

## Gibbs distributions and Markov Random Fields: Stat 775, 3/2/99

THE JOINT DISTRIBUTION OF A COLLECTION OF RANDOM VARIABLES IS RELATED TO THE COLLECTION OF FULL CONDITIONAL DISTRIBUTIONS. DEPENDENCE IS CHARACTERIZED WITH A CERTAIN UNDIRECTED GRAPH. THE HAMMERSLEY-CLIFFORD THEOREM ESTABLISHES A CONNECTION BETWEEN THE JOINT DISTRIBUTION AND THE FULL CONDITIONALS. THE GIBBS SAMPLER ENABLES MONTE CARLO SIMULATION.

**Graphs:** Let  $G = (V, E)$  denote an undirected graph on a finite index set  $V$ . For instance,  $V$  may be  $\{1, 2, \dots, m\}$  or perhaps the finite integer lattice. The edge set  $E$  contains some subset of unordered pairs  $\{u, v\}$  for  $u, v \in V$ . By convention we insist that there are no edges from a node to itself. If the cardinality of  $V$  is  $m$ , then the cardinality of  $E$  is at most  $\binom{m}{2}$ . A given graph will be used to represent dependence within a collection of random variables, but it is helpful to consider it as solitary for the moment. Graphs arising in statistical models usually have relatively small edge sets. Note that there are two key quantities determined by the graph. We say that a pair  $u, v \in V$  are neighbors, relative to  $G$  and written  $u \sim v$ , if the edge  $\{u, v\}$  is in  $E$ . Also, we have cliques in the graph, the set of which is denoted  $\mathcal{C}$ . That is,  $C \in \mathcal{C}$  if  $C \subset V$  and either  $C$  contains a single node or for every pair  $u, v \in C$  it must be that  $u \sim v$ . Cliques are maximally connected subgraphs. The regular integer lattice has cliques of size one and two only.

**Gibbs distributions:** We now consider a collection of random variables  $X = \{X_v : v \in V\}$ . For instance, if  $V$  denotes nodes of the integer lattice, we might have  $X_v \in \{-1, +1\}$  as in the Ising model of ferromagnetism. In fact, **any** finite collection of random variables may be under consideration. We say that the joint distribution of  $X$  is a Gibbs distribution relative to the graph  $G = (V, E)$  if it can be expressed as a product over cliques in  $G$ . That is, with notation  $x_C = \{x_v : v \in C\}$ , we require functions  $h_C$  such that

$$p(x) = \text{constant} \times \exp\left\{\sum_{C \in \mathcal{C}} h_C(x_C)\right\}.$$

As an example, consider the index set  $V = \{1, 2, \dots, m\}$  and an empty edge set, so that the cliques in  $G$  are only the singletons. This degenerate case corresponds to

$$p(x) = \text{constant} \times \prod_{i=1}^m \exp\{h_i(x_i)\},$$

in other words, mutual independence of all the  $X_i$ 's. More complex graphs encode more complex interactions among random variables. In the Ising model, for example, cliques are singletons and adjacent nodes on the lattice, and the joint distribution becomes a product

over adjacent nodes. Further inspection shows that being Gibbs is not such a special feature, since **any** joint distribution will be Gibbs with respect to *some* graph. The utility of this idea seems highest with relatively simple graphs; i.e. when the cliques are not too large.

**Markov Random Fields:** Consider a collection of random variables  $X$  as above, and note that the single joint distribution implies a collection of full conditional distributions  $p(x_v|x_{-v})$ . We say that  $X$  is a Markov random field relative to the graph  $G$  if

$$p(x_v|x_{-v}) = p(x_v|\{x_u : u \sim v\})$$

for all conditioning events. Thus to be a MRF relative to  $G$  is to have a certain limitation on your full conditional distributions. For example, we showed in class that the simple Ising model on the integer lattice is a Markov random field, and its full conditionals are all Rademacher distributions with probabilities determined by the sum of neighboring values. Inspection further shows that any collection of random variables will be an MRF with respect to some graph  $G$ . In fact, our first discussion of MRF graphs was exactly in that vein. Take a joint distribution; work out the full conditionals; observe for each variable those variables which it is not conditionally independent of, and include undirected edges to them in the graph. There was no restriction on the distribution we started with, and so we could without much work identify some graph with respect to which the variates form a Markov random field. We also showed how to identify this graph in the special case where we start with a DAG; i.e. by the process of moralization.

One reason why MRFs are important is because it may be easier to build a model for one-dimensional conditional distributions than it is to attempt to build an entire joint distribution all at once. This issue arose in the development of models for data distributed spatially. A second reason is in Bayesian computation where we use the full conditional distributions to make Monte Carlo samples from the joint distribution. You showed in homework 2 that under the positivity condition, knowing the set of full conditionals implies that you know the joint distribution exactly. By contrast, this is not true of marginal distributions, except if the random variables are independent. See also Gelman and Speed, 1993, *J. Roy. Statist. Soc. B*, **55**, pp. 185-188.

**Hammersley-Clifford Theorem:** *Under the positivity condition,  $X$  is Gibbs relative to  $G$  if and only if  $X$  is a Markov random field relative to  $G$ .*

We saw in class two joint distributions which did not satisfy the positivity condition but which had the same set of full conditional distributions. In that case, knowing the full conditionals does not allow us to know the joint distribution. Your homework 2 problem showed that under positivity the joint distribution is uniquely determined. The HC Theorem goes further by saying specifically how complicated can the joint distribution be if one knows the full conditionals. It cannot be more complicated than a product over clique functions on

the graph defining the neighbors. The pac-man principle is used to show one direction of the proof, namely that Gibbs implies MRF. A more complex argument is required to prove that MRF implies Gibbs. The first proof is given in Besag, 1974, *J. Roy. Statist. Soc. B*, **36**, pp. 192-236. This is a primary reference for models of spatial data. Positivity is a sufficient condition only, and may be relaxed. For example, see Besag, 1994, *Annals of Statistics*, **22**, pp. 1734-1741.

**Gibbs Sampler:** Suppose that modeling and inference considerations require us to calculate features of the joint distribution  $p(x)$  for a collection of random variables  $X = \{X_v : v \in V\}$ . We do so by having the computer realize a sequence of copies of  $X$ ,  $X^1, X^2, \dots, X^B$  for a large number  $B$ . Empirical averages will be used in place of theoretical expectations which we cannot calculate directly. If the joint distribution  $p(x)$  had a useful DAG factorization, we would use direct simulation methods to realize independent and identically distributed copies, but no such representation is available. We do, however, have the MRF representation, since we can take the joint distribution and apply the pac-man. Starting at some state  $X^1 = x$  having positive density, the Gibbs sampler successively substitutes the value  $x_v$  with a draw from the full conditional distribution of  $X_v$  given the current value of all other variables. The order by which we perform these updates is called the sweep strategy, and upon updating each coordinate we have completed a single scan of the Gibbs sampler. A systematic scan occurs if we consider a deterministic order through which to update the coordinates, and this is probably the simplest sweep strategy to use. After successively updating each coordinate variable, we arrive at a new, random state  $X_2$ . The process is repeated from this position, a large number of times. Each update uses a full conditional distribution, and thus we should be able to perform this Monte Carlo, since we may consider the updates to be one dimension at a time. (Recall our claim that we can get the computer to realize any one dimensional random variable.) In other words, the MRF representation is crucial for simulating the joint distribution under study. The key reference for Gibbs sampling is S. Geman and D. Geman, 1984, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **6**, pp. 721-741.

A basic distinction from direct Monte Carlo is that the random states  $X^1, X^2, \dots, X^B$  realized in the Gibbs sampler are dependent. In fact, they form a Markov chain, since the update after scan  $m$  depends on the previous states only through  $X^m$ . Nevertheless, a generalization of the law of large numbers holds for this Markov chain, and ensures that empirical averages accurately approximate features of the joint distribution. Monte Carlo standard error is somewhat more difficult to compute in this case than in the case of direct Monte Carlo. Whereas the standard error in approximating a probability  $p$  by direct MC is  $\sqrt{p(1-p)/B}$ , the standard error in the Gibbs sampler is inflated by a factor  $\tau > 1$  which depends on correlation among states. Time series methods may be used to approximate  $\tau$ . (Calculating  $\tau$  is beyond the scope of present discussion).

Another issue with Gibbs sampling is what authors call *burn-in*. The idea is that the distribution of  $X^2$  or  $X^3$ , etc., depends on the particular starting position  $X^1 = x$ . Burn-in is the number of scans it takes to *forget* the starting state. Burn-in is often not a big problem, as long as you make a good guess for the starting state. Ideally, one should take a point somewhere in the center of the target distribution. A poor starting state affects bias of the Monte Carlo approximations. Typically, one discards an initial segment of the Markov chain to reduce the bias.