Introduction to Graphical Models

STA 345: Multivariate Analysis
Department of Statistical Science
Duke University, Durham, NC, USA
Robert L. Wolpert

1 Conditional Dependence

Two real-valued or vector-valued random variables X, Y are *independent* for probability measure P (written: $X \perp\!\!\!\perp Y\ [P]$) if for all sets A and B,

$$\mathsf{P}[X \in A,\ Y \in B] = \mathsf{P}[X \in A] \cdot \mathsf{P}[Y \in B].$$

For jointly discrete or jointly continuous random variables this is equivalent to factoring of the joint probability mass function or probability density function, respectively. The variables X and Y are *conditionally* independent given a third random variable Z for probability distribution P (written: $X \perp\!\!\!\perp Y \mid Z\ [P]$) if the conditional pmf or pdf factors[1] in the form:

$$p(x, y \mid z) = p(x \mid z)\, p(y \mid z).$$

This relation arises frequently in Bayesian analysis and computation; we now explore it further. For nice discussions of conditional independence in statistical inference see (Dawid 1979a,b, 1980) and for a more advanced view (Dawid and Lauritzen 1993).

2 Gaussian Models

Let M be a square $(p+q) \times (p+q)$ matrix, partitioned in the form

$$M = \begin{bmatrix} A & B \\ C & D \end{bmatrix}$$

[1] More generally, $X \perp\!\!\!\perp Y \mid Z\ [P]$ if for each set A there exists a version of the conditional probability $\mathsf{P}[X \in A \mid Y, Z]$ which is a function only of Z

for some $(p \times p)$ A, $(p \times q)$ B, $(q \times p)$ C, and $(q \times q)$ D. Then you can verify by multiplying that the inverse and determinant are

$$\begin{aligned} M^{-1} &= \begin{bmatrix} (A - BD^{-1}C)^{-1} & -A^{-1}B(D - CA^{-1}B)^{-1} \\ -D^{-1}C(A - BD^{-1}C)^{-1} & (D - CA^{-1}B)^{-1} \end{bmatrix} \\ &= \begin{bmatrix} (A - BD^{-1}C)^{-1} & -(A - BD^{-1}C)^{-1}BD^{-1} \\ -(D - CA^{-1}B)^{-1}CA^{-1} & (D - CA^{-1}B)^{-1} \end{bmatrix} \\ \left|M\right| &= \left|A - BD^{-1}C\right| \times |D| \\ &= |A| \times \left|D - CA^{-1}B\right| \end{aligned}$$

(in each case both representations have the same value). Applying this identity to the covariance matrix Σ for jointly Gaussian random variables $\{X_i : i \in I \cup J\}$ with $I \cap J = \emptyset$, we find that the precision matrix for X_I is

$$\Lambda_{II} = \left(\Sigma_{II} - \Sigma_{IJ}\Sigma_{JJ}^{-1}\Sigma_{JI}\right)^{-1}$$

By an amazing coincidence, the conditional mean and variance of X_I given X_J are (as we've seen before):

$$\begin{aligned} \mathsf{E}[X_I \mid X_J] &= \mu_I + \Sigma_{IJ}\Sigma_{JJ}^{-1}(X_J - \mu_J) \\ \mathsf{V}[X_I \mid X_J] &= \Sigma_{II} - \Sigma_{IJ}\Sigma_{JJ}^{-1}\Sigma_{JI} \\ &= \Lambda_{II}^{-1} \end{aligned}$$

so if $I = \{\alpha, \beta\}$ and $J = \{\alpha, \beta\}^c$ then the conditional density for X_α and X_β, given X_J, will factor as the product of their individual conditional densities (and hence they will be conditionally independent), if and only if $\Lambda_{\alpha\beta} = 0$.

Thus it's easy to recognize or model conditional independence of Gaussian random variables— compute the covariance matrix, take its inverse, and look for zeros. Gaussian Markov chains will have tri-diagonal precision matrices, for example.

The conditional dependence structure for Gaussian random variables may be represented in the form of a graph with a circle for each of the $p + q$ variables, with a line connecting any pair (α, β) for which $\Lambda_{\alpha\beta} \neq 0$.

Life is more complicated outside the Gaussian world, where dependences involving more than two variables are possible, and where correlation does not characterize dependence.

3 Directed Acyclic Graphs

The joint distribution of any random vector $X \in \mathbb{R}^n$ can always be represented in "DAG" form

$$p(d\mathbf{x}) = \prod_{i=1}^{n} p\left(dx_i \mid \mathrm{pa}\left(i\right)\right)$$

of conditional distributions for each variable X_i given its "parents" $\{X_j : j \in \mathrm{pa}\left(i\right)\}$, its immediate predecessors in a graphical representation of the distribution. At worst we can set $\mathrm{pa}\left(1\right) = \emptyset$ and $\mathrm{pa}\left(i\right) = \{1, ..., i-1\}$ for $i > 1$, where $p\left(dx_1 \mid \emptyset\right)$ denotes the marginal distribution of X_1. This choice is "worst" because the resulting graph has $n(n-1)/2$ arrows pointing from parents to children, a rather large number; the most parsimonious representations for Markov chains or for iid random variables would have $n-1$ or zero arrows, respectively.

4 Undirected Graphical Representations

Motivated at least in part by the goal of finding convenient and tractable representations for spatial (especially lattice) random fields that have something like the familiar Markov property for processes indexed by a one-dimensional parameter, a flurry of activity in the 20th century led to the notions of Gibbs processes (or fields). Among the earliest was the "Ising Model"; in its simplest form, it consists of random variables X_s taking only the values $\{\pm 1\}$. The random variable are usually indexed by elements s of a lattice of integer points in $\mathbb{R}^d$, regarded as a "graph" with "edges" connecting any two "adjacent" points whose Euclidean distance from one another is exactly one; this relation is denoted "$s \sim t$" (one can construct Ising models on essentially any graph, but this is the earliest one).

Begin with some compact set K of indices (say, a cube with sides of length $L \in \mathbb{N}$), containing a finite number N_K of indices. For parameters $\alpha, \beta \in \mathbb{R}$, the Ising Model assigns each possible configuration $\{X_s\}_{s \in K}$ probability

$$p_K(\mathbf{x}) = \frac{1}{Z} \prod_{s \in K} e^{\alpha X_s + \beta \sum\{X_s \cdot X_t :\, t \sim s\}}, \tag{1}$$

where $Z = Z(\alpha, \beta)$ is chosen so that the sum of the probabilities of all 2^{N_K} possible states will be one. The parameter α influences the mean ($\alpha > 0$ if and only if each $\mathsf{E}[X_s] > 0$), while β determines the tendency for neighbors to

agree— as $\beta \to \infty$ the distribution becomes concentrated on configurations where every X_s has the same value, with expectation $\tanh(\alpha N_K)$ (so ± 1 are equally likely if $\alpha = 0$). The system shows interesting behaviour as $K \to \mathbb{Z}^d$. Its inventor Ising showed that there is a unique limiting measure as $K \to \mathbb{Z}$ in dimension $d = 1$, and conjectured that would also hold in dimensions $d \geq 2$. He was wrong (but still got the system named after himself!).

One interpretation of Equation (1) is that any specific configuration has "energy" $H(\mathbf{x}) = -\alpha \sum_s X_s - \beta \sum_{s \sim t} X_s X_t$, and tries to stay in "low energy states" (where neighboring X_s's agree) with high probability. Now 2β is the "energy per discordance" (often written $1/kT$, where T is interpreted as "absolute temperature" and k is called Boltzmann's constant), and α is viewed as the "external field".

The constant Z is "intractable" (very difficult to calculate numerically, even with huge computational resources) for even moderately-sized configurations, but still it's easy to *simulate* samples from this distribution. This is because of the particularly simple graphical structure of the distribution— the *conditional* distribution of any X_s, given *all* the other sites $\{X_t : t \neq s\}$, depends only on the $2d$ nearest-neighbors. It's trivial to draw a sample of X_s with this conditional distribution; and, over a hundred years ago, it was recognized that repeatedly replacing each X_s with a random draw from this conditional distribution (the original Gibbs sampler) would converge to a random draw from the Ising model.

Much of Graphical Models may be viewed as a program to repeat this success of the Ising model with more general distributions of the form $p(\mathbf{x}) \propto \exp\big(-H(\mathbf{x})\big)$, where $H(\mathbf{x}) = \sum U_c(x_c)$ is the sum of terms associated with subsets c of indices.

4.1 Hammersley-Clifford

In 1971 John Hammersley and Peter Clifford wrote but did not publish a seminal paper presenting a very general graph-theoretic characterization of joint probability distributions that generalizes the Ising model's structure. Besag (1974) is usually credited with the first published proof (which we follow below), but an alternate approach based on the Möbius Inversion Theorem was taken independently by Grimmett (1973) at about the same time.

Let $\mathbf{x} = (X_1, ..., X_n)$ be a random vector with a joint density function $p(\mathbf{x})$ with respect to some dominating measure (such as counting measure for discrete random variables or Lebesgue for continuous ones) and let $V =$

$\{v_1, ..., v_n\}$ be n distinct elements from any set (for example, they could be the integers 1...n). For $v_i \neq v_j \in V$ write "$i \sim j$" if "the distribution of X_i depends on X_j" in the sense that the conditional density

$$p(x_i \mid x_{(i)}) = \frac{p(\mathbf{x})}{\int p(\mathbf{x})\, dx_i}$$

depends on x_j, where "$x_{(i)}$" denotes the set of x_k for $k \neq i$. One can show that this is a reflexive relation, *i.e.*, that $i \sim j \Leftrightarrow j \sim i$, so the graph $\mathcal{G}$ with vertices V and edges $\mathcal{E} = \{(i,j) : i \sim j\}$ is undirected. For example: if $\{X_i\}$ are independent then $\mathcal{E} =$ and $\mathcal{G}$ is completely disconnected; if $\{X_j\}$ is a Markov chain then $\mathcal{E}$ consists only of the (undirected) edges $\{v_i, v_{i+1}\}$ for $1 \leq i < n$. We say that "$p(\cdot)$ is Markov for $\mathcal{G}$" or that "X is a Markov random field over $\mathcal{G}$."

For now let us assume or arrange that:

1. The set $\mathcal{X}$ of possible values of each X_i is a finite subset of $\mathbb{R}$;
2. The joint pmf $p(\mathbf{x})$ is strictly positive on all of $\Omega = \mathcal{X}^n$;
3. Zero is in $\mathcal{X}$.

4.1.1 Three Useful Equations

Let X have pdf $p(\mathbf{x})$ and let $Y = \{Y_1, ..., Y_n\}$ be another random variable taking values in the same set $\mathcal{X}$. In a minor abuse of notation we use the same letter $p(\cdot)$ for all distributions below, distinguished by their arguments; no suggestion of identical distributions is intended. By conditioning we may write:

$$p(\mathbf{x}) = p(x_n \mid x_1 \dots x_{n-1})\, p(x_1 \dots x_{n-1})$$

Now multiply by one in the form $1 = p(y_n \mid x_{(n)})/p(y_n \mid x_{(n)})$:

$$= \frac{p(x_n \mid x_1 \dots x_{n-1})}{p(y_n \mid x_1 \dots x_{n-1})}\, p(x_1 \dots x_{n-1}, y_n).$$

Similarly,

$$\begin{aligned} p(x_1 \dots x_{n-1}, y_n) &= p(x_{n-1} \mid x_1 \dots x_{n-2}, y_n)\, p(x_1 \dots x_{n-2}, y_n) \\ &= \frac{p(x_{n-1} \mid x_1 \dots x_{n-2}, y_n)}{p(y_{n-1} \mid x_1 \dots x_{n-2}, y_n)}\, p(x_1 \dots x_{n-2}, y_{n-1}, y_n). \end{aligned}$$

After n steps,

$$p(\mathbf{x}) = p(\mathbf{y}) \prod_{i=1}^{n} \frac{p(x_i \mid x_1, \ldots, x_{i-1}, y_{i+1}, \ldots, y_n)}{p(y_i \mid x_1, \ldots, x_{i-1}, y_{i+1}, \ldots, y_n)}. \tag{2.2}$$

Applying this at any fixed point $\mathbf{y}$ (such as $\mathbf{y} = \mathbf{0}$) shows that any finite-dimensional pmf $p(\mathbf{x})$ is determined uniquely by its complete conditionals $p(x_i \mid x_{(i)}$.

For $x \in \Omega$ set

$$Q(\mathbf{x}) \equiv \log \{p(\mathbf{x})/p(\mathbf{0})\}.$$

Hammersley and Clifford posed and answered the question:

> What is the most general form that $Q(\cdot)$ may have?

Introduce the notation

$$\mathbf{x}_i = (x_1, \ldots, x_{i-1}, 0, x_{i+1}, \ldots, x_n)$$

for the vector $\mathbf{x} \in \Omega$ with the ith component replaced by zero. Note

$$\exp \{Q(\mathbf{x}) - Q(\mathbf{x}_i)\} = \frac{p(\mathbf{x})}{p(\mathbf{x}_i)} = \frac{p(x_i \mid x_{(i)})}{p(0 \mid x_{(i)})} \tag{3.2}$$

depends only on the value of x_i and of its neighbors $\{x_j : j \sim i\}$. Thus an answer to the H-C question also gives the most general *conditional* distribution of the x_i's.

Besag's brilliant idea was to expand $Q(\mathbf{x})$ in the form:

$$\begin{aligned} Q(\mathbf{x}) = \sum_{1 \le i \le n} x_i G_i(x_i) + \sum_{1 \le i < j \le n} x_i x_j G_{ij}(x_i, x_j) + \\ + \ldots + x_1 x_2 \cdots x_n G_{123\ldots n}(x_1, \ldots, x_n). \end{aligned} \tag{3.3}$$

To see that this is possible, note from Eqn (3.3) that:

$$Q(0, \ldots, 0, x_i, 0, \ldots, 0) = x_i G_i(x_i)$$

so for each $1 \le i \le n$ and $x \in \mathcal{X}$,

$$G_i(x) = \frac{1}{x} \log \frac{p(0, \ldots, 0, x, 0, \ldots, 0)}{p(\mathbf{0})}$$

for $x \neq 0$ and $G_i(0) = 0$. By considering successively sequences with $2, 3, 4, \ldots$ non-zero x_i's we can solve for $G_{ij}(\cdot, \cdot)$ on $\mathcal{X}^2$, $G_{ijk}(\cdot, \cdot, \cdot)$ on $\mathcal{X}^3$, *etc.*

4.1.2 The Theorem

Now we can write the famous result as:

Theorem 1. (Hammersley-Clifford) *For each subset $c \subset \{1, \ldots, n\}$ the function $G_c(x_c)$ on $\mathcal{X}^{|c|}$ vanishes in representation (3.3) unless c is a* clique *of the graph $\mathcal{G} = (V, \mathcal{E})$. Subject to that constraint, the $G_c(x_c)$ are arbitrary provided that*

$$Z \equiv \int \exp\{Q(\mathbf{x})\}\, d\mathbf{x} < \infty. \tag{2}$$

Proof. Fix $i = 1$ (to simplify notation) and consider any $\ell \not\sim 1$. By Eqn (3.3),

$$\begin{aligned} Q(\mathbf{x}) - Q(\mathbf{x}_1) = x_1 \Big\{ & G_1(x_1) + \sum_{1<j\le n} x_j G_{1j}(x_1, x_j) \\ & + \sum_{1<j<k\le n} x_j x_k G_{1jk}(x_1, x_j, x_k) \\ & + \cdots + x_2 x_3 \cdots x_n G_{12\ldots n}(x_1, x_2, \ldots, x_n) \Big\} \end{aligned} \tag{3}$$

may depend only on $\{x_k : 1 \sim k\}$ and, in particular, must not depend on x_ℓ. If we take $\mathbf{x}$ such that $x_j = 0$ for all $j \notin \{1, \ell\}$, then we find that

$$Q(\mathbf{x}) - Q(\mathbf{x}_1) = x_1 \{G_1(x_1) + x_\ell G_{1\ell}(x_1, x_\ell)\}$$

must not depend on x_ℓ, so necessarily $G_{1\ell}(\cdot)$ must vanish. Similarly each $G_c(\cdot)$ must vanish for $|c| = 3, 4, \ldots, n$ if any $i, j \in c$ with $i \not\sim j$, *i.e.*, unless c is complete. Without losing generality we may arrange that $G_c = 0$ unless c is a *maximal* complete set, or *clique*, by accumulating all the terms with $G_\gamma(x_\gamma)$ for $\gamma \subset c$ into $G_c(x_c)$.

Conversely, any functions $G_c(\cdot)$ for which $Q(\cdot)$ satisfies Equation (2) lead to a probability distribution $p(\mathbf{x}) = Z^{-1} \exp\{Q(\mathbf{x})\}$ that is Markov for $\mathcal{G}$.

□

4.1.3 Examples

Besag (1974) introduces a number of "auto" examples from the exponential family— auto*logistic*, auto*normal*, auto*Poisson*, auto*binomial*, and more, based on the usual nearest-neighbor lattice graph for the integers $\mathbb{Z}^2$ in the plane (one of his discussants pleas for a hexagonal lattice instead— no

harder conceptually, but more programming effort). For example, in the autoPoisson model the conditional distribution of

$$X_i \mid X_{(i)} \sim \mathsf{Po}(\mu_i)$$

is Poisson with conditional mean

$$\mu_i = \exp\Big\{\alpha - \beta \sum_{j\sim i} X_j\Big\},$$

for some $\beta \geq 0$. Note that X_i and X_j are always negatively correlated. Even for $n = 2$ the marginal distributions are unavailable in closed form, but the distribution is easy to simulate using substitution sampling (Tanner and Wong 1987), a precursor of Gibbs Sampling (Gelfand and Smith 1990).

Much more general uses of the Hammersley-Clifford approach are now common, and graphical models have become a standard tool for modeling and for organizing the computation necessary for inference in high-dimensional problems. For an interesting example of trying to infer the graphical (*i.e.*, conditional independence) structure from data, see (Lunagomez et al. 2009).

5 Extensions

Lauritzen (1996, Ch. 3) considers a more general graphical structure in which vertices (corresponding to variables) are *marked* to indicate whether their distributions are *c*ontinuous (indicated graphically with a *c*ircle) or *d*iscrete (indicated with a *d*ot). He distinguishes several possible Markov properties a distribution might have with respect to a graph:

(P) Pairwise Markov, in which $x_i \perp\!\!\!\perp x_j \mid x_{(i,j)}$ (X_i and X_j are conditionally independent given all other variables) whenever $i \not\sim j$;

(L) Local Markov, if $X_i \perp\!\!\!\perp \{X_j : \ j \notin \mathrm{cl}\,(i)\} \mid \{X_j : \ j \in \mathrm{bd}\,(i)\}$;

(G) Global Markov, if $\{X_i : \ i \in A\} \perp\!\!\!\perp \{X_j : \ j \in B\} \mid \{X_k : \ k \in S\}$ whenever S separates A, B.

It's easy to show (G) $\Rightarrow$ (L) $\Rightarrow$ (P), and to construct examples where the other implications fail, but with enough regularity (for example, if the joint density has a positive continuous density with respect to a product measure) they all coincide with each other and with

(F) Factorisable, if the joint distribution has a density $p(\cdot)$ that factors as the product

$$p(\mathbf{x}) = \prod_{\text{cliques } c \subseteq \mathcal{G}} G_c(x_c)$$

Conditions 1 and 3 from Section (4.1) were only used to simplify the proofs, and are unnecessary; their wish to relax the positivity Condition 2 was the reason Hammersley and Clifford failed to publish their original result. As of Besag's 1974 paper the "question of positivity" was still open, but Moussouris (1974) gave a simple example with $n = 4$ variables and $\mathcal{G}$ a chordless square where positivity fails and $p(\mathbf{x})$ does *not* factor over cliques, so this condition is in fact necessary for the theorem (see Lauritzen (1996, Example 3.10) for a more accessible account).

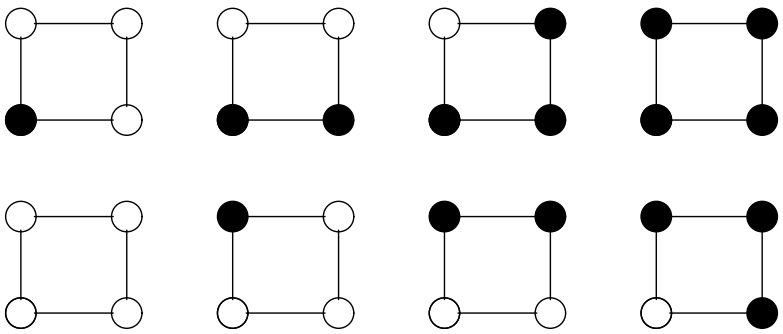

Figure 1: Eight equally-likely outcomes in Moussouris's example.

In the example the four variables X_i each take values $0, 1$ with proabability $1/2$ each. One of each diagonal pair has a degenerate conditional distribution given the variables on the other diagonal, so the distribution is Markov for the indicated graph, but one can show no H-C factorization applies.

5.0.4 Another Factorization

If $X \perp\!\!\!\perp Z \mid Y$ $[P]$ all have density functions, then multiply the defining relation $p(x, z \mid y) = p(x \mid y) \cdot p(z \mid y)$ by $1 = p(y)/p(y)$ to get the joint density

$$p(x, y, z) = \frac{p(x, y) \cdot p(y, z)}{p(y)}$$

as the product of marginal densities for the sets $\{x, y\}$ and $\{y, z\}$, on top, divided by the marginal density for their "separator", on the bottom. This "junction tree" factorization, which doesn't require positivity, can be extended to general graphical models over "decomposible" graphs. An undirected graph is decomposible if and only if it's *triangulated*, *i.e.*, if every cycle of length longer than three has a chord, so this doesn't apply to the distribution of Moussouris's example.

References

Besag, J. E. (1974), "Spatial interaction and the statistical analysis of lattice systems (with discussion)," *Journal of the Royal Statistical Society, Ser.* B*: Statistical Methodology*, 36, 192–236.

Clifford, P. (1990), "Markov Random Fields in Statistics," in *Disorder in Physical Systems: A Volume in Honour of John M. Hammersley*, eds. G. Grimmett and D. Welsh, Oxford, UK: Oxford University Press, pp. 19–32.

Dawid, A. P. (1979a), "Conditional independence in statistical theory (with discussion)," *Journal of the Royal Statistical Society, Ser.* B*: Statistical Methodology*, 41, 1–31.

Dawid, A. P. (1979b), "Some misleading arguments involving conditional independence," *Journal of the Royal Statistical Society, Ser.* B*: Statistical Methodology*, 41, 249–252.

Dawid, A. P. (1980), "Conditional independence for statistical operations," *Annals of Statistics*, 8, 598–617.

Dawid, A. P. and Lauritzen, S. L. (1993), "Hyper Markov Laws in the Statistical Analysis of Decomposable Graphical Models," *Annals of Statistics*, 21, 1272–1317.

Gelfand, A. E. and Smith, A. F. M. (1990), "Sampling-based approaches to calculating marginal densities," *Journal of the American Statistical Association*, 85, 398–409.

Grimmett, G. R. (1973), "A Theorem about Random Fields," *Bull. London Math. Soc.*, 5, 81–84.

Hammersley, J. M. and Clifford, P. (1971), "Markov fields on finite graphs and lattices," Unpublished; see however Clifford (1990).

Lauritzen, S. L. (1996), *Graphical Models*, *Oxfoord Statistical Science Series*, volume 17, New York, NY: Oxford University Press.

Lunagomez, S., Mukherjee, S., and Wolpert, R. L. (2009), "Geometric Representations of Hypergraphs for Prior Specification and Posterior Sampling," Discussion Paper 2009-01, Duke University Department of Statistical Science.

Moussouris, J. (1974), "Gibbs and Markov random systems with constraints," *Journal of Statistical Physics*, 10, 11–33.

Tanner, M. A. and Wong, W. H. (1987), "The Calculation of Posterior Distributions by Data Augmentation (with discussion)," *Journal of the American Statistical Association*, 82, 528–550.