trust in the instrument to begin with. Now it gives us nothing but positive reports: considering (ii), the instrument is likely to be a randomizer and so we become even more confident that the instrument is unreliable. But should this not be offset by the fact that we receive coherent test results in support of our hypothesis? No, since considering (iii), our tests are rather weak and these coherence effects count for little. Hence, when we get a second positive report, we become very confident that the instrument is unreliable and consequently our confidence in the hypothesis drops.

We turn to the question whether, *ceteris paribus*, the hypothesis receives more confirmation from a second positive report from one and the same LTFR instrument or from independent LTFR instruments. In the appendix, we show that

(3.3) $\Delta P = P'(HYP \mid REP1,REP2) - P(HYP \mid REP1,REP2) > 0$ iff

$$2a^2\bar{r} - 2pqr - a(p+q)(1-2r) > 0.$$

To evaluate this expression, we assume that the tests are reasonably strong by fixing p=.9 and q=.1 and construct a phase curve for values of (a,r) in figure 3.5. If the randomization parameter and the reliability parameter are set low, then one instrument tends to do better than two. Subsequently we assume mid-range values for the randomization and the reliability parameters (a=.5 and r=.5) and construct a phase curve for values of (p,q) in figure 3.6. We are interested in the area below the straight line where p>q. If the q-value is set high, i.e. if the testable consequences occur frequently also when the hypothesis is false, then one instrument tends to do better than two.

In the previous section, we explained why the consideration that our confidence in the reliability of a single instrument is boosted by coherent positive reports outweighs the consideration of the independence of multiple instruments for lower values of a and r. The same explanation can here be repeated. But why is this effect amplified for higher q-values? The higher the q-values, the more likely the testable consequences will hold true and so coherent positive reports will boost our confidence in the reliability of a single instrument even more. Hence higher q-values tend to favor a single over multiple instruments.

It is one of the textbook Bayesian success stories that an account can be provided of why variety of evidence is a good thing: It is shown that the increment of confirmation that the hypothesis receives from confirming test results becomes smaller and smaller as we run the same old test over and again. (E.g. Earman 1992, 77-79 and Howson and Urbach 1993, 119-123.) But what does it mean to run the same old test over and over again? Does it mean that we check the same old testable consequences rather than checking independent testable consequences of the hypothesis? Does it mean that we do our testing with the same old instrument rather than with independent

instruments? Presumably variety of evidence refers to multiple testable consequences as well as to multiple instruments. But notice that our investigation permits us to impose three caveats on the variety-of-evidence thesis. We argued in the last section that,

(i) if we are testing a single consequence, it is sometimes more beneficial to receive positive reports from the same instrument than from new instruments, *ceteris paribus*.

What we have seen in this section is that,

(ii) if we are testing different consequences, it is sometimes more beneficial to receive positive reports from the same instrument than from new instruments, *ceteris paribus*.

And there is still another conclusion to be drawn from our results. We saw in the previous section that it is always a good thing for the confirmation of the hypothesis to receive a second positive report from the same instrument about the same testable consequence. In this section, we saw that our confidence in the hypothesis may decrease as we receive a second positive report from the same instrument about a different testable consequence. Hence, we can add a third caveat:

(iii) If we are testing with a single instrument, it is sometimes more beneficial to receive positive reports about the same consequence rather than about different consequences, *ceteris paribus*.

4. Auxiliary Theories

Let us return to our basic model in section 2. In this model, the variable *REL* is a root node and we have assigned a probability value r which expresses the chance that the instrument is reliable. It is a common theme in contemporary philosophy of science that the workings of the instrument are themselves supported by an auxiliary theory of the instrument. If this is the case, then we should not model *REL* as a root node: whether the instrument is reliable or not is directly influenced by whether the auxiliary theory

holds or not (*AUX*). Just as we assigned a prior probability to the hypothesis, we also assign a prior probability t to the auxiliary theory. To keep matters simple, let us assume in this section that the instrument is reliable just in case the auxiliary theory is correct and that the testable consequence holds just in case the hypothesis is true. Our basic model is then expanded in the Bayesian Network in figure 4.1. In this Bayesian Network, *AUX* and *HYP* are still independent. This may or may not be a realistic assumption. Sometimes the auxiliary theory has no relation whatsoever to the hypothesis under test. But sometimes they are quite closely tied to each other: for instance, they may both be parts of a broader theory. We can model this positive relevance between *AUX* and *HYP* by connecting both variables in the Bayesian Network and by setting $P'(AUX \mid HYP) = t_h > t_{\bar{h}} = P'(AUX \mid \overline{HYP})$ in figure 4.2.

Here are some questions:

(i) *Ceteris paribus*, does the hypothesis receive more or less confirmation if the auxiliary theory that supports the reliability of the instrument is independent rather than positively relevant to the hypothesis under test?

(ii) Suppose that we receive a report from a LTFR instrument which provides confirmation to the hypothesis. We now appeal to an auxiliary theory which provides support for the reliability of the instruments, i.e. by bringing in an auxiliary theory we succeed in raising the reliability parameter r. Our question is the following: Is this, *ceteris paribus*, a successful strategy to increase the degree of confirmation for the hypothesis,

(a) if the auxiliary theory is independent of the hypothesis;

(b) if the auxiliary thesis is positively relevant to the hypothesis?

Let us first take up question (i). To respect the *ceteris paribus* clause we must make sure that the randomization parameter, the reliability parameter and the prior

probability of the hypothesis are fixed across both scenarios. To fix the reliability parameter, we must make sure $t = t_h h + t_{\bar{h}} \bar{h}$, since the instrument is reliable just in case the auxiliary theory is true. In the appendix, we have shown that

(4.1) $\Delta P = P(HYP|REP) - P'(HYP|REP) > 0$ iff $h + \bar{a}(ht_h + \bar{h}t_{\bar{h}} - 1) > 0$.

To evaluate this expression, we construct two graphs: in figure 4.3, we set $t_h = .8$ and $t_{\bar{h}} = .2$ and construct a phase curve for (a,h); in figure 4.4, we set a=1/3 and h=1/3 and construct a phase curve for $(t_h, t_{\bar{h}})$.

What we see in figure 4.3 is that a positively relevant auxiliary theory provides more of a boost to the degree of confirmation that the hypothesis receives from a positive test report than an independent auxiliary theory for lower prior probability values of the hypothesis and for lower values of the randomization parameter. In figure 4.4 we are only interested in the area below the line where $t_h > t_{\bar{h}}$. What we see is that for $t_h < 1/2$, a positively relevant auxiliary theory always provides more of a boost to the degree of confirmation of the hypothesis, while for $t_h > 1/2$, a positively relevant auxiliary theory provides more of a boost for values of $t_{\bar{h}}$ that are sufficiently smaller than t_h, in other words, for a theory that is sufficiently positively relevant to the hypothesis.

Can an intuitive account be given of these results? Let us first try to understand why a positively relevant auxiliary theory is favorable to the degree of confirmation for hypotheses with a lower prior probability. To say that the auxiliary theory and the hypothesis are *maximally* positively relevant is to say that there is maximal overlap between the theory and the hypothesis in the probability space. To say that the auxiliary theory and the hypothesis are *maximally* negatively relevant is to say that

there is minimal overlap between the auxiliary theory and the hypothesis in the probability space. The continuum from maximal positive to maximal negative relevance runs over the independence of the auxiliary theory and the hypothesis.

Let us start with an auxiliary theory and a hypothesis that both have a very low probability, say .1. It is easy to understand that maximal negative relevance would be worse for the confirmation of the hypothesis than maximal positive relevance in this case: if there is maximal negative relevance, then whatever gain is made in our confidence in the hypothesis is offset by the decrease in our confidence in the reliability of the instrument; if there is maximal positive relevance, then whatever gain is made in our confidence in the hypothesis is reinforced by the increase in our confidence in the reliability of the instrument. Assuming that the degree of confirmation for the hypothesis is a monotonically increasing function from maximal negative to maximal positive relevance, we can see that independence would score worse than positive relevance.

Let us now raise the probability of both the auxiliary theory and the hypothesis in unison and as much as possible under the constraint that there can be maximal negative relevance without overlap between the auxiliary theory and the hypothesis: then the maximal probability that we can assign to both the auxiliary theory and the hypothesis is .5. If there is maximal negative relevance, how could a positive report be received? Notice that we have a partition of the probability space: (i) either the hypothesis is true and the auxiliary theory is false (implying that the instrument is unreliable) or (ii) the hypothesis is false and the auxiliary theory is true (implying that the instrument is reliable). But option (ii) is impossible: a reliable instrument cannot give a positive report if the hypothesis is false. Only option (i) remains: hence the probability of the hypothesis rises to 1 and the probability of the auxiliary theory drops to 0 with a positive report. On the other hand, if there is maximal positive relevance,

the probability of the hypothesis and the auxiliary theory both rise in unison with a positive report, but they do not rise to 1. Assuming that the degree of confirmation is a monotonically increasing function from maximal positive to maximal negative relevance, we can see that positive relevance would score worse than independence.

Once this general scheme is understood, we can gain an intuitive understanding of the details in figure 4.3 and 4.4. In figure 4.3, we see that a high randomization parameter favors independence, while a low randomization parameter favors dependence. How can this be explained? Recall the argument that favors maximal negative relevance: randomization left room for a positive report from an unreliable instrument in support of a true hypothesis. Recall the argument that favors maximal positive relevance: the positive report more strongly confirms the hypothesis since its influence is reinforced by the increase in our confidence in the reliability of the instrument. This reinforcement effect is of course the more pronounced the less we have to deal with positive reports that are just due to randomization. So the argument for the virtues of maximal negative relevance rests on randomization, while the argument for the virtues of maximal positive relevance is offset by randomization. Since independence is on the continuum between maximal negative relevance and maximal positive relevance, it is no wonder that a high randomization parameter favors independence, while a low randomization parameter favors positive relevance. Figure 4.4 focuses on low (but not extremely low) values of a and h, such that positive relevance won't always (but may still) do better than independence. It is no surprise that it will still do better just in case it is a sufficiently strong form of positive relevance.

Let us now turn to our next question. We have received a report from a LTFR instrument to the effect that some testable consequence is true. Subsequently, we increase our confidence in the reliability of the instrument by appealing to an auxiliary

theory that is independent of the hypothesis under test. Is this a successful strategy to increase the degree of confirmation for our hypothesis?

It is easy to see that the answer to this question is univocally positive. Our basic model in figure 2.1 captures the situation before some auxiliary theory in support of our hypothesis has been spotted. The model in figure 4.1 captures the situation after some auxiliary theory has been spotted. We specify a probability distribution P for the Bayesian Network in figure 2.1 and P* for the Bayesian Network in figure 4.1. To respect the *ceteris paribus* clause, we specify the same values h, p, q and a for both distributions, but we choose the remainder of the probability values such that P(REL) < P*(REL). Then, as we show in the appendix,

(4.2) $\Delta P = P^*(HYP \mid REP) - P(HYP \mid REP) > 0.$

Matters are not as simple when we turn our attention to the last question. What happens if we increase our confidence in the reliability of the instrument by appealing to an auxiliary theory and the auxiliary theory and the hypothesis are positively relevant to one another? To investigate this question, we raise the reliability of the instrument by bringing in a positively relevant auxiliary theory: we construct a probability distribution P for our basic model in figure 2.1 and a probability distribution P* for the Bayesian Network in figure 4.2, carefully picking r, t_h and $t_{\bar{h}}$, so that r = P(REL) < P*(REL) = r* and, to respect the *ceteris paribus* clause, so that (ii) P(HYP) = P*(HYP). In the appendix we have shown that

(4.3) $\Delta P = P^*(HYP \mid REP) - P(HYP \mid REP) > 0$ iff $\bar{a}\bar{r}t_h + ar^* + \bar{a}rr^* - \bar{h}r - ht_h > 0.$

In figure 4.5, we set the values at h=.5, a=.4 and $t_h = .8$ and construct a phase curve for values of (r, r*). The part of the graph that interests us is the area above the line where r*>r. In the area above the phase curve a positively relevant auxiliary theory increases the degree of confirmation for the hypothesis. In the area underneath the phase curve a positively relevant auxiliary theory decreases the degree of confirmation for the hypothesis. Notice that there exists a region underneath the phase curve where r*>r. This is curious. Here is how the story goes for this region. We are about to test a hypothesis, but are not confident about the reliability of our instrument: we realize that our confidence in the hypothesis would not increase drastically even if we were to receive a positive report. We try to boost our confidence in the reliability of the instrument and consult an expert. The expert provides us with an auxiliary theory. The auxiliary theory is uncertain, but still boosts our confidence in the reliability of the instrument. It is positively relevant to the hypothesis, but the relevant probability values are such that the prior probability of the hypothesis remains unaffected. It turns out that we will now be less confident that the hypothesis is true after a positive test report comes in than had we not consulted the expert!

The phenomenon is definitely curious, but a moment's reflection will show that it was to be expected given our discussion of question (i) and question (ii.a). Suppose that we have no theoretical support for the reliability of our instrument and that the reliability parameter is set at r. From our discussion of question (ii.a), we know that the hypothesis receives precisely the same degree of confirmation when the reliability parameter has the same value r but rests on the support of some independent auxiliary theory. From our discussion of question (i), we also know that support from an independent as opposed to a dependent auxiliary theory can be better or worse for the degree of confirmation for a hypothesis, depending on the values of h, a, t_h and $t_{\bar{h}}$. So let us change the scenario slightly and assume that an independent auxiliary theory

raises the reliability parameter from r to r+ε, for some small ε. Then this increase will slightly raise the degree of confirmation for the hypothesis. But it is to be expected that this small raise would have been offset, if support had been sought from a dependent auxiliary theory yielding a reliability value of r+ε, at least for particular values of the relevant parameters! Hence, finding support in a dependent auxiliary theory for the reliability of the instrument may lower the degree of confirmation for the hypothesis.

The Duhem-Quine thesis notoriously states that if our experimental results are not in accordance with the hypothesis under investigation, there is no compelling reason to reject the hypothesis, since the blame could just as well fall on the auxiliary theories. One virtue of our model is that it gives a precise Bayesian account of how experimental results affect our confidence in the hypothesis and our confidence in the auxiliary theory. But there is also a more important lesson to be learned. In discussing the Duhem-Quine thesis, Bayesians typically assume that the auxiliary theory and the hypothesis are independent. (Cf. Howson and Urbach 1993, 139) This assumption certainly makes the calculations more manageable, but it does not square with the holism that is the inspiration for the Duhem-Quine thesis. Not only are experimental results determined by a hypothesis and auxiliary theories, they are determined by a hypothesis and auxiliary theories that are often hopelessly interconnected with each other. And these interconnections raise havoc in assessing the value of experimental results in testing hypotheses. There is always the fear that the hypothesis and the auxiliary theory really come out of the same deceitful family and that the lies of one reinforce the lies of the others. What our results show is that this fear is not entirely ungrounded: for hypotheses with a high prior probability, it is definitely better that the reliability of the instrument is supported by an independent auxiliary theory. But on the other hand, for hypotheses with a low prior probability, we should cast off such fears: hypotheses and auxiliary theories from the same family are very welcome, since

positive test reports provide stronger confirmation to the hypothesis under consideration.

5. Calibration

To raise the degree of confirmation for the hypothesis that a particular test result from a LTFR instrument has provided, we can try to increase our confidence in the LTFR instrument by calibrating it. Consider an example: we have a test result in our hands from a LTFR technique for dating artifacts in archeology. A simple form of calibration is to set the technique to work on some artifacts that have their dates chiseled into them (by a reliable stone mason) and to check whether the technique indeed provides the correct output. If so, then we can feel more confident that the technique is indeed reliable and subsequently that the test result and the hypothesis are correct. Let us model this simple form of calibration in a Bayesian Network before moving on to the more complex form in which the LTFR instrument is calibrated against test results from other LTFR instruments.

Suppose that we have a single report from a LTFR instrument and that the content of this report is a testable consequence of some hypothesis. This set up is captured by our basic model in section 2. Subsequently, we identify a series of data that are roughly of the same nature as the testable consequence in question but about which we are confident that they are true. The LTFR instrument is then calibrated by examining whether it yields correct values for these data. To keep things simple, we will model a case with two data (*DAT1* and *DAT2*). Following our heuristic, the reports about these data (*REPDAT1* and *REPDAT2*) are directly influenced by the reliability of the instrument in question and by whether the data are true or not. This yields the graph in figure 5.1.

We assign a probability value of 1 to DAT1 and DAT2 in line with our stipulation that we have chosen *certain* data. Nothing would prevent us of course from inserting lower degrees of confidence into our model. The graph displays a series of independences. One such independence is worth focusing on, since it does reflect a substantial simplification:

(5.1) $DATi \perp CON, DATj \quad \forall i, j = 1,2$ and $i \neq j$.

The data are independent of the testable consequence and are independent of one another. This is a plausible assumption for artifacts that are sufficiently heterogeneous: say, if they are not found on the same site, are not similar in style etc.

We can now turn to a more complex form of calibration which preserves the independence in (5.1). Quite often there are no clean data available against which to calibrate our instruments. Rather, we can do no better than calibrate our instrument against reports from a single or from multiple LTFR instruments about uncertain data. Let the LTFR instrument that is to be calibrated be the *calibratee* and the single or multiple LTFR instruments against whose reports the calibration takes place be the *calibrator*(*s*). If the calibratee yields the same reports as the calibrator(s) about these uncertain data, then we may be more confident that the calibratee is reliable and consequently that the testable consequence and the hypothesis is correct. We will model this more complex form of calibration for two uncertain data *DAT1* and *DAT2*. We receive test reports about these uncertain data from the calibratee (*REPEEDAT1* and *REPEEDAT2*) and from the calibrator(s) (*REPORDAT1* and *REPORDAT2*). Now we draw the following distinction: either we calibrate against the reports from a single calibrator, or we calibrate against the reports from multiple calibrators, one for each datum. In

accordance with this distinction, we can draw two graphs. The variable *RELCAL* expresses the reliability of the single calibrator in the graph in figure 5.2, while the variables *RELCAL1* and *RELCAL2* express the reliability of the respective calibrators in the graph in figure 5.3.

We can read off a series of independences from these graphs. As before, we have made the simplifying assumption that all the instruments are independent:

(5.2) $REPEEDATi \perp REPORDATi \mid DATi$.

We define a probability distribution over each graph and impose our usual *symmetry* conditions (within each distribution) and *ceteris paribus* conditions (between distributions). We assume that the calibrators are either fully reliable or fully unreliable; if they are fully unreliable, then they all are no better than randomizers with a common parameter a, which equals the parameter of the instrument to be tested. Let us also assume that we have the same degree of confidence in all the calibrators and the same degree of confidence in the data. P is the probability distribution for the graph in figure 5.2 and P' is the probability distribution for the graph in figure 5.3. Then, for i = 1,2

(5.3) $P(REPORDATi \mid RELCAL, DATi) = 1$ and $P(REPDATi \mid RELCAL, \overline{DATi}) = 0$

$P(REPORDATi \mid \overline{RELCAL}, DATi) = a$ for both values of *DATi*

$P'(REPEEDATi \mid RELCALi, DATi) = 1$ and $P'(REPDATi \mid RELCALi, \overline{DATi}) = 0$

$P'(REPEEDATi \mid \overline{RELCALi}, DATi) = a$ for both values of *DATi*

$P(RELCAL) = P'(RELCALi) = s$

$P(DATi) = P'(DATi) = f$.

It is reasonable to assume that if we are out to calibrate a LTFR instrument, then we will pick calibrators that we take to be more reliable than the calibratee, i.e. P(REL) = P'(REL) = r < s.

What needs to be investigated is under what conditions the strategy of calibrating against data from a single more reliable instrument as well as the strategy of calibrating against data from multiple more reliable instruments are successful strategies. We consider the point in time at which the hypothesis has received confirmation from a report about the testable consequence from the calibratee. Subsequently, we receive the additional information that the calibrator(s) provided the same reports about both data as the calibratee. In the appendix, we have shown under what conditions the additional information from a single calibrator raises the degree of confirmation for the hypothesis:

(5.4) $\Delta P = P(HYP \mid REP, REPEEDAT1, REPEEDAT2, REPORDAT1, REPORDAT2) -$

$$P(HYP \mid REP) > 0 \text{ iff } a^2 f^2 \bar{s} + \bar{a}(1+a) f^2 s - a^4 \bar{s} > 0$$

and under what conditions the additional information from multiple calibrators raises the degree of confirmation for the hypothesis:

(5.5) $\Delta P = P'(HYP \mid REP, REPEEDAT1, REPEEDAT2, REPORDAT1, REPORDAT2) -$

$$P'(HYP \mid REP) > 0 \text{ iff iff } af + fs - a^2 \bar{s} - 2afs > 0.$$

We plot phase curves for different values of the randomization parameter in the single-calibrator case in figure 5.4. What is going on here? Focus on the area where the data are improbable (i.e. where f is low) and the reliability parameter for the calibrator is low (i.e. where s is low): in this area $\Delta P < 0$, i.e. calibration decreases the degree of

confirmation that the hypothesis receives. This is to be expected: When we get calibration results from a calibrator that is likely to be unreliable and that in addition provides positive reports about implausible data, then we become even more suspicuous of the calibratee, since it yields the same odd results as the calibrator that is likely to be unreliable. And the more suspicuous we are of the calibratee, the less confirmation the hypothesis receives. Furthermore the higher we set the randomization parameter a, the stronger this effect will become, since positive reports are the more likely to come for unreliable instruments.

Subsequently, we are curious to know whether, *ceteris paribus*, the hypothesis receives more or less confirmation if we calibrate against data from a single rather than from multiple calibrators. Is there a general answer, or are there specific conditions under which it is better to calibrate against a single instrument and under which it is better to calibrate against multiple instruments? In the appendix we have shown that,

$$(5.6)\ \Delta P = P(HYP \mid REP, REPEEDAT1, REPEEDAT2, REPORDAT1, REPORDAT2) - P'(HYP \mid REP, REPEEDAT1, REPEEDAT2, REPORDAT1, REPORDAT2) > 0$$

$$\text{iff } 2\bar{a}fs + af + a - 2a\bar{a}s - 2a^2 > 0.$$

We plot phase curves for different values of the randomization parameter in figure 5.6. For all the values of s and f above these curves, $\Delta P>0$ and for all values of s and f underneath these curves, $\Delta P<0$. We see that for lower f, higher s and higher a, it is better to calibrate against two rather than one calibrator. In other words, as the data become less likely, as the calibrator(s) are more likely to be reliable and as the randomization parameter grows, it is better to calibrate against two rather than one calibrator.

How are we to interpret these results? There are two conflicting considerations at work in determining whether it is better to calibrate against a single as opposed to against multiple calibrators. On the one hand, we like to raise the probability that the calibrator is reliable by getting coherent reports from a single instrument. This effect will assert itself when we can assess highly plausible data, when the prior probability that the calibrator is reliable is still low, so that there is lots to be gained from the coherence of the reports, and when the randomization parameter is low, so that positive reports are unlikely to come from unreliable instruments. On the other hand, there is something to be gained from having independent calibrators to improve the reliability of the calibratee. This latter consideration gains the upper hand as the conditions which were favorable to the former consideration wear off: Coherent positive reports about *implausible* facts do not do much to boost the reliability of a single calibrator; if the single calibrator is already very likely to be reliable, then there is little to be gained anymore from coherent positive reports; and if the randomization parameter is set high, then coherent positive reports do not do much to convince us that the single calibrator is reliable, since they are likely to come from unreliable instruments. At this point more is to be gained from receiving independent reports from multiple calibrators.

Compare figure 2.3, 3.5 and 3.6 on the one hand with figure 5.6 on the other hand. In the former figures we compared whether it was better for the confirmation of the hypothesis to receive positive reports from one or from two instruments. Notice that two instruments do better than a single instrument for run-of-the-mill values, such as a=r=.5, p=.8 and q=.2. In the latter figure we compared whether it was better for the confirmation of the hypothesis to obtain agreement between the test instrument and a single or multiple calibrators. Notice that one calibrator does better than two calibrators for run-of-the-mill values such as a=f=.5 and s=.8 (which exceeds r), one calibrator does better than two calibrators. In modeling strategies to receive

confirmation from unreliable instruments with Bayesian Networks, it was this curious difference that first sparked our interest.

6. Concluding Remarks and Future Directions

Let us list some of the more striking results of our investigation:

(i) The standard strategies to deal with unreliable instruments are not always successful: for specific values of the relevant parameters, the degree of confirmation will drop rather than rise when we obtain (a) a positive report about an additional testable consequence from the same LTFR instrument, (b) support for our LTFR instrument from a dependent auxiliary theory, or (c) matching reports from the LTFR instrument and the calibrating instrument(s).

(ii) The variety-of-evidence thesis is not sacrosanct: positive reports from single rather than from multiple instruments and positive reports about a single rather than about multiple consequences may, *ceteris paribus*, provide more confirmation to a hypothesis.

(iii) The Duhem-Quine thesis is no reason to despair about confirmation. An appeal to an auxiliary theory in support of a LTFR instrument can improve the degree of confirmation for the hypothesis and the interdependency between the auxiliary theory and the hypothesis tends to favor the confirmation of initially less plausible hypotheses.

(iv) For run-of-the-mill values, positive reports from *multiple* instruments raise the degree of confirmation more than from a *single* instrument in repeated testing, while matching reports from a *single* calibrating instrument raise the degree of confirmation more than from *multiple* calibrating instruments.

We have taken the first steps in developing a new approach to thinking about confirmation and unreliable instruments. There are many directions to be explored. We conclude with a few suggestions for further research.

(I) The independency assumptions have substantially simplified our work, but may not always be warranted. Note that if an unreliable instrument is not an instrument that acts as a randomizer, but rather is an instrument that provides accurate measurements of other features than what it is supposed to measure, then the independence assumptions need to be relaxed.

(II) We have restricted our attention to two reports in our discussion of strategy 1 and 2. How does a *series* of positive reports from single *versus* multiple LTFR instruments affect the confirmation of the hypothesis? Similarly we have restricted ourselves to two reports on data in strategy 4. How does a *series* of matching reports from single *versus* multiple calibrators affect the confirmation of the hypothesis? Do we reach convergence and can a general characterization be given of the paths that lead towards convergence?

(III) In highly developed fields of science, there is often an intricate relationship between the hypothesis under investigation and the auxiliary theories. Consider the recent discovery of the top quark. This fundamental particle is suggested by the Standard Model of particle physics. But certain elements of this model also come in in the methods that were used to analyze the data collected by the instruments. These interrelationships are extremely complex and our model in strategy 3 is highly idealized. A case study in which a Bayesian Network is constructed that models the scientific process would lend support to our analysis.

(IV) We have restricted our attention to discrete binary variables. More often than not scientific experimentation deals with continuous variables. An extension of our approach in this direction could be rewarding.

(V) We have investigated how positive reports from LTFR instruments affect the degree of confirmation for the hypothesis under various strategies. But of course, in the beginning of the day, a researcher does not know whether positive or negative reports will be forthcoming. Even so, our approach can be turned into a decision procedure as to what strategy is to be preferred in a particular context. Consider a hypothesis which states that a patient has a particular disease and a policy that treatment will be started just in case the posterior probability of the hypothesis exceeds some critical value. We specify the utility values of treatment and abstention from treatment when the patient actually does and does not have the disease. We can then calculate the expected utility of a particular strategy of dealing with LTFR instruments at the beginning of the day and make recommendations accordingly. Leaning on decision-theoretic work in the theory of Bayesian Networks, a systematic study in a particular context may meet with genuine practical applications.

REFERENCES

DAWID, A. P. (1979) “Conditional Independence in Statistical Theory”, *Journal of the Royal Statistical Society* A41: 1-31.

EARMAN, J. (1992) *Bayes or Bust? A Critical Examination of Bayesian Confirmation Theory*. Cambridge [MA]: MIT Press.

FRANKLIN, A. (1986) *The Neglect of Experiment*. Cambridge: CUP.

HOWSON, C. and URBACH, P. (1993) *Scientific Reasoning - The Bayesian Approach*. (2nd ed.) Chicago: Open Court.

JENSEN, F.V. (1996) *An Introduction to Bayesian Networks*. Berlin: Springer.

NEAPOLITAN, R.E. (1990) *Probabilistic Reasoning in Expert Systems*. New York: Wiley.

PEARL, J. (1988) *Probabilistic Reasoning in Intelligent Systems*. San Mateo, [Ca].: Morgan Kaufmann.

SPOHN, W. (1980) “Stochastic Independence, Causal Independence, and Shieldability”, *Journal of Philosophical Logic* 9: 73-99.

APPENDIX

We provide the mathematical derivations of the equations that are presented in the text.

(2.2) We will lay out a standard procedure which will be repeated throughout the appendix. For any combination of values of the variables *HYP*, *REL*, *CON*, *REP1* and *REP2*, set Y(*HYP*, *REL*, *CON*, *REP1*, *REP2*) = P(*HYP*) P(*REL*) P(*CON* | *HYP*) P(*REP1* | *REL*, *CON*) P(*REP2* | *REL*, *CON*). For the Bayesian Network in figure 2.1,

$$\mathrm{P(HYP \mid REP1,REP2)} = \frac{\sum_{REL,CON} \mathrm{Y(HYP}, REL, CON, \mathrm{REP1, REP2)}}{\sum_{HYP,REL,CON} \mathrm{Y}(HYP, REL, CON, \mathrm{REP1, REP2})} \text{ and}$$

$$\mathrm{P(HYP \mid REP1)} = \frac{\sum_{REL,CON,REP2} \mathrm{Y(HYP}, REL, CON, \mathrm{REP1}, REP2)}{\sum_{HYP,REL,CON,REP2} \mathrm{Y}(HYP, REL, CON, \mathrm{REP1}, REP2)}.$$

Substituting the corresponding values and some algebraic manipulation yields

$$\Delta\mathrm{P=P(HYP \mid REP1,REP2) - P(HYP \mid REP1)} = \frac{a\bar{a}h\bar{h}r\bar{r}(p-q)}{(a\bar{r}+qr+hr(p-q))(a^2\bar{r}+qr+hr(p-q))}.$$

Since 0 < a, h, r < 1 and p > q, the expression is clearly greater than 0.

(2.3) Following our standard procedure for the Bayesian Network in figure 2.2, we calculate that

$$\Delta\mathrm{P=P'(HYP \mid REP1,REP2) - P'(HYP \mid REP1)} =$$

$$\frac{ah\bar{h}r\bar{r}(a\bar{r}+r)(p-q)}{(a\bar{r}+qr+hr(p-q))(a^2\bar{r}^2+qr^2+2aqr\bar{r}+hr(p-q)(r+2a\bar{r}))}.$$

Since 0 < a, h, r < 1 and p > q, the expression is clearly greater than 0.

(3.1) Following our standard procedure for the Bayesian Network in figure 3.2, we calculate that

$\Delta P = P'(HYP \mid REP1, REP2) - P'(HYP \mid REP1) =$

$$\frac{h\bar{h}r(p-q)(rp+a\bar{r})(rq+a\bar{r})}{(hr(p-q)+rq+a\bar{r})(hr^2(p^2-q^2)+2ahr\bar{r}(p-q)+2ahpr+2aqr\bar{r}+q^2r^2)}.$$

Since $0 < a, h, r < 1$ and $p > q$, the expression is clearly greater than 0.

(3.2) Following our standard procedure for the Bayesian Network in figure 3.1, we calculate that

$$\Delta P = P(HYP \mid REP1, REP2) - P(HYP \mid REP1) = \frac{h\bar{h}r(a\bar{r}(p+q-a)+pqr)}{(hr(p-q)+a\bar{r}+qr)(hr(p^2-q^2)+a^2\bar{r}+q^2r)}.$$

Since $0 < p\text{-}q < 1$ and $0 < h, a, r < 1$, $\Delta P > 0$ iff $a\bar{r}(p+q-a)+pqr > 0$

(3.3) P(HYP|REP1,REP2) and P'(HYP|REP1,REP2) were calculated in respectively (3.1) and (3.2). We now calculate the difference:

$\Delta P = P'(HYP \mid REP1, REP2) - P(HYP \mid REP1, REP2) =$

$$\frac{ah\bar{h}(p-q)\bar{r}(2a^2\bar{r}-2pqr-a(p+q)(1-2r))}{(hr(p^2-q^2)+q^2r+a^2\bar{r})(hr(p^2-q^2)+q^2r^2+(2ahr(p-q)+2aqr+a^2)\bar{r})}.$$

Since $0 < p\text{-}q < 1$ and $0 < h, a, r < 1$, $\Delta P > 0$ iff $2a^2\bar{r}-2pqr-a(p+q)(1-2r) > 0$.

(4.1) Consider the Bayesian Network in figure 4.1. Add an arrow from *HYP* to *AUX* and define a new probability distribution P* over this new Bayesian Network. Since *AUX* is no longer a root node, we delete P(AUX) = t and fill in $P^*(AUX \mid HYP) = t_h = t$ and $P^*(AUX \mid \overline{HYP}) = t_{\bar{h}} = t$. For all other probability values in the Bayesian Network, P = P*. It is easy to show that this adapted Bayesian Network expresses precisely the same probability distribution as the Bayesian Network in figure 4.1. We follow the standard procedure for the adapted Bayesian Network and calculate P*(HYP|REP) which is equal to P(HYP|REP). Subsequently we follow the standard procedure for the Bayesian Network in figure 4.2 and calculate P'(HYP|REP). We now construct the difference:

$$\Delta P = P(HYP|REP) - P'(HYP|REP) = \frac{ah\bar{h}(t_h - t_{\bar{h}})(h + \bar{a}(ht_h + \bar{h}t_{\bar{h}} - 1))}{(aht_{\bar{h}} + \bar{a}ht_h + a\bar{t}_{\bar{h}})(a + (h - a)(ht_h + \bar{h}t_{\bar{h}}))}.$$

Since $0 < (t_h - t_{\bar{h}}) < 1$ and $0 < a, h < 1$, $\Delta P > 0$ iff $h + \bar{a}(ht_h + \bar{h}t_{\bar{h}} - 1) > 0$.

(4.2) Construct a probability distribution P# for the Bayesian Network in figure 2.1 which is just like P, except that P#(REL) = P*(REL) > P(REL). It is easy to show that P#(HYP|REP) = P*(HYP|REP). Hence, to prove that (4.2), it is sufficient to prove that for any properly constrained probability distribution **P** for the Bayesian Network in figure 2.1, **P**(HYP|REP) is a positively increasing function of r. By our standard procedure, we calculate for the Bayesian Network in figure 2.1 that **P**(HYP|REP) $= h(r + \bar{r}a)/(hrp + \bar{r}a)$ and we differentiate towards r: $d\mathbf{P}(HYP|REP)/dr = ah\bar{h}/(a\bar{r} + hr)^2$. Since $0 < a, h, r < 1$, this expression is greater than 0 and hence **P**(HYP|REP) is a positively increasing function of r.

(4.3) By our standard procedure, we calculate P(HYP|REP) for the Bayesian Network in figure 2.1 and P*(HYP|REP) for the Bayesian Network in figure 4.2. Since $r_t = 1$ and $r_{\bar{t}} = 0$, $r^* = t_h h + t_{\bar{h}}\bar{h}$. Hence we can substitute $t_{\bar{h}}$ for $(r^* - ht_h)/\bar{h}$ in P*(HYP|REP). We calculate:

$$\Delta P = P^*(HYP|REP) - P(HYP|REP) = \frac{ah(\bar{a}\bar{r}t_h + ar^* + \bar{a}rr^* - \bar{h}r - ht_h)}{(a\bar{r} + hr)(a\bar{r^*} + ht_h)}.$$

Since $0 < a, h, r, r^* < 1$, $\Delta P > 0$ iff $\bar{a}\bar{r}t_h + ar^* + \bar{a}rr^* - \bar{h}r - ht_h > 0$.

(5.4) and **(5.5)** Following our standard procedure for the Bayesian Networks in figures 5.2 and 5.3, we calculate that

ΔP = P(HYP|REP, REPEEDAT1,REPEEDAT2,REPORDAT1,REPORDAT2) - P(HYP|REP) =

$$\frac{a\bar{h}r\bar{r}(p-q)(a^2f^2\bar{s}+\bar{a}(1+a)f^2s-a^4\bar{s})}{(hr(p-q)+a\bar{r}+q\bar{r})((a^2\bar{s}+s)f^2hr(p-q)+a^5\bar{r}\bar{s}+a^3f^2(\bar{r}s+qr\bar{s})}.$$

Since $0 < a, f, h, r, s, q < 1$ and $p > q$, $\Delta P > 0$ iff $a^2f^2\bar{s}+\bar{a}(1+a)f^2s-a^4\bar{s} > 0$.

Furthermore,

ΔP = P'(HYP | REP,REPEEDAT1,REPEEDAT2,REPORDAT1,REPORDAT2) - P'(HYP | REP) =

$$\frac{1}{N_1N_2}a\bar{h}r\bar{r}(p-q)(af+fs+a^2\bar{s})(af+fs-a^2\bar{s}-2afs) \text{ with}$$

$N_1 = hr(p-q)+rq+a\bar{r}$

and

$N_2 = (a^2\bar{s}^2+as+s^2)f^2hr(p-q)+a^5\bar{r}\bar{s}^2+2a^4\bar{r}s\bar{s}+a^3f^2\bar{r}s^2+a^2f^2qr\bar{s}^2+2af^2qrs\bar{s}+f^2qrs^2.$

Since $0 < a, f, h, r, s, q, < 1$ and $p > q$, $\Delta P > 0$ iff $af+fs-a^2\bar{s}-2afs > 0$.

(5.6) Leaning on (5.4) and (5.5), we calculate:

ΔP = P(HYP | REP, REPEEDAT1, REPEEDAT2, REPORDAT1, REPORDAT2) -

P'(HYP | REP, REPEEDAT1, REPEEDAT2, REPORDAT1, REPORDAT2) =

$$\frac{1}{N_1N_2}a^4f^2\bar{f}h\bar{h}r\bar{r}s\bar{s}(p-q)(2\bar{a}fs+af+a-2a\bar{a}s-2a^2) \text{ with}$$

$N_1 = (a^2\bar{s}^2+s^2\bar{a}^2)f^2hr(p-q)+a^5\bar{r}\bar{s}^2+2a^4f\bar{r}s\bar{s}+a^3f^2s^2\bar{r}+(2a\bar{a}s+(a-s)^2)f^2qr$

and

$N_2 = (a^2\bar{s}+s)f^2hr(p-q)+a^5\bar{r}\bar{s}+a^3f^2\bar{r}s+a^2f^2qr\bar{s}+f^2qr.$

Since $0 < a, f, h, r, s, q, < 1$ and $p > q$, $\Delta P > 0$ iff $(2\bar{a}fs+af+a-2a\bar{a}s-2a^2)>0$.

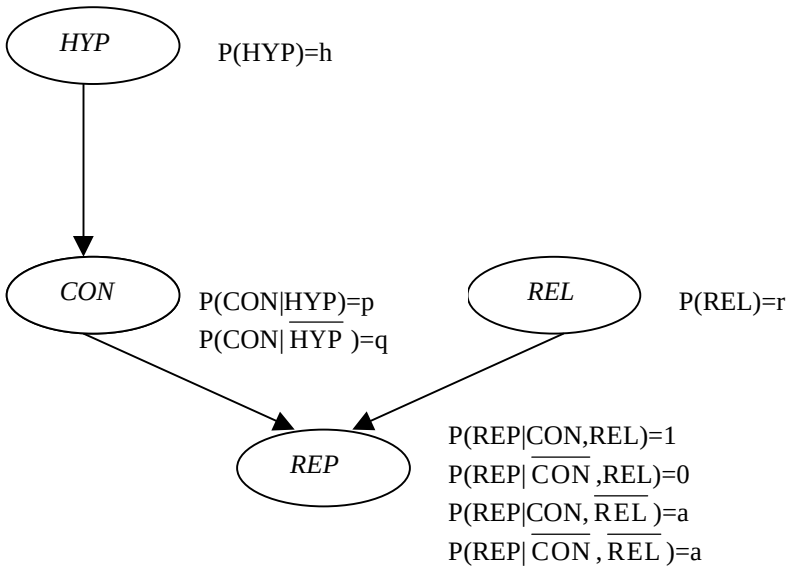
HYP
P(HYP)=h
CON
P(CON|HYP)=p
P(CON|$\overline{\text{HYP}}$)=q
REL
P(REL)=r
REP
P(REP|CON,REL)=1
P(REP|$\overline{\text{CON}}$,REL)=0
P(REP|CON,$\overline{\text{REL}}$)=a
P(REP|$\overline{\text{CON}}$,$\overline{\text{REL}}$)=a

Figure 1.1

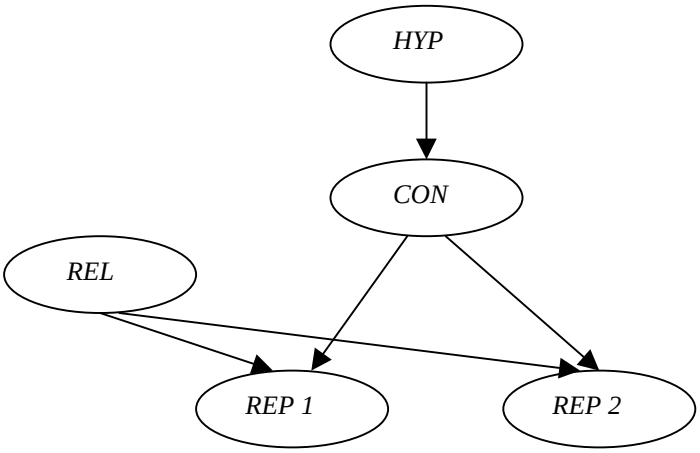

Figure 2.1

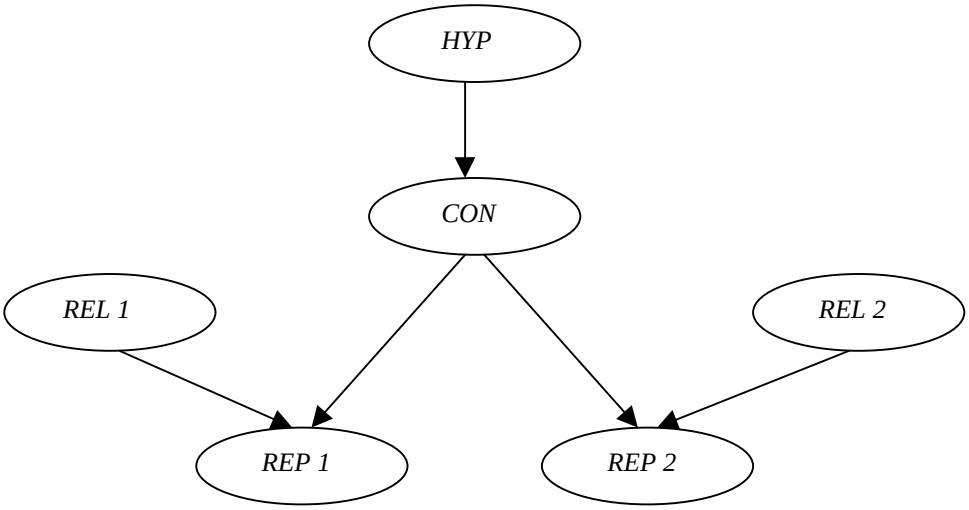

Figure 2.2

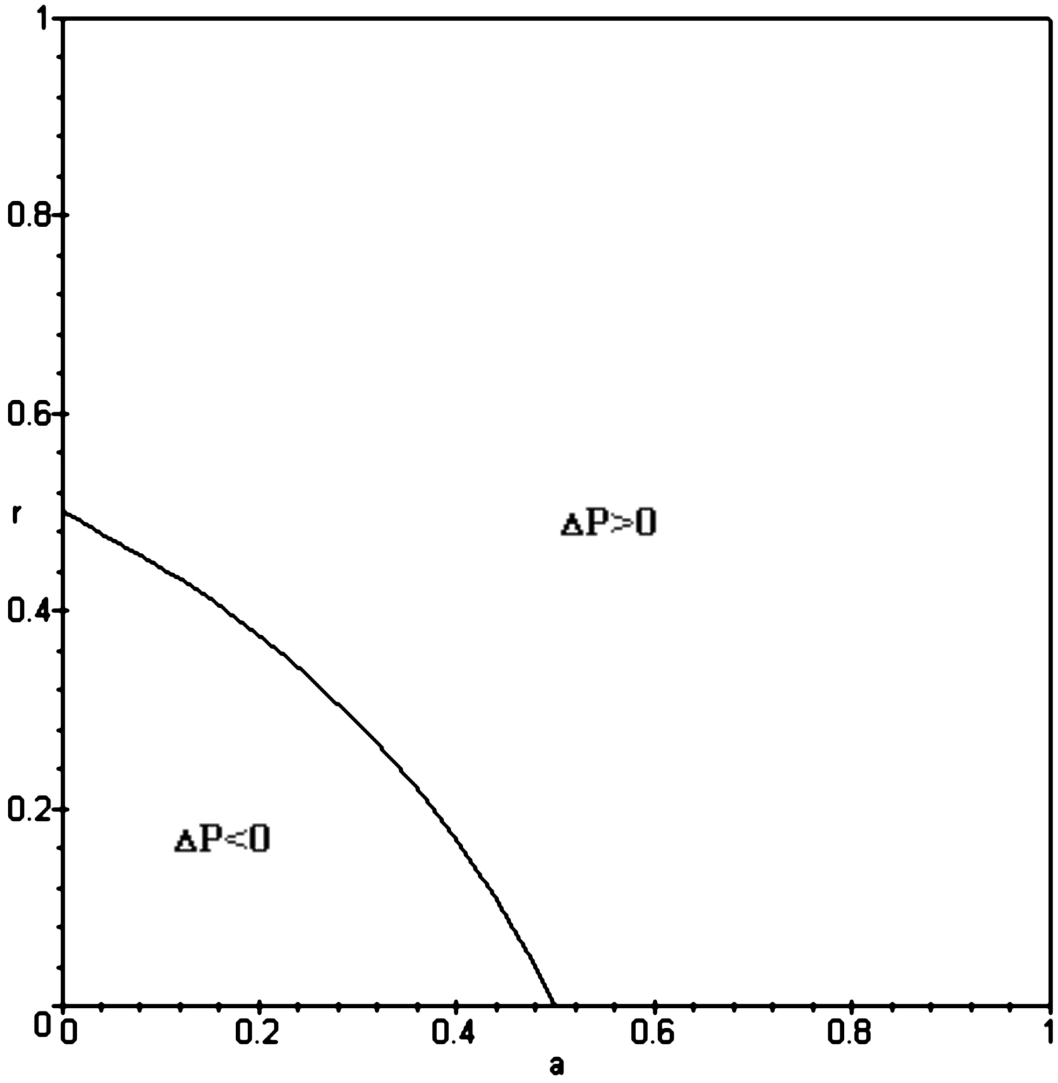

Figure 2.3

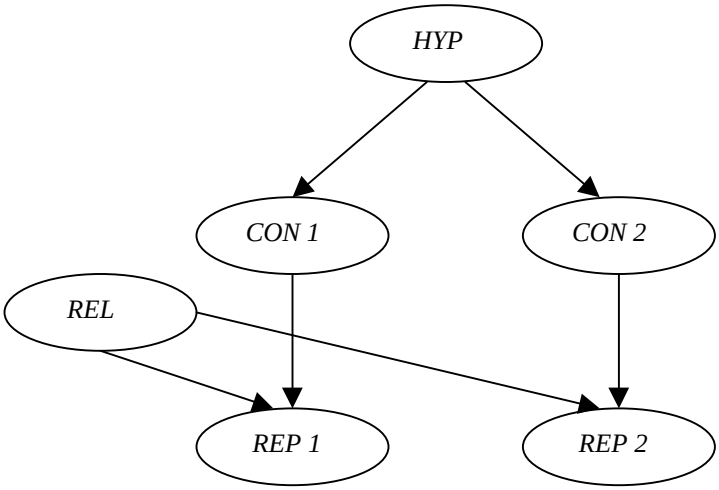

Figure 3.1

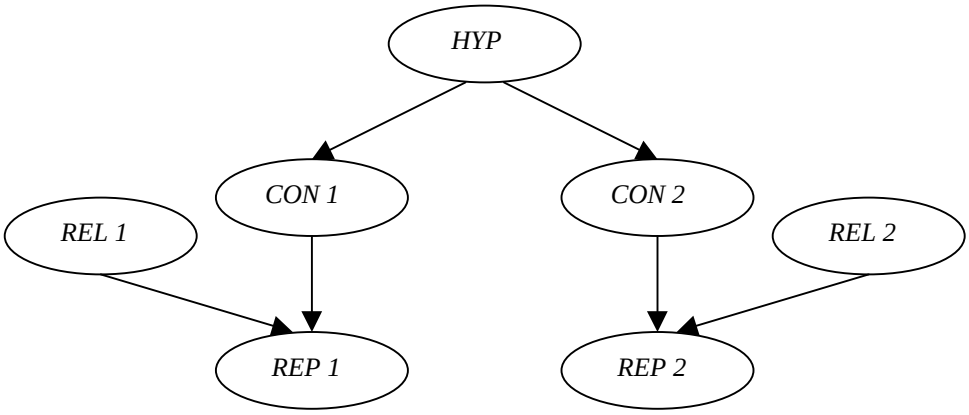

Figure 3.2

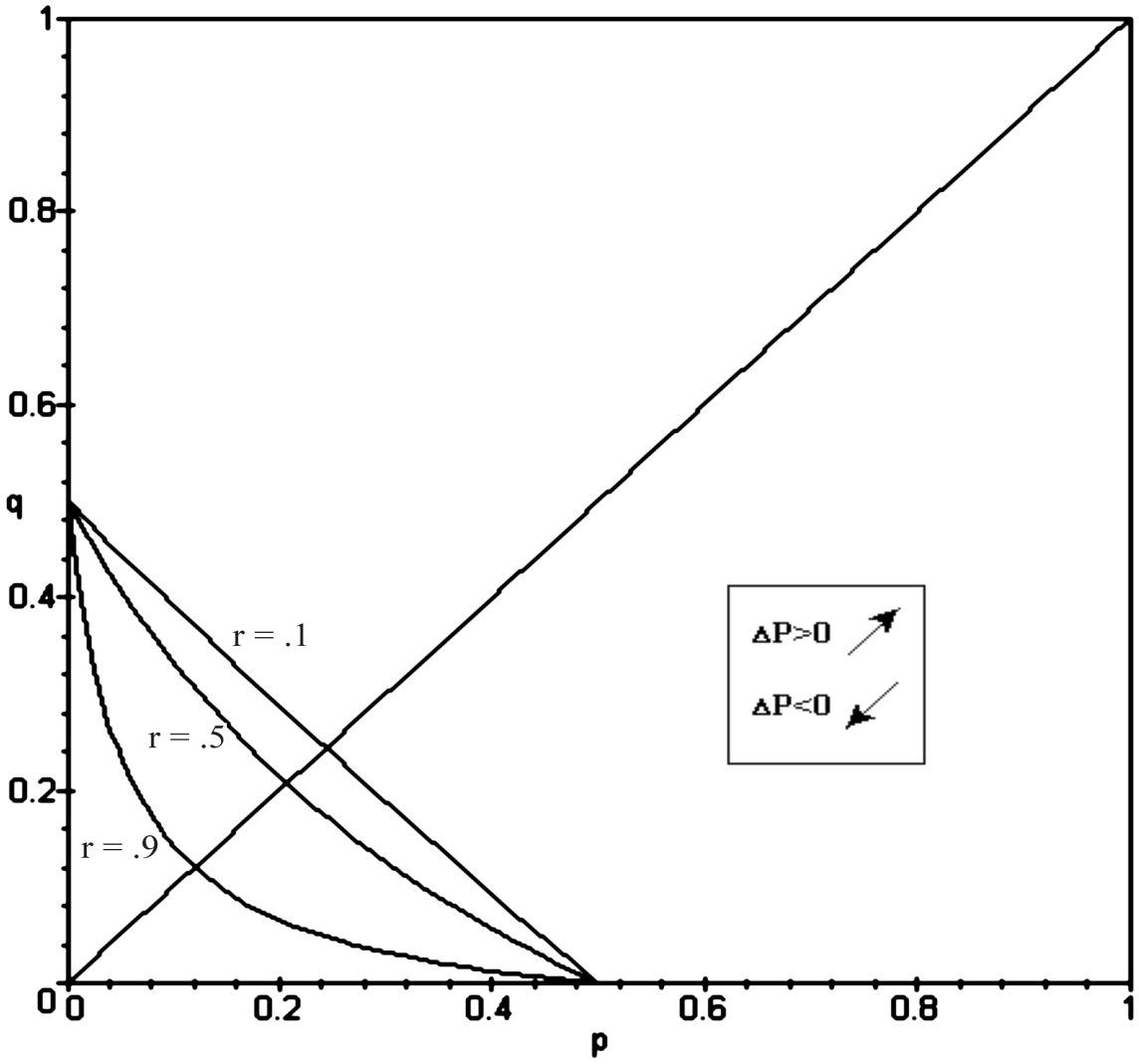

Figure 3.3

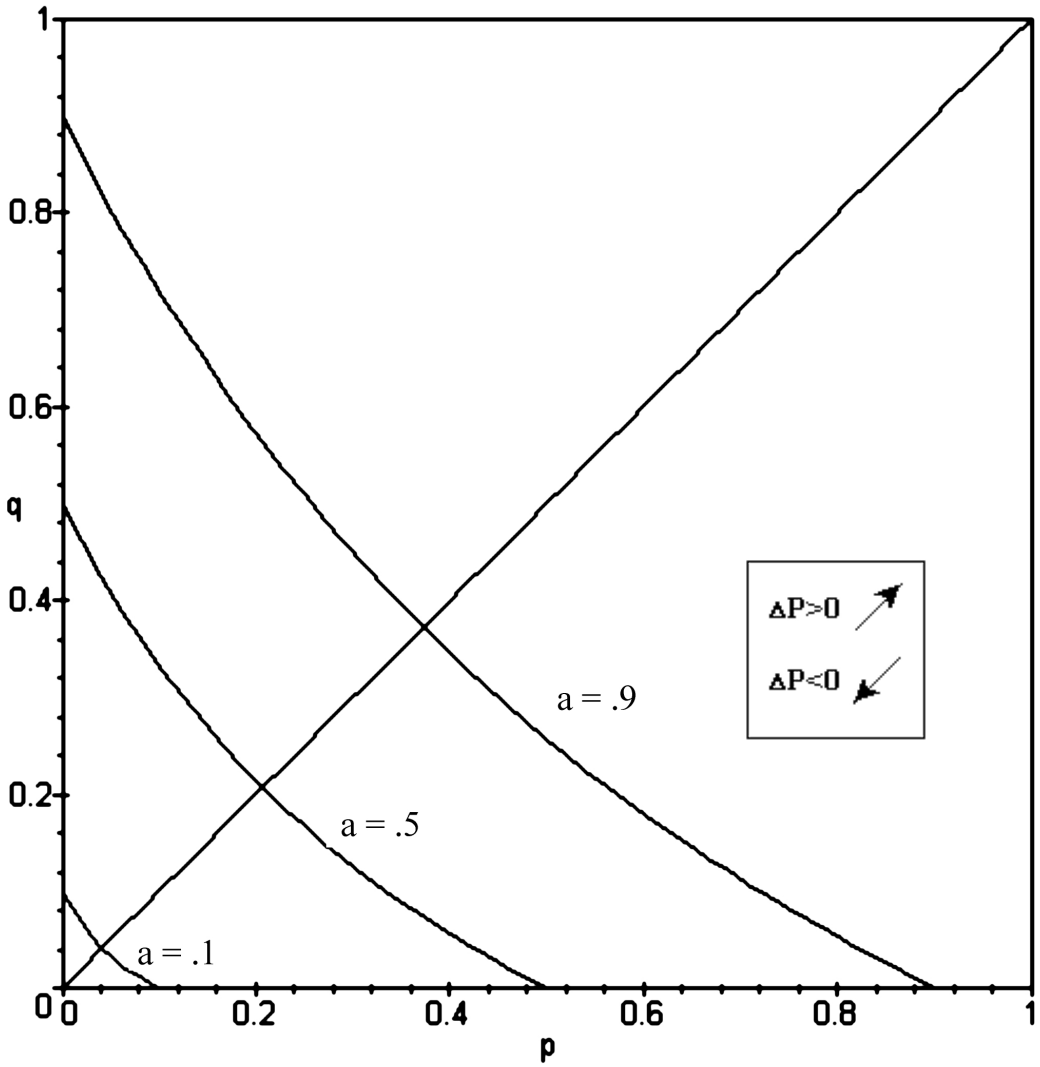

Figure 3.4

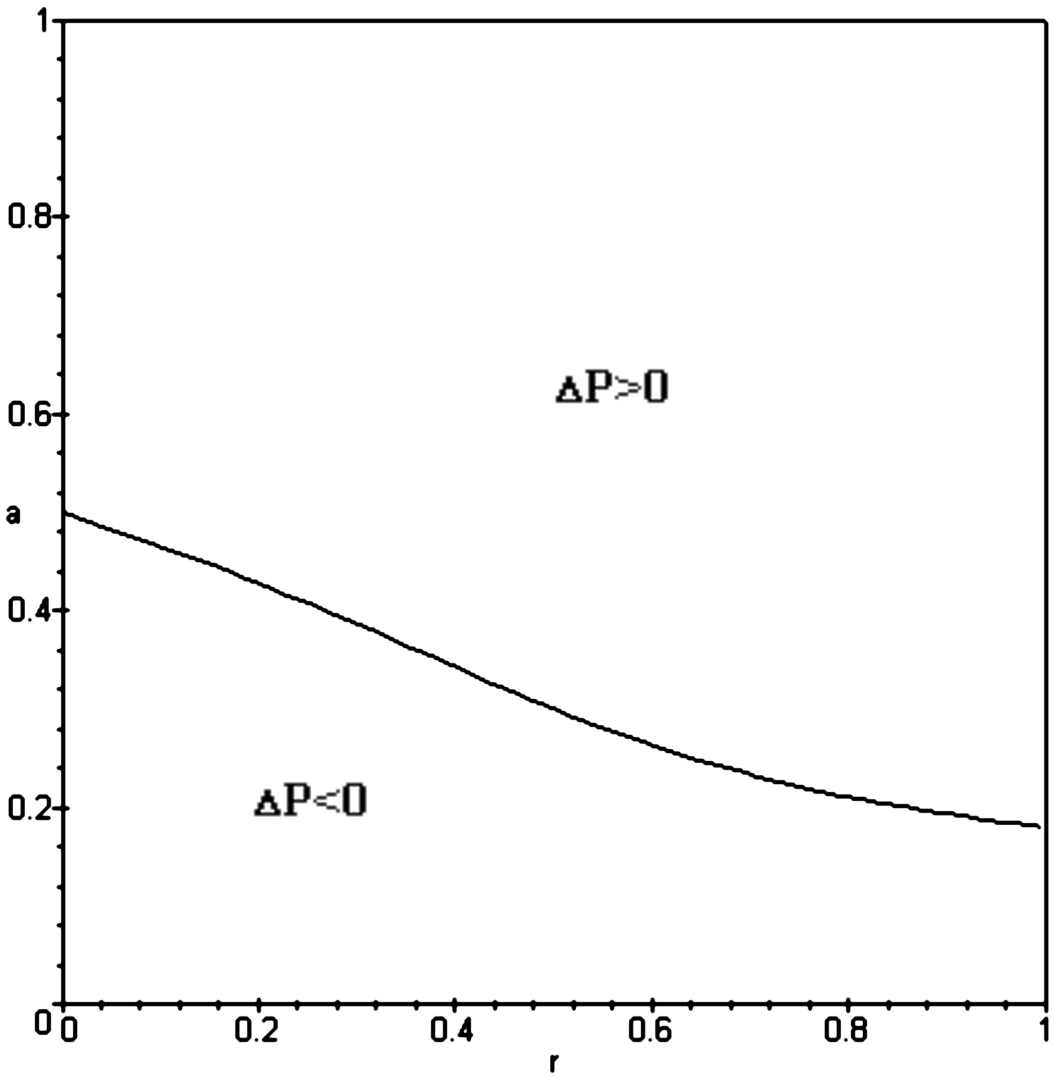

Figure 3.5

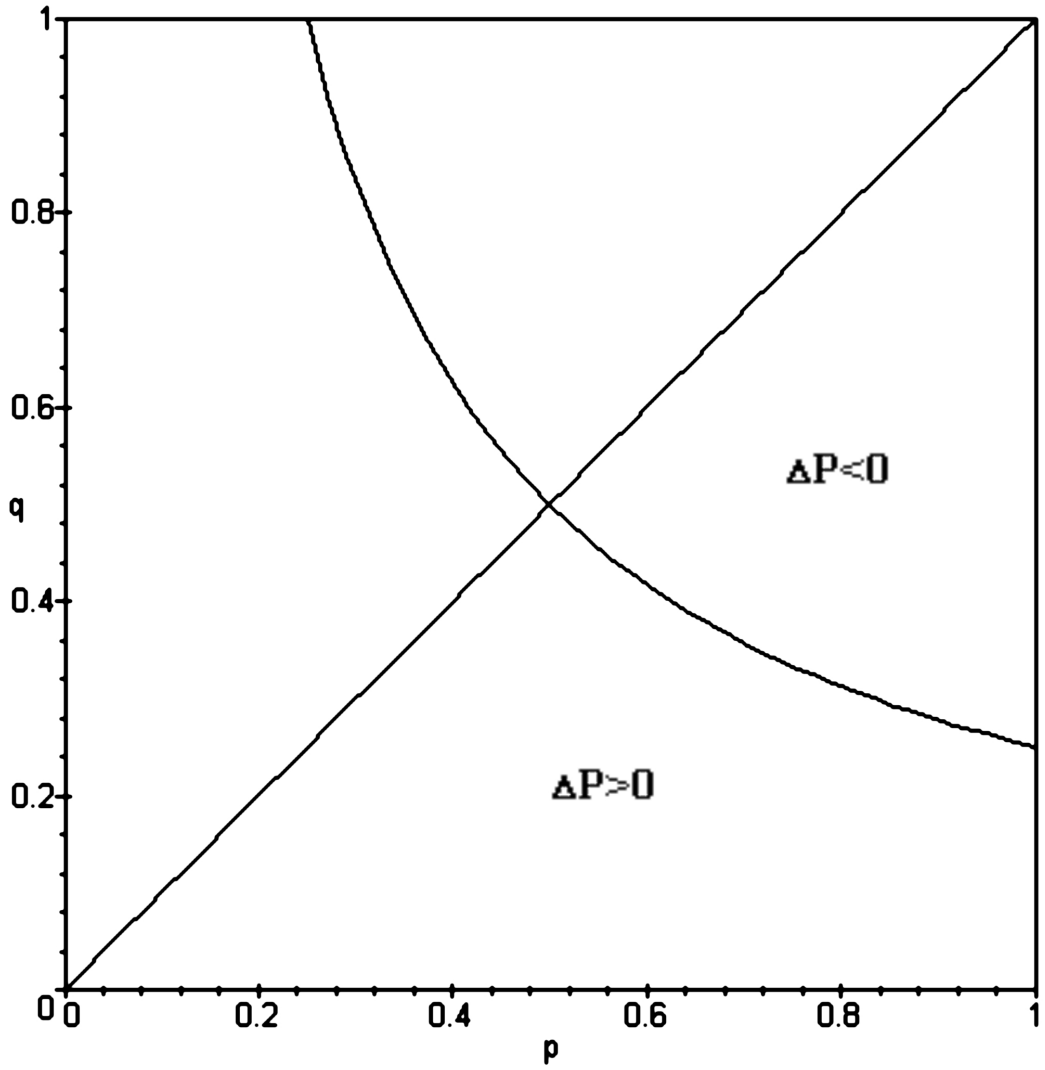

Figure 3.6

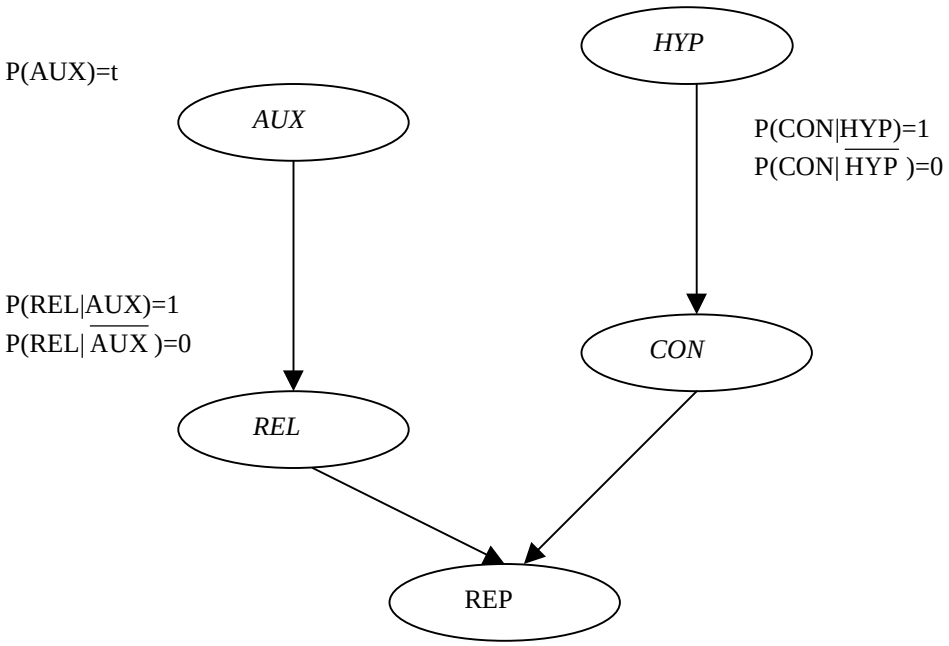
P(AUX)=t
AUX
HYP
P(CON|HYP)=1
P(CON| $\overline{\text{HYP}}$)=0
P(REL|AUX)=1
P(REL| $\overline{\text{AUX}}$)=0
CON
REL
REP

Figure 4.1

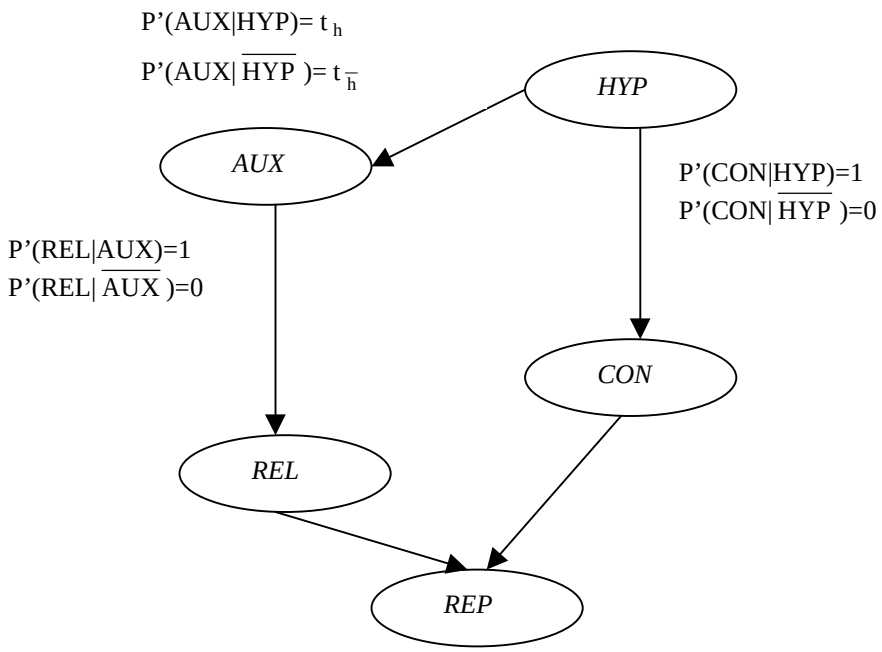
P’(AUX|HYP)= t h
P’(AUX| HYP)= t h
HYP
AUX
P’(CON|HYP)=1
P’(CON| HYP)=0
P’(REL|AUX)=1
P’(REL| AUX)=0
CON
REL
REP

Figure 4.2

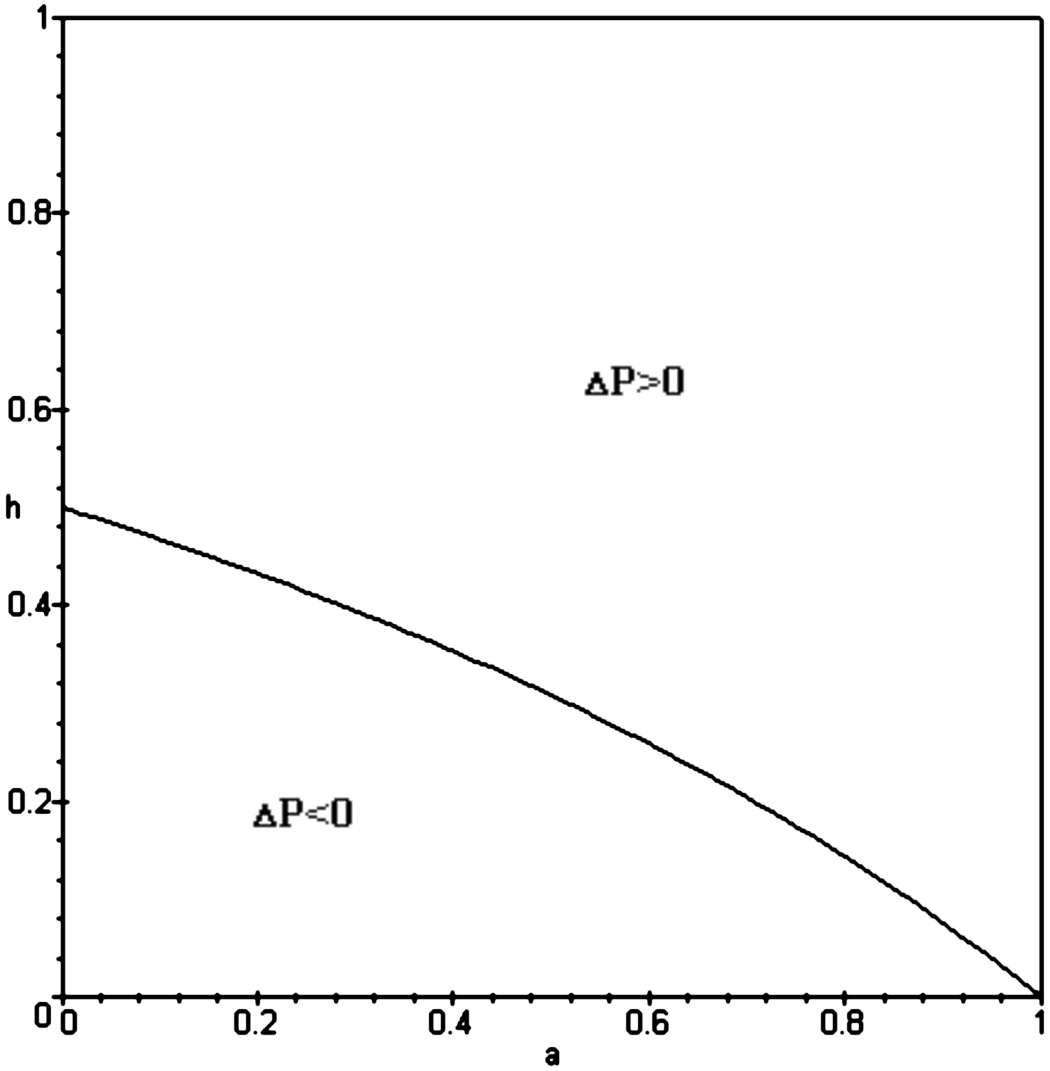

Figure 4.3

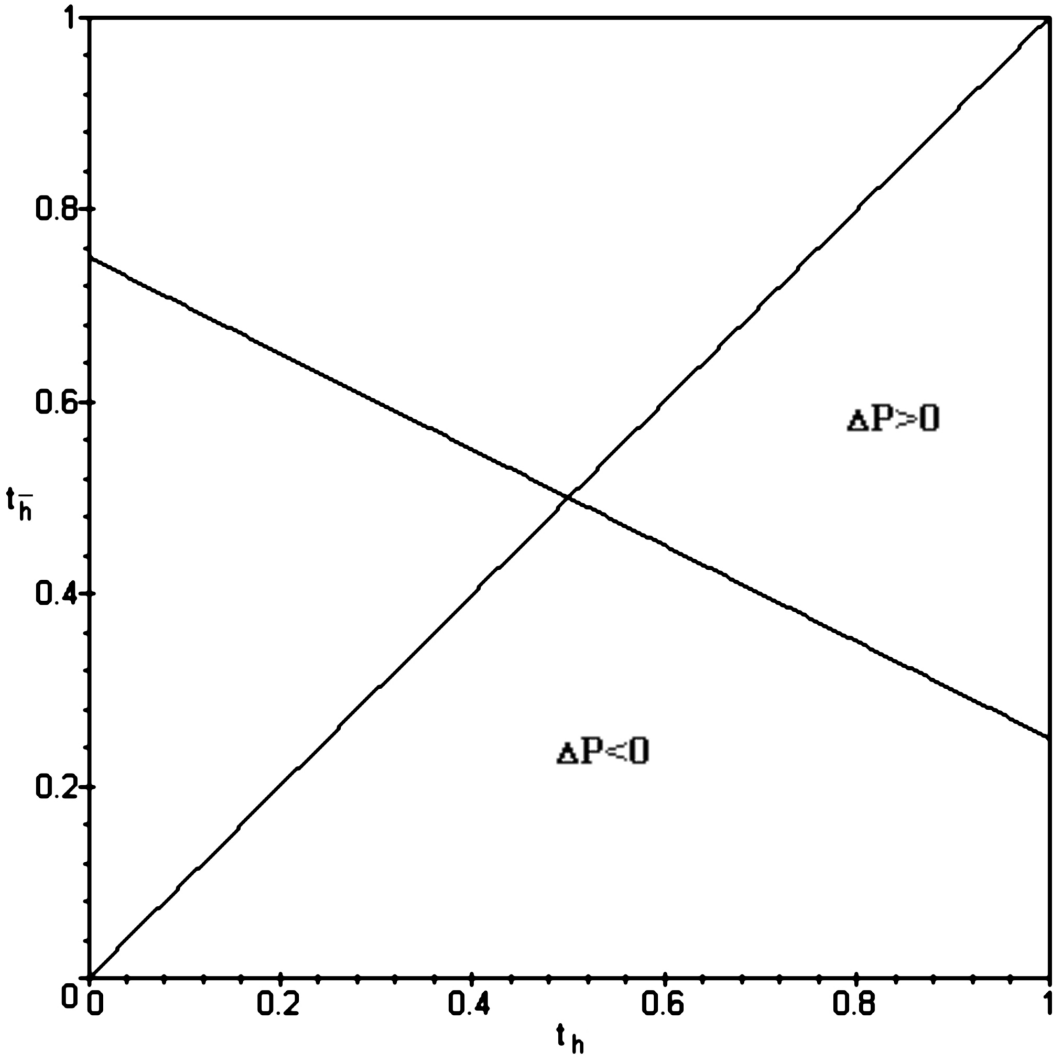

Figure 4.4

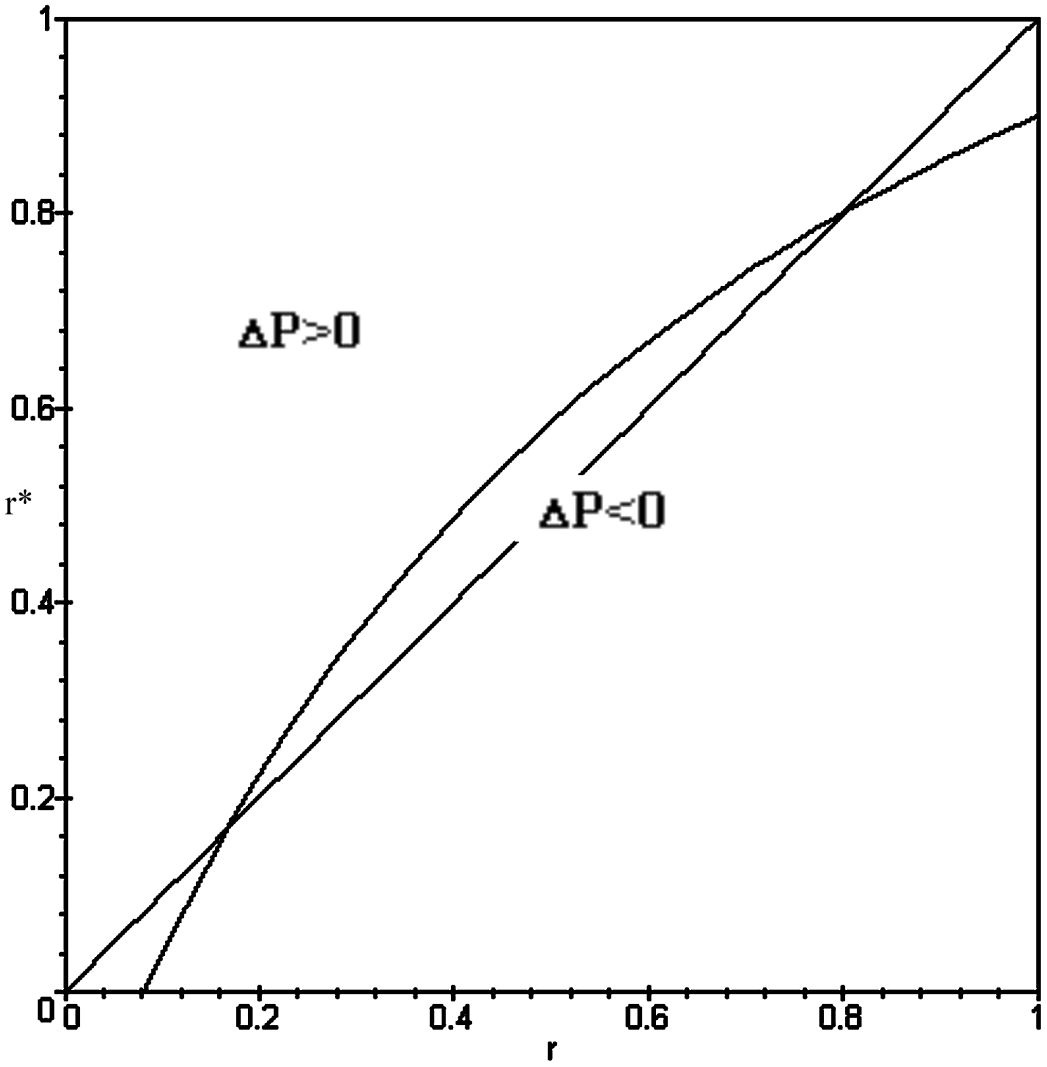

Figure 4.5

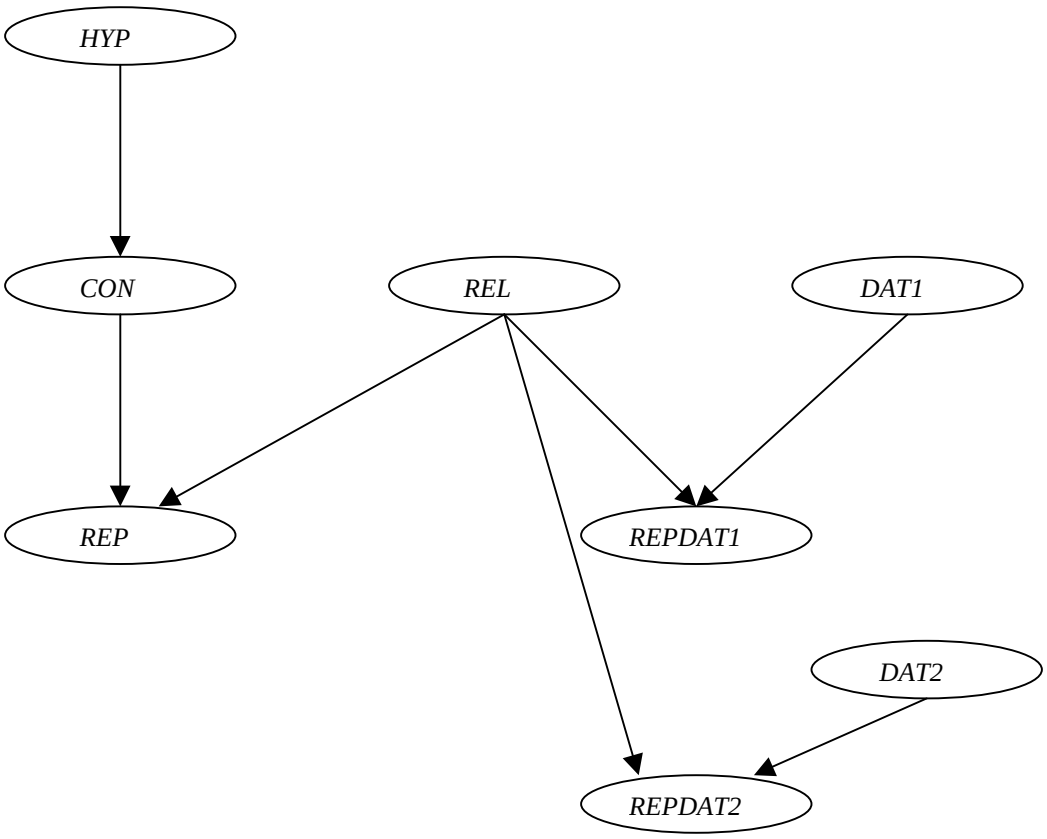

Figure 5.1

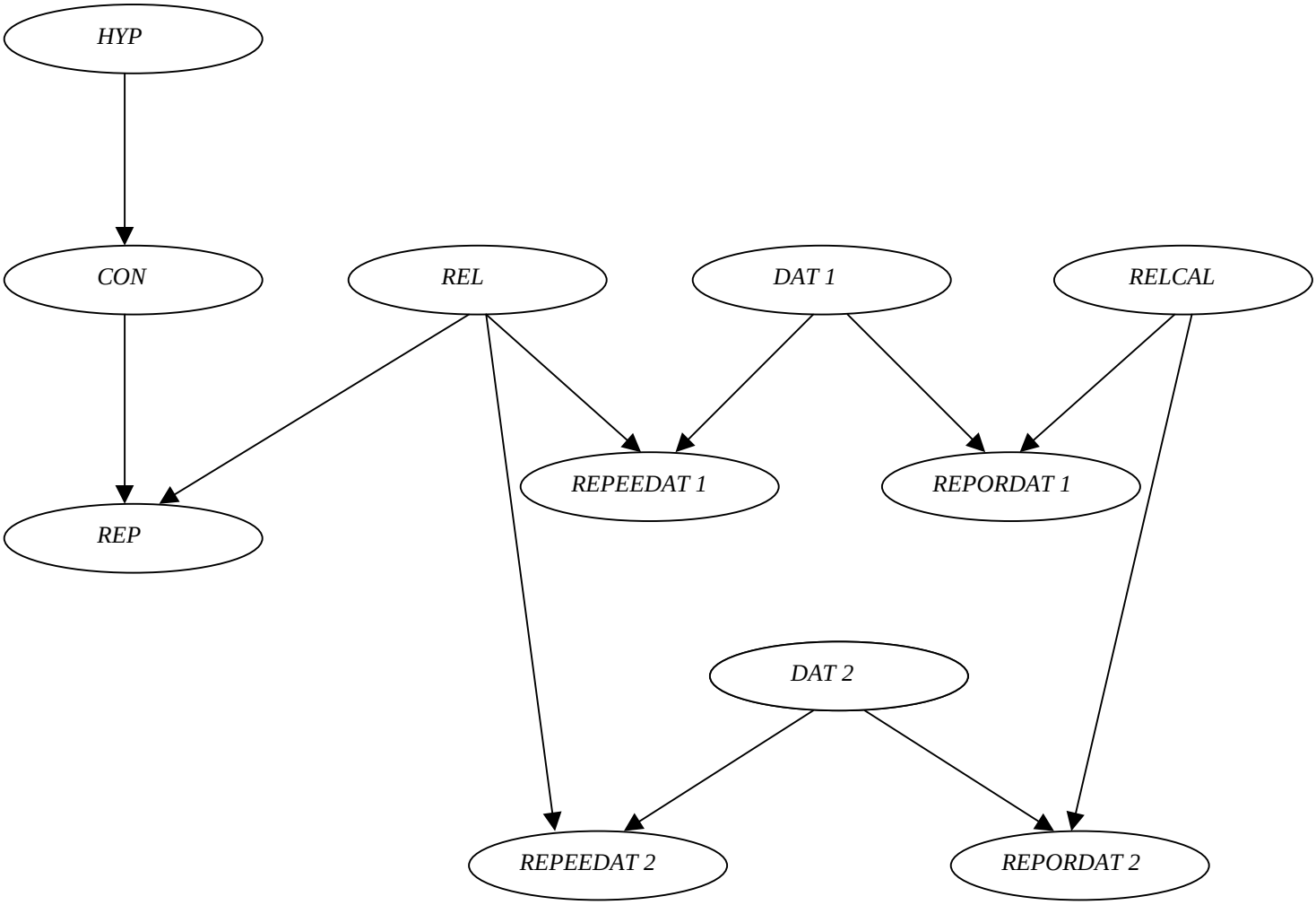
HYP
CON
REL
DAT 1
RELCAL
REPEEDAT 1
REPORDAT 1
REP
DAT 2
REPEEDAT 2
REPORDAT 2

Figure 5.2

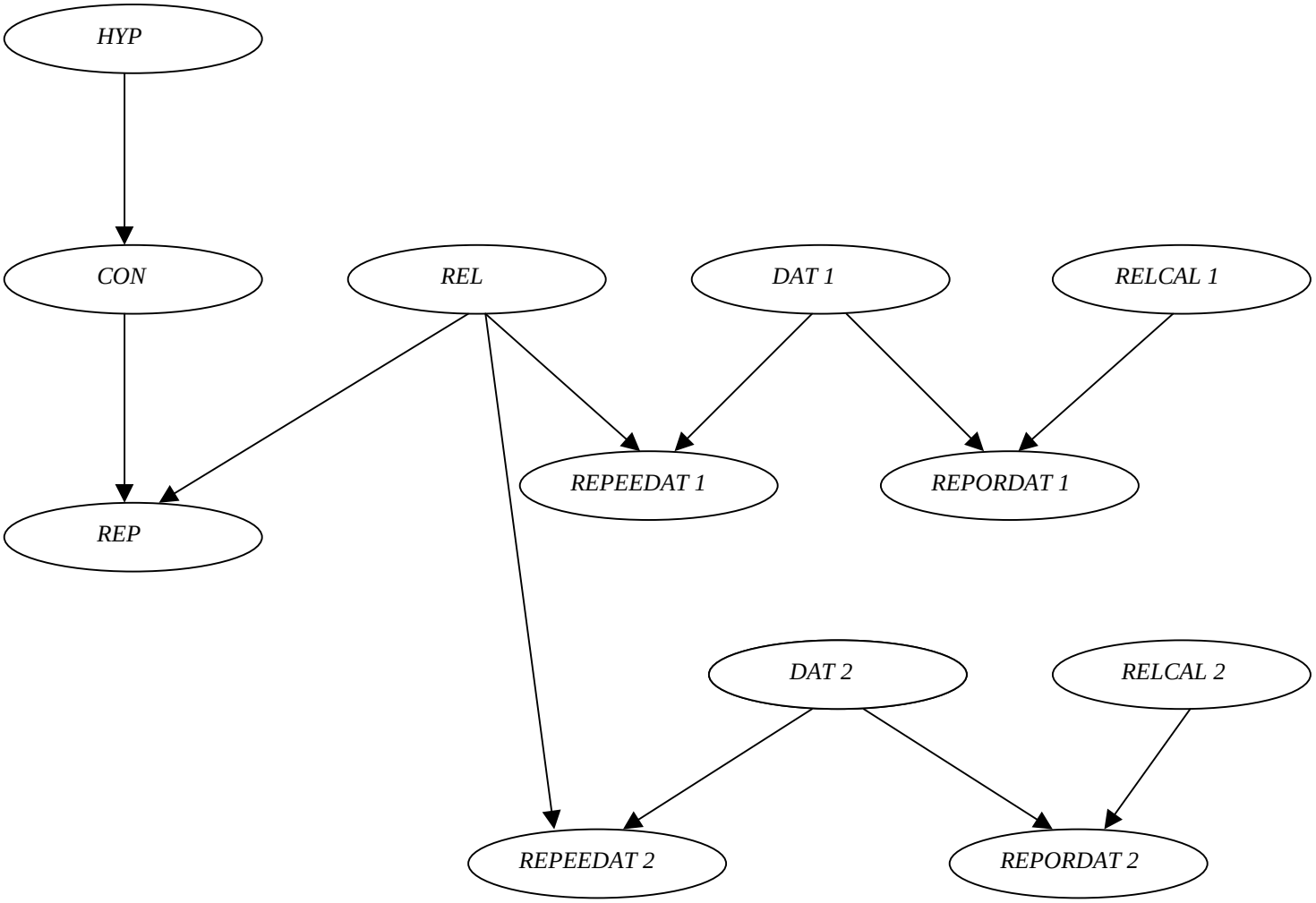
HYP
CON
REL
DAT 1
RELCAL 1
REPEEDAT 1
REPORDAT 1
REP
DAT 2
RELCAL 2
REPEEDAT 2
REPORDAT 2

Figure 5.3

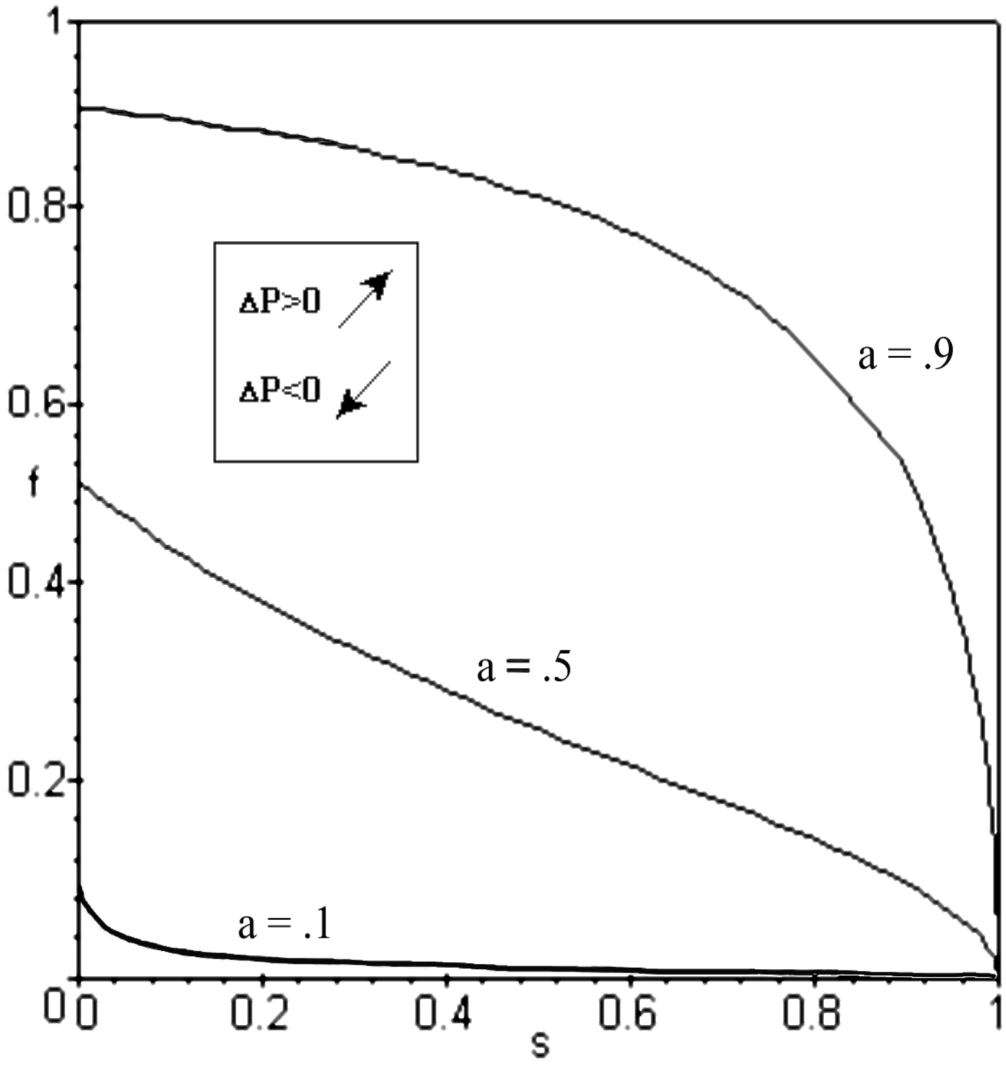

Figure 5.4

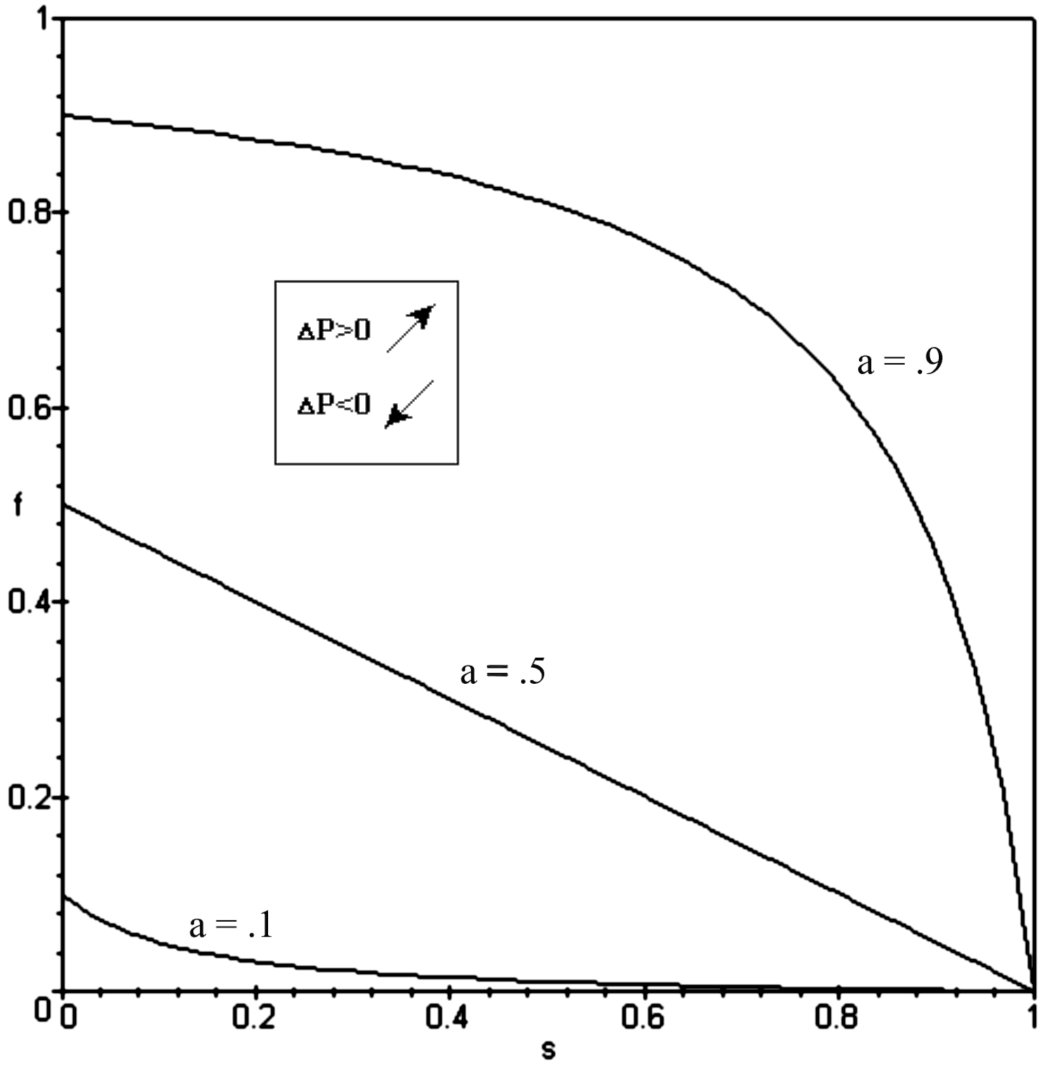

Figure 5.5

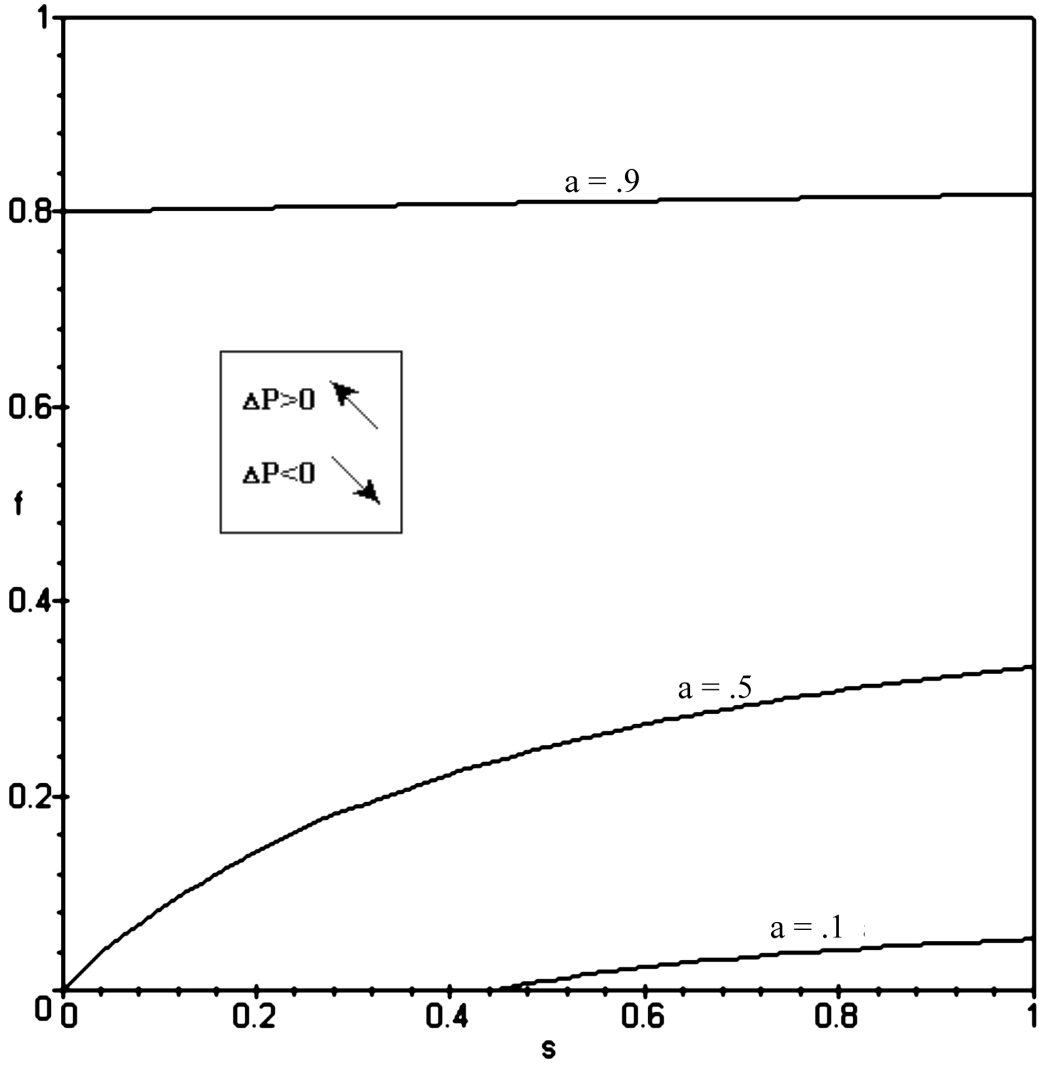

Figure 5.6