LATTICE DECOMPOSITIONS OF MULTIVARIATE PROBABILITY DENSITY FUNCTIONS EXTENDING THE ICA MODEL TO INCORPORATE MORE GENERAL SOURCE STRUCTURES AND DEPENDENCIES

Juan K. Lin

Department of Statistics
Rutgers University
Piscataway, NJ 08854

ABSTRACT

The classical ICA model assumes that observations are all linear combinations of statistically independent scalar sources. This, as well as prior assumptions on the number of sources and their distributions are often seen as the weakest aspects of the ICA source model. In this paper, we present the mathematical structure necessary for extending ICA to more flexible models of real-world data.

1. INTRODUCTION

Applications of ICA to real-world data such as MRI data have often noted that while their algorithms find some scalar components that are visually and intuitively independent, the remaining components are not as well separated. From a generating model perspective, starting from linear transformations of statistically independent scalar sources, subsequent non-linear transformations or projection transformations will almost surely result in subspaces with dimensionality greater than one which cannot be linearly separated by ICA. The notion of higher dimensional "vector" sources were introduced by Cardoso and Lin, and was also addressed by Hyvarinan et. al. [1-3]. The information geometry of stochastic interaction decompositions was presented in Amari [4].

From a source modeling perspective, we seek to address the basic question of whether or not sources are fundamental. Can the same observations be generated by different sets of sources? More generally, can a joint probability density function be equally well described by different sets of conditional independencies? Or is there a optimal source-interaction structure?

In the field of Graphical models and probabilistic networks, causal relationships between random variables are inferred through the conditional independence structure in their joint probability density function. This structure is represented as an undirected graph over a set of nodes mapping to the random variables, and is manifested in the decomposition of the joint density into a product of lower dimensional functions. In this paper we explicitly focus on the decomposition of multivariate functions into a product of lower dimensional functions in a lattice theoretic framework (see e.g. Stanley [5]; Birkhoff [6]). In a probabilistic setting, conditional independence is found to be insufficient for fully describing the decomposition. Characterizations of higher order interactions between the random variables beyond conditional independence is discussed. We generalize some decomposition results for categorical random variables (see e.g. Cowell, Dawid, Lauritzen and Spiegelhalter [7]; Bishop, Fienberg and Holland [8]) to continuous random variables, and explicitly describe the basic building blocks (meet irreducibles) of the decomposition structure. Intuitively appealing results are presented for logically combining models based on the interactions present in both models, thereby proving that there is an optimal interaction model for continuous random variables in a fixed coordinate frame. We conclude with extensions to affine transformations of continuous random variables, and its implications for sources modeling.

An outline of this paper is as follows. Section two defines the antichain structure of the decompositions being investigated. Sections three and four describe a lattice of antichains and derives the meet irreducible decomposition of an antichain. These sections can be skimmed on first reading. Much of the motivation and intuition behind the described decomposition is captured in Figure 1. The main functional decomposition results are contained in section five. An information propagation interpretation of functional decomposition is presented in section six, along with some examples and extensions to affine transformations.

2. DECOMPOSITIONS OF INTEREST

Let $S = \{x_1, ..., x_n\}$ be a set of n variables, and let $A = \{a_1, ..., a_l\}$, $a_i \subset S$, be a set of subsets of S. In shorthand notation, we denote a function of the variables as

$$f(x_1, ..., x_n) = f(S).$$

From a probabilistic perspective, we are interested in decompositions of the joint probability density function into a product of functions of subsets of the variables. For simplicity, we will consider additive decompositions of the form

$$f(S) = \sum_{i=1}^{l} \alpha_i(a_i),$$

for some functions of subsets $\alpha_i(a_i)$. Thus, in a probabilistic setting, this would correspond to an additive decomposition of the log–likelihood function.

In general, the functions α_i are not unique. Furthermore, if $a_i \subset a_j$, we can simplify the decomposition by subsuming $\alpha_i(a_i)$and $\alpha_j(a_j)$ into one function:

$$\tilde{\alpha}_j(a_j) = \alpha_i(a_i) + \alpha_j(a_j).$$

We therefore restrict ourselves to a set of subsets $A = \{a_1, ..., a_l\}$, where $a_i \not\subseteq a_j$ if $i \neq j$. In a lattice theoretic description, A is an antichain in the lattice of subsets of S ordered by inclusion, often denoted B_n (Stanley 1986, Birkhoff 1967).

3. LATTICE OF ANTICHAINS

Recall that a lattice consists of a set endowed with a greatest lower bound (meet) and least upper bound (join) for any pair of elements. In order to be able to combine different ways of decomposing functions, we need to define a proper lattice of antichains. This lattice is isomorphic to the lattice of order ideals of B_n, where the mapping simply takes an antichain to the order ideal generated by the elements of the antichain.

For completeness, we will define the lattice of antichains explicitly. Again, let $S = \{x_1, ..., x_n\}$ be a set of n elements.

Definition 3.1 *Given $A = \{a_1, a_2, ...a_l\}$, and $B = \{b_1, b_2, ...b_m\}$, $a_i, b_i \subseteq A$, two sets of subsets of S. Let $\lceil A \rceil$ and $\lfloor A \rfloor$ be, respectively, the antichains consisting of the set of maximal and minimal elements of A ordered by inclusion. Denote*

$$A \bigcup B = \{a_1, a_2, ..., a_l, b_1, b_2, ..., b_m\}$$

$$A \bigcap_{\odot} B = \{a_i \cap b_j : a_i \in A, b_j \in B\}$$

Definition 3.2 *Let A, and B be two antichains in B_n, the lattice of subsets of S. Define the meet $\wedge$ and join $\vee$ of two antichains as follows:*

$$A \vee B = \lceil A \bigcup B \rceil$$

$$A \wedge B = \lceil A \bigcap_{\odot} B \rceil.$$

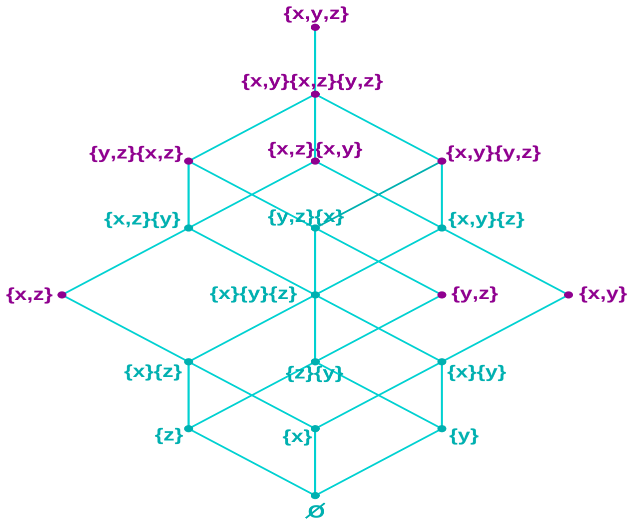

Fig. 1. Lattice of antichains in B_3. The elements labeled in darker text correspond to the meet irreducibles of the lattice. Notice that all elements in the lattice can be decomposed into the meet of meet irreducibles.

It is readily verified that The set of antichains of B_n together with the given defined meet and join operations define a lattice. Recall that the definitions of the meet and join define a partial ordering of the antichains by setting $A \leq B$ *if and only if* $A \wedge B = A$. The lattice of antichains in the lattice of subsets of three elements B_3 is shown in Figure 1. Each element in the lattice of antichains corresponds to a class of functions of three variables which admit a specific "dependency" decomposition. Our goal is to be able to combine different classes using lattice meets and joins.

4. MEET IRREDUCIBLES OF THE ANTICHAIN LATTICE

Definition 4.1 *Given an antichain A, let $\varphi(A) = \lfloor \{w \subseteq S : w \not\subseteq a_i \ for\ any\ a_i \in A\} \rfloor$ be the antichain consisting of the set of minimal elements of the set of all elements in the lattice B_n that are not less than or equal to an element in the antichain A.*

In other words, $\varphi(A)$ is constructed by first taking all elements in B_n that are above the antichain A, and then taking the set of minimal elements.

Lemma 4.2 *Given antichains $A = \{a_1, ..., a_l\}$ and $B = \{b_1, ..., b_k\}$, Suppose $B \leq A$, then for any b_i, there exists an element $a_k \in A$ such that $b_i \subseteq a_k$.*

Proof. Since B is an antichain, $b_i \not\subseteq b_j$ for any $j \neq i$. So $b_j \cap a_k \neq b_i$ for all $j \neq i$. So b_i must be contained in a_k, $b_i \subseteq a_k$ for some k.

Lemma 4.3 *Given an antichain A, and an element $w \in \varphi(A)$, any proper subset $c \subset w$ must be contained in some $a_i \in A$.*

Proof. Any proper subset c of an element $w \in \varphi(A)$ must be contained in some $a_i \in A$, otherwise, since $c \subset w$, w cannot be an element of the antichain $\varphi(A)$.

Lemma 4.4 *Given antichain A, and an element $w \in \varphi(A)$, $A \vee w$ covers A, $A \vee w \succ A$. That is, $A \vee w$ is immediately above A.*

Proof. Let $A = \{a_1, ..., a_l\}$, and without loss of generality, suppose $A \vee w = \{a_1, ..., a_k, w\}$, where $k \leq l$. So $w \not\subseteq a_1, ..., a_k$, $a_1, ..., a_k \not\subseteq w$, and $a_1, ..., a_k \subset w$ – they have to be proper subset of w since $w \in \varphi(A)$. The special case where $A \vee w = w$ also follows from this proof. Suppose there exists an antichain C such that $A \leq C \leq A \vee w$. From lemma, there exists an element $c_1 \in C$ such that $a_1 \subseteq c_1$. (without loss of generality, relabel the elements of C) Since $C < A \vee w$, applying the lemma again, there must be an element $d \in A \vee w$ such that $c_1 \subseteq d$. Combining, $a_1 \subseteq c_1 \subseteq d$, and since $a_1 \not\subseteq w$, d can only be a_1. Thus $c_1 = a_1$. Proceeding in the same way, we must require $c_2 = a_2, ..., c_k = a_k$. Continuing, there must be an element of $c_{k+1} \in C$ such that $a_{k+1} \subseteq c_{k+1}$ On the other side, $c_{k+1} \subseteq e$ for some element $e \in A \vee w$. The only possibility is $e = w$, resulting in the relation $a_{k+1} \subseteq c_{k+1} \subseteq w$. Thus C must contain all the elements $a_1, ..., a_k$. Any subsequent elements in C must be contained in w. There are two cases at this point.

Case 1 $c_{k+1} = w$.
Suppose there exists another element $c_{k+2} \in C$. Since all the elements of $A \vee w$ already appear in C, this is impossible since C is an antichain. So C can only have $k + 1$ elements, and $C = A \vee w$ in this case.

Case 2 $c_{k+1} \subset w$.
If c_{k+1} is a proper subset of w, by lemma 4.3, $c_{k+1} \subseteq g$ for some $g \in A$. Since c_{k+1} already contains a_{k+1}, and A is an antichain, we must have $g = a_{k+1}$, and consequently, $c_{k+1} = a_{k+1}$. Continuing, there must be an element $c_{k+2} \in C$ such that $a_{k+2} \subseteq c_{k+2} \subseteq w$ (must be w, C already contains $a_1, ..., a_k$). Since $c_{k+1} \subset w$, c_{k+2} cannot be equal to w, so $c_{k+2} \subset w$. Again, applying lemma 4.3, $c_{k+2} = a_{k+2}$. Proceeding, $c_{k+3} = a_{k+3}, ..., c_l = a_l$. So all the elements of the antichain A are in the antichain C. Now, there cannot be an element $c_{l+1} \in C$. This is seen as follows. First, we must have $c_{l+1} \subseteq w$. Second, $c_{l+1} \neq w$ since $a_{k+1}, ..., a_l \subset w$. So c_{l+1} must be a proper subset of w. By Lemma 4.3 c_{l+1} must be contained in an element in A, which is impossible since C is an antichain. Thus, there cannot be more than l elements in C, and we have $C = A$ in this case.

This proves that there are no antichains between A and $A \vee w$.

Lemma 4.5 *Given $\alpha \in \varphi(A)$. If $\beta \in \varphi(A) - \alpha$, then $\beta \in \varphi(A \vee \alpha)$.*

Proof. We prove this by contradiction. Assume there exists an element $\beta \in \varphi(A) - \alpha$ with $\beta \notin \varphi(A \vee \alpha)$. Since $\beta \in \varphi(A) - \alpha$, $\beta \not\subseteq a_i$ for all $a_i \in A$. Since $\alpha \in \varphi(A)$, and $\varphi(A)$ is an antichain, $\beta \not\subseteq \alpha$. Therefore $\beta \not\subseteq w$ for all $w \in (A \vee \alpha)$. Since $\beta \notin \varphi(A \vee \alpha)$, there must be a subset $\gamma \subset \beta$ such that $\gamma \in \varphi(A \vee \alpha)$. This means that γ is not contained in any element of $A \vee \alpha = \lceil a_1, ..., a_l, \alpha \rceil$. The elements of $\{a_1, ..., a_l\}$ which are not in the antichain $A \vee \alpha$ are contained in α, and since $\gamma \not\subseteq \alpha$, $\gamma \not\subseteq a_i$ for all $a_i \in A$. But since $\beta \in \varphi(A)$, we cannot have a subset $\gamma \subset \beta$ which is not contained in all the elements of A. This is the desired contradiction. The result of the lemma follows.

Lemma 4.6 *Given $\alpha \in \varphi(A)$. If $\beta \in \varphi(A \vee \alpha)$, then $\beta \in (\varphi(A) - \beta) \cup g_1 \cup ... \cup g_j$, where $\alpha \subset g_i$ for all g_i.*

Proof. Given $\beta \in \varphi(A \vee \alpha)$, this means, $\beta \not\subseteq a_1, ..., a_k, \alpha$, with $a_{k+1}, ..., a_l \subset \alpha$. This implies that $\beta \not\subseteq a_{k+1}, ..., a_l$, and hence there must exist a subset $\gamma \subset \beta$ such that $\gamma \in \varphi(A)$. Recall, $\alpha \in \varphi(A)$ also. Now we have two cases:

Case 1 $\gamma = \alpha$. So $\alpha \subseteq \beta$. Since $\beta \not\subseteq \alpha$, this implies that $\alpha \subset \beta$.

Case 2 $\gamma \neq \alpha$, so $\gamma \not\subseteq \alpha$. But since $\gamma \subseteq \beta$, and $\beta \in \varphi(A \vee \alpha)$, we must have $\gamma = \beta$, and so $\beta \in \varphi(A)$. Since $\beta \not\subseteq \alpha$, we must have $\beta \in \varphi(A) - \alpha$.

This proves our desired lemma.

These two lemmas establish that $\varphi(A \vee w)$ contains all the elements of $\varphi(A) - w$, and the only other elements it contains must be proper supersets of w.

Definition 4.7 *We will use the following notation. Given $w \in \varphi(A)$, with $w = \{w_1, ..., w_j\}$. Let $\tilde{w} = \{S - w_1, S - w_2, ..., S - w_j\}$. Let $\eta(A) = \{\tilde{w} | w \in \varphi(A)\}$. We will prove below that the elements of $\eta(A)$ are the meet irreducibles of the antichain A.*

From the previous two lemmas, we have $\eta(A \vee w) = \eta(A) - \tilde{w} + \{\tilde{g_j}\}$, where $\tilde{g_j} \wedge w = w$.

Lemma 4.8 *Given $w \in \varphi(A)$, with $w = \{w_1, ..., w_j\}$, $A \leq \tilde{w}$.*

Proof. For any $a_i \in A$, $w \not\subseteq a_i$, thus there exists an element $w_k \in w$ such that $w_k \notin a_i$. Thus $a_i \subseteq S - w_k$. Consequently, $A \wedge \tilde{w} = A$, so $A \leq \tilde{w}$.

Lemma 4.9 $(A \vee w) \wedge \tilde{w} = A$.

Proof. First we have $(A \vee w) \wedge \tilde{w} \leq A \vee w$. Second, no element in the antichain $\tilde{w}$ can contain all of w, so no element in $(A \vee w) \wedge \tilde{w}$ can contain all of w. Since $w \in A \vee w$, so $(A \vee w) \wedge \tilde{w} \neq A \vee w$, thus $(A \vee w) \wedge \tilde{w} < A \vee w$. Finally, since $\tilde{w} \geq A$ (from previous lemma), and $A \vee w > A$ (it actually covers A), thus $(A \vee w) \wedge \tilde{w} \geq A$ Combining, and using the result from lemma 4.4 that $A \vee w \succ A$, we necessarily have $(A \vee w) \wedge \tilde{w} = A$.

Theorem 4.10 *Given an antichain A, $A = \bigwedge \eta(A)$ for all $A \neq S$.*

Proof. Let $\mathcal{X} = \{antichain\ A | A \neq \bigwedge \eta(A)\}$, let $\mathcal{Y} = \lceil \mathcal{X} \rceil$, that is, the set of maximal elements in $\mathcal{X}$, which is an antichain in the lattice of antichains. Let $A \in \mathcal{Y}$. First, if $\varphi(A) = \emptyset$, then there do not exist any elements in the lattice B_n above the antichain A. This implies that $A = S$.

Since we are not allowing A to be the full set S, $\varphi(A)$ cannot be the empty set, so there exists an element $w \in \varphi(A)$. Consider the antichain $A \vee w$. Since $A \in \mathcal{Y}$, we must have $A \vee w = \bigwedge \eta(A \vee w)$.

By the lemmas 4.5, 4.6 and 4.9, we have

$$\begin{aligned} A &= (A \vee w) \wedge \tilde{w} \\ &= (\bigwedge \eta(A \vee w)) \wedge \tilde{w} \\ &= (\bigwedge \eta(A) - \tilde{w}) \bigwedge_j \tilde{g}_j \wedge \tilde{w} \\ &= (\bigwedge \eta(A) - \tilde{w}) \wedge \tilde{w} \\ &= \bigwedge \eta(A). \end{aligned}$$

This is the desired contradiction since $A \in \mathcal{Y}$, and A cannot be equal to $\bigwedge \eta(A)$. The result of the theorem follows. It is easy to see that elements of $\eta(A)$ are meet irreducibles in the lattice of antichains, thus $A = \bigwedge \eta(A)$ is the decomposition of any antichain into the meet of meet irreducibles.

5. FUNCTIONAL DECOMPOSITION

Here we present our main results on functional decomposition. For $S = \{x_1, ..., x_n\}$, we will use the shorthand notation $f(S) = f(x_1, ..., x_n)$. Given a specific point $\bar{\mathbf{x}} = (\bar{x_1}, ..., \bar{x_n})$, let $f(S|x_1) = f(x_1 = \bar{x_1}, x_2, ..., x_n)$, and more generally for $w \subseteq S$, $f(S|w)$ to be the function f, with the variables in w fixed to the corresponding values in the vector $\bar{\mathbf{x}}$.

Definition 5.1 *Given an antichain $A = \{a_1, ..., a_l\}$, let*

$$\mathcal{F}_A = \{f : \Re^n \to \Re \mid f(S) = \sum_{i=1}^{l} \alpha_i(a_i)\},$$

where $\alpha_i(a_i)$ can be any function of the subset of variables a_i. We will say that the function f admits a decomposition according to antichain A.

Theorem 5.2 *If $f \in \mathcal{F}_A$, and $A \leq B$, then $f \in \mathcal{F}_B$.*

Proof. Using the fact that $f \in \mathcal{F}_A$, we can write $f(S) = \sum_i \alpha_i(a_i)$. Since $A \leq B$, for every $a_i \in A$, there is an element $b_j \in B$, such that $a_i \subseteq b_j$. Now, by including the functions of the smaller subsets into a single function when needed

$$\beta_j(b_j) = \sum_{a_i \subseteq b_j; a_i \not\subseteq b_1, ..., b_{j-1}} \alpha_i(a_i),$$

we explicitly construct a functional decomposition of $f(S)$ in accordance with the antichain B.

Theorem 5.3 *If either $f \in \mathcal{F}_A$ or $f \in \mathcal{F}_{\mathcal{B}}$, then $f \in \mathcal{F}_{A \vee B}$.*

Proof. This follows from the previous theorem. An alternative version of this theorem is: Given $f \in \mathcal{F}_A$ and $g \in \mathcal{F}_{\mathcal{B}}$, then a linear superposition $af + bg \in \mathcal{F}_{A \vee B}$.

Theorem 5.4 *If $f \in \mathcal{F}_A$ and $f \in \mathcal{F}_{\mathcal{B}}$, then $f \in \mathcal{F}_{A \wedge B}$.*

First we need a lemma.

Lemma 5.5 *Given $f \in \mathcal{F}_A$ and $w \in \varphi(A)$,*

$$(-1)^{|w|} f(S|w) + (-1)^{|w|-1} \sum_{b \subseteq w, |b|=1} f(S|w-b) + ... + f(S) = 0.$$

Or, more compactly

$$\sum_{i=0}^{|w|} (-1)^i \sum_{b \subseteq w, |b|=i} f(S|w-b) = 0. \tag{1}$$

Proof. Since $A \leq \tilde{w}$, from Theorem 5.2, $f \in \mathcal{F}_{\tilde{w}}$. This means we can write

$$f(S) = \sum_{i=1}^{|w|} \alpha_i(S - w_i),$$

for some functions $\alpha_i(S - w_i)$. Using this expansion for all the terms in Eqn. 1, we see that every term $\alpha_i(S - w_i | w - b)$ with $w_i \not\subseteq b$ from $f(S|w-b)$ is paired with another identical term $\alpha_i(S - w_i | w - b + w_i)$ of the opposite sign from $f(S|w - b + w_i)$. Thus all the terms sum to zero. This equation states that the sum of the function, evaluated with alternating signs at all the vertices of the $|w|$–hypercube defined by the opposing vertices $\mathbf{x}$ and $\bar{\mathbf{x}}$ is equal to zero.

This lemma essentially describes how information propagates along the joint probability density function, and generalizes the equation $p(x, y) = p(x)p(y)$ when X and Y are independent random variables.

Proof. We are now ready to prove our main theorem. For every element in $\varphi(A)$ and $\varphi(B)$, there is an expansion of $f(S)$ in terms of the function itself evaluated at all the remaining vertices of a hypercube. Combining all the expansions of $f(S)$ from $\varphi(A)$ and $\varphi(B)$, and using our theorem that $A = \bigwedge \eta(A)$ for all $A \in \mathcal{A}, A \neq S$, we obtain explicitly the desired expansion of $f(S)$.
Finally, as a simple corollary to Theorem 5.4, we have the following important result.

Corollary 5.6 Let

$$\mathcal{M}(f) = \{antichain\ A \mid f \in \mathcal{F}_A\},$$

be the set of antichains for which the function f admits the corresponding decomposition. From the theorem, given any two antichain decompositions of a function f, we can decompose f according to the (more compact) meet of the two antichains. Thus, $\mathcal{M}(f) = \{antichain\ A \mid A \geq A_0\}$ is a *principal filter*, in other words, it consists of all antichains greater than or equal to a single antichain A_0. So we've proven the nice result that there is an optimal decomposition $\mathcal{F}_{A_0}$ of any function f.

6. INFORMATION PROPAGATION

As stated before, lemma 5.5 has a nice interpretation in terms of information propagation along the joint p.d.f. We present an example which elucidates the nature of higher order causal relations beyond conditional independence. Consider functions of three variables. Let $A = \{y,z\}\{x,z\}\{x,y\}$. We do not need to decompose the antichain A into the meet of meet irreducibles since A is already meet irreducible. In particular, $\varphi(A) = \{x,y,z\}$, and with $w = \{x,y,z\}$, Lemma 5.5 states that

$$\sum_{i=0}^{|w|}(-1)^i \sum_{b \subseteq w, |b|=i} f(S|w-b) = 0.$$

Expanding out the equation, we have

$$\begin{aligned} & f(x_0,y_0,z_0) \\ - \quad & f(x,y_0,z_0) - f(x_0,y,z_0) - f(x_0,y_0,z) \\ + \quad & f(x_0,y,z) + f(x,y_0,z) + f(x,y,z_0) \\ - \quad & f(x,y,z) \\ = \quad & 0, \end{aligned}$$

A geometric visualization of this information propagation equation is shown in Figure 2, where signs reflect respective weightings of the function evaluated at the vertices.

More generally, if the meet irreducible corresponds to a subset $w \subset S$ with m elements, the lemma defines an information propagation scheme along the vertices of an

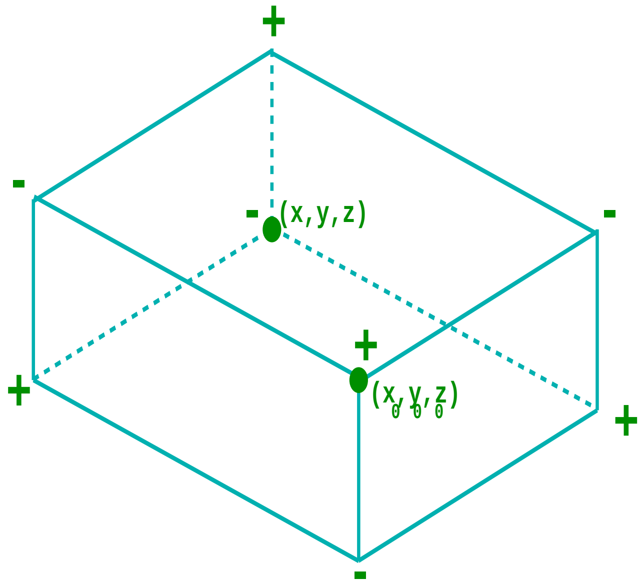

Fig. 2. Three dimensional information propagation along the vertices of a cube.

$m-hypercube$. Thus, in a probabilistic setting, the meet irreducibles in the lattice of antichains have an intuitive probabilistic interpretation. The set of meet irreducibles with corresponding $|w| = 1$ correspond to variables for which there is no dependence. The set of meet irreducibles with corresponding $|w| = 2$ lists the conditional independence structure contained in the joint probability density function, and higher order causal structure are reflected in the meet irreducibles with corresponding $|w| > 2$.

Returning to our example, a function $f \in \mathcal{F}_A$, admits the decomposition $f(x,y,z) = \alpha_1(y,z)+\alpha_2(x,z)+\alpha_3(x,y)$. In a probabilistic setting, the corresponding product decomposition is

$$p(x,y,z) = exp[f(x,y,z)] = p_1(y,z)p_2(x,z)p_3(x,y).$$

This decomposition is the simplest example of a causal relationship which cannot be described by conditional independence. An example of a joint probability density function which has this higher order causal structure is the following multivariate normal distribution

$$p(x,y,z) \propto exp[-(y+z)^2 - (x+z)^2 - (x+y)^2].$$

The ICA reader will immediately suggest a coordinate transformation which will result in an independent source description of the random vector. This again brings up our original motivating question - how fundamental and unique are the sources? Sources are defined through statistical independence, and statistical independence is shown to consist of pairwise conditional independencies. We have already fully investigated how sets of generalized conditional

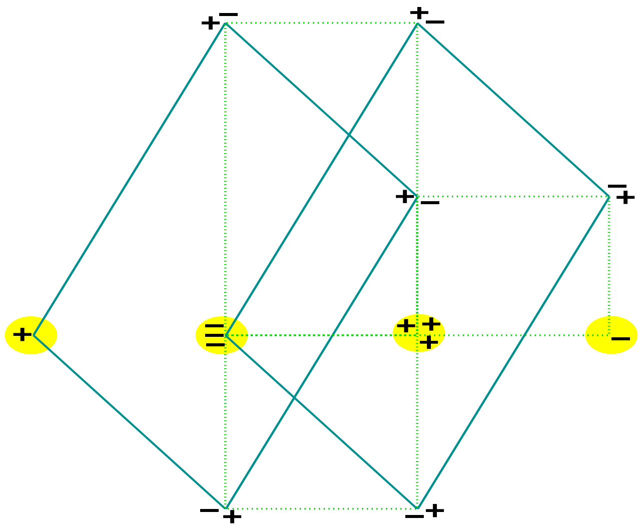

Fig. 3. Combining conditional independence decompositions along two pairs of directions in the same plane.

independences can be combined to yield the most optimal functional decomposition in a fixed coordinate system. Now the challenge is to be able to combine generalized conditional independencies in various linearly related coordinate systems.

The covariant aspect of the information propagation hypercubes means that when linearly transformed, the information will simply propagate along oblique hypercubes. In Figure 3, we geometrically combine conditional indendence structures in two dimensions along two linear coordinate systems, with a non-singular transformation relating the two. Without loss of generality, one coordinate system is taken to be the orthogonal coordinate system. As seen in the figure, by proper alignment of the information propagation oblique and orthogonal quadrilaterals summed with alternating signs along the vertices, the coefficients of many vertices can be made to cancel. The combination of the two information propagation schemes results in the following one-dimensional information propatation equation (finite difference equation)

$$f(x,y) - 3f(x+a,y) + 3f(x+2a,y) - f(x+3a,y) = 0,$$

for any x, y and scale a. Thus, $f(x,y)$ has to be quadratic in x. Similar geometric construction and reasoning forces $f(x,y)$ to be quadratic in y as well. Since $f(x,y)$ is the log-likelihood function, normalization considerations force us to conclude that $p(x,y) = exp(f(x,y))$ is bivariate Gaussian. Thus we have a geometric proof of that if X, Y are independent, and $aX + bY, cX + dY$ are also independent, with $ad - bc \neq 0$, then both X and Y are normally distributed. (see e.g. Linnik and Ostrovskii [9])

7. DISCUSSION

The generalization of functional decomposition from the graphical description used in the graphical modeling community to a lattice description described in this paper is a natural and powerful one. Causal relationships between the random variables are described by an antichain structure, and are manifested in a set of ways information propagates for the joint probability density function. These correspond to basic building blocks (meet irreducibles) of the decomposition of the joint *p.d.f.* The final theorem and associated corollary contain the main results of this paper. It answers the question posed in an earlier work regarding uniqueness and optimality of "factorizations" of multivariate functions in a fixed coordinate system.

Affine transformations significantly complicates the issue since sources and interactions are effectively coupled. However, the linearly transformed information propagation structures will still be basic building blocks of the decomposition. A challenge to the ICA community will be to discover these generalized conditional independence structures from data, and to combine them into a source–interaction description of the data. We presented a geometric proof that two distinct conditional independence direction pairs in the same plane forces the density to be bivariate normal in those variables, however, the existence of an optimal source–interaction description remains an open question.

References

[1] J.-F. Cardoso, *Multidimensional independent component analysis* in Proceedings of ICASSP (1998).

[2] J.K. Lin, *Factorizing multivariate function classes*, Advances in Neural and Information Processing Systems **10** (1998), 563–569.

[3] Aapo Hyvrinen, Patrik O. Hoyer, and Mika Inki *Topographic Independent Component Analysis*, Neural Computation 2001 13: 1527-1558

[4] S. Amari, *Information Geometry on Hierarchical Decomposition of Stochastic Interactions*, submitted, IEEE Transactions on Information Theory. (1999)

[5] R. Stanley, *Enumerative Combinatorics*, Wadsworth and Brooks/Cole, Monterey, CA, 1986.

[6] G. Birkhoff, *Lattice Theory*, third edition. Amer. Math. Soc., Providence, RI, 1967.

[7] R. Cowell, A. Dawid, S. Lauritzen and D. Spiegelhalter, *Probabilistic Networks and Expert Systems*, Springer Verlag, 1999.

[8] Y. Bishop, S. Fienberg, and P. Holland *Discrete Multivariate Analysis: Theory and Practice* MIT Press, Cambridge, MA, 1975.

[9] J. Linnik and I. Ostrovskii, *Decompositions of Random Variables and Vectors*, Amer. Math. Soc., Providence, RI, 1977.