Synergy, suppression and immorality: forward differences of the entropy function

Joe Whittaker\textsuperscript{1}, Florian Martin\textsuperscript{2} and Yang Xiang\textsuperscript{2}

\textsuperscript{1} Mathematics and Statistics, Lancaster University, UK, joe.whittaker@lancaster.ac.uk,
\textsuperscript{2} Biological Sciences, PMI Research and Development, Neuchatel, Switzerland.

Abstract

Conditional mutual information is important in the selection and interpretation of graphical models. Its empirical version is well known as a generalised likelihood ratio test and that it may be represented as a difference in entropy. We consider the forward difference expansion of the entropy function defined on all subsets of the variables under study. The elements of this expansion are invariant to permutation of their suffices and relate higher order mutual informations to lower order ones. The third order difference is expressible as an, apparently asymmetric, difference between a marginal and a conditional mutual information. Its role in the decomposition for explained information provides a technical definition for synergy between three random variables. Positive values occur when two variables provide alternative explanations for a third; negative values, termed synergies, occur when the sum of explained information is greater than the sum of its parts. Synergies tend to be infrequent; they connect the seemingly unrelated concepts of suppressor variables in regression, on the one hand, and unshielded colliders in Bayes networks (immoralities), on the other. We give novel characterizations of these phenomena that generalise to categorical variables and to higher dimensions. We propose an algorithm for systematically computing low order differences from a given graph. Examples from small scale real-life studies indicate the potential of these techniques for empirical statistical analysis.

Keywords: Bayes network; Conditional mutual information; Imsets; Mobius inversion; Suppressor variable; Unshielded collider.

1 Introduction

The independence of two random variables is denoted by $X_1 \perp \perp X_2$, \cite{Dawid}, and the conditional independence of these two, given a third, by $X_1 \perp \perp X_2 \mid X_3$. The marginal mutual information of two random variables and the conditional mutual information of two variables given a third are, in terms of the joint probability density or mass function,

$$I_{12} = \inf(X_1 \perp \perp X_2) = E \log \frac{f_{12}}{f_1 f_2} \quad \text{and} \quad I_{12|3} = \inf(X_1 \perp \perp X_2 \mid X_3) = E \log \frac{f_{12|3}}{f_{1|3} f_{2|3}}.$$ (1)
The difference in these measures is
\[
\inf(X_1 \perp X_2) - \inf(X_1 \perp X_2 \mid X_3) = -E \log \frac{f_{123} f_{12} f_{13} f_{23}}{f_{12} f_{13} f_{23}}. \tag{2}
\]

The key property is that the right hand side is symmetric to any permutation of suffices 1, 2, 3 even though the left does not appear to be. Define \( \delta_{123} \) by the right hand side expression.

Conditional independence and mutual information lie at the foundations of graphical models, for texts see Koller and Friedman (2009); Lauritzen (1996); Whittaker (1990). The seminal citation for the separation properties of undirected graphical models is Darroch et al. (1980). Pearl (1988) made the big step in establishing acyclic directed graphs and the concept of d-separation for Bayes networks. In the class of these directed graphs certain subsets are probabilistically (Markov) equivalent, and an important theorem is that the skeleton and its unshielded colliders (or immoralities) specify the equivalence classes. The criterion for this and the generalisation to chain graphs was independently established by Verma and Pearl (1990) and Frydenberg (1990), and later developed by Andrsen et al. (1997).

Suppressor effects in multiple regression were first elucidated by Horst (1941). The phenomenon arises when the dependent variable has a smaller prediction error by including an additional explanatory variable that has no (or little) explanatory effect when used by itself; often this is manifest in enhanced regression coefficients. There is a social science literature concentrated in educational and psychological testing theory that has an interest in suppression because of its concern to design experiments that make predictions more precise. Ludlow and Klein (2014) gives a substantial review of this area. and so we just mention a few other references: McNemar (1945), Voyer (1996), Maassen and Bakker (2001), Shieh (2006). The well known structural equations text Kline (2011) cites suppression among one of the fundamental concepts of regression.

The technical literature on alternative ways to define and explain suppression includes Velicer (1978), Bertrand and Held (1988), Smith et al. (1992), MacKinnon et al. (2000), Shieh (2001). More recently Friedman and Wall (2005), give a survey and point out that the term synergism follows a suggestion of Hamilton (1988). This literature distinguishes several types of suppression. Classical suppression: \( X_2 \), say, is the suppressor variable, it is uncorrelated (or nearly so) with \( Y \), but adds to the predictive power of the regression \( Y \) on \( X_1 \) when both included. Negative suppression: both \( X_1 \), \( X_2 \) have a positive zero-order correlations with \( Y \), and correlate positively with each other, but one has a negative coefficient in the regression of \( Y \) on both. Reciprocal suppression: both variables are marginally good predictors of \( Y \), but are negatively correlated.

There are several seemingly different indicators of suppression, which are varyingly described by conditions on the correlations (marginal, multiple, partial), or in terms of regression and correlation coefficients, or in terms of explained variance, or even in terms of a rather confusing semi-partial correlation introduced by Velicer. All authors give conditions for three variable regression scenario, some attempt to generalise to \( p \)-
variables, and some explanations are geometric; most however reduce to conditions on correlation coefficients. That suppression is usually presented as a three dimensional correlation phenomenon does not make clear how to measure its strength, or how to generalise to higher dimensions; or how to generalise to other distributions.

Our contribution is to show the 3rd-order forward difference of (2) relates the seemingly unrelated topics of immorality and suppression in a natural way. The condition for suppression is that \( \delta_{123} < 0 \); noting the phrase ‘the whole regression can be greater than the sum of its parts’ in the title of Bertrand and Holder (1988) suggests that synergy is a good synonym for the triple 123. The condition for an unshielded collider (immorality) at 3 is that \( \delta_{123} < 0 \) and \( \delta_{12} = 0 \).

To set this within a wider framework we write down forward difference expansion for the entropy function, and use Mobius inversion to calculate the differences given the entropies. All forward differences are invariant to permutation of their suffices. Marginal mutual informations are second order differences and conditional measures have additive expressions in terms of the second and higher order forward differences. Higher order differences, are made more tractable by defining conditional forward differences.

We interpret the negative third order forward differences as synergies. Classic examples of graphical models in low dimensions illustrate the role of forward differences in interpretation of the model. A computing scheme for 3rd-order elements from a given graph based on cluster nodes is used to investigate empirical data for synergies.

The forward differences of the entropy provides a wider framework to explore suppression and immorality. This setting explains why the essence of both phenomena concerns exactly three variables; and why suppression is symmetric. It distinguishes suppression from both mediation and confounding where \( \delta \) is positive. It generalises the notion of suppressor variables to higher dimensions and to other distributions, for instance to categorical data. It gives an alternative characterization of immoralities (unshielded colliders).

**Plan of the paper:** In Section 2 we define the forward difference expansion of the entropy function and elaborate its properties. In Section 3 we make the connection to suppressor variables in regression and immoralities in Bayes networks, and give alternative characterizations of these phenomena. In Section 4 we consider more detailed applications to the categorical and continuous data and examples from small scale empirical studies. Proofs are collected in the Appendix. An algorithm for systematically computing low order differences from a given graph is provided in Supplementary Material.

## 2 Forward differences of the entropy

### 2.1 Preliminaries

The nodes in \( P = \{1, 2, \ldots, p\} \) correspond to random variables \( X_1, X_2, \ldots, X_p \) having a joint distribution. For subsets \( A, B, C \) of the power set \( P \) and conditional independence statements of the form \( X_A \perp \perp X_B \mid X_C \) where \( X_A \) refers to the
vector \((X_i; i \in A)\) simplify the dependency structure of \(X(\equiv X_P)\).

The entropy function, \(h\), is defined on \(\mathcal{P}\) by \(h_A = -E \log f_A(X_A)\), where \(f\) is the derivative of the joint probability measure. Without loss of generality we assume this is always well defined for, if not, we may replace it by the (negative) relative entropy \(-E \log f_A(X_A)/\prod_{i \in A} f_i(X_i)\), termed the multi-information by Studeny (2005). For a \(p\)-dimensional mass point distribution the entropy is \(h_A = -\sum_{x_A} p_A(x_A) \log p_A(x_A)\) where \(p_A\) is the mass function on the margin determined by \(A\). This is always non-negative. For a \(p\)-dimensional multivariate Normal distribution with mean zero and correlation matrix \(\Sigma\) the entropy is \(h_A = 1/2 \log \det(\Sigma_{AA})\). Any additive term, constant with respect to \(A\), may be ignored since our concern is with entropy differences. Note that \(h_\emptyset = 0\) and that the notation \(h_A\) presumes invariance to any permutation of the subscripts, justified because the underlying distribution is invariant.

In reporting numerical values of the entropy, or more usually differences in entropy, \(h\) is scaled to millibits by multiplying by the factor \(2^{10}/\log(2)\). The upper limit for the mutual information against independence for two binary variables with equi-probable margins is 1024mbits, attained when the variables always take the same value. For two Gaussian variables with correlation 0.5 the measure is 212.5mbits, but there is no upper limit.

For disjoint sets \(A, B, C \in \mathcal{P}\) the conditional mutual information is

\[
I_{AB|C} = -h_{A \cup B \cup C} + h_{A \cup C} + h_{B \cup C} - h_C.
\] (3)

It is useful to retain both the \(\inf\) and \(I\) notations for this measure. The marginal information is \(I_{ij}\) where the conditioning set is empty. We require the well known lemma that

\[
I_{AB|C} = 0 \iff X_A \perp \!\!\!\perp X_B \mid X_C.
\] (4)

The proof uses the non-negativity of the Kullback-Liebler divergence between the joint distribution and the distribution factorised according to the independence statement.

### 2.2 Entropy function expansion

The entropy function \(h\) is defined on the power set of the nodes, \(\{h_A; A \in \mathcal{P}\}\). The forward differences \(\{\delta_A; A \in \mathcal{P}\}\) of the entropy are defined by the additivity relations

\[
h_A = \sum_{B \subseteq A} \delta_B \quad \text{for} \quad A \in \mathcal{P}.
\] (5)

Solving by Mobius inversion, Rota (1964), gives

\[
\delta_A = \sum_{B \subseteq A} (-1)^{|A| - |B|} h_B \quad \text{for} \quad A \in \mathcal{P}.
\] (6)
**Theorem 2.2.1 Symmetry of the forward differences.** *The forward differences are symmetric, that is, any* \( \delta_A \) *is invariant to any permutation of the indices within the set* \( A \in \mathcal{P} \).

A detailed proof is given in the Appendix.

We need the following lemma in a later section.

**Lemma 2.1 Additivity.** *When the entropy is additive, so that* \( h_{a \cup b} = h_a + h_b \) *for all* \( a \subseteq A, b \subseteq B, \) *non-empty* \( a \) *and* \( b, \) *and disjoint* \( A, B \subseteq \mathcal{P}, \) *then* \( \delta_{a \cup b} = 0. \) *The converse also holds.*

The proof, given in the Appendix, essentially invokes a triangular elimination scheme.

### 2.3 Conditional mutual information and forward differences

From (3) the conditional mutual information of two random variables \( X_i, X_j \) given a subset of others, \( X_A, \) is

\[
I_{ij|A} = -h_{Ai} + h_Ai + h_Aj - h_A,
\]

where \( Ai \) is shorthand for \( A \cup \{i, j\}. \) The right hand side is the elementary imset representation, for pairwise conditional independence, Studeny (2005). It is the scalar product of the entropy function with the imset (integer valued multi-set) \((\ldots,-1,1,1,-1\ldots)\) of length \( 2^p \) that has zeros in the appropriate places. It is elementary because it represents a single conditional independence statement.

**Theorem 2.3.1 Conditional mutual information and forward differences.** *The conditional mutual information can be expressed in terms of the forward differences, \( \{\delta\}, \) of the entropy function by*

\[
I_{ij|A} = -\sum_{ij \subseteq B \subseteq Ai} \delta_B \quad \text{for} \quad i, j \in P, A \in \mathcal{P}.
\]

The subset \( \{i, j\} \) occurs in every term on the right of (8). The first term is the marginal mutual information \( I_{ij}. \) Each \( \delta \) term on the right is invariant to permutation of its suffices. If the conditioning set \( A \) is of moderate size then there are only a moderate number of terms in the summation.

**Corollary 2.3.1 Third order forward differences.** *When* \( A = \{k\} \) *consists of a single element*

\[
\delta_{ijk} = I_{ij} - I_{ij|k}, \quad \text{and} \quad h_{ijk} - h_{ij} - h_{ik} - h_{jk} + h_i + h_j + h_k - h_\emptyset.
\]
This follows because setting $A = \phi$ in (8) gives $I_{ij} = -\delta_{ij}$, and $A = \{k\}$ gives $I_{ij|k} = -(\delta_{ij} + \delta_{ijk})$. Subtraction gives (9). The second statement is just the inversion formula (6) for $\delta_{ijk}$.

This corollary locates the identity introduced at (2) within a wider framework. The key property is the difference $\delta$ is symmetric in permutation of suffices $i, j, k$, as in (10), while intuitively the right hand side of (9) is not.

### 2.4 Forward differences of the conditional entropy

The conditional entropy function \{h_{A|B}; A \in \mathcal{P}(P\setminus B)\} is defined on the restricted power set that excludes $B$ where $h_{A|B} = -E\log f_{A|B}(X_A|X_B)$. The corresponding conditional forward differences are defined by (11) and (12) giving \{\delta_{A|B}; A \in \mathcal{P}(P\setminus B)\}. The set notation in $\delta_{A|B}$ makes evident the symmetry of the differences.

**Theorem 2.4.1** A recursion for conditional forward differences. For $k \in P$, $B \subseteq P\setminus k$ and $A \in \mathcal{P}(P\setminus (B \cup k))$ the conditional forward differences satisfy

$$
\delta_{A|Bk} = \delta_{Ak|B} + \delta_{A|B}. \tag{11}
$$

When $B$ is empty, the identity (11) shows that the higher order forward difference is the difference between a conditional and a marginal forward difference:

$$
\delta_{Ak} = \delta_{A|k} - \delta_A.
$$

The size of the higher order term $\delta_{Ak}$ is useful in assessing how much $\delta_A$ might change by conditioning on a further variable. This is invariant to permutation of the set $Ak \equiv A \cup \{k\}$. To illustrate with $|A| = 3$, $\delta_{1234} = \delta_{123|4} - \delta_{123} = \delta_{124|3} - \delta_{124}$ and so on.

The identity (11) generalises to express a conditional forward difference as sums of conditional forward differences conditioning on a lower order:

$$
\delta_{A|B\cup C} = \sum_{D \subseteq C} \delta_{A\cup D|B}. \tag{12}
$$

**Theorem 2.4.2** Separation and the forward difference. Whenever $C$ separates $A$ and $B$ in the conditional independence graph $\delta_{A\cup B|C} = 0$.

The value of this result is that it allows easy interpretations of marginal forward differences in examples. There is a converse to this theorem if the condition on the conditional forward differences is strengthened.
2.5 Non-collapsibility of mutual information

Collapsibility is important in statistical inference because it elucidates which properties of a joint distribution can be inferred from a margin. Simpson’s paradox (Simpson (1951)) refers to a violation of collapsibility; other references are Bishop et al. (1975), Whittemore (1978), Whittaker (1990), Greenland et al. (1999), among others.

Consider three variables with \( I_{ik|j} = 0 \) and corresponding independence graph

![Graph with variables i, j, k]

with one missing edge. The strength of the relationship between \( X_i \) and \( X_j \) is measured in two dimensions by \( I_{ij} \) and in three dimensions by \( I_{ij|k} \). If \( I \) were collapsible then \( I_{ij} - I_{ij|k} = 0 \). But this difference is \(-\delta_{ij} + \delta_{ij|k} = \delta_{ijk}\) by (9), the 3rd-order difference. By symmetry \( \delta_{ijk} \) is also equal to \( \delta_{ik|j} - \delta_{ik} \) and so \( \delta_{ijk} = 0 \) together with \( \delta_{ik|j} = 0 \) would imply \( \delta_{ik} = 0 \); which is false in general. The premiss that the measures are equal is untenable.

Large values of \( \delta_{ijk} \) indicate that conditioning on \( X_k \) modifies the strength of the relationship between \( X_i \) and \( X_j \); even though it is a symmetric measure this does not imply that that subgraph be complete.

More generally requiring the collapsibility of \( \delta_A \) in the space \( A \cup B \) requires \( \delta_{A|B} = \delta_A \); by Theorem 2.4.2 this is equivalent to \( X_A \perp \perp X_B \).

3 Synergy, suppression and immorality

3.1 Synergy

The information against the independence of two variables is synonymous with the information explained in one variable by predicting from the other.

**Theorem 3.1.1 Explained information.** The explained information in one variable expressed in terms of the marginal mutual information of others is

\[
\inf(X_k \perp X_A) = \sum_{i \in A} \inf(X_k \perp X_i) - \sum_{B \subseteq A, |B| > 1} \delta_{Bk},
\]

(12)

where the last summation is over subsets \( B \) with at least 2 elements. In particular

\[
\inf(X_k \perp (X_i, X_j)) = \inf(X_k \perp X_i) + \inf(X_k \perp X_j) - \delta_{ijk}.
\]

(13)
The proof is included in the Appendix. When \( \delta_{ijk} < 0 \) the triple \( \{i, j, k\} \) is called a synergy, as the total information explained exceeds the sum of the marginal informations taken alone. It is appropriate to label the triple a synergy, rather than the variable \( k \), since \( (13) \) is invariant to permutation of the indices.

**Corollary 3.1.1** Partially explained information. The explained information in one variable expressed in terms of the marginal mutual informations of variables in \( A \) adjusted for variables in \( B \) is

\[
\inf(X_k \perp \perp X_A \mid X_B) = \sum_{i \in A} \inf(X_k \perp \perp X_i \mid X_B) - \sum_{C \subseteq A, |C| > 1} \delta_{Ck|B}. \tag{14}
\]

When there are just two variables in \( A \)

\[
\inf(X_k \perp \perp (X_i, X_j) \mid X_B) = \inf(X_k \perp \perp X_i \mid X_B) + \inf(X_k \perp \perp X_j \mid X_B) - \delta_{ijk|B}, \tag{15}
\]

the sum of the parts adjusted by the conditional 3rd-order difference.

The proof follows the previous argument and is straightforward. When \( \delta_{ijk|B} < 0 \) the triple \( \{i, j, k\} \) is also called a synergy though a conditional or partial synergy is more specific.

### 3.2 Suppression

The term suppressor variable is used in regression applications where there is an contextual asymmetry between the dependent and the explanatory variable, see the introductory section. The suppressor variable describes a third variable which is uncorrelated (or nearly so) with the dependent variable, but adds to the predictive power of the regression on both; this is technically described by \( \delta_{ijk} < 0 \) from \( (13) \) together with \( I_{ij} = 0 \). The corresponding Bayes network is displayed in Figure 1.

Figure 1: Suppression and immorality; in a suppressor regression context \( j \) is the dependent variable, \( k \) the explanatory variable, and \( i \) the suppressor variable.

The diagram makes clear that suppression is symmetric in the sense that the variables \( i \) and \( j \) are interchangeable. Elaboration of this condition in terms of correlations is the content of Theorem 1.1 in the Applications Section.

Expressing the criterion in the more general framework of information theory, extends the idea of suppression in linear regression to variables measured on other scales with well defined information measures, including categorical data. Examples are given in the next section.
Expressing suppression in the general terms of information, clarifies the issues when more than two explanatory variables are involved. Recognition of a synergy in a particular context could just reduce to calculating the conditional 3rd-order difference $\delta_{ijk|B}$. Screening for synergies or partial synergies involves repeated calculations, there are many triples as ways of choosing the two explanatory variables from the candidate set.

An alternative direction is to develop (15). For instance with three explanatory variables and variable $k = 1$ dependent the synergy criterion becomes

$$\delta_{123} + \delta_{124} + \delta_{134} + \delta_{1234} < 0.$$  

If, as well, only $\delta_{1234} < 0$ suppression is truly a function of the three explanatory variables taken as a whole.

### 3.3 Immorality

The concept of an unshielded collider, [Pearl (1988)](http://example.com), is key for understanding d-separation in Bayes networks on acyclic directed graphs. Lauritzen and Spiegelhalter (1988) refer to the same concept as an immorality. Bayes networks with different directions may be probabilistically equivalent and an important result in this area, [Frydenberg (1990)](http://example.com), [Verma and Pearl (1990)](http://example.com), is that the equivalence class is characterized by the skeleton of the graph and its unshielded colliders. An unshielded collider is displayed in Figure 1 for three variables, where the absence of an arrow joining $i$ and $j$ indicates $X_i \perp \perp X_j$. Consequently $I_{ij} = 0$, but $I_{ijk|k} > 0$ so that $\delta_{ijk} < 0$ by (9), the same condition as for suppression.

In a Bayes network with additional antecedent variables the condition for an unshielded collider requires that $k$ of Figure 1 is not in the separation set $A$ for which $I_{ijk|A} = 0$. This translates to

**Theorem 3.3.1 Characterization of an unshielded collider.** In a Bayes network the condition

$$I_{ij|A} = 0 \quad \text{and} \quad \delta_{ijk|A} \leq 0$$

where $\delta_{ijk|A}$ is the 3rd-order conditional forward difference is necessary and sufficient for an unshielded collider at $k$.

The proof is just a rephrasing of the definition of unshielded collider.

This result gives an interpretation for negative conditional 3rd-order forward differences, and suggests a method of identifying immoralities in a Bayes network.

### 3.4 Systematic computation of low order forward differences

The forward differences of the entropy offer low dimensional summaries of the data. Consider which differences to compute. For a small number of variables, all are possible, but for moderate and large numbers this is a formidable task. Furthermore the
differences required may be different for the analysis of regression and suppression, for contingency table analysis and collapsibility, and for direction determination in Bayes networks.

In the analysis of a candidate graph second order differences are routinely computed as marginal informations for all pairs. Third order differences complement the information from the nested pairs and may flag suppression and immorality, which are interesting because they are infrequent. We propose these are computed for a subset of the triples of a given graph. Fourth order difference show changes in third order differences and hence may flag conditional synergies; we suggest these are only computed in relation to specific triples of interest.

Requirements: The subset of triples are required to cover the graph without unnecessary computing. In particular previously computed differences should not be recomputed nor should redundant ones that have an priori zero value with respect to the graph. Because higher order conditional mutual informations are additive in lower order differences, see (8), it is desirable to require nested subsets, so that for example if a fourth order difference is computed its corresponding lower order differences are available.

Redundancy: For a given conditional independence graph certain forward differences are either identically zero, or reduce to linear combinations of lower order differences. For instance if \( i \) and \( j \) belong to separate connected components of the graph \( \delta_A = 0 \) whenever \( A \) includes both \( i \) and \( j \). Consequently when a putative graph describing the dependencies in the data is given, not all forward differences are interesting. Note that \( \delta_{ijk} = 0 \) whenever the subgraph of the triple is not connected. The proof is straightforward: the subgraph is not connected when \( \inf(X_k, \perp \perp (X_i, X_j)) = 0 \). Consequently both \( I_{ik} = 0 \) and \( I_{jk} = 0 \) and so \( \delta_{ijk} = 0 \) by (13). Only connected triples have interesting forward differences. Restricting attention to subsets that have complete subgraphs with respect to a given graph satisfies the nesting criterion, but would disallow computation of a third order difference on the chain in Figure 1.

Node clusters: We suggest that a subset of nodes in which one node has an edge to every other node is a configuration for which it is appropriate to compute forward differences of an order up to the subset size. Node clusters of this form have an approximate nesting structure: all but one subsets of the cluster are clusters themselves, and if any one edge is dropped from the cluster, it leaves a cluster. The complete subgraph on any number of nodes is a node cluster where any one of the vertices may take the role of the cluster node. Certain configurations are eliminated, for instance, a chain or a chordless cycle on four variables. A subset of size 3 forms a node cluster if one node is adjacent to both others.

An algorithm based on this concept is included in the Supplementary Material.

Collider colouring of the synergies: Synergies are infrequent and so of interest. They are a property of a triple and so more difficult to portray than a node. However the node opposite the weakest edge may be singled out as a collider, generalising the term used in Bayes networks, Pearl (1988). When the weakest edge has zero mutual information, so that its nodes are marginally independent, then the resulting configuration is an unshielded collider (immorality) and the two notions coincide.

The collider may be indicated by a colour (red, say) and the other two nodes yellow,
Colouring the two edges adjacent to the collider indicate the other elements of the triple. Additional rules are needed for overlapping synergies; for instance, any node tagged as a collider is overprinted as a collider. This can lose some detail of the synergy in the graph.

4 Applications of forward differences

We give some low dimensional examples of forward differences of the entropy, both theoretical and empirical, for categorical and continuous data. In three dimensions the third order differences quantify the difference in mutual information between two variables with and without conditioning on a third. We compute and display these differences from some known standard models numerically and, where possible, give an analytic condition for a synergy. A difference is measured in millibits, the same units that measure entropy. For continuous data, we elaborate the conditions for suppression for a theoretical variance matrix with a known graph structure, and give some simple examples. For categorical data we illustrate synergy with examples of binary data in three dimensions, and relate these to the issue of collapsibility. We elucidate examples of four dimensions continuous models that are interesting in the context of Bayes networks.

Higher dimensional examples discuss 3rd-order forward differences and synergies using the skeleton of the Bayes network, known or postulated to have generated the data. Firstly from an artificial tree averaging process, which establishes why the skeleton rather than the moral graph is the right graph to determine which differences need to be computed. Secondly the real-life example of wine quality data is analysed and the synergies suggest that a chain graph model might represent the structure of the variables well. The analysis of the carcass data leads to similar conclusions, but is included because it is easily accessible through R.

4.1 Three dimensional correlations

The lower off-diagonal elements of the correlation matrix $\Sigma$ are $\rho_{12}, \rho_{13}, \rho_{23}$, constrained by requiring $\Sigma$ to be positive definite. The Gaussian entropy function is given in the preliminary remarks to Section 2. The power set has $2^3$ elements, the entropy of the singleton sets are standardised to zero and all others are negative. The 2nd-order forward differences are negatives of the marginal mutual informations, so that the information against the independence of $X_1$ and $X_2$ is $\delta_{12} = -I_{12} = \log(1 - \rho_{12}^2)/2$.

**Theorem 4.1.1** Synergy with three Gaussian variables. The 3rd-order forward difference is

$$\delta_{123} = \frac{1}{2} \log \frac{1 - \rho_{12}^2 - \rho_{13}^2 - \rho_{23}^2 + 2\rho_{12}\rho_{13}\rho_{23}}{(1 - \rho_{12}^2)(1 - \rho_{13}^2)(1 - \rho_{23}^2)},$$

(17)

$$= \frac{1}{2} \log \frac{1 - \rho_{123}^2}{1 - \rho_{12}^2},$$

(18)
and the condition for a synergy, $\delta_{123} < 0$, is that one marginal correlation coefficient is smaller than its corresponding partial in absolute value, for instance $|\rho_{12}| < |\rho_{12|3}|$.

The proof is in the Appendix.

**Corollary 4.1.1 Synergy and negative correlation.** A synergy occurs whenever exactly one marginal correlation is negative.

There are only two cases of correlation matrix to consider: one where all coefficients are positive and the other where exactly one is negative. The corollary deals with the second. It follows because if one correlation is negative then the corresponding partial, say $\rho_{12|3} = (\rho_{12} - \rho_{13}\rho_{23})\{(1 - \rho_{13}^2)(1 - \rho_{23}^2)\}^{-\frac{1}{2}}$, exceeds the marginal in terms of absolute value since the numerator is inflated and the denominator deflated.

In regression scenarios the condition that one marginal correlation is negative may be subdivided by whether the correlation is between a response and an explanatory variable, or between two explanatory variables. This corresponds to the classification of suppression into type: negative or reciprocal, occurring in the literature on suppression and briefly reviewed in the Introduction. The special case $\rho_{12} = 0$ corresponds to classical suppression.

Of interest to us is that a synergy does not occur when $\rho_{12|3} = 0$, and the inequality condition $|\rho_{12}| < |\rho_{12|3}|$ is invariant to permuting indices.

**Example 1.** (Numerical): For a numerical illustration the forward differences are displayed using $\Sigma$ specified by its lower triangle 1.0, 0.2, 1.0, 0.7, 0.5, 1.0, and for comparison, of the same $\Sigma$ with 0.2 replaced by −0.2. The forward differences are, respectively,

| subset | $\phi$ | 1 | 2 | 3 | 12 | 13 | 23 | 123 |
|--------|-------|---|---|---|----|----|----|----|
| fwd.diff($\rho_{12} = 0.2$) | 0 | 0 | 0 | 0 | -30.15 | -497.4 | -212.5 | -14.63 |
| fwd.diff($\rho_{12} = -0.2$) | 0 | 0 | 0 | 0 | -30.15 | -497.4 | -212.5 | -1126.0 |

The values are reported in millibits, see the preliminaries to Section 2. In both the information against $X_1 \perp X_2$ is 30.15mbits. In the first instance the 3rd-order difference $\delta_{123}$ is $-14.63$mbits so that the information against $X_1 \perp X_2 \mid X_3$ is $30.15 + 14.63 = 44.78$mbits. In the second instance $\delta_{123} = -1126$mbits indicating a much more substantial synergy.

The result (18) generalises easily to give a condition for partial synergy.

**Corollary 4.1.2 Partial synergy with three Gaussian variables.** The 3rd-order conditional forward difference for three Gaussian variables given a set $A$ of other such variables is

$$
\delta_{123|A} = \frac{1}{2} \log \frac{1 - \rho_{12|A}^2}{1 - \rho_{12|3}^2}
$$

(19)

and the condition for a partial synergy, $\delta_{123|A} < 0$, is that one marginal correlation coefficient is smaller than its corresponding partial in absolute value, that is $|\rho_{12|A}| < |\rho_{12|A,3}|$. 

12
4.2 Three dimensional contingency tables

While the value of the 3rd-order difference clearly flags the phenomenon of suppression in regression it does not give a definitive answer to non-collapsibility in three way tables. We consider three examples related to Simpson’s paradox: the first is an archetypal loglinear model, the second is numerical and the third is real-life.

Example 2. (Analytic): The first example of a $2^3$-table is analytic where each margin shows independence but the three variables are dependent. A priori the 3rd-order forward difference must be negative.

The log-linear is expansion of $p_{123}$ on $\{0, 1\}^3$ is

$$\log(\alpha) + (x_1 + x_2 + x_3) \log(\beta/\alpha) - 2(x_1x_2 + x_1x_3 + x_2x_3) \log(\beta/\alpha) + 4x_1x_2x_3 \log(\beta/\alpha), \quad (20)$$

where $x_1, x_2, x_3$ take values 0, 1; and parameterised by $\alpha \in (0, 1/4)$ with $\beta = 1/4 - \alpha$. In standard order the joint probabilities are $(\alpha, \beta, \beta, \alpha, \beta, \alpha, \alpha, \beta)$. It illustrates non-collapsibility because every margin has equi-probability entries so that $\inf(X_i \perp \perp X_j) = 0$, while any two variables contribute positively to the prediction of the third.

By direct evaluation,

$$\delta_{123} = -4(\alpha \log(\alpha) + \beta \log(\beta)) - 3 \log(2).$$

This is zero when $\beta = \alpha$, but otherwise negative.

Example 3. (Kidney stones): This is taken from [Julious and Mullee (1994)] has previously been used as a real-life instance of Simpson’s paradox. There are two factors (Treatment, Size), each with two levels (A/B, small/large stones respectively). Outcomes (81/87, 234/270, 192/263, 55/80) are recorded as the success/total count in the four groups, in Treatment within Size order.

The entropy function and its forward differences are displayed here

| subset | $\phi$ | O   | T   | S    | OT   | OS   | TS   | OTS   |
|--------|-------|-----|-----|------|------|------|------|-------|
| entropy | 0.0   | 733.4 | 1024.0 | 1023.7 | 1754.9 | 1725.8 | 1835.3 | 2533.7 |
| fwd.diff. $\delta$ | 0 | 733.4 | 1024.0 | 1023.7 | -2.443 | -31.31 | -212.4 | -1.198 |

with values in millibits. The $T$ margin is exactly balanced (1024mbits is the maximum), and the $S$ margin almost so, but the $T \times S$ table is not (the mutual information is 212.4mbits and far from zero). The value of $\delta_{OTS} = -1.198$mbits is negative; it is also negligible so that marginal and conditional independence measures are approximately the same. The independence graph approximating these data is

$\circarrowleft O \circarrowright S \circarrowright T$

Here Simpson’s paradox occurs when comparing the $OT$ interaction conditionally on $S$, with its value marginalised over $S$, and arises because of the large imbalance in the $T \times S$ table. The value of $\delta_{OTS}$ does not signal the paradox.
It is easy to construct examples where the paradox (log odds ratio in the marginal and in the conditional tables are of opposite sign) goes with a negative and examples with a positive third order difference.

### 4.3 Four dimensional correlation matrices

**Example 4.** (Analytic): The forward differences of the entropy are calculated from the theoretical correlation matrix of various four dimensional graphical models. We compute forward differences all orders, though report only the most salient features to illustrate what may be expected if data is generated from such models. The graphical models, characterized by the graphs in Figure 2, include the so-called cluster model, a chain, a decomposable model, the 4-cycle, Bayes networks with one, two and three unshielded colliders.

The interpretation of their differences derives from the separation properties of the graph translating to a statement of the form \( \delta_{A\cup B|C} = 0 \) in Theorem 2.4.2. This in turn leads to one or more linear relationships using Theorem 2.4.1. These results are summarised in Table 1.

(a) Cluster: This node cluster is a sparse configuration sufficiently complex that \( \delta_{1234} \) is not zero. The variables \( X_1, X_3, X_4 \) are mutually independent given the cluster node \( X_2 \), consequently three of the four 3rd-order differences involving the cluster node \( X_2 \) are positive as the information conditioned on \( X_2 \) is zero. The term \( \delta_{134} \) is necessarily positive, for instance, because \( X_3 \perp X_4 | X_2 \), and \( X_1 \) is a predictor of \( X_2 \), so that \( I_{34} > I_{34|1} > I_{34|2} = 0 \) (or equivalently \( \delta_{34} < \delta_{34|1} < \delta_{34|2} = 0 \)). As \( 0 = \delta_{134|2} = \delta_{134} + \delta_{1234}, \delta_{1234} \) is always negative.

(b) Chain: \( X_1 \perp X_3 | X_2 \) implies \( \delta_{123} > 0 \), similarly all other triples have a positive forward difference. That \( X_1 \perp X_{34} | X_2 \) implies \( \delta_{134|2} = 0 \); this, together with the identity \( \delta_{134|2} = \delta_{134} + \delta_{1234} \) involving the fourth order difference, implies \( \delta_{1234} < 0 \). The dependence structure of this graph is characterized by the values of \( \{\delta_{12}, \delta_{23}, \delta_{34}, \delta_{123}, \delta_{234}\} \).

(c) Decomposable: \( X_1 \perp X_3 | X_2 \) implies \( \delta_{123} > 0 \), similarly \( \delta_{124} > 0 \). Because \( \delta_{134|2} = 0 = \delta_{134} + \delta_{1234} \) they are of opposite sign, but otherwise arbitrary.

(d) 4-cycle: There are two independences leading to two zero linear combinations: \( X_1 \perp X_3 | X_2 \) translates to \( \delta_{13|24} = 0 = \delta_{13} + \delta_{123} + \delta_{134} + \delta_{1234} \), and \( X_2 \perp X_4 | X_3 \) translates to \( \delta_{24|13} = 0 = \delta_{24} + \delta_{124} + \delta_{124} + \delta_{1234} \). We argue that \( \delta_{13} < \delta_{13|24} = 0 \) because the information decreases as the conditioning set is enlarged. Consequently \( \delta_{123} > 0 \), and symmetry shows the other 3rd-order differences are positive. Also \( 0 < \delta_{134|2} = \delta_{134} + \delta_{1234} \), so that \( \delta_{1234} < 0 \).

(e) bayesNetA: There are two independences manifest: firstly \( X_2 \not\perp X_4 | X_1 \) translates to \( \delta_{24|1} = 0 = \delta_{24} + \delta_{124} \); secondly \( X_1 \perp X_3 | X_2 \) translates to \( \delta_{13|24} = 0 = \delta_{13} + \delta_{123} + \delta_{134} + \delta_{1234} \). The 3rd-order difference \( \delta_{124} \) is positive and there is a partial synergy at 3 as \( \delta_{234|1} < 0 \).

(f) bayesNetB: There are two independences: \( X_2 \not\perp X_4 \) implies \( \delta_{24} = 0 \); secondly \( X_1 \not\perp X_3 | X_2 \) is again \( \delta_{13|24} = 0 = \delta_{13} + \delta_{123} + \delta_{134} + \delta_{1234} \). There are two marginal synergies at 1 and at 3 so \( \delta_{124} < 0 \) and \( \delta_{234} < 0 \).
Figure 2: Four dimensional configurations of independence graphs (undirected and directed).

Table 1: Summary of four dimensional forward differences for examples in Figure 2

|   |   |   |   |   |
|---|---|---|---|---|
|   |   |   |   |   |
|   |   |   |   |   |
|   |   |   |   |   |
|   |   |   |   |   |
|   |   |   |   |   |
|   |   |   |   |   |
|   |   |   |   |   |
|   |   |   |   |   |
|   |   |   |   |   |
|   |   |   |   |   |
|   |   |   |   |   |
|   |   |   |   |   |
|   |   |   |   |   |
|   |   |   |   |   |
|   |   |   |   |   |
|   |   |   |   |   |
|   |   |   |   |   |
|   |   |   |   |   |
|   |   |   |   |   |
|   |   |   |   |   |
|   |   |   |   |   |
|   |   |   |   |   |
|   |   |   |   |   |
|   |   |   |   |   |
|   |   |   |   |   |
|   |   |   |   |   |
|   |   |   |   |   |
|   |   |   |   |   |
|   |   |   |   |   |
|   |   |   |   |   |
|   |   |   |   |   |
|   |   |   |   |   |
|   |   |   |   |   |
|   |   |   |   |   |
|   |   |   |   |   |
|   |   |   |   |   |
|   |   |   |   |   |
|   |   |   |   |   |
|   |   |   |   |   |
|   |   |   |   |   |
|   |   |   |   |   |
|   |   |   |   |   |
|   |   |   |   |   |
|   |   |   |   |   |
|   |   |   |   |   |
|   |   |   |   |   |
|   |   |   |   |   |
|   |   |   |   |   |
|   |   |   |   |   |
|   |   |   |   |   |
|   |   |   |   |   |
|   |   |   |   |   |
|   |   |   |   |   |
|   |   |   |   |   |
|   |   |   |   |   |
|   |   |   |   |   |
|   |   |   |   |   |
|   |   |   |   |   |
|   |   |   |   |   |
|   |   |   |   |   |
|   |   |   |   |   |
|   |   |   |   |   |
|   |   |   |   |   |
|   |   |   |   |   |
|   |   |   |   |   |
|   |   |   |   |   |
|   |   |   |   |   |
|   |   |   |   |   |
|   |   |   |   |   |
|   |   |   |   |   |
|   |   |   |   |   |
|   |   |   |   |   |
|   |   |   |   |   |
|   |   |   |   |   |
|   |   |   |   |   |
|   |   |   |   |   |
|   |   |   |   |   |
|   |   |   |   |   |
|   |   |   |   |   |
|   |   |   |   |   |
|   |   |   |   |   |
|   |   |   |   |   |
|   |   |   |   |   |
|   |   |   |   |   |
|   |   |   |   |   |
|   |   |   |   |   |
|   |   |   |   |   |
|   |   |   |   |   |
|   |   |   |   |   |
|   |   |   |   |   |
|   |   |   |   |   |
|   |   |   |   |   |
|   |   |   |   |   |
|   |   |   |   |   |
|   |   |   |   |   |
|   |   |   |   |   |
|   |   |   |   |   |
|   |   |   |   |   |
|   |   |   |   |   |
|   |   |   |   |   |
|   |   |   |   |   |
|   |   |   |   |   |
|   |   |   |   |   |
|   |   |   |   |   |
|   |   |   |   |   |
|   |   |   |   |   |
|   |   |   |   |   |
|   |   |   |   |   |
|   |   |   |   |   |
|   |   |   |   |   |
|   |   |   |   |   |
|   |   |   |   |   |
|   |   |   |   |   |
|   |   |   |   |   |
|   |   |   |   |   |
|   |   |   |   |   |
|   |   |   |   |   |
|   |   |   |   |   |
|   |   |   |   |   |
|   |   |   |   |   |
|   |   |   |   |   |
|   |   |   |   |   |
|   |   |   |   |   |
|   |   |   |   |   |
|   |   |   |   |   |
|   |   |   |   |   |
|   |   |   |   |   |
|   |   |   |   |   |
|   |   |   |   |   |
|   |   |   |   |   |
|   |   |   |   |   |
|   |   |   |   |   |
|   |   |   |   |   |
|   |   |   |   |   |
|   |   |   |   |   |
|   |   |   |   |   |
|   |   |   |   |   |
|   |   |   |   |   |
|   |   |   |   |   |
|   |   |   |   |   |
|   |   |   |   |   |
|   |   |   |   |   |
|   |   |   |   |   |
|   |   |   |   |   |
BayesNetC: There are three marginal independences between pairs of $X_1, X_3, X_4$ with corresponding 2nd-order differences being zero; and as these variables are mutually independent the 3rd-order difference is zero too. There are three marginal synergies at 2, with $\delta_{123} < 0$, $\delta_{124} < 0$ and $\delta_{234} < 0$.

**Example 5.** (GP burn-out): This example was used by Maassen and Bakker (2001) to illustrate suppression in the context of path analysis. We use it to illustrate forward differences of the entropy in four dimensions. Surprisingly we find that there are no synergies in any of the three dimensional margins nor any partial synergy in four dimensions, and consequently no colliders.

A two wave study of burnout among 207 general practitioners measured levels of the lack of job satisfaction and of burn-out. The variables here are denoted by $js_1, js_2, bo_1, bo_2$, with the numeral denoting the wave. The correlation matrix, reported in supplementary material, shows all marginal correlations to be positive. Forward differences of the entropy higher than the first are

| subset | js1:bo1 | js1:js2 | bo1:js2 | js1:bo2 | bo1:bo2 | js2:bo2 |
|--------|--------|--------|--------|--------|--------|--------|
| 2nd-order fwd.diff | -191.6 | -98.9 | -123.9 | -114.5 | -356.9 | -259.2 |
| subset | js1:bo1:js2 | js1:bo1:bo2 | js1:js2:bo2 | bo1:js2:bo2 | js1:bo1:js2:bo2 |
| 3,4-orders fwd.diff | 66.96 | 103.7 | 71.63 | 118.5 | -61.96 |

The 2nd-order differences (pairwise MIs) are all substantial; the 3rd-order differences are all positive, so clearly there are no synergies in any three dimensional marginal. There are two (approximate) linear relations corresponding to the 2nd-order statements

$js_1 \perp \perp bo_2 | \{js_2, bo_2\}: \delta_{js_1:bo_2} + \delta_{js_1:js_2:bo_2} + \delta_{js_1:bo_1:bo_2} + \delta_{js_1:bo_1:js_2:bo_2} = -1.13\text{mbits}$; and

$bo_1 \perp \perp js_2 | \{js_1, bo_2\}: \delta_{bo_1:js_2} + \delta_{js_1:bo_1:js_2} + \delta_{bo_1:js_2:bo_2} + \delta_{js_1:bo_1:js_2:bo_2} = -0.40\text{mbits}$.

This suggests the 4-cycle with graph

```
  js1 ---- bo1
    |      |   |
    j s 2 ---- bo 2
```

Standard model fitting using the R-packages pcalg, gRim, or ggm gives the same independence graph.

The context suggests that synergies might be found at one or both of the second wave nodes: for $js_2$, $\delta_{js_1:js_2:bo_2|bo_1} = \delta_{js_1:js_2:bo_2} + \delta_{js_1:bo_1:js_2:bo_2} = 9.67\text{mbits}$ and for $bo_2$, $\delta_{bo_1:js_2:bo_2} + \delta_{js_1:bo_1:js_2:bo_2} = 56.55\text{mbits}$. However both are positive indicating that this is not the case, and we conclude there are no suppression effects manifest in the observed data.

### 4.4 Higher dimensions

**Example 6.** (A tree averaging structure): This artificial tree averaging process provides
an example of computing third order forward differences with respect to a given graph. The process starts with a founding generation of independent Gaussian random variables. Pairs of these are parents to a single child, giving a new generation of half the size; and the process repeats until only one successor is left. The parent-child relation is specified by the parameter $\alpha$ in

$$X_{\text{child}} = \alpha (X_{\text{par}_1} + X_{\text{par}_2}) + \epsilon$$

where $\epsilon$ are independent standard Normal. The correlation matrix is determined by the parameter $\alpha$. With 8 founders there are $p = 15$ variables; so that in principle there are 455 subsets of size 3 to examine.

The Bayes network generating the process is displayed in Figure 3.

![Figure 3: The Bayes network generating the tree averaging process.](image)

Emulating a data processing exercise with observations on this process would lead to the skeleton with 19 triples or to the moral graph with $19 + 12 = 31$ triples. Recall that a triple with respect to a graph, is a subset of size 3 with (at least) one node adjacent to the two others.

In the moral graph there are seven synergies (negative third order forward differences) that exactly correspond to the seven immoralities in the graph. With $\alpha = 0.6$, the strongest synergy is at the apex of the pyramid (-164.02mbits), followed by two in next tier (-116.78mbits) and the four weaker ones at the bottom tier (-53.66mbits). The positive differences each correspond to a child–parent–grandparent conditional independence. The four stronger ones (65.90mbits) are at the apex of the tree and involve the final survivor $X_1$; the other eight positive ones (44.06mbits) involve a founder node.

There are exactly twelve differences that are identically zero corresponding to moralisation: applying d-separation, for instance to the 2,3,4 triple, marginally $X_{24} \perp \perp X_3$, so that both $I_{23}$ and $I_{234}$ are zero. In large graphs it is more efficient to compute low order forward differences from the skeleton rather than from the moralised graph of a Bayes network.

Example 7. (Carcass data): A well known data set is the so-called carcass data available from the R-package gRim, Højsgaard et al. (2012); the correlation matrix is reproduced in supplementary material. It consists of 7 nutritional content measurements on 374 pigs(?). The skeleton is found using the pcalg R-package, Kalisch et al. (2012), with
standard settings and a 5% significance level for edge testing, gives the graph of the skeleton as the left diagram in Figure 4. With this graph there are exactly nine node clusters of order 3 (triples). The corresponding forward differences are listed in Table 2. Most of entries are positive, and quite a few are large indicating large duplication of effects, especially within the Fat measures and within the Meat measures. Strikingly there are two overlapping synergies (the negative differences). They have the same collider LeanMeat, and the nodes of the synergies are coloured in the graph on the right using the colouring rule above. Reading from the graph Fat11 and Meat13

Table 2: Third order forward differences for the carcass data based on the graph.

| Nodes          | δ      |
|----------------|--------|
| Fat11 Meat13   | LeanMeat 78.13 |
| Fat12 Meat13   | LeanMeat 76.55 |
| Meat12 Meat13  | LeanMeat 32.32 |
| Meat11 Meat13  | LeanMeat 50.54 |
| Fat12 Fat13    | LeanMeat 458.18 |
| Fat11 Fat13    | LeanMeat 460.97 |
| Fat11 Fat12    | LeanMeat 514.10 |
| Fat11 Meat12   | Fat13 694.67 |
| Meat11 Meat12  | Meat13 894.42 |

are marginally independent and together enhance LeanMeat more than their separate effects would warrant. The same is true of the effect of Fat12 and Meat13 on LeanMeat. Both of these synergies suggest that the data be modelled as a chain graph, Wermuth and Lauritzen (1990), with LeanMeat as the single outcome variable.

Figure 4: Skeleton of the carcass data (left) with two overlapping coloured synergies (right).

Going from the coloured graph may be misleading without access to the corresponding table of synergies; for instance the graph might be taken to indicate that \{Fat11,Fat12,LeanMeat\} is a synergy when it is not.

Example 8. (Wine quality data): We consider a regression example of wine quality taken from the machine learning data set repository at UCI (archive.ics.uci.edu/ml/datasets/
There are 4898 observations on 11 physico-chemical properties and a sensory quality variable for the white Portuguese Vinho Verde wine reported by Cortez et al. (2009). The red wine data was used as one of the test sets in Elidan (2010). The quality outcome is an ordered categorical response, the other variables are continuous. Our objective is to find and display any synergies in the explanatory variables so leading to a better understanding of the data set.

An exploratory analysis reveals transformations are required to establish linearity and normality. The simple approach of taking the normal scores, based on ranking each variable, produces pairs plots for the bivariate margins that are now almost all uniformly ovaloid.

The skeleton is found using the `pcalg` R-package, Kalisch et al. (2012), with standard settings and a 1% significance level for edge testing. There are 21 edges and 51 triples in the skeleton compared to 55 and 165, respectively, in the complete graph. The empirical cumulative distribution function of the corresponding 3rd-order forward differences is displayed on the left in Figure 5. The majority of the 3rd-order differences are near zero. There is clearly one large synergy, four others of some size, three large positive forward differences, and five more of some size. The detail is given in Table 3.

The skeleton on the right of the Figure is coloured with the five synergies in the Table that are stronger than $-15$mbits. This has the effect of classifying the explanatory variables into red, yellow and white. Each red node belongs to one or more synergistic triples and is defined as the node opposite the weakest edge. There are just two red nodes: density, which occurs in four synergies, and total.sulfur.dioxide occurring once. Three of the four synergies including density are immoralities and make density an unshielded collider. The context of this physico-chemical data set suggests a causal mechanism in which density and total.sulfur.dioxide are responses directly affected by the yellow nodes in a synergistic relation. Interestingly density and total.sulfur.dioxide
Table 3: Larger synergies for the wine data, with the node to be coloured red indicated by the asterisk.

| triple                     | 3rd-order diff. |
|---------------------------|-----------------|
| residual.sugar density*   | -61.74          |
| volatile.acidity          |                 |
| free.sulfur.dioxide       | -25.49          |
| total.sulfur.dioxide*     |                 |
| fixed.acidity             |                 |
| residual.sugar density*   | -23.29          |
| fixed.acidity             |                 |
| density*                  | -19.64          |
| residual.sugar            |                 |
| chlorides                 | -18.50          |
| total.sulfur.dioxide      |                 |
| density*                  | -79.46          |
| chlorides                 |                 |
| total.sulfur.dioxide      | 84.40           |
| density                   | 139.82          |
| residual.sugar            |                 |
| total.sulfur.dioxide      |                 |
| density                   | 158.78          |
| chlorides                 |                 |
| density                   | 184.43          |

are associated, but not synergistically, occurring together in the three of the largest five positive forward differences. The white nodes are not members of any synergistic triple. This coloured classification of nodes suggests further fitting the data as variables in a chain graph, Wermuth and Lauritzen (1990).

5 Discussion

To turn forward difference estimation into a practical statistical tool requires a reliable method of assessing sampling errors. This is clearly necessary in empirical estimation though perhaps less so in the testing scenarios of graphical model search. For now we make two remarks. Firstly, it is probable that lower order conditional mutual informations have smaller sampling errors than any associated higher order measures, as these have fewer additional terms in their expansion. Secondly the sampling error of the highest order order term in the forward difference expansion of information is probably of the same order of variability as the information itself. However in the absence of good approximations to sampling errors the parametric bootstrap should work well.

Forward differences of the entropy function promise a productive vein of research related to graphical models. For instance the additive expansion of the conditional mutual information statistic in terms of 3rd-order differences give a particularly simple proof of the so-called information inequality. The potential efficiency gains in graphical model constraint based search might be leveraged to attain or surpass that of current algorithms such as pcalg mentioned above. A difficult problem is to locate and evaluate higher dimensional synergies of the form $\delta_{ijk|A} < 0$ where the subset $A$ is arbitrary. A possible line of research is investigation if synergies for shielded colliders have a role to play in understanding causal graphs.

Parallel to forward differences are backward differences generated by inverting the lattice of entropies and taking $h_P$ as the minimal and $h_\phi$ as the maximal elements respectively. A better way to study this might be to take the forward differences of the conditional entropy function $h_{P|A}(P|A)$ on $\{A \in \mathcal{P}\}$. It is quickly seen that the 2nd-order differences are pair-wise mutual informations conditioned on the all other variables, and 3rd-order differences are $\delta_{ijk|P\setminus ijk}$. 

20
Acknowledgement

We thank Peter van der Heijden for pointing out the connection with suppressor variables. The work of two authors (FM, YX) was supported by Philip Morris International Research and Development.

Appendix Proofs

Proof of Theorem 2.2.1. The notation $\delta_A$ is shorthand for $\delta_{(i; i \in A)}$ where the round brackets indicate an ordered sequence. We wish to show $\delta_{\pi(A)} = \delta_A$ for any permutation $\pi$.

We argue by enumeration on $|A|$. For $|A| = 1$ there is nothing to show. For $|A| = 2$ with $A = \{i, j\}$ say $h_A = \delta_\phi + \delta_i + \delta_j + \delta_{ij}$ and $h_{\pi(A)} = \delta_\phi + \delta_i + \delta_j + \delta_{\pi(ij)}$. As the entropies are equal, subtraction shows $\delta_{\pi(ij)} = \delta_{ij}$ so that the 2nd-order forward differences are symmetric. For $|A| = 3$ a similar result is attained using the symmetry of the 2nd-order terms. The argument continues until $|A| = p$.

Proof of Lemma 2.1. Note that $\delta_\phi = 0$ but the term is included to preserve symmetry. The forward difference expansion (5) of $h_{a \cup b}$ is

$$h_{a \cup b} = \sum_{c \subseteq a \cup b} \delta_c = -\delta_\phi + h_a + h_b + \sum_{c : |a \cap c| > 1 , |b \cap c| > 1} \delta_c.$$ 

The last summation on the right is the sum over terms with at least one element from $a$ and one from $b$. By hypothesis it is 0.

Direct enumeration on the elements $(|a|, |b|) \in \{1, 2, \ldots , |A|\} \times \{1, 2, \ldots , |B|\}$ shows that every $\delta_c$ in this summation is 0. Start with singletons $a = \{i\}$, $b = \{j\}$. The only term is $\delta_{ij}$ and so it is 0. Repeat this over all pairs $ij$. A similar argument applied to $a = \{i\}$, $b = \{j, k\}$ and using $\delta_{ij} = 0$ establishes $\delta_{ijk} = 0$, for all $k$. Repeating this argument establishes $\delta_{ib} = 0$ for any nonempty $b \subseteq B$. A similar enumeration on $|a|$ then gives the result.

The proof of the converse follows immediately from the expansion of $h_{a \cup b}$.

Proof of Theorem 2.3.1. Take $A$ disjoint from $\{i, j\}$ and note

$$h_{Ai} = h_A + \sum_{j \subseteq B \subseteq A} \delta_B.$$ 

This additivity recurrence follows directly from the definition of the forward differences at (5). Now use this in the elementary imset representation at (3)

$$-I_{ij|A} = (h_{Aij} - h_{Ai}) - (h_{Aj} - h_A)$$

$$= \sum_{j \subseteq B \subseteq A} \delta_B^j - \sum_{j \subseteq B \subseteq A} \delta_B,$$ using (21),

$$= \sum_{ij \subseteq B \subseteq A} \delta_B.$$
Cancellation leaves only those terms with both \(i\) and \(j\) in the subscript, as required.

**Proof of Theorem 2.4.1** Firstly, we show \(\delta_{12|3} = \delta_{123} - \delta_{12}\) as the structure of the proof is contained in this special case. Take the definition of a conditional forward difference

\[
\delta_{12|3} = h_{12|3} - h_{1|3} - h_{2|3} + h_{\emptyset|3},
\]

\[
= h_{123} - h_{13} - h_{23} + h_{3},
\]

simplified by applying \(h_{A|B} = h_{A\cup B} - h_{B}\) repeatedly and noting that the four \(h_{B} = h_{3}\) terms cancel. The term \(\delta_{123}\) is a sum over the \(2^3\) elements of the power set \(\mathcal{P}(\{1, 2, 3\})\) of the signed function \(h\). Partition this into the sum of those elements that contain 3 and those that do not, then

\[
\delta_{123} = \sum_{C \subseteq \{1, 2\}} (-1)^{2^{1+\cdot|C|}} h_{C|3} + \sum_{C \subseteq \{1, 2\}} (-1)^{2^{1+\cdot|C|}} h_{C|3},
\]

taking care with the signs, and as required.

More generally consider \((\ast)\); from the definition of conditional forward differences

\[
\delta_{A|Bk} = \sum_{C \subseteq A} (-1)^{|A| - |C|} h_{C|Bk}
\]

\[
= \sum_{C \subseteq A} (-1)^{|A| - |C|} h_{Ck|B},
\]

where the \(h_{k|B}\) terms cancel. The sum for \(\delta_{Ak|B}\) is partitioned into the sum over the power sets including and excluding \(k\):

\[
\delta_{Ak|B} = \sum_{C \subseteq A} (-1)^{|Ak| - |Ck|} h_{Ck|B} + \sum_{C \subseteq A} (-1)^{|Ak| - |Ck|} h_{C|B}
\]

\[
= \delta_{A|Bk} - \delta_{A|B},
\]

as required.

**Proof of Theorem 2.4.2** When \(C\) separates \(A\) and \(B\) then as a consequence of the Markov properties of the graph \(X_A \perp \perp X_B \mid X_C\); consequently in turn \(h_{A\cup B|C} = h_{A|C} + h_{B|C}\). By a small generalisation of the the additivity Lemma 2.1 to incorporate conditioning, the result follows.

**Proof of Theorem 3.1.1** Endow the set \(A\) with a total ordering so that for \(i \neq j \in A\) either \(i < j\) or \(j < i\). Apply the information identity, Cover and Thomas (2006), to get

\[
\inf(X_k \perp \perp X_A) = \sum_{j \in A} \inf(X_k \perp \perp X_j \mid X_{\{i : i < j\}}). \quad (22)
\]

Use \((\ast)\) of Theorem 2.3.1 to express the conditional mutual informations in terms of the forward differences, so

\[
\inf(X_k \perp \perp X_A) = \sum_{j \in A} \left[ - \sum_{B \subseteq \{i : i < j\}} \delta_{BkJ} \right].
\]
Isolate the 2nd-order differences from sum and rearrange the index of summation gives the result:

\[
\inf(X_k \perp \perp X_A) = - \sum_{j \in A} \delta_{kj} - \sum_{j \in A} \sum_{B \subseteq \{i, i < j\} : |B| > 1} \delta_{Bkj} \\
= \sum_{j \in A} \inf(X_j \perp \perp X_k) - \sum_{B \subseteq A : |B| > 1} \delta_{Bk}.
\]

**Proof of Theorem 4.1.1.** The expression (17) may be derived directly by evaluating the determinants in the Gaussian entropy. The second statement follows from (9) and the fact that \( I_{12|3} = - \log(1 - \rho_{12|3}^2)/2 \).

\[\square\]

**References**

Andersson, S. A., Madigan, D., and Perlman, M. D. (1997). A characterization of Markov equivalence classes for acyclic digraphs. *The Annals of Statistics*, 25(2):505–541.

Bertrand, P. V. and Holder, R. L. (1988). A quirk in multiple regression: the whole regression can be greater than the sum of its parts. *The Statistician*, pages 371–374.

Bishop, Y. M., Fienberg, S. E., and Holland, P. W. (1975). *Discrete Multivariate Analysis: Theory and Practice*. MIT Press.

Cortez, P., Cerdeira, A., Almeida, F., Matos, T., and Reis, J. (2009). Modeling wine preferences by data mining from physicochemical properties. *Decision Support Systems*, 47(4):547–553.

Cover, T. M. and Thomas, J. A. (2006). *Elements of Information Theory*. Wiley-interscience, 2nd edition.

Darroch, J. N., Lauritzen, S. L., and Speed, T. (1980). Markov fields and log-linear interaction models for contingency tables. *The Annals of Statistics*, 8:522–539.

Dawid, A. P. (1979). Conditional independence in statistical theory. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 1–31.

Elidan, G. (2010). Copula bayesian networks. In Lafferty, J., Williams, C., Shawe-Taylor, J., Zemel, R., and Culotta, A., editors, *Advances in neural information processing systems*, pages 559–567.

Friedman, L. and Wall, M. (2005). Graphical views of suppression and multicollinearity in multiple linear regression. *The American Statistician*, 59(2):127–136.

Frydenberg, M. (1990). The chain graph Markov property. *Scand. J. Statist.*, 17:333–353.

Greenland, S., Robins, J. M., and Pearl, J. (1999). Confounding and collapsibility in causal inference. *Statistical Science*, pages 29–46.
Hamilton, D. (1988). When is $R^2 > r_{y1}^2 + r_{y2}^2$? (revisited). *The American Statistician*, 42(1):89–91.

Højsgaard, S., Edwards, D., and Lauritzen, S. (2012). *Graphical models with R*. Springer.

Horst, P. (1941). The role of predictor variables which are independent of the criterion. *Social Science Research Council*, 48:431–436.

Julious, S. A. and Mullee, M. A. (1994). Confounding and Simpson’s paradox. *BMJ*, 309(6967):1480–1481.

Kalisch, M., Mächler, M., Colombo, D., Maathuis, M. H., and Bühlmann, P. (2012). Causal inference using graphical models with the R package pcalg. *Journal of Statistical Software*, 47(11):1–26.

Kline, R. B. (2011). *Principles and Practice of Structural Equation Modeling*. Guilford press.

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

Lauritzen, S. L. (1996). *Graphical Models*. Oxford University Press.

Lauritzen, S. L. and Spiegelhalter, D. J. (1988). Local computations with probabilities on graphical structures and their application to expert systems. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 157–224.

Ludlow, L. and Klein, K. (2014). Suppressor variables: The difference between is versus acting as. *Journal of Statistics Education*, 22(2).

Maassen, G. H. and Bakker, A. B. (2001). Suppressor variables in path models: definitions and interpretations. *Sociological Methods & Research*, 30(2):241–270.

MacKinnon, D. P., Krull, J. L., and Lockwood, C. M. (2000). Equivalence of the mediation, confounding and suppression effect. *Prevention Science*, 1(4):173–181.

McNemar, Q. (1945). The mode of operation of suppressant variables. *The American Journal of Psychology*, pages 554–555.

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems*. Morgan and Kaufman, San Mateo.

Rota, G.-C. (1964). On the foundations of combinatorial theory I. Theory of Möbius functions. *Probability Theory and Related Fields*, 2(4):340–368.

Shieh, G. (2001). The inequality between the coefficient of determination and the sum of squared simple correlation coefficients. *The American Statistician*, 55(2):121–124.

Shieh, G. (2006). Suppression situations in multiple linear regression. *Educational and psychological measurement*, 66(3):435–447.

Simpson, E. H. (1951). The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 238–241.
Smith, R. L., Ager, J. W., and Williams, D. L. (1992). Suppressor variables in multiple regression/correlation. *Educational and Psychological Measurement*, 52(1):17–29.

Studený, M. (2005). *Probabilistic Conditional Independence Structures*. Springer.

Velicer, W. F. (1978). Suppressor variables and the semipartial correlation coefficient. *Educational and Psychological Measurement*, 38(4):953–958.

Verma, T. and Pearl, J. (1990). Equivalence and synthesis of causal models. In Henrion, M. et al., editors, *Uncertainty in Artificial Intelligence: Proceedings of the Sixth Conference*, pages 220–227, San Francisco. Morgan Kaufman.

Voyer, D. (1996). The relation between mathematical achievement and gender differences in spatial abilities: A suppression effect. *Journal of Educational Psychology*, 88(3):563.

Wermuth, N. and Lauritzen, S. L. (1990). On substantive research hypotheses, conditional independence graphs and graphical chain models. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 21–50.

Whittaker, J. (1990). *Graphical Models in Applied Multivariate Statistics*. Wiley, Chichester.

Whittemore, A. S. (1978). Collapsibility of multidimensional contingency tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 328–340.
Supplementary material to Synergy, suppression and immorality

Jan 17, 2015

A node cluster algorithm for computing forward differences

We are given an undirected (simple) graph on \( p \) nodes and wish to compute forward differences of order \( \kappa \) or less, for that graph. For \( \kappa = 3 \) the difference \( \delta_{ijk} \) is evaluated, if in the graph, the node \( i \), say, has two neighbours \( j \) and \( k \), so that this triple forms a node cluster. More generally, a subset (of any order) is viable if there is one node that is a neighbour to all other nodes.

Examples focus on low order differences so we adopt a breadth first computation. We resolve orderings by choosing the weakest candidates based on the marginal mutual information \( \{I_{ij}; \ i, j \in P\} \). This makes sense when adapting the algorithm to discard edges.

Algorithm 1 A node cluster algorithm

Increment \( \kappa \): starting with \( \kappa = 3 \).

- LOOP on nodes: to pass through whole graph.
  Choose node with maximum degree, not yet visited.
- LOOP on all tuples (length \( \kappa - 1 \)) of its neighbours:
  visit weakest tuple first, via sum MIs.
  Check tuple forms a node cluster,
  put subset=(node,tuple),
  if subset is new store.
  Evaluate the entropy of the subset, store.
- LOOP on all sub-subsets of the subset:
  evaluate forward difference, using stored entropies.
  If a relevant sub-subset unvisited,
  compute entropy, store,
UNTIL all sub-subsets, tuples, and nodes visited.
Correlation matrices

GP burn-out data

|       | SAT11 | B011 | SAT12 | B012 |
|-------|-------|------|-------|------|
| SAT11 | 1.000 | 0.478| 0.354 | 0.379|
| B011  | 0.478 | 1.000| 0.393 | 0.619|
| SAT12 | 0.354 | 0.393| 1.000 | 0.544|
| B012  | 0.379 | 0.619| 0.544 | 1.000|

Wine data

archive.ics.uci.edu/ml/datasets/Wine+Quality
dim(wine)
4898  12
qnrank=function(x){
  n = length(x)
  qn = qnorm(seq(1:n)/(n+1))
  return(qn[ rank(x ,ties ="random")])
}
xqn = apply(wine,2,qnrank)
data = xqn[,1:11]
exclude quality as categorical
noquote(colnames(data))
[1] fixed.acidity volatile.acidity citric.acid
[4] residual.sugar chlorides free.sulfur.dioxide
[7] total.sulfur.dioxide density pH
[10] sulphates alcohol
colnames(data)=NULL
cor(data)

|       | [,1]       | [,2]      | [,3]     | [,4]    | [,5]     | [,6]    | [,7]          |
|-------|------------|-----------|----------|---------|----------|---------|---------------|
| [,1]  | 1.000000   | -0.030966 | 0.31742  | 0.09673 | 0.08979  | -0.037524| 0.10117       |
| [,2]  | -0.03097   | 1.000000  | -0.16975 | 0.10579 | 0.01394  | -0.084837| 0.11634       |
| [,3]  | 0.31742    | -0.169748 | 1.000000 | 0.04799 | 0.04927  | 0.089613 | 0.10153       |
| [,4]  | 0.09673    | 0.105791  | 0.04799  | 1.00000 | 0.20868  | 0.319827 | 0.41631       |
| [,5]  | 0.08979    | 0.013938  | 0.04927  | 0.20868 | 1.00000  | 0.162736 | 0.35118       |
| [,6]  | -0.03752   | -0.084837 | 0.08961  | 0.31983 | 0.16274  | 1.000000| 0.62336       |
| [,7]  | 0.10117    | 0.116342  | 0.10153  | 0.41631 | 0.35118  | 0.623356| 1.000000      |
| [,8]  | 0.29980    | 0.002518  | 0.12049  | 0.74793 | 0.47804  | 0.299242| 0.53097       |
| [,9]  | -0.43610   | -0.045856 | -0.15785 | -0.16324| -0.04935 | 0.094488| 0.01014       |
| [10]  | -0.01824   | -0.034665 | 0.07433  | 0.01986 | 0.09323  | 0.068181| 0.16226       |
| [11]  | -0.12775   | 0.054468  | -0.05880 | -0.41334| -0.53105 | -0.258037| -0.44020      |

|       | [,8]       | [,9]      | [,10]    | [,11]   |
|-------|------------|-----------|----------|---------|
| [,8]  | 0.299804   | -0.436098 | -0.01824 | -0.12775|
| [,9]  | 0.002518   | -0.045856 | -0.03467 | 0.05447 |
| [,10] | 0.120485   | -0.157846 | 0.07433  | -0.05880|
| [,11] | 0.747925   | -0.163236 | 0.01986  | -0.41334|
| [1]   | 0.478040   | -0.049348 | 0.09323  | -0.53105|
| [2]   | 0.299242   | 0.009448  | 0.06818  | -0.25804|
|    | 0.530971 | 0.010136 | 0.16226 | -0.44020 |
|----|----------|----------|---------|----------|
| 8. | 1.000000 | -0.097515| 0.11178 | -0.80751 |
| 9. | -0.097515| 1.000000 | 0.15959 | 0.15053  |
|10. | 0.111781 | 0.159591 | 1.00000 | -0.03972 |
|11. | -0.807506| 0.150530 | -0.03972| 1.00000  |