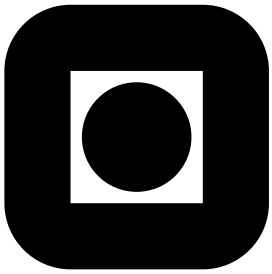
NORGES TEKNISK-NATURVITENSKAPELIGE
UNIVERSITET

Information Gathering in Bayesian Networks

by

Marie Lilleborge and Ragnar Hauge and Jo Eidsvik

PREPRINT
STATISTICS NO. 1/2015



NORWEGIAN UNIVERSITY OF SCIENCE AND
TECHNOLOGY
TRONDHEIM, NORWAY

This report has URL
http://www.math.ntnu.no/preprint/statistics/2015/S1-2015.pdf
Jo Eidsvik has homepage: http://www.math.ntnu.no/∼joeid
E-mail: joeid@stat.ntnu.no
Address: Department of Mathematical Sciences, Norwegian University of Science and Technology, N-7491 Trondheim, Norway.

Information Gathering in Bayesian Networks

Marie Lilleborge[1], Ragnar Hauge[1] and Jo Eidsvik[2]
1) Norwegian Computing Center, Norway.
2) Department of Mathematical Sciences, NTNU, Norway.

Abstract

The optimal design of data acquisition is not obvious in Bayesian Network models. The dependency structure may vary dramatically, which makes learning and information evaluation complicated and sometimes non-intuitive. Our application, and the motivation for working on this topic, is prospect selection for petroleum exploration in the North Sea. Here, the data gathering is often carried out during seasonal campaigns, and it is useful to plan the experimentation and to understand which data are likely to be most informative. We use information measures to compare possible future observation sets.

Four information measures are studied: Shannon Entropy, sum of Variances, Node-wise Entropy and overall Prediction Error. The Shannon Entropy is commonly considered the standard measure of information, and the Node-wise Entropy measure can be interpreted as an approximation to the former. The Variance measure links uncertainty and variance. The Prediction Error measure is tied to decision making rules.

The results lead to new insight about prospect selection. For example, the Node-wise Entropy and the Variance measure behaves similarly, and the optimal observation set of Shannon Entropy does not correspond to what we intuitively would consider as minimizing unknown information in this case.

1 Introduction

A collection of petroleum prospects and their probabilistic dependencies can be modelled as a bayesian network (BN), see e.g. Wees et al. (2008) and Martinelli et al. (2011). In this paper we define and discuss various information measures for BNs. These should be useful for evaluating petroleum exploration strategies. The BN models are among the key inventions from

statistics the last 25 years. They are convenient for modeling complex dependencies between random variables, and allow the construction of intuitive and modular probability statements at the local level. In principle, these models can also account for any correlation structure within the variables. This leaves a wealth of modeling opportunities, but this flexibility often makes the interpretation of data conditioning and the evaluation of information gathering harder than for a simpler model. BNs are used a lot in various applications, see e.g. Jensen and Nielsen (2007) for an overview, or Heavlin (2003) and Mortera et al. (2013). Despite a large interest in such models, there has not been much work on designing experiments for BNs.

Our goal with this paper is to evaluate and compare various information gathering schemes for BNs in the context of prospect selection in the North Sea. We assume the joint probability structure to be known, and study how information at selected nodes influence the probability structures at the non-selected nodes. Typical questions include; where should we drill exploration wells? What is a natural information measure to use for BNs? We study a BN with 25 prospects, and aim to design a strategy for selecting the best subsets of prospects for information gathering. This setting is relevant for a petroleum company which plans for seasonal drilling campaigns. Via this application, we attempt to develop new approaches for data gathering schemes for BNs.

Ginebra (2007) studies how to measure information in a statistical experimental design setting and discusses what is a valid measure of information in an experiment. Not aiming for the same level of generality as in Ginebra (2007), we focus on a special type of information gathering in a BN. While one in the general experimental setting strives to learn the unknown index θ of the possible probability distributions $\{P_\theta\}$ driving the experiment, we try to reduce the combined uncertainty in a collection $\{X_i\}$ of dependent random variables. We will study four information measures in this paper, each of which can be related to some examples and the general theory of Ginebra (2007).

The expected reduction in Shannon Entropy was introduced as a measure of information by Lindley (1956), and examples of applications can be found in Ko et al. (1995) and Bueso et al. (1998). Shewry and Wynn (1987), Royle (2002) and Le and Zidek (2006) successfully apply the Shannon Entropy criterion to spatial models. Although a BN is a convenient tool to model Gaussian variables, the dependency structure in a general BN is often not as homogeneous as in e.g. Gaussian Random Field models. This calls

for studying different information measures. Based on the results for the prospect selection case presented in this paper, we will advocate the use of a Variance criteria, a Node-wise Entropy criteria, or a Prediction Error criteria for BN models of similar application.

In some situations the variables of interest $\{X_i\}$ are tied to decisions. If costs and revenues for the underlying decision problem are well known, the value of information approach is perhaps the most natural information measure. It represents how much a risk-neutral person should be willing to pay for a given observation. The resulting values can then be compared in order to figure out which observations are optimal on average. Krause and Guestrin (2009) and Bhattacharjya et al. (2010) provide examples of value of information analysis. Martinelli et al. (2011) perform value of information analysis for the BN we study here. When there is ambiguity in the underlying decision problem, we need to compare possible observations without reference to any monetary values, and our suggested measures could be useful here. For instance, value of information analysis for the oil exploration case requires monetary values for the future price for oil.

In Sect. 2 we present the basic notation used to describe BNs and information gathering schemes in our context of prospect selection. We define four information measures and discuss their properties in Sect. 3. Section 4 demonstrates properties of our information measures on illustrative examples. In Sect. 5, we apply the information measures on the case study with North Sea petroleum prospects. In Sect. 6, we summarize our findings and provide guidelines for the choice of information measure. Section 7 points to future work.

2 Background and notation

Figure 1 shows the BN with 42 nodes from Martinelli et al. (2011). The black circles represent petroleum prospects, where we could choose to collect data. Because there is dependence in the network, information gathered at one prospect will propagate to the other prospects. Martinelli et al. (2011) illustrate probability updating in this BN to see the effect of an observation. A question is where to collect data? Another is which criteria should the selection of prospects be based on? In this section we present the notation required to study various information gathering schemes for BNs.

For an introduction to BNs, see e.g. Lauritzen and Spiegelhalter (1988);

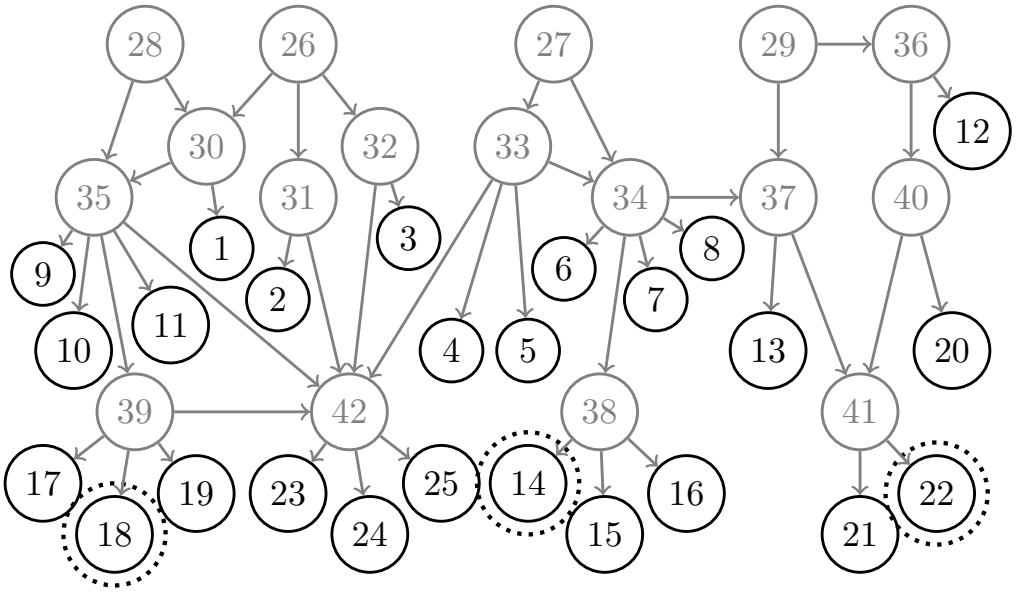


Figure 1: Illustration of the Bayesian Network model with 42 nodes. The black nodes, numbered from 1 to 25, represent petroleum prospects in the North Sea where it is possible to collect information. The gray nodes have a geological interpretation, but are not directly observable. We also marked a possible size 3 observation set $\{14, 18, 22\}$ by dotted circles. The BN was originally presented in Martinelli et al. (2011).

Jensen and Nielsen (2007); Cowell et al. (2007) or Koller and Friedman (2009). Assume a collection of n binary random variables X_i, $i \in V = \{1, 2, \cdots, n\}$. The set V is the vertex set of the network, and its elements are called nodes. The conditional dependency structure among the random variables is described by edges. Figure 1 visualizes each node by a circle, and edges $e = (i, j) \in E$ are shown as arrows from a node i to another node j. The edge set E is a collection of ordered pairs of elements in V, and the pair (V, E) is a directed graph. We say that node i is a parent of node j for each edge $e = (i, j) \in E$, and denote the set of parents of node j by $\mathrm{Pa}(j)$. For a collection of indexes $A = \{i_1 < i_2 < \cdots < i_m\} \subseteq V$, we let $X_A = (X_{i_1}, \cdots, X_{i_m})$. Then, X_V is a vector with all random variables as entries, and $X_{\mathrm{Pa}(i)}$ is a vector of the random variables corresponding to the parents of a node i.

Following Russell and Norvig (2003), we define a BN as a directed acyclic graph (DAG), that is, a directed graph where the edge set does not contain any directed cycle. In addition, each random variable X_i has a local probability distribution $P\left(X_i = x_i | X_{\mathrm{Pa}(i)} = x_{\mathrm{Pa}(i)}\right)$ associated with it. Note that from here on, we will only include the assignment to a random variable in our notation in the cases we believe it is required to clarify the mathematical understanding. The joint probability distribution for the BN is given by

$$\mathbb{P}\left(X_1, \cdots, X_n\right) = \prod_{i=1}^{n} P\left(X_i | X_{\mathrm{Pa}(i)}\right),$$

where the parent set is empty for the top nodes (roots) of the network. The joint model is in this way fully specified by the edge structure and conditional probability statements. These are easy to visualize, as in Fig. 1. BNs have shown to be useful models for capturing complex dependency. They further enjoy the efficiency of established inference algorithms, see e.g. Lauritzen and Spiegelhalter (1988) and Cowell et al. (2007).

We assume a known joint probability distribution, but uncertain outcomes X_V of the random variables. The goal is to collect data at a subset of the nodes to learn as much as possible about the variables of interest. In this way we attempt to contribute to the design of experiments for networks. Let $L \subseteq V$ denote the set of observable nodes. We want to select an observation set $B \subset L$ in order to gain as much information as possible about all realizations in L. We are not interested in the latent variables in $V \setminus L$. Therefore we only (directly) value information about the realizations in L,

which means that X_L also plays the role as our scoring variables. In Fig. 1, the L set are the black nodes $1-25$, and the latent variables $V \setminus L$ are nodes numbered $26-42$, marked in gray. The latent nodes are important to model the geological mechanisms, but it is not possible to observe any of these node variables. We study measures of information associated with such an observation set B. Note that in this paper, L plays both the role of our observable set and the set in which we want to minimize the unknown information.

It is natural to require that for any $A \subset B$, the information gained when observing X_A is less than or equal to the information we would gain by observing X_B. Thus, we limit scope to the optimal observation set $B_m \subset L$ of size m. We define

$$\mathcal{B}_k = \{B \subseteq L \ : \ |B| = k\}, \qquad k = 1, \ldots, |L|,$$

so the candidates for B_m are exactly the elements of $\mathcal{B}_m$. Note that the number of candidates for B_m is $\binom{|L|}{m}$, which is of order $|L|^m$ when $m << |L|$. We focus on small levels m in this paper; $m = 1, \ldots, 6$ for the prospect selection case in Sect. 5. Figure 1 shows an observation set of size $m = 3$ in dotted circles.

We will evaluate information measures before the data is acquired. To achieve this task, one needs to take expected values over the observation set, and conditional expectations over the nodes of interest. The expected value of a function $f(X_A)$ over some set of binary random variables X_A is given by

$$\mathbb{E}_{[X_A]} f(X_A) = \sum_{X_A = x_A \in \{0,1\}^{|A|}} f(X_A) \mathbb{P}\left(X_A\right).$$

Similarly, we define the conditional expectation by

$$\mathbb{E}_{[X_A | X_{L \setminus A}]} f(X_L) = \sum_{X_A = x_A \in \{0,1\}^{|A|}} f(X_L) \mathbb{P}\left(X_A | X_{L \setminus A}\right),$$

where some conditional assignment $X_{L \setminus A} = x_{L \setminus A}$ is implicit. The evaluation of expected values and conditional expectations require marginalization and conditioning in the joint distribution defined by the BN. This must be done many times when computing the information measures, and it is crucial to use fast routines, such as the Junction Tree Algorithm of Lauritzen and Spiegelhalter (1988). The function $f(\cdot)$ will differ between the various information measures presented in the next section.

3 Measures of Information

Measures of information can be divided into two distinct types; one type is based purely on a reduction of uncertainty within the probability distribution, and the other on the monetary value of information. The first type allows comparison of experimental designs without reference to any specific decision problem with associated costs or revenues, since the information measure is a function of the probability distribution alone. The second type enjoys decision theoretic advantages as it is tied to the monetary values of the underlying decision problem. In this paper, we study the first type of measure, hence only assuming that the probability distribution is known. Our focus will be on our information measures being able to evaluate observation sets in a BN with a complex dependency structure, e.g. as in the dependent prospect situation of Fig. 1.

3.1 Definitions

The interpretation of the Shannon Entropy measure is tied to the log likelihood, where we now take the expected value over the non-observed random variables. The expected remaining Shannon Entropy is as follows:

Definition 1. *Shannon's Entropy measure*

$$\mu_{ShE}(B) = -\mathbb{E}_{[X_B]}\left[\mathbb{E}_{[X_{L\setminus B}|X_B]}\left[\log \mathbb{P}\left(X_{L\setminus B}|X_B\right)\right]\right].$$

The entropy is larger when we are more uncertain about the outcomes of the random variables.

We are going to compare this Shannon Entropy measure to three other candidates; a Prediction Error measure, a Node-wise Entropy measure and a Variance measure, defined in the following.

Definition 2. *Expected number of prediction errors*

$$\mu_{PrE}(B) = \sum_{i\in L} \mathbb{E}_{[X_B]}\left[1 - \max_{x\in\{0,1\}}\left\{\mathbb{P}\left(X_i = x|X_B\right)\right\}\right].$$

The motivation for the Prediction Error measure is that data X_B should on average improve our ability to classify each X_i correctly. Thus, reducing the uncertainty about the random vector X_L is connected to getting as few prediction errors as possible. We will see that this approach is equal to the value of information approach when we assume equal costs and revenues for all nodes.

For statisticians, it is natural to interpret the uncertainty of a variable as its marginal variance, and here we look at the sum of conditional variances over the random variables in L, averaged over the possible observations for the B set:

Definition 3. *Variance measure*

$$\mu_{Var}(B) = \sum_{i \in L} \mathbb{E}_{[X_B]} \left[\mathbb{V}ar_{[X_i|X_B]} \left[X_i \right] \right].$$

The last measure is inspired by the Shannon Entropy measure, as a version with reduced time complexity for calculations. It is a node-wise sum over terms corresponding to the remaining Shannon Entropy of single nodes.

Definition 4. *Node-wise Entropy measure*

$$\mu_{NwE}(B) = -\sum_{i \in L} \mathbb{E}_{[X_B]} \left[\mathbb{E}_{[X_i|X_B]} \left[\log \mathbb{P} \left(X_i \mid X_B \right) \right] \right].$$

We will see that the Node-wise Entropy measure is more related to the Variance measure than to the Shannon Entropy measure.

3.2 Properties of our measures

Complex BN models can incorporate dependency structures of a less uniform type than most other well-known models. Thus, probability updates would not be as intuitive as in for instance a spatial Gaussian model where correlation between a pair of variables just depends on the physical distance between them. Therefore, our information measures should be able to see a wide range of dependency structures. We want the optimal observation set to tell as much as possible, not only about the observed variables, but also about the ones left unobserved. In our North Sea prospect application, this

will be very important, since we consider to observe just 6 (or fewer) of the 25 prospects.

The four measures in this paper can be related to Ginebra (2007). The focus in Ginebra (2007) is on learning the unknown index θ of the distribution driving an experiment, while we try to learn about the realization in a collection of correlated random variables. However, if we compare θ to a collection containing only one Random Variable X_i, we observe a direct similarity to Table 2 p.32 in Ginebra (2007); between our Variance measure and Ginebra's example 2, between our Prediction Error measure and Ginebra's example 3 and also between our Node-wise Entropy measure and Ginebra's example 4. When we deal with the combined uncertainty of several random variables, each of these three measures evolve into a sum of terms with the corresponding similarity. The Shannon Entropy measure arises from Ginebra's Example 4 when θ is viewed as the vector X_L.

Three of our measure definitions are of similar form; a sum over the node set of interest and an outer expectation over the observation set B. For a random variable $X_i \in \{0, 1\}$, the inner terms in these three measures can be described by the following concave functions

$$\begin{aligned} f_{PrE} &: [0,1] \to \mathbb{R}, \qquad f_{PrE}(p) = \min\{p, 1-p\}, \\ f_{Var} &: [0,1] \to \mathbb{R}, \qquad f_{Var}(p) = p \cdot (1-p), \\ f_{NwE} &: [0,1] \to \mathbb{R}, \qquad f_{NwE}(p) = -p \cdot \log(p) - (1-p) \cdot \log(1-p). \end{aligned}$$

We can then write

$$\mu_T(B) = \sum_{i \in L} \mathbb{E}_{[X_B]} \left[f_T \left(\mathbb{P} \left(X_i = 1 | X_B \right) \right) \right], \tag{1}$$

for each subscript $T \in \{PrE, Var, NwE\}$. This formulation illustrates the similarities between the information measures $\mu_{PrE}(\cdot)$, $\mu_{Var}(\cdot)$, $\mu_{NwE}(\cdot)$, and will be used to prove the following results.

Theorem 1. *Given $A \subset B \subseteq L$, the information measures $\mu_{ShE}(\cdot)$, $\mu_{PrE}(\cdot)$, $\mu_{Var}(\cdot)$ and $\mu_{NwE}(\cdot)$ are all non-increasing as the input set is increased from A to $B \supset A$.*

- *The decrease in any of our four measures is positive unless $X_{B \setminus A}$ is deterministically given by X_A.*

- *For each subscript $T \in \{PrE, Var, NwE\}$, the information measure μ_T consists of terms $\mu_T^i \equiv \mathbb{E}_{[X_B]}\left[f_T\left(\mathbb{P}\left(X_i = 1 | X_B\right)\right)\right]$ which are separately non-increasing as the observation set is increased from A to B. More specifically:*

  1. *For each subscript $T \in \{Var, NwE\}$, each term μ_T^i has a zero-valued decrease if and only if $X_i \perp X_{B\setminus A} | X_A$.*
  2. *The decrease in μ_{PrE}^i is non-zero if and only if the outcome of $X_{B\setminus A}$ can change the prediction of X_i when X_A is already known.*

The proofs and closed form solutions connected with these results are provided in the Appendix.

For any $i \in A \subset L$ and $T \in \{PrE, Var, NwE\}$, we have $\mu_T^i(A) = 0$. When comparing $\mu_T(B)$ to $\mu_T(A)$ for some $B \supseteq A$, we see $|B \setminus A|$ more terms $\mu_T^i(B)$ which evaluate to zero for the same reason. This will be referred to as the *self-effect* of the additional nodes in $B \setminus A$, and we define the self-effect of these additional nodes to have value $\sum_{i \in B\setminus A} \mu_T^i(A)$. We split the measure value reduction $\mu_T(A) - \mu_T(B)$ into the self-effect of the additional nodes in $B \setminus A$, and their effect $\sum_{i \in L\setminus B} \left(\mu_T^i(A) - \mu_T^i(B)\right)$ on the unobserved nodes in $L \setminus B$ through correlations. Unless $X_{B\setminus A}$ is deterministically given by X_A, we have a positive self-effect of increasing the observation set from A to B.

In Fig. 2, we see the shapes of the base functions $f_{PrE}(p), f_{NwE}(p)$ and $f_{Var}(p)$. We observe that $f_{NwE}(p)$ and $f_{Var}(p)$ have very similar shapes, and that both are strictly concave, while f_{PrE} is concave. The concavity of the functions $f_{PrE}(p), f_{NwE}(p)$ and $f_{Var}(p)$ is what ensures non-negative effects for the unobserved nodes in $L \setminus B$ through correlations with the additional nodes in $B \setminus A$. For each unobserved X_i which is dependent on $X_{B\setminus A}$ (conditional on X_A), we have a positive $\mu_T^i(A) - \mu_T^i(B)$ for $T \in \{Var, NwE\}$, while $\mu_{PrE}^i(\cdot)$ additionally requires the information from $X_{B\setminus A}$ to have potential to switch the maximal probability state for X_i.

The similar shapes of $f_{NwE}(p)$ and $f_{Var}(p)$, visualized in Fig. 2, implies that an ordering of candidate prospect nodes according to μ_{NwE} would be very similar to the prospect node ordering according to μ_{Var}. In fact, Taylor

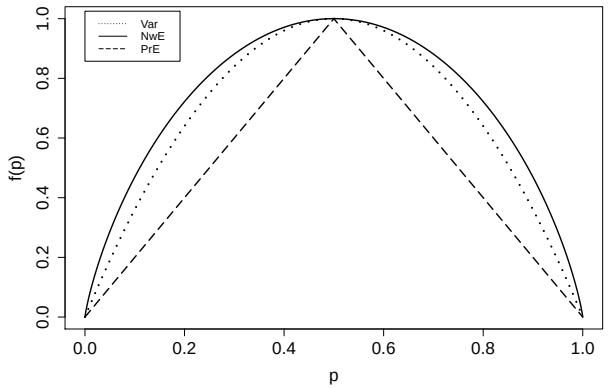


Figure 2: Base functions f_{PrE} (dashed), f_{NwE} (solid) and f_{Var} (dotted) for the measures μ_{PrE}, μ_{NwE} and μ_{Var}, each scaled to have maximum value equal to 1.

series expansions centred at $x = 0$ and $x = 1$ for the function $\log(x)$ give

$$f_{NwE}(p) = 4\log(2)\left(f_{Var}(p) + \frac{p(1-p)}{\log(2)}\sum_{j=1}^{\infty} s_j \left(2p-1\right)^{2j}\right),$$
$$s_j \equiv \sum_{n=2j}^{\infty} \binom{n}{2j} \frac{1}{(n+1)\cdot 2^{n+1}},$$

where each s_j is a constant defined by a convergent series (ratio test). This means that $f_{NwE}(p)$ in the interior of its domain can be written as $4\log(2)f_{Var}(p)$ plus an infinite polynomial which is zero-valued for $p \in \{0, \frac{1}{2}, 1\}$. Because of the similarity to the Variance measure, we will not list the results for the Node-wise Entropy measure in our examples below.

The formulation in Eq. (1) gives the information measures μ_{PrE}, μ_{Var} and μ_{NwE} common properties in the way they evaluate correlation structures in the network, and in terms of complexity for computations. Assuming, calculations of $f_T\left(\mathbb{P}\left(X_i = 1|X_B\right)\right)$ has approximately constant time-complexity given an assignment to X_B, we see that calculating $\mu_T(B)$ consists of adding $L \setminus B$ terms which again consists of $2^{|B|}$ terms of constant complexity. That yields a complexity of $\mathcal{O}\left(|L \setminus B| \cdot 2^{|B|}\right)$ for calculating $\mu_T(B)$ when $T \in \{PrE, Var, NwE\}$. However, if we want to find the optimal $B_m \in \mathcal{B}_m$,

we have to compare the values of $\binom{|L|}{m}$ candidates, and we end up with a total time complexity of $\mathcal{O}\left(\binom{|L|}{m+1} \cdot m \cdot 2^m\right)$.

The time-complexity for calculating the value $\mu_{ShE}(B)$ for an observation set B is exponential in the size of the full set L of observable nodes. When comparing different candidates B, this complexity can be reduced. The measure is inspired by the Shannon Entropy H. The Shannon Entropy can be written as a telescoping sum over the nodes in B,

$$H(X_B) = -\mathbb{E}_{[X_B]}\left[\log \mathbb{P}\left(X_B\right)\right] = -\sum_{i \in B} \mathbb{E}_{[X_{\{j \in B:\ j \leq i\}}]}\left[\log \mathbb{P}\left(X_i | X_{\{j \in B:\ j < i\}}\right)\right].$$

From the telescope sum formula we have

$$\mu_{ShE}(B) = H(X_L) - H(X_B),$$

which means that comparing $\mu_{ShE}(B)$ for different observation sets B, effectively is the same as comparing only on $H(X_B)$, since $H(X_L)$ is constant. That is,

$$\arg\min_{B \in \mathcal{B}_m} \left\{\mu_{ShE}(B)\right\} = \arg\max_{B \in \mathcal{B}_m} \left\{H(X_B)\right\}. \tag{2}$$

This means that the Shannon Entropy measure performs its evaluation based on the probabilistic properties within the marginal distribution for the B-set. We use the dependence structure between X_B and the unobserved $X_{L \setminus B}$ to calculate the distribution of X_B as a marginal of the distribution of X_V. However, when evaluating the observation set B, the Shannon Entropy measure is indifferent on whether single node-probabilities and correlations within B are induced by scoring variables outside of B or not. In this sense, the dependence structure between X_B and the unobserved $X_{L \setminus B}$ is only implicitly taken into account. Further, this means the Shannon Entropy measure does not give credit for probability updates for the unobserved variables. The other measures explicitly incorporate the effect of the observation set on each scoring variable.

Computing $H(X_B)$ has time complexity $\mathcal{O}\left(2^{|B|}\right)$. Since we repeat this for all $B \in \mathcal{B}_m$, we end up with a total time complexity of $\mathcal{O}\left(\binom{|L|}{m} \cdot 2^m\right)$ for finding the optimal B_m. We have seen that when $|L|$ and m are large, finding the optimal observation set B_m by comparing all candidates in $\mathcal{B}_m$ is computationally infeasible, for all of our measures. However, in this paper, we have m small enough to be able to compare all $\binom{|L|}{m}$ candidates for each measure.

We now explore the relation between value of information and our Prediction Error measure μ_{PrE}. We focus on the random variable $X_i \in \{0,1\}$ and want to figure out which action $a_i \in \mathcal{A} = \{0,1\}$ to perform on node i, e.g. we compare two possible actions. Any utility function on X_i and a_i can be written as

$$u_i \ : \ \{0,1\} \times \{0,1\} \to \mathbb{R}, \qquad u_i(x,\alpha) = \eta_i + \beta_i x - \gamma_i \alpha + \delta_i x \alpha,$$

where $u(X_i, a_i)$ is the utility realized after performing action a_i and the realization in node i was X_i. Here $a_i = 1$ is the most expensive action, which could be interpreted as the possibility to pay an additional $\gamma_i > 0$ before the experiment in order to increase your income by $\delta_i > \gamma_i$ whenever $X_i = 1$. For the prospect selection case, action a_i would be related to reservoir development. The constants $\eta_i, \beta_i, \gamma_i$ and δ_i connects to revenues and costs, and may be hard to assign in general. The prior value of node i is the expected utility for the optimal action.

$$\mathrm{PV}_i = \max_{a_i \in \{0,1\}} \left\{ \mathbb{E}_{[X_i]} u(X_i, a_i) \right\} = \eta_i + \beta_i \mathbb{P}\left(X_i = 1\right) + \max\left\{ \delta_i \mathbb{P}\left(X_i = 1\right) - \gamma_i, 0 \right\},$$

where the last term is included if and only if it pays off apriori to set $a_i = 1$. The terms with α_i and β_i are not influenced by our action.

Correspondingly, after observing the outcome in some observation set B, node i has posterior value

$$\begin{aligned} \mathrm{PoV}_i = \mathrm{PV}_i + \mathbb{E}_{[X_B]} \left[\max\left\{ \delta_i \mathbb{P}\left(X_i = 1 | X_B\right) - \gamma_i, 0 \right\} \right] \\ - \max\left\{ \delta_i \mathbb{P}\left(X_i = 1\right) - \gamma_i, 0 \right\}. \end{aligned}$$

The total value of information of observing X_B is defined by the sum of the difference between posterior value and prior value over all scoring nodes. We get

$$\begin{aligned} \mathrm{VoI}(B) = \sum_{i \in L} \Bigg(& \mathbb{E}_{[X_B]} \left[\max\left\{ \delta_i \mathbb{P}\left(X_i = 1 | X_B\right) - \gamma_i, 0 \right\} \right] \\ & - \max\left\{ \delta_i \mathbb{P}\left(X_i = 1\right) - \gamma_i, 0 \right\} \Bigg). \end{aligned}$$

In fact, if we let μ_{WPrE} be the expected weighted number of prediction errors, where a false predicted success of node i is given weight γ_i and a false predicted failure is given weight $\delta_i - \gamma_i$, the above equation implies that

$$\mathrm{VoI}(B) = \mu_{WPrE}(\emptyset) - \mu_{WPrE}(B).$$

This proves that minimizing the Prediction Error measure μ_{PrE} is equivalent to maximizing the value of information whenever $\delta_i = 2\gamma_i$. This corresponds to a utility function where the decision $a_i = 1$ has a gain $\delta_i - \gamma_i$ for $X_i = 1$ which equals the penalty γ_i for $X_i = 0$.

4 Illustrative examples

This section provides three simple BNs to illustrate properties of the measures presented in the previous section. The BNs are made to prepare the analysis of the North Sea prospect case in Sect. 5. There are similarities in network structure, and we build understanding to better interpret the effects we observe in our main application in Sect. 5. Section 4.1 shows the main difference between the Shannon Entropy measure and the others. Section 4.2 illustrates how the information in a node is evaluated differently depending on the other variables in the observation set. In Sect. 4.3, we study how the optimal single node observation in a success propagating chain changes with the success probability parameter.

4.1 The blind spot of the Shannon Entropy measure

Assume $N \geq 3$ variables in a BN where nodes 1 and 2 are roots and all other nodes has node 2 as a single parent, like in Fig. 3. Here,

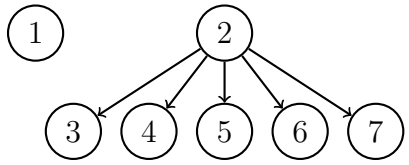


Figure 3: A BN with two uncorrelated parts.

$$\mathbb{P}(X_1 = 1) = q, \quad \mathbb{P}(X_2 = 1) = p, \quad \mathbb{P}(X_i = 1|X_2) = \alpha_i X_2, \ i \geq 3,$$

with

$$p > \frac{1}{2} \quad \text{and} \quad \frac{1-p}{p} \leq \alpha_N \leq \cdots \leq \alpha_3 \leq \frac{1}{2p},$$

so that X_3 has marginal success probability closest to $1/2$, possibly except X_1.

When the observation set B consists of a single node, the marginal success probability $\mathbb{P}(X_i = 1)$ is the only probabilistic property within the observation set. Thus, Eq. (2) in Sect. 3.2 simplifies to

$$\arg\min_i \mu_{ShE}(\{X_i\}) = \arg\min_i \left\{ \left| \mathbb{P}(X_i = 1) - \frac{1}{2} \right| \right\},$$

and we chose to observe the node with marginal closest to 1/2.

When $q = p\alpha_3$, the Shannon Entropy measure is indifferent between observing node 1 and node 3, even though when observing node 3, we simultaneously learn about nodes 2 and $4, \cdots, n$. The expected remaining Shannon Entropy is smallest when observing the node with marginal probability closest to 0.5. That is, the Shannon Entropy does not account for the possible information propagated to dependent variables. If we further slightly increased q to $p\alpha_3 + \epsilon$, node 1 turns strictly optimal, even though its marginal uncertainty is barely larger than in node 3 and we have no learning for other nodes. Note that both the Prediction Error measure and the Variance measure rates node 1 as suboptimal for small ϵ. This example is designed to illustrate how the Shannon Entropy differs from the other measures for single node observations. The example could easily be expanded to comparing two larger observation sets: Assume their corresponding vectors are (close to) independent copies, but only one of the observation sets is correlated with other scoring variables.

4.2 The information value of a node

Assume three random variables, in a BN as in Fig. 4. Let the joint probability

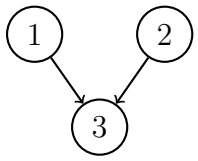


Figure 4: Two independent parents of a single child.

distribution be determined by

$$\mathbb{P}(X_1 = 1) = \mathbb{P}(X_2 = 1) = p, \qquad \text{and} \qquad X_3 = \min\{X_1, X_2\}.$$

Knowing the realization of X_1 and X_2 deterministically gives the value of X_3, so obviously, the optimal observation set of size 2 is $B_2 = \{1, 2\}$. But what is the optimal B_1? The value of X_1 alone gives node 1 as well as some indication on node 3. A similar argument holds for node 2. However, the value of X_3 gives indications on both X_1 and X_2, and node 3 will be the optimal choice as long as the success probability is high enough. This happens through the implicit effect on the marginal of the observation node in consideration for Shannon Entropy. Additionally, we see an explicit effect on the observation nodes' ability to better predict the other variables for the other measures. The thresholds for each measure can be found in Table 1. That is, if $p > p_{Var}$,

p_{ShE}	p_{Var}	p_{PrE}
$\frac{\sqrt{5}-1}{2} \approx 0.62$	$\sqrt{2} - 1 \approx 0.41$	$\frac{1}{3} \approx 0.33$

Table 1: The smallest success probability p_T that makes $B_1 = 3$ for the measure $\mu_T(\cdot)$.

the variance measure μ_{Var} chooses node 3 as the optimal single node observation. Note that $p_{ShE} > p_{Var} > p_{PrE}$, so whenever $p > 0.62$, all measures agree on $B_1 = 3$ in this example, and we have $B_1 \not\subset B_2$. This illustrates how the information from a node depends on whether it is accompanied by information from other sources. Getting the same information twice does not have double information value. Dependent variables give information about each other, since the realization in one node updates the probability distribution on all correlated nodes. Observing two correlated variables is likely to give some of the same information twice, which means that the information value for the pair is less than the sum of the information values separately in each of the two nodes. Shannon Entropy has a larger threshold for p, since it does not account for the potential for probability updates. The Prediction Error has a smaller threshold, since $p < 1/2$ results in just the self effect for observing node 1 (or 2) here, while the same does not hold for node 3.

4.3 Comparing size sets in a chain network

Assume we have four random variables in a success propagating chain as in Fig. 5. Let the joint probability distribution be determined by the single parameter p, according to

$$\mathbb{P}(X_1 = 1) = p, \qquad \text{and} \qquad \mathbb{P}(X_i = 1 | X_{i-1}) = pX_{i-1}, \quad i \geq 2.$$

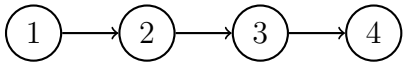


Figure 5: A four node success propagating chain

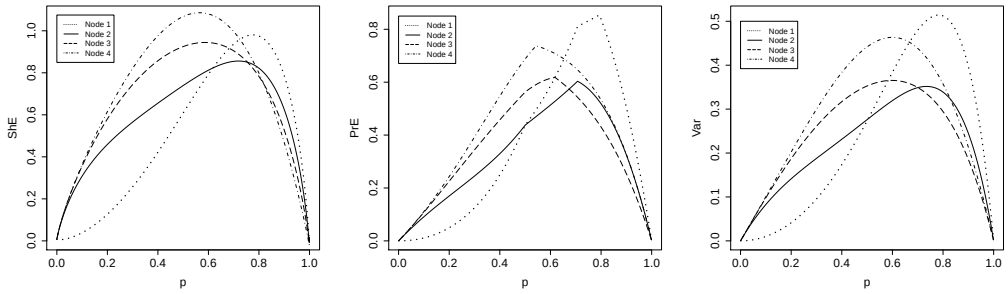


Figure 6: Evaluation of single node observation sets for the Shannon Entropy measure (left window), the Prediction Error measure (middle window) and the Variance measure (right window) as a function of the parameter p.

Observe that the probabilistic relationship between node 1 and node 2 in this example is exactly the same as between node 1 and node 3 in the previous (three node) example. Thus, we know that for a single node observation, the Shannon Entropy measure would prefer node 2 to node 1 when $p > \frac{\sqrt{5}-1}{2}$. Correspondingly, if the success (propagation) probability p is even higher, the measure μ_{ShE} could rate node 3 even higher, and for p very close to 1, the success is most likely to propagate throughout the whole chain, and node 4 would give most information about the whole network. In Fig. 6, we see how the Shannon Entropy, the Prediction Error and the Variance measures rate the information from each node as a function of p. Recall that for Shannon Entropy the optimal single node only depends on how far from 1/2 the corresponding marginal success probability for each node is.

If we set $p = 0.65$, we know that the optimal single node to observe is node 2, and we can see how the optimal observation set B_m changes as we increase the observation size m. All the information measures agree on the following sequence

$$B_1 = \{2\}, \qquad B_2 = \{1, 3\}, \qquad B_3 = \{1, 2, 4\},$$

which interestingly has $B_1 \subset B_3$, but also $B_1 \not\subset B_2 \not\subset B_3$. In applications with a large number of variables, forward search approximations are popular

to determine a sequence of candidates $\tilde{B}_m$, $m = 1, 2, \cdots$. That is, one starts with finding the true optimal $\tilde{B}_1 = B_1$ for measure $\mu_T(\cdot)$, and continues by adding one node at the time such that

$$\tilde{B}_m = \tilde{B}_{m-1} \cup \arg\min_{\{i\}} \left\{ \mu_T(\tilde{B}_{m-1} \cup \{i\}) \right\}.$$

Both in the previous three-node example (Sect. 4.2) and in this section, we have situations where a forward search would fail. For $p = 0.65$, the forward search approximation for Shannon Entropy gives

$$\tilde{B}_1 = \{2\}, \qquad \tilde{B}_2 = \{1, 2\}, \qquad \tilde{B}_3 = \{1, 2, 4\},$$

which coincides with a backward search where one starts with the full set of observable nodes and remove one at the time. In fact, since the two searches start at different end points, one could hope that their agreement would indicate that the approximation is in fact optimal. However, this four node-chain provides a counterexample. Our implementation also gives agreeing forward and backward sequences for Prediction Error, and equal to Shannon Entropy except for $\tilde{B}_2 = \{2, 4\}$. Note that this is neither the unique forward nor the unique backward sequence, since $\mu_{PrE}(\{2, 4\}) = \mu_{PrE}(\{2, 3\})$ and $\mu_{PrE}(\{1, 2, 4\}) = \mu_{PrE}(\{1, 2, 3\})$.

5 Application with 25 North Sea petroleum prospects

We now turn to the case study of 25 petroleum prospects in the Norwegian part of the North Sea. Figure 1 in Sect. 2 shows the nodes and edge structure of the network. Altogether, the network consists of 42 nodes, but only 25 of these represent actual locations where data may be acquired. The remaining 17 nodes (numbered from 26 to 42 in Fig. 1) are required to build the dependency model expected by geologists with expert knowledge about the formation of hydrocarbon (HC) in this region of the North Sea. Thus, as in many network applications, the nodes here have a clear physical meaning. The top nodes are so-called kitchens where the HC was created. Directed edges are indicative of causal mechanisms, which in our situation relate to the migration of HC over geologic time. They point from kitchens to regional prospect areas, and to the 25 prospect locations which are bottom nodes

(numbered from 1 to 25). In this BN the dependency structure for the bottom nodes of primary interest is incorporated via the directed graph.

The HC at any of the 25 prospects is assumed to be a binary {success, failure} variable. The marginal success probabilities of HC at the prospects are reported in Fig. 7. The conditional probabilities involved in the BN

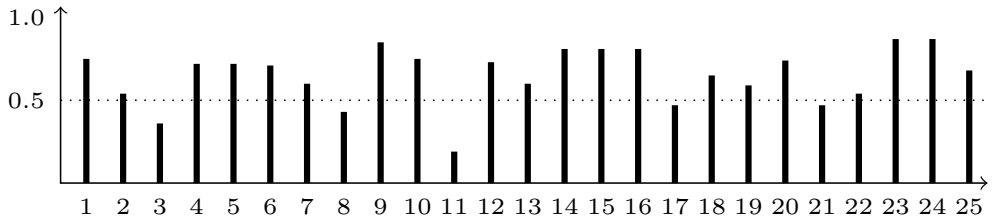


Figure 7: Marginal probabilities of success at the 25 prospect nodes.

are as indicated previously, defined from expert knowledge and a series of constraints. Some of these are similar to the ones in Sect. 4. In particular, a strict limitation is enforced on the propagation of HC : If a parent node is not a success (dry), the child node will for sure not be a success, i.e. $P(X_i = 0|X_{\mathrm{Pa}(i)} = 0) = 1$. With multiple parents, all must be dry for this to hold; otherwise HC could still flow from one of the parent nodes containing HC.

Martinelli et al. (2011) performed value of information analysis of exploration wells for this network, Martinelli et al. (2013) studied various strategies for sequential decision making, while Brown and Smith (2013) looked at clustering strategies for estimating the optimal value of sequential strategies with upper bounds. See also Martinelli and Eidsvik (2014) who looked at clustering approaches for optimal sequences based on entropy reduction or expected profit optimization. In this paper we use the same BN to demonstrate and evaluate the suggested information measures for fixed-size (non-sequential) data gathering schemes. Unlike Martinelli et al. (2011) and the other references mentioned above, we impose no cost or revenue levels to the prospects. Instead we perform the information assessments based on the probability model alone. In fact, determining the costs and revenues is not straightforward for this case, since there are several shared costs between prospects as well as large uncertainties associated with the future production costs, recovery rates and price level of petroleum resources. As in Brown and Smith (2013), we assume kitchen uncertainty, as we let the kitchens (top nodes)

be producing independently with probability 0.9. Compared with the original paper by Martinelli et al. (2011), we further slightly altered a couple of the conditional prospect probabilities (14/15/16 and 4/5) to avoid having siblings which are independent copies given their parent.

First, we look at data gathering in a single prospect. Which node is the most informative? Figure 8 shows the top ten ranked single prospects for each of the three information measures. Thus, if we can gather data in one

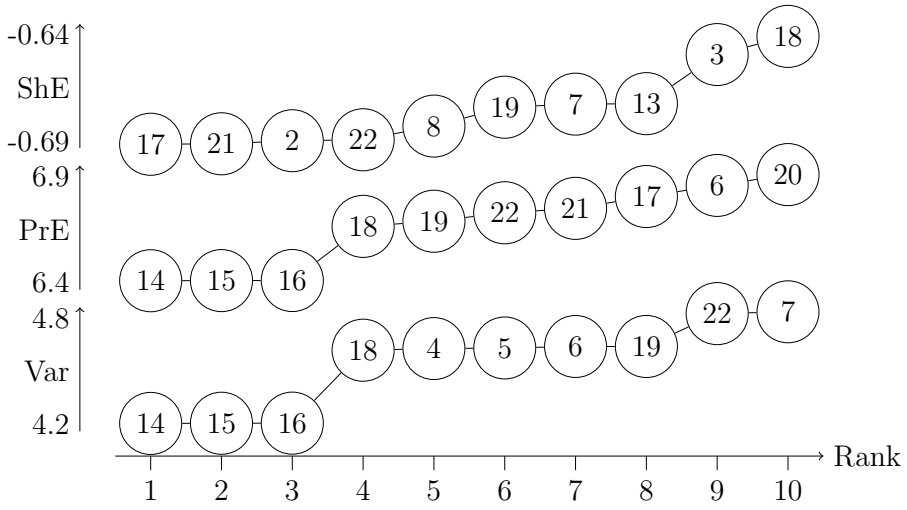


Figure 8: Ranking of the best single prospects for the three information criteria; the leftmost prospect is the optimum. The center of each node is placed at the corresponding measure value according to the axis on the left.

node, the Shannon Entropy criteria guides us to prospect number 17. The Variance criteria and the Prediction Error criteria agree on node 14.

There is a large difference in the rankings for Shannon Entropy and the other measures. For instance, prospect 14 is not even on the top-ten list for the Shannon Entropy measure. We see that Shannon Entropy gives nodes $2, 17, 21$ and 22 very similar measure values, and also very close to the theoretical maximum of $\log 2 \approx 0.6931$ for a binary variable. Figure 7 confirms that these four nodes have marginal probabilities very close to $1/2$. The Shannon Entropy criteria makes its single node choices based on these marginal probabilities alone, otherwise ignoring the correlation structure of the network. Thus, the Shannon Entropy criterion does not guide us towards any strategic parts with high correlations in the network. Both the Variance measure and

the Prediction Error measure have nodes $14, 15, 16$ as their top three choices. They are all tied to parent node 38. Roughly speaking, nodes on one side of Fig. 1 give information about nodes on the same side of the graph. Central nodes can be characterized as giving information in both directions. This possibly means that the central nodes propagate information to the whole network to a larger extent. We observe from Fig. 1 that most of the top ten single choices for the Variance criterion are prospects located in the central part of the graph. The same holds for the Prediction Error measure.

We go on to data gathering at multiple nodes. Here, we consider various sizes m of the set B_m defined above. Figure 9 shows the best subsets of various sizes for each of the three information measures. We note a strong

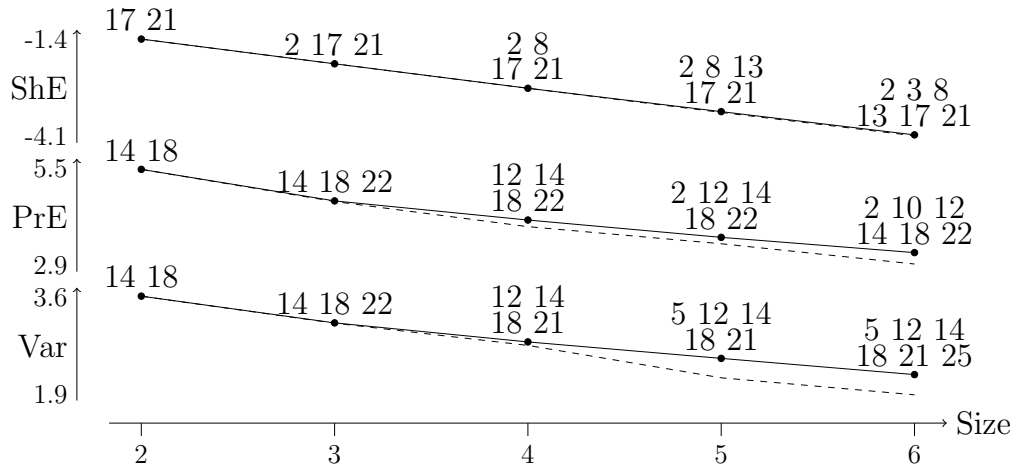


Figure 9: Optimal subsets B_m of size $m = 2, 3, 4, 5, 6$ for each of the three information criteria. The solid line shows the corresponding measure value, while the dashed line shows an independent information approximation. If the two lines are close for B_m, it means that the different observation nodes in B_m give close to independent information, i.e. the measure ensures that two observation nodes do not (partly) tell the same.

link with the ranking of single nodes in Fig. 8, since e.g. $\{17, 21\}$ is the best pair for Shannon Entropy and these are also ranked first and second in the single prospect list. However, the subset selection is not just going down this list; it also accounts for the dependence between prospects, as indicated by the inclusion of prospect 8 and prospect 13 in the size four and size five subsets for the Shannon Entropy measure. Again, this could be interpreted

by nodes on the left giving information about nodes on the left side of the graph. Note for instance how the edge from 33 to 42 connects left to center, and the edge from 34 to 37 connects center to right. For the Prediction Error and Variance criteria, the three top ranked single nodes do not follow each other in the best B_3 set, because 14, 15 and 16 are tied to the same parent node (38), and the criteria prefer to explore new parts of the network instead. The same holds for the selection of 22. Node 22 is selected before 19, which has a common parent with the included node 18.

To study the actual measure values, first note that $\mu_{Var}(\emptyset) = 4.99$. After observing node 14 independently, the sum of variances have decreased $\mu_{Var}(\emptyset) - \mu_{Var}(\{14\}) = 0.79$ units on expectation, and correspondingly for node 18 we get an expected 0.58 decrease. If 14 and 18 gave close to independent information, we could assume that they (mainly) shrink the variance in disjoint variable sets, and

$$\mu_{Var}(\{14, 18\}) \approx 4.99 - 0.79 - 0.58 = \mu_{Var}(\{14\}) + \mu_{Var}(\{18\}) - \mu_{Var}(\emptyset) = 3.62.$$

The calculations for the Prediction Error are done correspondingly, while for Shannon Entropy, we compare $H(B_i)$ to $\sum_{j \in B_i} H(\{j\})$, since these quantities are equal if and only if B_i consists of independent nodes. These numerical experiments are referred to as the individual information approximation in Fig. 9. We see a strong agreement with the true measure values which fully accounts for the dependency structure, and this illustrates how the measures actually have sought to collect information from near independent sources. Shannon Entropy has nearly perfect agreement with the true measure value, and this indicates that gaining independent information is more important for the Shannon Entropy measure than it is for the other information measures.

Recall that each Prediction Error measure term μ^i_{PrE} has $\mu^i_{PrE}(B) > \mu^i_{PrE}(B \cup \{j\})$ if and only if there exist an assignment to $X_{B \cup \{j\}}$ (of positive probability) which yields a different prediction for the node X_i compared to a situation where we only saw the corresponding X_B. In the North Sea petroleum prospect BN, we always see a positive self-effect ($i = j$), since no two prospects are fully dependent and thus $\mu^j_{PrE}(B) \neq 0$ whenever $j \notin B$. We can expect that the Prediction Error measure would pick js simply according to the self-effect $\mu^j_{PrE}(B)$ more often than the Variance measure would, since the Prediction Error measure is more restrictive on which other μ^i_{PrE} terms that experience a reduction in value. The Prediction Error has $B_m \subset B_{m+1}$ for all observation set sizes m listed in Fig. 9, so it makes sense to investigate

which new observation nodes that has potential to change the prediction of some unobserved nodes. Calculations show that when the Prediction Error criterion is obtaining B_{m+1} by adding a node j to B_m for $m = 1, 2, 3, 5$, there is a reduction in μ^i_{PrE} for two or three nodes i lying close to j in the graph. However, when going to B_5 by including node 2 in the observation set B_4, this is due to a pure self-effect, as there are no $i \in L \setminus \{2\}$ such that $\mu^i_{PrE}(B_4) \neq \mu^i_{PrE}(B_5)$.

For homogeneous spatial models, we know that it is optimal to spread out the observations. Studying the BN in Fig. 1, we could expect the measures to choose observation nodes B of different parents, and again so that the parents are from different parts of the network. A spread out observation set indicates observation nodes with little internal dependence, so that we do not get similar information from multiple sources. It also indicates that we try to learn from several parts of the network, so that the information obtained propagates to the whole network. As an example, we could expect B_6 to be something like $1, 7, 12, 18, 22, 24$ just from looking at the structure of the BN (the DAG). This observation set has some nodes close to the kitchens, some further out, and it also covers the left, the right and the central parts of the network.

The information measures evaluate the structure together with the conditional probabilities, as the probabilities determine the strength of dependence. In Fig. 10, we see how all three measures actually spread out the observation nodes in order to get information from different parts of the network. This is illustrated for $m = 6$. All three measures spread out their observations. They all select nodes in the left, center and right parts. The Prediction Error criteria is the only one which chooses no prospects connected to node 33 or 34, which bridge nodes left and right in the network. Learning about the realization in 33 or 34 helps split the network in two, and their evidence spread both to the right and left part of the BN.

Note from Fig. 9 that both Prediction Error and Variance first select a node in the center (14), then add a node to the left (18), then a node on the lower right (22) and then a node on the upper right (12). On their fifth and sixth choice their strategies separate, as the Variance criterion samples more in the central area (5 and 25), while Prediction Error goes for the pure self-effect (2) and then goes left (10). The strategies of the Variance criterion and the Prediction Error separate as they value the dependence over the bridge nodes differently. The Shannon Entropy ends up following the single node observation ranking in Fig. 8, except never adding a sibling to a node

already in the observation set. In Fig. 10, we see this as a well spread out observation set, while Fig. 7 illustrates how the marginal uncertainty in each node also plays a major part.

6 Discussion and guidelines

The motivation for this work is to evaluate each information measure's ability to see the wide range of dependency structures present in BNs such as the North Sea prospect example. For the last 25 years, BN models have made it easier to model dependency structures of a less uniform type. It is not obvious whether the widely endorsed Shannon entropy measure would be as successful as for homogeneous models. We also wanted to interpret what characteristic properties of the Shannon entropy measure meant in a prospect selection case. BNs give a wealth of opportunities in modelling, but the flexibility can make the interpretation and evaluation of data conditioning harder. We also find that this makes the choice of measure more subjective and dependent on the goal of the analysis. For the North Sea example, all measures try to get information from a set of sites with little internal correlation, in order to avoid getting similar information from multiple sources. In the oil exploration case, we care about the number of successes and getting more certain about as many outcomes as possible.

For our application, Shannon Entropy built the sequence B_1 to B_6 of optimal observation sets by including the unobserved node with marginal success probability closest to $1/2$ if this is not a sibling to a node in the smaller set. This selection strategy appears unnatural for petroleum prospect selection. There are two main reasons the Shannon entropy chooses differently from the other measures. One is that the distribution of the BN is heterogeneous, letting observations have different impact on their neighborhoods. The Shannon entropy measure does not take into account the impact on unobserved nodes, while the other measures do. Also, the other measures are explicitly by design equally interested in the outcome in all nodes, regardless of their correlation.

The difference between the Shannon entropy measure and the others is most clearly illustrated in the case where B consists of one node. Given the marginal probabilities for the number of nodes in L, we can now compute the Shannon Entropy reduction without knowing anything about the network structure. For all the possible Bayesian networks that have the same

number of nodes with the same marginal probabilities, we get the same entropy reduction, regardless of network structure. The Shannon entropy does not see the network structure, it only sees the marginal properties of B. This is illustrated in Sect. 4.1.

The balance between searching large self-effect or other probability updates is nearly equivalent for the Variance measure and the Node-wise Entropy measure. The Prediction Error measure just counts the probability update part whenever it has potential to be large enough to change the prediction of the unobserved nodes. That is, it acknowledges some and ignores the updates which are too small to shift a prediction. For the petroleum example, these three measures make similar choices for the smallest observation sizes, $m = 1, 2, 3, 4$, while at $m = 5$, the Prediction Error only sees self-effects of adding a node, and thus evaluates more like the Shannon Entropy criterion at this point. The Variance measure, or the Node-wise Entropy measure, seems to give the best balance in choosing an observation set with small internal correlation and simultaneously valuing all information obtained through probability updates.

For many applications such as spatial statistics, Shannon Entropy has been successfully applied to monitor environmental variables. In those models, the marginal probability distributions are very similar for the observable variables, which eliminates the issue of putting too much weight on the individual behavior of a variable. The learning structure is also often homogeneous, as the probability updates for neighboring variables is similar for all observable variables. That is, in most of those models, the main concern is to avoid overlapping information from correlated sources, a task the Shannon Entropy handles very well. However, for BNs with complex correlation structure, one of our contributions is that Shannon Entropy has limitations partly because it ends up not valuing probability updates outside the current observation set. Hence, it aims to select the observation set we are most uncertain about in itself.

When just seeking to remove as much uncertainty as possible from the joint distribution as a whole, not really worrying about each variable in itself, the Shannon entropy is a good choice, and has largely appreciated theoretical properties. In this paper we emphasise that in the Shannon Entropy setting, two very correlated variables combined are viewed very similarly as one of them alone. We acknowledge that this could be a practical view in other applications.

We consider applications similar to the oil exploration case presented in

Sect. 5 when we evaluate the information measures. We consider a set of observable nodes, of which we are restricted to observe a subset. After the observations, we want to minimize the combined uncertainty of all observable nodes, e.g. in order to make an optimal decision for each of the observable nodes. Oil exploration has several phases, and the step considered in this paper is an initial exploration phase where we try to learn more about the whole area covered in the BN. Subsequent phases would also include more exploration wells, as well as appraisal wells for areas where oil is found. We recommend using a Variance measure, a Node-wise Entropy measure, or a Prediction Error measure for cases where we care about each of the observable variables after the observations are made. The Prediction Error could safely be used in cases where a 0/1-loss function makes sense. This is the closest we get to when the associated costs and incomes for the decision problem is well known, and the optimal would be a Value of Information analysis.

7 Future work

This paper assumes a set $L \subseteq V$ to represent both the observable nodes and the scoring nodes. A future study of cases where these two sets differ would be interesting. In some applications, the desired effect of the observations might be to make a more informed choice of action or strategy to influence the realization of a future random variable which also is correlated with the observable variables.

In this paper, we only deal with small observation set sizes m for the prospects. This makes the comparison of all possible observation sets computationally feasible. We wrote all our code in `C++` using the Junction Tree Algorithm for fast probability updates. In order to work with larger networks and observation sets, one needs useful approximations or maybe a sophisticated stochastic search.

Appendix: Proof of Theorem 1

Proof. Let $\emptyset \subseteq A \subset B \subseteq L$.

The Shannon Entropy measure has

$$\begin{aligned}\mu_{ShE}(A) &= -\mathbb{E}_{[X_L]}\left[\log \mathbb{P}\left(X_{L\setminus A}|X_A\right)\right] \\ &= -\mathbb{E}_{[X_L]}\left[\log\left(\mathbb{P}\left(X_{L\setminus B}|X_B\right)\mathbb{P}\left(X_{B\setminus A}|X_A\right)\right)\right] \\ &= -\mathbb{E}_{[X_L]}\left[\log \mathbb{P}\left(X_{L\setminus B}|X_B\right)\right] - \mathbb{E}_{[X_L]}\left[\log \mathbb{P}\left(X_{B\setminus A}|X_A\right)\right] \\ &\geq -\mathbb{E}_{[X_L]}\left[\log \mathbb{P}\left(X_{L\setminus B}|X_B\right)\right] \\ &= \mu_{ShE}(B),\end{aligned}$$

with equality if and only if the distribution $\mathbb{P}\left(X_{B\setminus A}|X_A\right)$ is trivial for each assignment to X_A.

Observe that f_{NwE} and f_{Var} are strictly concave, since

$$f''_{NwE} = \frac{-1}{p(1-p)} \quad \text{and} \quad f''_{Var} = -2 \qquad \forall p \in \langle 0,1\rangle,$$

while f_{PrE} is concave on $[0,1]$ and linear on $[0,1/2]$ and on $[1/2,1]$. Fix an $i \in L$, and assume a measure term of the form

$$\mu^i(B) = \mathbb{E}_{[X_B]} f\left(\mathbb{P}\left(X_i = 1|X_B\right)\right),$$

for a concave function $f:\ [0,1] \to \mathbb{R}$ with $f^{-1}(0) = \{0,1\}$. For a given assignment $X_A = x_A$ to the random variables in A, $\mathbb{P}\left(X_i = 1|X_B\right)$ is a function of $X_{B\setminus A}$ and thus a Random Variable. By Jensen's inequality,

$$f(\mathbb{P}\left(X_i = 1|X_A\right)) \geq \mathbb{E}_{[X_{B\setminus A}|X_A]} f(\mathbb{P}\left(X_i = 1|X_B\right)),$$

with equality if and only if f is linear on $\left[\min_{x_{B\setminus A}}\{\mathbb{P}\left(X_i = 1|X_B\right)\}, \max_{x_{B\setminus A}}\{\mathbb{P}\left(X_i = 1|X_B\right)\}\right]$. If $i \in B$, the right hand side of the inequality is zero-valued, and we have equality if and only if $\mathbb{P}\left(X_i|X_A\right)$ is trivial as well. If $i \in L \setminus B$ and f is strictly concave, the inequality is strict unless

$$\mathbb{P}\left(X_i = 1|X_B\right) \equiv \mathbb{P}\left(X_i = 1|X_A\right).$$

Since the assignment $X_A = x_A$ was arbitrary,

$$\mathbb{E}_{[X_A]}\left[f(\mathbb{P}\left(X_i = 1|X_A\right))\right] \geq \mathbb{E}_{[X_B]}\left[f(\mathbb{P}\left(X_i = 1|X_B\right))\right],$$

and the claims follow. □

References

Bhattacharjya, D., J. Eidsvik, and T. Mukerji (2010). The Value of Information in Spatial Decision Making. *Mathematical Geosciences 42*(2), 141–163.

Brown, D. and J. Smith (2013). Optimal Sequential Exploration: Bandits, Clairvoyants, and Wildcats. *Operations Research 61*(3), 644–665.

Bueso, M., J. Angulo, and F. Alonso (1998). A State-Space Model approach to Optimum Spatial Sampling Design based on Entropy. *Environmental and Ecological Statistics 5*, 29–44.

Cowell, R., P. Dawid, S. Lauritzen, and D. Spiegelhalter (2007). *Probabilistic Networks and Expert Systems: Exact Computational Methods for Bayesian Networks*. Statistics for Engineering and Information Science Series. Springer.

Ginebra, J. (2007). On the Measure of the Information in a Statistical Experiment. *Bayesian Analysis 2*(1), 167–212.

Heavlin, W. D. (2003, May). Designing experiments for causal networks. *j-TECHNOMETRICS 45*(2), 115–129.

Jensen, F. V. and T. D. Nielsen (2007). *Bayesian Networks and Decision Graphs* (2nd ed.). Springer Publishing Company, Incorporated.

Ko, C. W., J. Lee, and M. Queyranne (1995, July). An Exact Algorithm for Maximum Entropy Sampling. *Operations Research 43*(4), 684–691.

Koller, D. and N. Friedman (2009). *Probabilistic Graphical Models: Principles and Techniques*. MIT Press.

Krause, A. and C. Guestrin (2009). Optimal Value of Information in Graphical Models. *Journal of Artificial Intelligence Research 35*, 557–591.

Lauritzen, S. L. and D. J. Spiegelhalter (1988). Local Computation with Probabilities on Graphical Structures and their Application to Expert Systems (with discussion). *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 50*(2), 157–224.

Le, N. and J. Zidek (2006). *Statistical Analysis of Environmental Space-Time Processes*. Springer Series in Statistics. Springer.

Lindley, D. V. (1956). On a Measure of the Information provided by an Experiment. *Annals of Mathematical Statistics 27*(4), 986–1005.

Martinelli, G. and J. Eidsvik (2014). Dynamic Exploration Designs for Graphical Models using Clustering with Applications to Petroleum Exploration . *Knowledge-Based Systems 58*, 113–126.

Martinelli, G., J. Eidsvik, and R. Hauge (2013). Dynamic Decision Making for Graphical Models applied to Oil Exploration. *European Journal of Operational Research 230*, 688–702.

Martinelli, G., J. Eidsvik, R. Hauge, and M. D. Førland (2011). Bayesian Networks for Prospect Analysis in the North Sea. *AAPG Bulletin 95*(8), 1423–1442.

Mortera, J., P. Vicard, and C. Vergari (2013). Object-Oriented Bayesian Networks for a Decision Support System for Antitrust Enforcement. *Annals of Applied Statistics 7*(2), 714–738.

Royle, J. A. (2002, February). Exchange Algorithms for constructing large Spatial Designs. *Journal of Statistical Planning and Inference 100*(2), 121–134.

Russell, S. and P. Norvig (2003). *Artificial Intelligence: A Modern Approach* (2nd ed.). Prentice Hall.

Shewry, M. C. and H. P. Wynn (1987). Maximum Entropy Sampling. *Journal of Applied Statistics 14*(2), 165–170.

Wees, J.-D. V., H. Mijnlieff, J. Lutgert, J. Breunese, C. Bos, P. Rosenkranz, and F. Neele (2008). A Bayesian belief network approach for assessing the impact of exploration prospect interdependency: An application to predict gas discoveries in the Netherlands. *AAPG Bulletin 92*, 1315–1336.

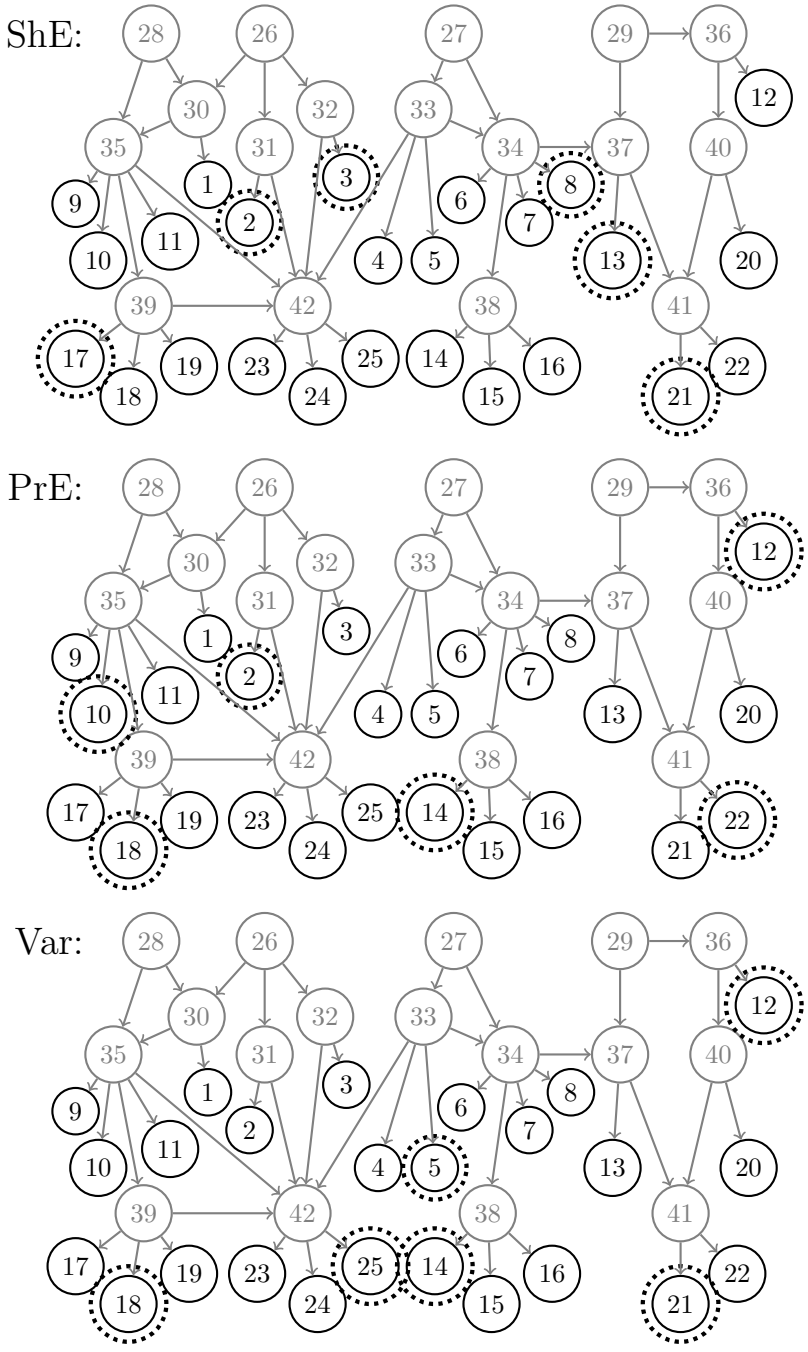


Figure 10: The optimal size 6 observation set marked with dotted circles; for the Shannon Entropy measure (top), the Variance measure (middle) and the Prediction Error measure (bottom). All three measures spread their observation set .