Mixed Markov Models

Artur Fridman
Applied Computer Science and Mathematics
Merck Research Laboratories
Rahway, NJ 07065
Arthur_Fridman@merck.com

July 21, 2002

Abstract

Markov random fields can encode complex probabilistic relationships involving multiple variables, and admit efficient procedures for probabilistic inference. However, from a knowledge engineering point of view, these models suffer from a serious limitation. The graph of a Markov field must connect all pairs of variables that are conditionally dependent even for a single choice of values of the other variables. This makes it hard to encode interactions that occur only in some (perhaps only one!) context and are absent in all others. Furthermore, the requirement that two variables be connected unless *always* conditionally independent may lead to excessively dense graphs, obscuring the independencies present among the variables and leading to computationally prohibitive inference algorithms.

Mumford [1, 2] proposed an alternative modeling framework where the graph needn't be rigid and completely determined a-priori. *Mixed Markov models* contain node-valued random variables that, when instantiated, augment the graph by a set of *transient* edges. A single joint probability distribution relates the values of regular and node-valued variables. In this paper, we study the analytical and computational properties of mixed Markov models. In particular, we show that positive mixed models have a local Markov property that is equivalent to their global factorization. We also describe a computationally efficient procedure for answering probabilistic queries in mixed Markov models.

Keywords: mixed Markov models, Markov random fields, graphical models, probabilistic inference, Gibbs sampling.

1 Introduction

Graphical models such as Markov random fields [3] and Bayesian networks [4] are powerful tools for representing complex multivariable distributions using the adjacency structure of a graph. A Markov field is a probability distribution on an undirected graph whose edges connect those variables that are directly dependent, i.e., remain dependent even after all other variables have been instantiated

Specifying a Markov field requires identifying in advance and connecting every pair of variables that could *ever* be conditionally dependent, even for a single choice of values of the other variables. This may lead to graphs that are excessively dense, hiding potentially relevant independencies

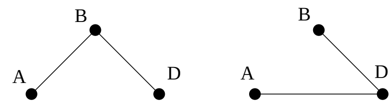

Figure 1: Left: when $X_C = 0$, X_D is independent of X_A given X_B. Right: when $X_C = 1$, X_D is directly dependent on both X_A and X_B.

from the human interpreter and rendering intractable those inference algorithms that are specified automatically once the structure of the graph is determined [5]. As a simple example of this limitation, consider the following gene regulatory network [6]:

> The protein product pA of gene A induces expression of gene B. Compound C, when present, binds to pA. The resulting dimer induces expression of gene D. Genes B and D are co-repressive (pB inhibits expression of D; pD inhibits expression of B).

Denote by X_A, X_B, and X_D the expression levels of the corresponding three genes, quantized into two levels, 1 (on) or 0 (off). The fourth variable, X_C, also binary, indicates if compound C is present.

Due to the stochastic nature of gene expression, the relationships "upregulates" or "inhibits" are not absolute [6]. For example, it is possible for X_D to be *off* when X_A is *on* and compound C is present. It is natural then to represent the interactions in this network as a stochastic process. Let us determine its graph. Suppose we learn that X_A is *on*. Would this help predict, say, X_D *if* we already knew the states of the other two variables? The answer, of course, depends on the actual value of X_C. If $X_C = 0$ (C is absent), A affects D only indirectly, through B, so given X_B, X_D is independent of X_A, cf. Fig. 1, left. If, however, $X_C = 1$ (C is present), the states of X_A and X_B are both predictive of that of X_D, cf. Fig. 1, right.

This behavior, when whether or not two variables are directly dependent is determined by the value of a third variable, is difficult to model within the limits imposed by the rigid topology of a Markov field. In fact, it is easy to verify that *every* pair of variables in this example is directly dependent, i.e.,

$$P(x_1, x_2|x_3, x_4) \neq P(x_1|x_3, x_4) \cdot P(x_2|x_3, x_4),^1$$

for any permutation (X_1, X_2, X_3, X_4) of (X_A, X_B, X_C, X_D) and at least one choice of values x_1 through x_4. Thus the graph of this network when modeled as a Markov field is *fully connected*!

The simple example above motivates us to consider graphical models in which the Markov boundary of a variable is at least in part random, determined by instantiating certain variables. These "context" variables have the ability to create direct dependencies between other variables when assigned certain values and destroy them when assigned others. A single joint probability distribution will relate the values of the "regular" and "context" variables. This is the basic setup of mixed Markov models [1, 7]. We describe them more formally in the following section.

2 Definition

Let $G = (V, E)$ be an undirected graph and $\mathbf{X}$ a random vector indexed by the vertices in V.[2] There are two types of variables in $\mathbf{X}$ corresponding to two types of nodes in V:

$$\mathbf{X} = (\mathbf{X}_I, \mathbf{X}_A).$$

[1] We will write $P(x_i|x_j)$ in place of $P(X_i = x_i|X_j = x_j)$.

[2] Random variables are capitalized: X_i, $\mathbf{X}$; their assignments are not: x_i, $\mathbf{x}$. Sets of random variables and their assignments are shown in bold face.

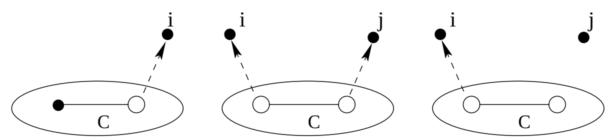

Figure 2: Mixed potential function V_C is pulling back the values of X_i (left), X_i and X_j (center), and X_i only (right).

The random variables in $\mathbf{X}_I = (X_i)_{i \in I}$ are the standard variables found in Markov fields; they may encode, for example, pixel intensities, degrees of belief, or gene expression levels. In this paper, we assume they are finite (hence discrete). The X_i's are called *regular variables*, and the vertices in I *regular nodes*.

The variables in $\mathbf{X}_A$, on the other hand, are not real valued; rather, they are pointers to regular nodes and take values in the set $I \cup \{\text{nil}\}$. The X_a's are called *address variables* or *pointers* and the vertices in A *address nodes*. Note that by definition, address variables cannot have address nodes as values; no pointers to pointers are allowed. On the other hand, there is no restriction on the graph topology: the edges in G can link any two nodes.

We will define the new model the same way Markov fields are defined, i.e., as a probability distribution that is factorizable into a product of local terms. The difference will come when defining what "local" means, in the presence of address variables.

Let $\mathfrak{C}$ denote the set of cliques[3] in G. A clique may contain both regular and address nodes.

Definition 2.1 A family of nonnegative functions

$$\{V_C : \ C \in \mathfrak{C}\}$$

is called a *mixed neighbor potential* if for any pair of configurations $\mathbf{x}$ and $\mathbf{y}$,

$$V_C(\mathbf{x}) = V_C(\mathbf{y})$$

whenever

$$\mathbf{x}_C = \mathbf{y}_C \text{ and } x_{x_a} = y_{y_a} \forall a \in A \cap C.$$

Thus a mixed potential function V_C depends only on the values of (a) variables within C, and (b) variables pointed to by an address variable within C. Compared to the potential functions comprising Markov fields, mixed potential functions V_C can be thought of as having extra slots in their argument list, up to the number of address nodes in C, and filling them in by "pulling back" the values of the regular variables pointed to from within C, cf. Fig. 2.

We can now define the new model as follows:

Definition 2.2 A probability distribution P on a graph $G = (I \cup A, E)$ is called a *mixed Markov model* if P is factorizable into a product of mixed potential functions over the cliques:

$$P(\mathbf{x}) = \alpha \prod_{C \in \mathfrak{C}} V_C(\mathbf{x}_C, \mathbf{x}_{\mathbf{x}_C}), \tag{1}$$

where $\mathbf{x}_{\mathbf{x}_C}$ is the vector of states of those regular variables pointed to from within C. (1) is a Gibbs representation of a mixed Markov model. Note that in the absence of address variables, (1) defines a classical Markov random field.

The gene regulatory network above has a natural representation as a mixed Markov model. Redefine X_C to be an address variable whose value indicates which gene is being directly regulated by

[3] A clique is a fully connected set of nodes.

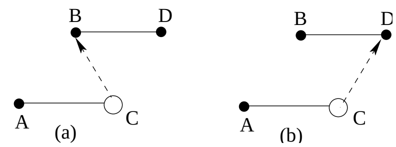

Figure 3: The gene regulatory network represented with a mixed model. (a) $X_C = B$, (b) $X_C = D$.

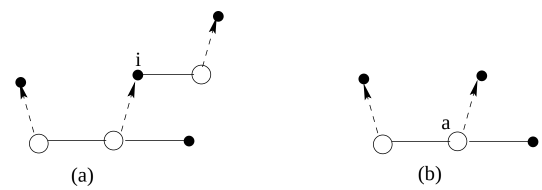

Figure 4: The Markov boundary of a regular variable (left) and address variable (right) in a mixed model.

gene A. The range of X_C is the union of two nodes B or D, cf. Fig. 3. The joint distribution is given by

$$P(x_A, x_B, x_C, x_D) = \alpha\ V_{AC}(x_A, x_C, x_{x_C}) V_{BD}(x_B, x_D),$$

with the mixed potential functions over the maximum cliques subsuming all others. The first factor, V_{AC}, encodes the constraint that $x_{x_C} = 1$ when $x_A = 1$; the second enforces $x_B = 0$ if and only if $x_D = 1$. Either constraint may be soft, of course.

3 Markov Property

While mixed models are more flexible than Markov fields from a knowledge engineering point of view, it is unclear if they admit efficient procedures for probabilistic inference. To that end, we will establish in this section the Markov property of mixed models and its equivalence to their Gibbs characterization (1). We will then use this equivalence to prove the computational feasibility of stochastic relaxation algorithms such as Gibbs sampling for inference in mixed models.

In Markov fields, the Markov property refers to the conditional distribution of any variable given all others (its local characteristic) being a *local* quantity, i.e., a function of only the neighbors of that variable [8]. For example, in a Markov chain, the local characteristic of the present is a function only of the immediate future and immediate past. The equivalence of the Markov property and the Gibbs characterization for positive Markov fields over arbitrary undirected graphs was first established by Hammersley and Clifford [9]; see also [10]. The following result generalizes the Hammersley-Clifford theorem to mixed Markov models.

Theorem 1 *A positive distribution P is a mixed Markov model with graph G if and only if*

(a) the local characteristic of a regular variable X_i,

$$P(x_i | \mathbf{x} \setminus x_i) \tag{2}$$

is a function only of

1. *X_i's neighbors,*
2. *variables pointed to by address variables among X_i's neighbors,*
3. *variables pointing to i, and their neighbors, and*

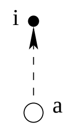

Figure 5: The local characteristic of an address variable in a mixed Markov model may not be a local quantity.

4. variables pointed to by address variables among those in 3, cf. Fig. 4a; and

(b) the normalized *local characteristic of an address variable X_a,*

$$\frac{P(x_a|\mathbf{x} \setminus x_a)}{P(nil|\mathbf{x} \setminus x_a)} \tag{3}$$

is a function only of

1. X_a's neighbors, and

2. variables pointed to by X_a and the address variables among X_a's neighbors, cf. Fig. 4b.

The normalization in (3) is necessary since the conditional distribution $P(x_a|\mathbf{x} \setminus x_a)$ may not be a local quantity. Consider, for example, a simple mixed model with two regular and one address variable and a graph as shown in Fig. 5. The joint distribution is given by

$$P(x_i, x_j; x_a) = \alpha\, V_a(x_a, x_{x_a}) V_i(x_i) V_j(x_j).$$

The local characteristic $P(x_a = i|x_i, x_j)$ depends on x_j as well as x_i, whereas the ratio $P(x_a = i|x_i, x_j)/P(x_a = \text{nil}|x_i, x_j)$ doesn't:

$$P(x_a = i|x_i, x_j) = \frac{V_a(i; x_i)}{V_a(i; x_i) + V_a(j; x_j) + V_a(\text{nil})}$$

$$\frac{P(x_a = i|x_i, x_j)}{P(x_a = \text{nil}|x_i, x_j)} = \frac{V_a(i; x_i)}{V_a(\text{nil})}$$

Fortunately, the locality of the normalized local characteristic suffices to make tractable probabilistic inference based on Gibbs sampling, as we show in the next section.

The only-if part of Theorem 1 determines the mechanism by which address variables enable context-specific relationships in a mixed model. Consider what happens to the graph when an address variable is instantiated.

Theorem 2 *Let P be a mixed Markov model. The conditional distribution $P(\mathbf{X}_{V-a} \,|\, X_a = x_a)$ is again a mixed Markov model, with a graph that is derived from that of P by replacing each edge $<v, a>$ incident to node a with the edge $<v, x_a>$.*

Assigning a value to an address variable X_a causes the edges incident to it to be pushed along the transient directed edge (a, x_a), cf. Fig. 6. Consequently, the overall connectivity of the graph does not increase: if the graph of a mixed model is sparse to begin with, it will remain sparse and thus interpretable and computable after some, or all, of the address variables are instantiated. Recall, for example, the regulatory network of Section 1. When the joint distribution is modeled as a Markov field, its graph is fully connected, cf. Fig. 7, upper left. When address variable X_C is assigned a value, say, B, the graph of the conditional distribution $P(X_A, X_B, X_D|X_C = B)$ is, according to a generic conditioning algorithm for Markov random fields[4], a fully connected graph

[4] which is to remove from the graph the instantiated node and all edges incident to it.

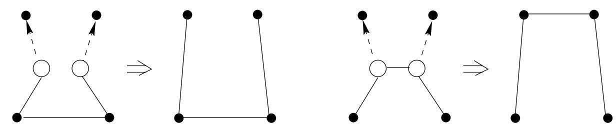

Figure 6: The graph of a mixed Markov model conditioned on the values of address variables is obtained by pushing the incident edges forward, over the transient directed edges.

on the remaining variables (Fig. 7, upper right).

On the other hand, suppose the joint distribution is represented with a mixed model (Fig. 7, lower left). Theorem 2 tells us how to obtain the graph of the conditional distribution $P(X_A, X_B, X_D|X_C = B)$ (Fig. 7, lower right). We know this distribution is a Markov field since the only address variable has been instantiated. Thus we have found a sparser graph that nonetheless respects the distribution $P(X_A, X_B, X_D|X_C = B)$.

4 Probabilistic Inference

Once a graphical model is defined or estimated, a procedure must be identified for obtaining quantities of interest, e.g., posterior marginals on the unobserved variables, highly likely states, or the probability of the observed data. In Markov fields, hence in mixed models as well, exact computation of these quantities is in general intractable [11]. An alternative is to seek approximate solutions by a stochastic sampling procedure such as a Gibbs sampler [12] below:

Let P be a mixed model (or a Markov field or a Bayesian network) with a graph $G = (V, E)$. Let $s : \{1, \ldots, |V|\} \to V$ be an enumeration of the nodes (often referred to as a visiting scheme). Let $v = s(1)$ be the first node in the visiting scheme. Select an initial configuration $\mathbf{x}$.

1) sample from the local characteristic $P(X_v|\mathbf{x}_{V\setminus v})$
2) if y_v is the state drawn, set $x_v \leftarrow y_v$
3) set v to be the next vertex in the visiting scheme, or $s(1)$ if the current node is last in the visiting scheme
4) return to step 1.

By simulating a Markov chain that converges to the joint distribution P, Gibbs sampler gives approximate marginals of P or, by clamping observed variables to their observed values, approximate posterior marginals on the unobserved variables. Simple modification of Gibbs sampler allow computing local and, at least theoretically, global maxima of P and its conditionals (Besag's ICM [13] and Simulated Annealing [12], respectively). It is unclear, however, if sampling from the local characteristic is at all tractable in the mixed model. After all, as we noted before, if X_a is an address variable, $P(x_a|\mathbf{x}_{V\setminus a})$ may not be a local quantity.

As it turns out, there is a way around this problem. Since X_a takes values in a subset of $I \cup \{\text{nil}\}$, we can think of the local characteristic $P(x_a|\mathbf{x}_{V\setminus a})$ as a vector with at most $|I| + 1$ components, each corresponding to a different state of X_a. In order to sample from the local characteristic, it suffices to know this vector up to a multiplicative constant. When this constant is $1/P(X_a = \text{nil}|\mathbf{x}_{V\setminus a})$, the ratios

$$\frac{P(X_a = x_a|\mathbf{x}_{V\setminus a})}{P(X_a = \text{nil}|\mathbf{x}_{V\setminus a})}$$

are in fact computable by Theorem 1. Assuming that each ratio can be evaluated in constant time, the computational complexity of sampling from the local characteristic of an address variable is $O(|I|)$.

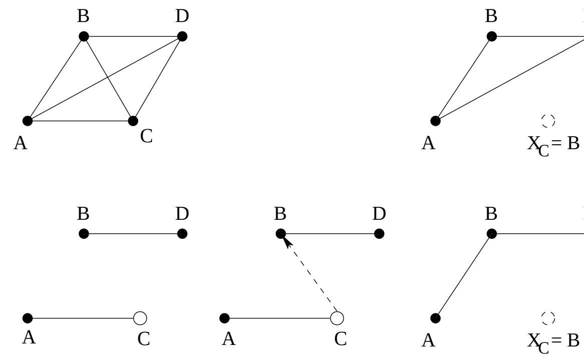

Figure 7: The graph of a sparse mixed model conditioned on an address variable remains sparse.

5 Examples

The flexibility of mixed Markov models is particularly useful when modeling global structures such as *templates* within Markov fields [1]. A template replaces the local statistics at its location with those specific for that template. As a simple example, consider n independent, identically distributed Bernoulli random variables $X_1, \ldots, X_n$ labeled by nodes 1 through n. Let $P(X_i = 1) = p$, $1 \leq i \leq n$. Now imagine a template A that, when placed at location i, changes the probability of 1 at that location from p to q. Let random variable X_A taking values in $I = \{1, \ldots, n\}$ specify the location of the template.

When modeled as a Markov field, the joint distribution of $(X_1, \ldots, X_n, X_A)$ has a star-shaped graph, cf. Fig. 8, left. The fact that the X_i's are independent given the template's location is easily inferred from the graph.

Now suppose there are two templates, their locations specified by a pair of random variables (X_A, X_B). Allow the templates to interact spatially, attracting or repelling each other or perhaps exhibiting a more complex pattern of interaction, encoded in a joint distribution $P(X_A, X_B)$. Let us further assume that the templates force the same values on the X_i's at their locations. I.e., if $X_A = i$ and $X_B = j$, then, with high probability, $X_i = X_j$. For example, if A and B are left and right eye templates, we would probably want to force the pixel colors at their locations in the image to be the same.

Now if we represent the joint distribution of $(X_1, \ldots, X_n, X_A, X_B)$ as a Markov field, we cannot do any better than a fully connected graph. In fact, X_i clearly depends directly on both X_A and X_B; furthermore, X_i and X_j are conditionally dependent on each other since

$$\begin{aligned} & P(X_i = 1 | X_j = 1, \mathbf{X}_{I \setminus \{i,j\}} = \mathbf{x}, X_A = i, X_B = j) \\ \ll\ & P(X_i = 1 | X_j = 0, \mathbf{X}_{I \setminus \{i,j\}} = \mathbf{x}, X_A = i, X_B = j) \end{aligned} \tag{4}$$

But in fact both sides of (4) are almost always equal, except for a single choice of values of X_A and X_B! This independence is completely lost in the graph in Fig. 8, center. On the other hand, the graph of the same process represented with a mixed model (Fig. 8, right) clearly shows that X_i and X_j, $i \neq j$, are conditionally independent except when each containing one of the templates. Furthermore, the set of (current) neighbors of each variable is easily obtainable from the graph by Theorem 1.

Adding more templates to the model is straightforward, spatial relationships between the templates and between each template and the variables at its location represented with a set of local constrains that are reflected in the graph of the mixed model. Flexible templates [14, 15] arise as a special case. One can also imagine having templates of templates, leading to mixed models with flexible hierarchical structure. See [7] for further discussion and more examples.

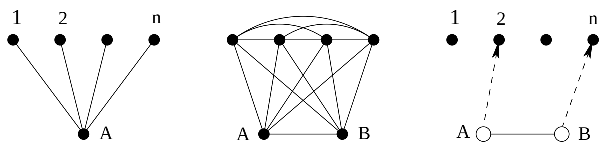

Figure 8: Left: a star-shaped graph adequately represents this one-template process. Center: when modeled as a Markov field, the graph of the two-template process is fully connected. Right: the same process represented with a mixed model.

6 Conclusion

In this paper, we have analyzed the analytical and computational properties of *mixed Markov models* [1, 2, 7]. In these graphical models, the graph needn't be rigid and completely determined a-priori. Instead, *address variables*, node-valued random variables, are allowed to modify the graph by augmenting it with a set of transient edges. A single joint probability distribution relates the values of regular and address variables.

The introduction of address variables allows sparser graphs that are more easily interpretable by human experts. Positive mixed Markov models have a local Markov property that is equivalent to their global factorization (Theorem 1). Sparse mixed models will remain sparse after some, or all, of the address variables are instantiated (Theorem 2). A Gibbs sampler [12] is a computationally efficient means of performing probabilistic inference in mixed Markov models.

Appendix: Proofs of Theorems

Theorem 1: We begin with some notation. $N(v)$ will denote the set of neighbors (adjacent vertices) of a node v; also, $\overline{N(v)} = N(v) \cup \{v\}$. For a configuration $\mathbf{x}$ and a set of vertices D, define

$$\begin{aligned} D_* &= \{i \in I : \ i = x_a \text{ for some } a \in A \cap D\} \\ D^* &= \{a \in A : \ x_a \in D\} \end{aligned}$$

D_* is the set of regular nodes pointed to by an address variable in D. D^* is the set of address nodes whose corresponding variables are pointing to a node in D.

Suppose first that P has the Gibbs characterization (1). The normalized local characteristic of an address variable X_a is

$$\begin{aligned} \frac{P(x_a|\mathbf{x} \setminus x_a)}{P(\text{nil}|\mathbf{x} \setminus x_a)} = \prod_{C:\ a \in C} \frac{V_C(x_a \mathbf{x}_{C \setminus a}; \mathbf{x}_{C_*})}{V_C(\text{nil}_a \mathbf{x}_{C \setminus a}; \mathbf{x}_{C_*})} \\ = f(\mathbf{x}_{N(a) \cup \overline{N(a)}_*}). \end{aligned}$$

In fact, if a node is not a neighbor of a, it cannot be in the same clique as a, and the only way it can then influence one of the terms in the product above is by being pointed to from within a clique containing a.

Next, the local characteristic of a regular variable X_i is

$$P(x_i|\mathbf{x} \setminus x_i) = \frac{P(x_i \mathbf{x}_{(I \cup A) \setminus i})}{\sum_{z_i} P(z_i \mathbf{x}_{(I \cup A) \setminus i})}.$$

Its reciprocal is equal to

$$\begin{aligned} \sum_{z_i} \prod_{C:\ i \in C} \frac{V_C(z_i \mathbf{x}_{C \setminus i}; \mathbf{x}_{C_*})}{V_C(x_i \mathbf{x}_{C \setminus i}; \mathbf{x}_{C_*})} \cdot \prod_{C:\ i \in C_*} \frac{V_C(\mathbf{x}_C; z_i \mathbf{x}_{C_* \setminus i})}{V_C(\mathbf{x}_C; x_i \mathbf{x}_{C_* \setminus i})} \\ = f(\mathbf{x}_{N(i) \cup N(i)_* \cup \overline{N(i^*)} \cup \overline{N(i^*)}_*}). \end{aligned}$$

Turning to the if part, fix a reference configuration $\mathbf{o} = (\mathbf{o}_I, \mathbf{nil}_A)$. By the Möbius inversion formula [8],

$$
\begin{aligned}
\frac{P(\mathbf{x})}{P(\mathbf{o})} &= \prod_{D \subset I \cup A} U_D(\mathbf{x}_D), \quad \text{where} \qquad (5)\\
U_D(\mathbf{x}_D) &= \prod_{E \subset D} \left[\frac{P(\mathbf{x}_E \mathbf{o}_{(I\cup A)\setminus E})}{P(\mathbf{o})} \right]^{(-1)^{|D\setminus E|}}
\end{aligned}
$$

We will construct mixed potential functions V_C on the cliques of G in such a way that $\prod_C V_C = \prod_D U_D$. Given a set $D \subset I \cup A$ and a configuration $\mathbf{x}_D$ on D, set

$$
C = D \setminus D_* \qquad (6)
$$

C contains all of the address nodes of D, plus those regular nodes in D *not* pointed to from within D. Suppose $U_D(\mathbf{x}_D) \neq 1$. We assert that

1. C is a clique
2. $C \subset D \subset C \cup C_*$ and $C \cap C_* = \emptyset$
3. C as given by (6) is the only clique satisfying condition 2.

Assume for the moment that 1–3 hold. Set the mixed potential functions as follows:

$$
V_C(\mathbf{x}_C; \mathbf{x}_{C_*}) = \begin{cases} \displaystyle\prod_{D:\ C \subset D \subset C \cup C_*} U_D(\mathbf{x}_D), & \text{if } C \cap C_* = \emptyset \\ 1, & \text{otherwise.} \end{cases}
$$

By exchanging the order of multiplication,

$$
\begin{aligned}
\prod_{C \in \mathfrak{C}} V_C(\mathbf{x}_C; \mathbf{x}_{C_*}) &= \prod_{D \subset I \cup A} \prod_{C \in \mathfrak{A}(D)} U_D(\mathbf{x}_D) \\
&= \prod_{D \subset I \cup A} U_D(\mathbf{x}_D),
\end{aligned}
$$

where $\mathfrak{A}(D) = \{C \in \mathfrak{C} : C \subset D \subset C \cup C_*,\ C \cap C_* = \emptyset\}$ contains a single clique, $D \setminus D_*$, whenever $U_D(\mathbf{x}_D) \neq 1$. From (5) and the last chain of equalities,

$$
P(\mathbf{x}) = P(\mathbf{o}) \prod_{C \in \mathfrak{C}} V_C(\mathbf{x}_C; \mathbf{x}_{C_*}),
$$

the Gibbs characterization of a mixed model, Q.E.D.

Of the three assertions, only claim 1 is non-trivial; it is also the only one requiring the Markov property. We must show that if $U_D(\mathbf{x}_D) \neq 1$ then any two nodes u and v in $D \setminus D_*$ are neighbors. Consider two cases. If at least one of the two vertices (say, u) is an address node then $U_D(\mathbf{x}_D) =$

$$
\prod_{E \subset D \setminus \{u,v\}} \left[\frac{P(x_u | x_v \mathbf{x}_E \mathbf{o}_{(I\cup A)\setminus(E\cup\{u,v\})})}{P(\mathrm{nil}_u | x_v \mathbf{x}_E \mathbf{o}_{(I\cup A)\setminus(E\cup\{u,v\})})} \Big/ \frac{P(x_u | o_v \mathbf{x}_E \mathbf{o}_{(I\cup A)\setminus(E\cup\{u,v\})})}{P(\mathrm{nil}_u | o_v \mathbf{x}_E \mathbf{o}_{(I\cup A)\setminus(E\cup\{u,v\})})} \right]^{(-1)^{|D-E|}} \qquad (7)
$$

By the Markov property of address variables, $U_D(\mathbf{x}_D)$ is 1 unless either $v \in N(u)$ or $v \in \overline{N(u)}_*$, meaning $v = x_a$ for some address node $a \in \overline{N(u)}$. But all address variables in (7) are set to nil,

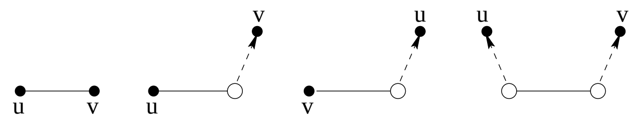

Figure 9: If $U_D(x_D) \neq 1$ and u and v two regular nodes in D, one of the possibilities above must hold.

except those in D. Hence $v \in D_*$, contradicting $v \in D \setminus D_*$.

Suppose now that u and v are both regular nodes. Express $U_D(\mathbf{x}_D)$ as in (7) with o_u in place of nil_u. The Markov property of regular variables gives us $U_D(\mathbf{x}_D) = 1$ unless $v \in N(u)$, or $v \in N(u)_*$, or $u \in N(v)_*$, or $u \in N(v^*)_*$, see Fig. 9.

In all but the first case, however, either u or v is the value of an address variable residing in D; otherwise it would be set to nil in (7). It follows that either u or v is in D_*, a contradiction. Thus $v \in N(u)$.

Claims 2 and 3 are simple set-theoretical exercises; both follow from the requirement that address variables cannot point to address nodes. □

Theorem 2: Let G' denote the modified graph in which all edges $<v, a>$ incident to a have been replaced with $<v, x_a>$. Let N and M be the neighborhood relations in G and G', respectively. We must show that the local characteristics of all regular variables and all address variables other than a are local with respect to M, in the sense of Theorem 1.

Let $b \neq a$ be an address node. Since P is a mixed model, the normalized local characteristic of X_b is a function of $x_{N(b) \cup \overline{N(b)}_*}$. Now $M(b) \cup \overline{M(b)}_* = N(b) \cup \overline{N(b)}_* \setminus \{a\}$. In fact, if $a \notin N(b)$, then $N(b) = M(b)$ and $\overline{N(b)}_* = \overline{M(b)}_*$ and neither set includes a. If $a \in N(b)$ then $\overline{N(b)}_*$ includes one extra regular node, x_a, compared to $\overline{M(b)}_*$; however, that node is now in $M(b)$, so the union of the two sets is still the same except for the node a.

That the local characteristic of a regular variable X_i is local with respect to M is shown in a similar fashion, by establishing $M(i) \cup M(i)_* \cup \overline{M(i^*)} \cup \overline{M(i^*)}_* = N(i) \cup N(i)_* \cup \overline{N(i^*)} \cup \overline{N(i^*)}_* \setminus \{a\}$. □

References

[1] D. Mumford, *The Statistical Description of Visual Signals*, ICIAM 95, Akademie Verlag, 1996

[2] D. Mumford, *The Mathematical Modeling of Cortical Functioning and Thought*, Proc. Norbert Wiener Centennial Conf., AMS, 1997

[3] R. Kindermann and J. Snell, *Markov Random Fields and Their Applications*, Contemporary Mathematics, vol. 1, AMS, Providence, RI, 1980

[4] F. Jensen, *An Introduction to Bayesian Networks*, University College London Press, London, 1996

[5] P. Smyth, *Belief Networks, Hidden Markov Models, and Markov Random Fields: a Unifying View*, Pattern Recognition Letters, 1998

[6] J. Hasty, D. McMillen, F. Isaacs, J. Collins, *Computational Studies of Gene Regulatory Networks:* In Numero *Molecular Biology*, Nature Genetics, **2**, 268–279, 2001

[7] A. Fridman, *Mixed Markov Fields*, Ph.D. Thesis, Brown University, 2000

[8] G. Winkler, *Image Analysis, Random Fields and Dynamic Monte Carlo Methods*, Springer, 1995

[9] J. Hammersley and P. Clifford, *Markov Fields on Finite Graphs and Lattices*, Preprint, Univ. of Cal., Berkeley, 1968

[10] G. Grimmett, *A Theorem about Markov Fields*, Bull. Lond. Math. Soc., **5**, 81–84, 1973

[11] S. Lauritzen, *Graphical Models*, Oxford University Press, Oxford, 1996

[12] S. Geman and D. Geman, *Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images*, IEEE Trans. PAMI **6**, 721–741, 1984

[13] J. Besag, *On the Statistical Analysis of Dirty Pictures (with discussion)*, J. R. Stat. Soc. B, **48**, 259–302, 1986

[14] M. Fischler and R. Elschlager, *The Representation and Matching of Pictorial Structures*, IEEE Trans. on Computers, **22**, 1973

[15] A. Yuille, *Generalized Deformable Templates, Statistical Physics, and Matching Problems*, Neural Computation, **2**, 1–24, 1990