Mixed Markov models

Arthur Fridman*

Applied Computer Science and Mathematics, Merck & Co., Inc., Rahway, NJ 07065

Communicated by David Mumford, Brown University, Providence, RI, March 30, 2003 (received for review July 21, 2002)

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 a certain 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 [Mumford, D. (1996) in *ICIAM 95*, eds. Kirchgassner, K., Marenholtz, O. & Mennicken, R. (Akademie Verlag, Berlin), pp. 233–256] proposed an alternative modeling framework where the graph need not 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 article, 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.

Graphical models such as Markov random fields (1) and Bayesian networks (2) are powerful tools for representing complex multivariate 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 from the human interpreter and rendering intractable those inference algorithms that are specified automatically once the structure of the graph is determined (3). As a simple example of this limitation, consider the following gene regulatory network (4).

> The protein product pA of gene A induces expression of gene B. Compound C, when present, modifies pA. The modified protein is unable to regulate B directly but can induce expression of gene D. Genes B and D are corepressive (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). Another binary variable X_C will indicate whether compound C is present.

Due to the stochastic nature of gene expression, the relationships "upregulates" or "inhibits" are not absolute (4). For example, it is possible for gene D not to be expressed even when A is upregulated and 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 $X_A = 1$, i.e., gene A is expressed. 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 $X_C = 0$ (compound C is absent), X_A affects X_D only indirectly, through X_B. In other words, given X_B, X_D is independent of X_A (see 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 (see Fig. 1 *Right*).

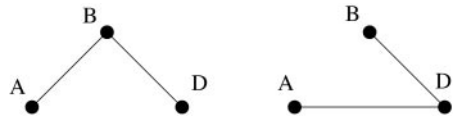

Fig. 1. (*Left*) When $X_C = 0$, X_D is independent of X_A given X_B. (*Right*) When $X_C = 1$, X_D is directly dependent on both X_A and X_B.

This behavior, when whether or not two variables are directly dependent is determined by the value of a third variable, is difficult to capture 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.,

$$P(x_1, x_2|x_3, x_4) \neq P(x_1|x_3, x_4)\cdot P(x_2|x_3, x_4),^{\dagger}$$

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 (5, 6). We describe them more formally in the following section.

Definition

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

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

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 article, 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

*E-mail: arthur_fridman@merck.com.

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

‡Random variables are indicated by a capital letter: X_i, $\mathbf{X}$, and their values by the same letter in lower case: x_i, $\mathbf{x}$. Sets of random variables and their values are shown in boldface.

www.pnas.org/cgi/doi/10.1073/pnas.0731829100

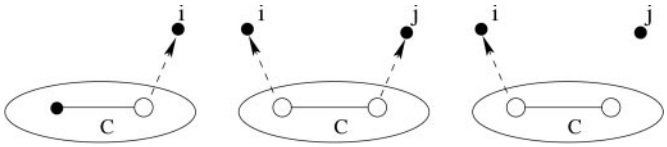

Fig. 2. Mixed interaction function V_C pulling back the values of X_i (*Left*), X_i and X_j (*Center*), and X_i only (*Right*).

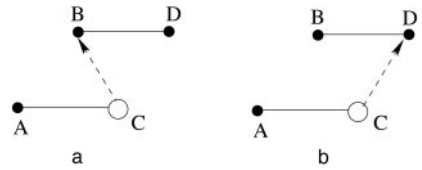

Fig. 3. The gene regulatory network represented with a mixed model. (*a*) $X_C = B$. (*b*) $X_C = D$.

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 $\mathbf{C}$ denote the set of cliques[§] in G and C a clique in $\mathbf{C}$. Note that cliques may contain both regular and address nodes.

Definition 2.1: A nonnegative function V_C on $\mathbf{X}$ is a *mixed interaction function* if for any configurations $\mathbf{x}$ and $\mathbf{y}$,

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

whenever

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

Thus V_C depends only on the values of (*i*) variables within C, and (*ii*) variables pointed to by an address variable within C. Mixed interaction functions can be thought of as having extra slots in their argument list, up to the number of address nodes in the corresponding clique, and filling them in by "pulling back" the values of the regular variables pointed to from within that clique (see 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 interaction functions over the cliques

$$P(\mathbf{x}) = \alpha \prod_{C \in \mathbf{C}} V_C(\mathbf{x}_C, \mathbf{x}_{\mathbf{x}_C}), \qquad [1]$$

where $\mathbf{x}_{\mathbf{x}_C}$ is the vector of states of those regular variables pointed to from within C. Note that in the absence of address variables, Eq. **1** defines a classical Markov field with neighbor potential $\{-\log V_C: C \in \mathbf{C}\}$ (7).

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 gene A. The range of X_C is the union of two nodes B or D (see 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 mixed interaction 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.

Markov Property

While mixed models are more flexible than Markov fields from a knowledge engineering point of view, it is unclear whether 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 (Eq. **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 term Markov property refers to the fact that the conditional distribution of any variable given all others (its *local characteristic*) is a local quantity, i.e., a function of only the neighbors of that variable (7). 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 (ref. 9; see also ref. 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*

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

$$P(X_i|\mathbf{x}\backslash x_i) \qquad [2]$$

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*
4. *variables pointed to by address variables among those in 3 (see Fig. 4*a*): and*

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

$$\frac{P(X_a|\mathbf{x}\backslash x_a)}{P(X_a = \text{nil}|\mathbf{x}\backslash x_a)} \qquad [3]$$

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 (see Fig.* 4b).

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

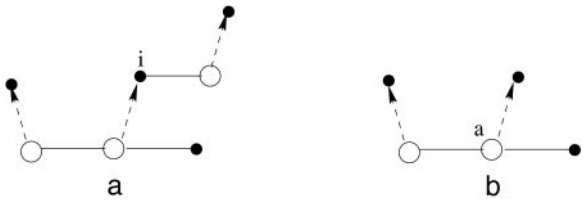

Fig. 4. The Markov boundary of a regular variable (*a*) and address variable (*b*) in a mixed model.

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

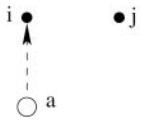

Fig. 5. The local characteristic of an address variable in a mixed Markov model may not be a local quantity.

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

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 is sufficient 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 model, with a graph that is derived from that of P by replacing each edge $\langle v, a\rangle$ incident to node a with the edge $\langle v, x_a\rangle$.*

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) (see 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 described in the Introduction. When the joint distribution is modeled as a Markov field, its graph is fully connected (see Fig. 7 *Upper Left*). When address variable X_C is assigned a value, say, B, the graph of the conditional distribution $P(X_A, X_B, X_D|X_! = B)$ is, according to a generic conditioning algorithm for Markov fields,[¶] a fully connected graph 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 $P(X_A, X_B, X_D|X_C = B)$ (Fig. 7 *Lower Right*). We know this distribution is a Markov field because the only address variable has been instantiated. Thus we have found a sparser graph that respects the distribution $P(X_A, X_B, X_D|X_C = B)$.

[¶]A generic conditioning algorithm for Markov fields is to remove from the graph the instantiated node and all edges incident to it.

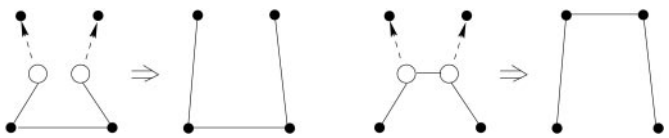

Fig. 6. The graph of a mixed Markov model conditioned on the values of address variables is obtained by pushing the incident edges forward, along the transient directed edges.

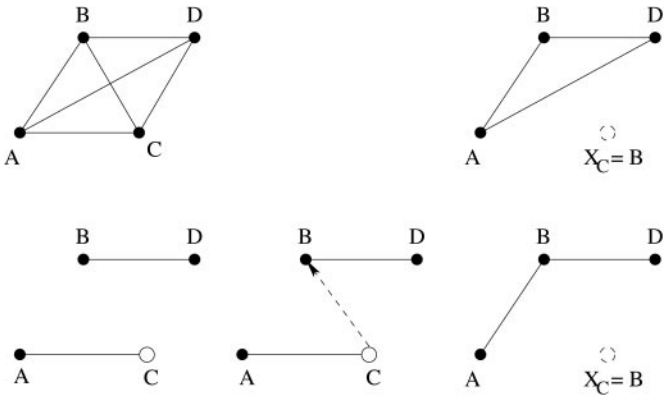

Fig. 7. The graph of a sparse mixed model conditioned on an address variable remains sparse.

It is instructive to compare mixed models and mixture models such as Bayesian multinets (10). A mixed model can be regarded as a mixture of Markov fields, with address variables acting as latent variables:

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

Note that the mixture components of a mixed model need not have the same graph. In this respect, mixed models are similar to Bayesian multinets. However, the mixture components of a multinet may have arbitrary, unrelated graphs. In contrast, the graphs of the mixture components of a mixed model are all related and are obtained from the graph of the joint distribution by the algorithm in *Theorem 2.* This compromise permits the existence of context-dependent relationships while simplifying the modeling and machine-learning tasks.

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 Bayes 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 iterated conditional mode (ICM; ref. 13) and simulated annealing (12), respectively]. It is not clear, however,

[‖]In the case of the mixed model, a visiting scheme will include both 1 and address nodes. Gibbs sampler does not differentiate between the two types of nodes.

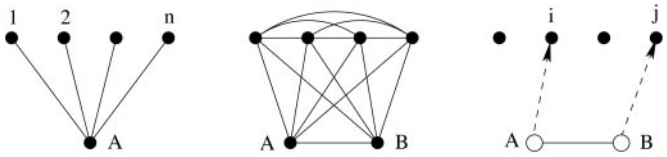

Fig. 8. (*Left*) A star-shaped graph adequately represents this single-template process. (*Center*) When modeled as a Markov field, the graph of the two-template process is fully connected. (*Right*) The same process represented with a mixed model.

whether in the mixed model, the sampling from a local characteristic required in step 1 is computationally feasible, the way it is in a Markov field or a Bayes network. 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 complication. 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 $O(|I|)$.

Examples

The flexibility of mixed Markov models is particularly useful when modeling global structures such as *templates* (5) within Markov fields. 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 $P(X_i = 1) = p$, $1 \le i \le n$. 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 (see 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 because

$$P(X_i = 1|X_j = 1, \mathbf{X}_{I\setminus\{i,j\}} = \mathbf{x}, X_A = i, X_B = j) \gg P(X_i = 1|X_j = 0, \mathbf{X}_{I\setminus\{i,j\}} = \mathbf{x}, X_A = i, X_B = j). \quad [4]$$

But in fact both sides of Eq. **4** are almost always equal: $X_i \perp X_j|(X_A, X_B)$ unless $X_A = i$ and $X_B = j$ or $X_A = j$ and $X_B = i$. This independence is completely lost in the graph in Fig. 8 *Center.* On the other hand, when the same process is represented as a mixed model (Fig. 8 *Right*), one can see that X_i and X_j are conditionally independent except when both pointed to from within the clique $\{A, B\}$. In general, the neighbors of any variable can be easily identified from this graph by using *Theorem 1.*

The same framework can be used to design models containing multiple templates, each possibly consisting of several interacting elements. Flexible templates (14, 15) arise as a special case. One can also imagine having templates of templates, leading to mixed models that have a flexible hierarchical structure such as multilayer texton maps (16). In every case, spatial interactions between templates, between the elements of a template, and between a template and its location are encoded with an sparse set of local constraints reflected in the graph of the mixed model.

Conclusion

In this article, we have analyzed the analytical and computational properties of mixed Markov models (5, 6). In these graphical models, the graph needn't be rigid and completely determined *a priori.* Instead, node-valued *address 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 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 have been instantiated (*Theorem 2*). A Gibbs sampler (12) is a computationally efficient means of performing probabilistic inference in mixed Markov models.

Appendix: Proofs of Theorems

Proof of Theorem 1*:* We begin with some notation. Let $N(v)$ denote the set of neighbors of a node v (i.e., those nodes incident to v); also, let $\overline{N(v)} = N(v) \cup \{v\}$. For a configuration $\mathbf{x}$ and a set of vertices D, define

$$D_* = \{i \in I: i = x_a \text{ for some } a \in A \cap D\},$$

$$D^* = \{a \in A: x_a \in D\}.$$

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

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

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

$$\sum_{z_i} \prod_{C:\, i\in C} \frac{V_C(z_i\mathbf{x}_{C\setminus i};\, \mathbf{x}_{C_*})}{V_C(x_i\mathbf{x}_{C\setminus i};\, \mathbf{x}_{C_*})} \cdot \prod_{C:\, i\in C_*} \frac{V_C(\mathbf{x}_C;\, z_i\mathbf{x}_{C_*\setminus i})}{V_C(\mathbf{x}_C;\, x_i\mathbf{x}_{C_*\setminus i})}$$

$$= f(\mathbf{x}_{N(i)\cup N(i)_*\cup\overline{N(i^*)}\cup\overline{N(i^*)}_*}).$$

Turning to the *if* part, fix a reference configuration $\mathbf{o} = (\mathbf{o}_I, \mathbf{nil}_A)$. By the Möbius inversion formula (7),

$$\frac{P(\mathbf{x})}{P(\mathbf{o})} = \prod_{D\subset I\cup A} U_D(\mathbf{x}_D), \quad \text{where} \qquad \textbf{[5]}$$

$$U_D(\mathbf{x}_D) = \prod_{E\subset D} \left(\frac{P(\mathbf{x}_E\mathbf{o}_{(I\cup A)\setminus E})}{P(\mathbf{o})}\right)^{(-1)^{|D\setminus E|}}$$

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

$$C = D\backslash D_* \qquad \textbf{[6]}$$

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 Eq. **6** is the only clique satisfying condition 2.

Assume for the moment that 1–3 hold. Set the mixed interaction 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,

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

$$= \prod_{D\subset I\cup A} U_D(\mathbf{x}_D),$$

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

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

the Gibbs characterization of a mixed model, q.E.d.

Of the three assertions, only claim 1 is nontrivial; it is also the only one requiring the Markov property. We must show that if

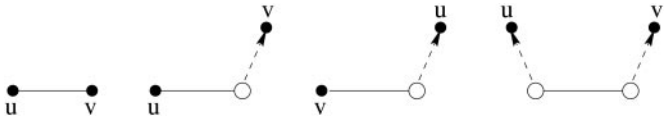

Fig. 9. If $U_D(x_D) \neq 1$ and u and v two regular nodes in D, one of the possibilities above must hold.

$U_D(\mathbf{x}_D) \neq 1$ then any two nodes u and v in $D\backslash D_*$ are neighbors. Consider two cases. If at least one of the two vertices (say, u) is an address node then $U_D(\mathbf{x}_D) =$

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

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

Suppose now that u and v are both regular nodes. Express $U_D(\mathbf{x}_D)$ as in Eq. **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 Eq. **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. □

Proof of Theorem 2*:* Let G' denote the modified graph in which all edges $\langle v, a\rangle$ incident to a have been replaced with $\langle v, x_a\rangle$. Let N and M be the neighborhood relations in G and G', respectively. We must show that the local characteristics of the remaining variables 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 $M(b) \cup \overline{M(b)}_* = \overline{N(b)} \cup \overline{N(b)}_*\backslash\{a\}$. 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 $a \in \overline{N(b)}$ 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 $\overline{M(i)} \cup \overline{M(i)}_* \cup \overline{M(i^*)} \cup \overline{M(i^*)}_* = \overline{N(i)} \cup \overline{N(i)}_* \cup \overline{N(i^*)} \cup \overline{N(i^*)}_*\backslash\{a\}$.

I am grateful to Song-Chun Zhu and Michael Isard for critical review of this manuscript. Thanks also go to Jeff Sachs for helpful discussions and for suggesting ref. 4.

1. Kindermann, R. & Snell, J. (1980) *Markov Random Fields and Their Applications* (Amer. Math. Soc., Providence, RI).
2. Jensen, F. (1996) *An Introduction to Bayesian Networks* (University College Press, London).
3. Smyth, P. (1997) *Pattern Recognit. Lett.* **18,** 1261–1268.
4. Hasty, J., McMillen, D., Isaacs, F. & Collins, J. (2001) *Nat. Genetics* **2,** 268–279.
5. Mumford, D. (1996) in *ICIAM 95*, eds. Kirchgassner, K., Mahrenholtz, O. & Mennicken, R. (Akademie Verlag, Berlin), pp. 233–256.
6. Mumford, D. (1997) in *The Legacy of Norbert Wiener: A Centennial Symposium*, eds. Jerison, D., Singer, I. & Stroock, D. (Amer. Math. Soc., Providence, RI), pp. 235–260.
7. Winkler, G. (1995) *Image Analysis, Random Fields and Dynamic Monte Carlo Methods* (Springer, Berlin).
8. Hammersley, J. & Clifford, P. (1968) *Markov Fields on Finite Graphs and Lattices*, preprint.
9. Grimmett, G. (1973) *Bull. Lond. Math. Soc.* **5,** 81–84.
10. Geiger, D. & Heckerman, D. (1996) *Artificial Intelligence* **82,** 45–74.
11. Lauritzen, S. (1996) *Graphical Models* (Clarendon, Oxford).
12. Geman, S. & Geman, D. (1984) *IEEE Trans. Pattern Anal. Machine Intell.* **6,** 721–741.
13. Besag, J. (1986) *J. R. Stat. Soc. B* **48,** 259–302.
14. Fischler, M. & Elschlager, R. (1973) *IEEE Trans. Computers* **22,** 67–92.
15. Yuille, A. (1990) *Neural Comput.* **2,** 1–24.
16. Guo, C. E., Zhu, S. C. & Wu, Y. N. (2003) *Int. J. Computer Vision* **53,** 5–29.