The Hammersley-Clifford Theorem and its Impact on Modern Statistics

Helge Langseth

Department of Mathematical Sciences

Norwegian University of Science and Technology

Outline

→ Historical review

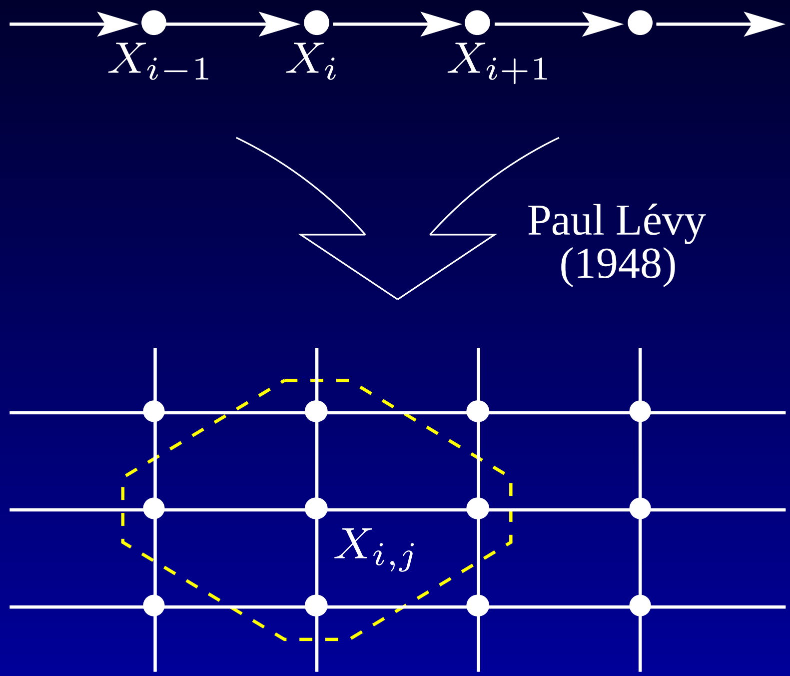
→ Hammersley-Clifford's theorem

→ Usage in

- Spatial models on a lattice
- Point processes
- Graphical models
- Markov Chain Monte Carlo

→ Conclusion

Markov chains in higher dimensions



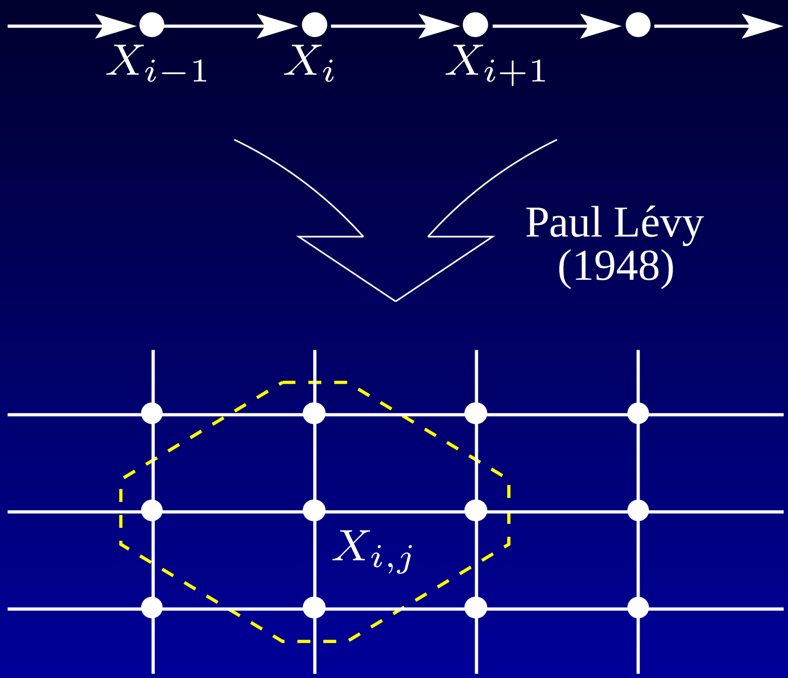
→ Define neighbouring set in the 2D-model:

$\mathcal{N}(x_{i,j}) = \{x_{i-1,j}, x_{i+1,j}, x_{i,j-1}, x_{i,j+1}\}$

→ Sought independence relations:

$$p\big(x_{i,j} | \boldsymbol{x} \setminus \{x_{i,j}\}\big) = p\big(x_{i,j} | \mathcal{N}(x_{i,j})\big)$$

Markov chains in higher dimensions



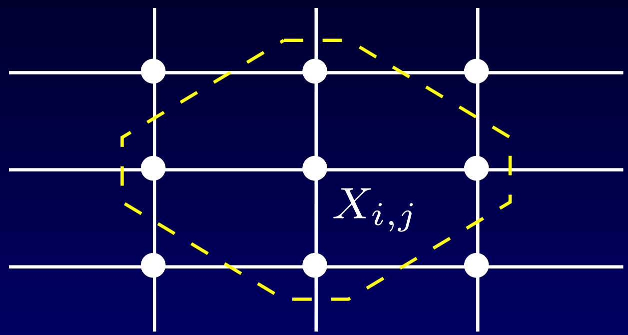
Example: The Ising model (Ising, 1925):

→ Model for ferromagnetism

→ $X_{i,j} \in \{-1, 1\}$, $E_{i,j}(\boldsymbol{x}) = \frac{-1}{kT} \sum_{x_{\ell,m} \in \mathcal{N}(x_{i,j})} x_{i,j} \cdot x_{\ell,m}$

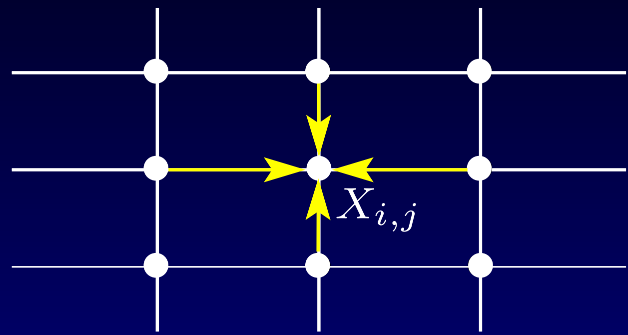
→ $p(\boldsymbol{x}) = \frac{1}{Z} \cdot \exp(-\sum_{i,j} E_{i,j}(\boldsymbol{x}))$

Defining the Markov models in two dimensions



$p(\boldsymbol{x}) = \prod_{i,j} \Psi_{i,j} \left(x_{i,j}, \mathcal{N}(x_{i,j}) \right)$

Joint model (Whittle, 1963)



$p(x_{i,j} | \boldsymbol{x} \setminus \{x_{i,j}\}) = p\left(x_{i,j} | \mathcal{N}(x_{i,j}) \right)$

Conditional model (Bartlett, 1966)

→ For *Nearest neighbour systems*: The class of joint models contains the class of conditional models (Brook, 1964)

→ Not known (at the time) how to define the full joint distribution from the conditional distributions

→ Severe constraints in Bartlett's model

Besag (1972) on nearest neighbour systems

What is the most general form of the conditional probability functions that define a coherent joint function? And what will the joint look like?

→ Assume $p(\boldsymbol{x}) > 0$, and define

$$Q\left(x_{i,j} | x_{i-1,j}, x_{i+1,j}, x_{i,j-1}, x_{i,j+1}\right) = \log\left\{\frac{p\left(x_{i,j}|\mathcal{N}(x_{i,j})\right)}{p\left(0|\mathcal{N}(x_{i,j})\right)}\right\}.$$

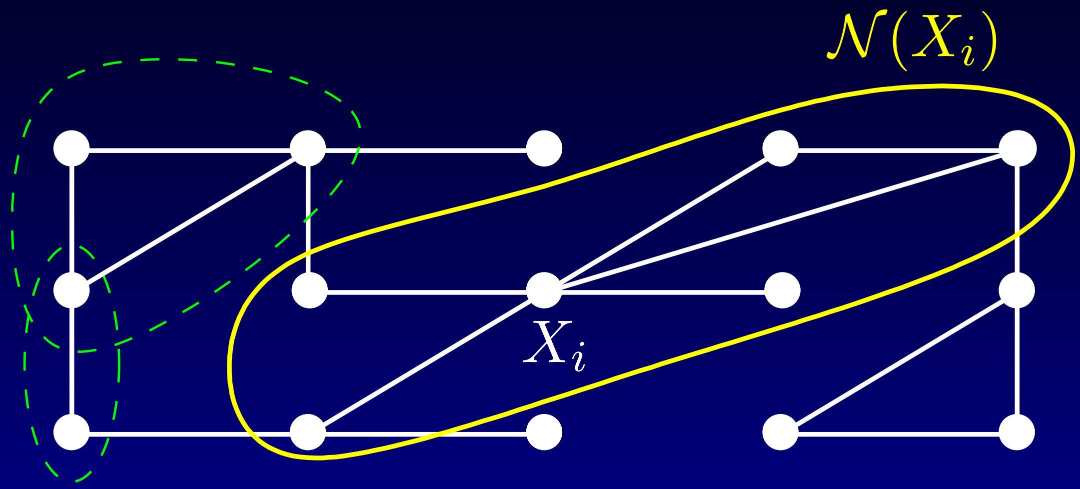
→ $Q(x \,|\, t, u, v, w) \equiv$

$$x\{\psi_0(x)+t\psi_1(x,t)+u\psi_1(u,x)+v\psi_2(x,v)+w\psi_2(w,x)\}$$

→ Let $\boldsymbol{x}_B$ be the boundary, and $\boldsymbol{x}_I = \boldsymbol{x} \setminus \boldsymbol{x}_B$.

$$p(\boldsymbol{x}_I|\boldsymbol{x}_B = 0) = \frac{1}{Z} \cdot \exp\left(\sum_{i,j} x_{i,j}\left\{\psi_0(x_{i,j})+ x_{i-1,j}\psi_1(x_{i,j}, x_{i-1,j}) + x_{i,j-1}\psi_2(x_{i,j}, x_{i,j-1})\right\}\right)$$

Hammersley-Clifford's theorem - Notation



→ Define a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, s.t. $\mathcal{V} = \{X_1, \ \ldots, \ X_n\}$ and $\{X_i, X_j\} \in \mathcal{E}$ iff

$$p(x_i \,|\, \{x_1, \ldots, x_n\} \setminus \{x_i\}) \neq p(x_i \,|\, \{x_1, \ldots, x_n\} \setminus \{x_i, x_j\})$$

→ Define $\mathcal{N}(X_i)$ s.t. $X_j \in \mathcal{N}(X_i)$ iff $\{X_i, X_j\} \in \mathcal{E}$

→ $C \subseteq \mathcal{V}$ is a clique iff $C \subseteq \{X, N(X)\} \, \forall X \in C$.

Hammersley-Clifford's theorem - Result

Assume that $p(x_1, \ldots, x_n) > 0$ (*positivity condition*). Then,

$$p(\boldsymbol{x}) = \frac{1}{Z} \prod_{C \in \mathrm{cl}(\mathcal{G})} \phi_C(\boldsymbol{x}_C)$$

Thus, the following are equivalent (given the positivity condition):

Local Markov property: $p\big(x_i \,|\, \boldsymbol{x} \setminus \{x_i\}\big) = p\big(x_i \,|\, \mathcal{N}(x_i)\big)$

Factorization property: The probability factorizes according to the cliques of the graph

Global Markov property: $p(\boldsymbol{x}_A \,|\, \boldsymbol{x}_B, \boldsymbol{x}_S) = p(\boldsymbol{x}_A \,|\, \boldsymbol{x}_S)$ whenever $\boldsymbol{x}_A$ and $\boldsymbol{x}_B$ are separated by $\boldsymbol{x}_S$ in $\mathcal{G}$

Hammersley-Clifford's theorem - Proof

Line of proof due to Besag (1974), who clarified the original "circuitous" proof by Hammersley & Clifford

- → Assume the *positivity condition* to be correct
- → Let $Q(\boldsymbol{x}) = \log\left[\, p(\boldsymbol{x})/p(\mathbf{0})\,\right]$
- → There exists a unique expansion of $Q(\boldsymbol{x})$,

$$\begin{aligned} Q(\boldsymbol{x}) &= \sum_{1\leq i\leq n} x_i G_i(x_i) + \sum_{1\leq i<j\leq n} x_i x_j G_{i,j}(x_i, x_j) + \cdots \\ &+ x_1 x_2 \ldots x_n G_{1,2,\ldots,n}(x_1, x_2, \ldots, x_n) \end{aligned}$$

- → $G_{i,j,\ldots,s}(x_i, x_j, \ldots, x_s) \neq 0$ only if $\{i, j, \ldots, s\} \in \mathrm{cl}(\mathcal{G})$

Positivity condition: Historical implications

→ Hammersley & Clifford (1971) base their proof on the *positivity condition*:

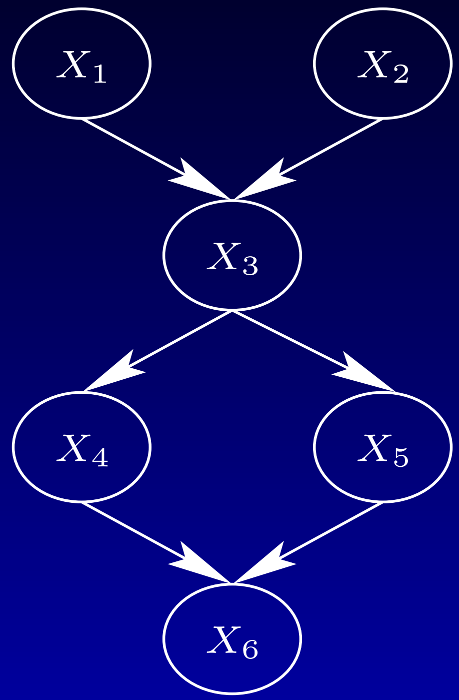
$$p(x_1, \ldots, x_n) > 0$$

→ They find the positivity condition *unnatural*, and postpones publication in hope of relaxing it

→ They are thereby preceded by Besag (1974) in publishing the theorem

→ Moussouris (1974) shows by a counter-example involving only four variables that the positivity condition is *required*

Markov properties on DAGs



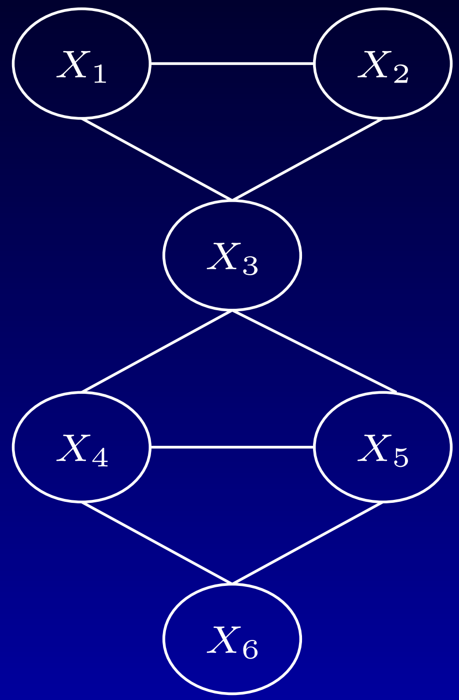
Define a DAG $\mathcal{G}^{\rightarrow} = (\mathcal{V}, \mathcal{E}^{\rightarrow})$ for a well-ordering $X_1 \prec X_2 \prec \cdots \prec X_n$ s.t.

- $\rightarrow$ $\mathcal{V} = \{X_1, \ \ldots, \ X_n\}$ (as before)
- $\rightarrow$ Assume $X_j \prec X_i$. Then $(X_j, X_i) \in \mathcal{E}^{\rightarrow}$ (i.e., $X_j \rightarrow X_i$ in $\mathcal{G}^{\rightarrow}$) iff $p(x_i \,|\, x_1, \ \ldots, \ x_{i-1}) \neq p(x_i \,|\, x_1, \ \ldots, \ x_{j-1}, x_{j+1}, \ \ldots, \ x_{i-1})$

Define the parents of X_i as $\mathrm{pa}(X_i) = \{X_j : (X_j, X_i) \in \mathcal{E}^{\rightarrow}\}$

Directed factorization property: $p(\boldsymbol{x})$ factorizes according to $\mathcal{G}^{\rightarrow}$ iff $p(\boldsymbol{x}) = \prod_i p\big(x_i \,|\, \mathrm{pa}(x_i)\big)$

Markov properties on DAGs (cont'd)



→ Define *moral graph* $\mathcal{G}^m = (\mathcal{V}, \mathcal{E}^m)$ from $\mathcal{G}^{\to}$ by connecting parents and dropping edge directions

→ Note that $\{X_i, \mathrm{pa}(X_i)\} \in \mathrm{cl}(\mathcal{G}^m)$, *i.e.*, factorization relates to $\mathrm{cl}(\mathcal{G}^m)$

→ *Local* and *Global* Markov properties defined "as usual"

The following are equivalent *even without the positivity condition* (Lauritzen *et al.*, 1990):

→ Factorization property

→ *Local* Markov property

→ *Global* Markov property

Spatial statistics

The theorem has had major implications in many areas of spatial statistics. Application areas include:

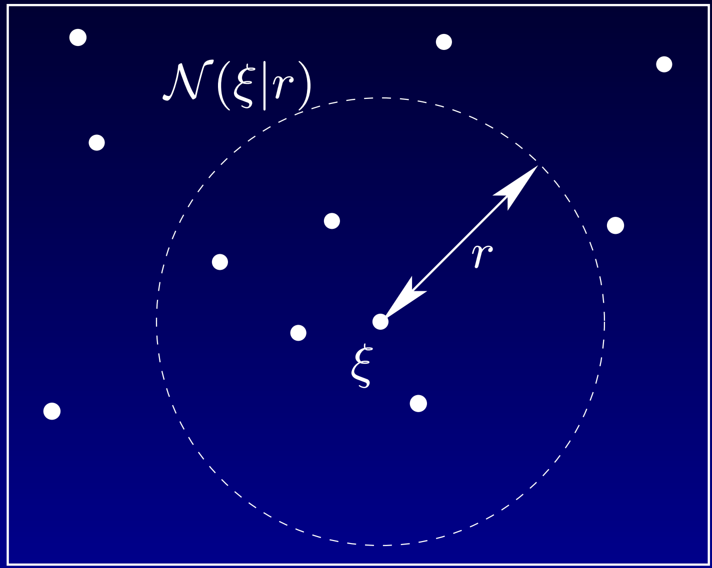
- → Quantitative geography (*e.g*, Besag, 1975)
- → Geographical analysis of the spread of diseases (*e.g*, Clayton & Kaldor,1987)
- → Image analysis (*e.g*, Geman & Geman, 1984)

Markov Point Processes

→ Consider a point process on *e.g.* $\mathbb{R}^n$

→ Let $\boldsymbol{x} = \{x_1, x_2, \ldots, x_m\}$ be the observed points

→ Define the neighbour set as $\mathcal{N}(\xi|r) = \{x_i : ||\xi - x_i|| \leq r\}$



→ A density function f is Markov if $f(\xi \mid \boldsymbol{x})$ depends only on ξ and $\mathcal{N}(\xi) \cap \boldsymbol{x}$

→ Ripley&Kelly (1977): $f(\boldsymbol{x})$ is a Markov function iff there exist functions ϕ_C s.t. $f(\boldsymbol{x}) = \frac{1}{Z} \prod_{C \in \mathrm{cl}(\mathcal{G})} \phi_C(\boldsymbol{x}_C)$

Log-linear models

- → The analysis of *contingency tables* set into the framework of *log-linear* models in the 70's
- → $\log p(\boldsymbol{x}) = u_\phi + \sum_i u_i(x_i) + \cdots + u_{1\ldots n}(x_1, \ldots, x_n)$

Log-linear models

→ The analysis of *contingency tables* set into the framework of *log-linear* models in the 70's

→ $\log p(\boldsymbol{x}) = u_\phi + \sum_i u_i(x_i) + \cdots + u_{1\ldots n}(x_1, \ldots, x_n)$

→ Connection with Hammersley & Clifford's theorem made by Darroch *et al.* (1980):

- $\mathcal{G}$ is defined s.t. X_i and X_j are only connected if $u_{ij} \neq 0$ (with consistency assumptions)
- A Hammersley & Clifford theorem can be proven for this structure
- Representational benefits follows for the class of graphical models

MCMC and the Gibbs sampler

→ Metropolis-Hastings algorithm: Define a Markov chain which has a desired distribution $\pi(\cdot)$ as its unique stationary distribution

Algorithm:

1. Initialization: $\boldsymbol{x}^{(0)} \leftarrow$ fixed value
2. For $i = 1, 2, \ldots$:

   $i)$ Sample $\boldsymbol{y}$ from $q(\boldsymbol{y} \,|\, \boldsymbol{x}^{(i-1)})$

   $ii)$ Define $$\alpha_{\boldsymbol{y}} \leftarrow \frac{\pi(\boldsymbol{y}) \cdot q(\boldsymbol{x}^{(i-1)} \,|\, \boldsymbol{y})}{\pi(\boldsymbol{x}^{(i-1)}) \cdot q(\boldsymbol{y} \,|\, \boldsymbol{x}^{(i-1)})}$$

   $iii)$ $$\boldsymbol{x}^{(i)} \leftarrow \begin{cases} \boldsymbol{y} & \text{with } p = \min\{1, \alpha_{\boldsymbol{y}}\} \\ \boldsymbol{x}^{(i-1)} & \text{with } p = \max\{0, 1 - \alpha_{\boldsymbol{y}}\} \end{cases}$$

MCMC and the Gibbs sampler (cont'd)

- → Geman & Geman (1984): Metropolis Hastings for high-dimensional $\boldsymbol{x}$
- → Problem: How to sample $\boldsymbol{y}$ and calculate $\alpha_{\boldsymbol{y}}$ efficiently?
- → Solution: Flip only *one* element $x_j^{(i)}$ at a time:
  $\boldsymbol{x}^{(i+1)} = \left(x_1^{(i)}, \ldots, x_{j-1}^{(i)}, x_j^{(i+1)}, x_{j+1}^{(i)}, \ldots, x_n^{(i)}\right)$
- → $q\left(\boldsymbol{y} \,|\, \boldsymbol{x}^{(i)}\right)$ is defined by the conditional probability $p\left(x_j \,|\, \boldsymbol{x}^{(i)}\right)$:

$$p\left(x_j^{(i+1)} \,|\, \boldsymbol{x}^{(i)}\right) = \frac{1}{Z_j} \prod_{C:X_j \in C} \phi_C\left(\boldsymbol{x}_C^{(i)}\right)$$

- → $\alpha_{\boldsymbol{y}} = 1$ for the Gibbs sampler
- → An algorithm of *constant time* complexity **can** be designed!

Too much of a good thing?

→ Global properties from local models:

- Model error dominates (*e.g.* Rue and Tjelmeland, 2002)
- The critical temperature of the Ising model

"Beware — Gibbs sampling can be dangerous!"

Spiegelhalter *et al.* (1995): The BUGS v0.5 manual, p. 1

→ Alternative representations:

- Bayesian networks (*e.g.* Pearl, 1988)
- Vines (*e.g.* Bedford and Cooke, 2001)

Clifford's (MCMC) conclusion

"... from now on we can compare our data with the model we actually want to use rather than with a model which has some mathematical convenient form. This is surely a revolution."

Dr. Peter Clifford (1993),
The Royal Statistical Society meeting on the Gibbs sampler and other statistical Markov Chain Monte Carlo methods

Journal of the Royal Statistical Society, *Series* **B**, **55**(1), p. 53

References

I have benefited from getting the opinion of Peter Clifford, A. Philip Dawid, Steffen L. Lauritzen, David J. Spiegelhalter and Håvard Rue on these issues.

- ⟶ Adrian Baddeley and Jesper Møller (1989): Nearest-Neighbour Markov Point Processes and Random Sets. International Statistical Review, 57, pp. 89–121.
- ⟶ Tim J. Bedford and Roger M. Cooke (2001): Probability density decomposition for conditionally dependent random variables modelled by vines. Annals of Mathematics and AI, 32, 245–268.
- ⟶ Julian Besag (1972): Nearest-neighbour Systems and the Auto-logistic Model for Binary data. Journal of the Royal Statistical Society, Series B, 34, pp. 75–83.
- ⟶ Julian Besag (1974): Spatial Interaction and the Statistical Analysis of Lattice Systems. Journal of the Royal Statistical Society, Series B, 36, pp. 192–236.
- ⟶ Julian Besag (1975): Statistical Analysis of Non-lattice Data. The Statistician, 24, pp. 179–195.
- ⟶ Julian Besag (1991): Spatial Statistics in the Analysis of Agricultural Field Experiments. In: Spatial statistics and digital image analysis. Washington, D.C.: National Academy Press.

References (cont'd)

- → Peter Clifford (1990): Markov Random Fields in Statistics. In: Geoffrey Grimmett and Domnic Welsh (Eds.), Disorder in Physical Systems: A Volume in Honour of John M. Hammersley, pp. 19–32. Oxford University Press.
- → Peter Clifford (1993): Discussion on the meeting on the Gibbs sampler and other statistical Markov Chain Monte Carlo methods. Journal of the Royal Statistical Society, Series B, 55, pp. 53–102.
- → John N. Darroch, Steffen L. Lauritzen, and Terry P. Speed (1980): Markov fields and log-linear interaction models for contingency tables. Annals of Statistics, 8, pp. 522–539.
- → Stuart Geman and Donald Geman (1984): Stochastic Relaxation, Gibbs distribution, and the Bayesian restoration of images. IEEE Transactions on Pattern Analysis and Machine Intelligence, 6, pp. 721–741.
- → John M. Hammersley and Peter Clifford (1971): Markov fields on finite graphs and lattices. Unpublished.
- → S.L. Lauritzen, A.P. Dawid, B.N. Larsen and H.-G. Leimer (1990): Independence Properties of Directed Markov Fields. Networks, 20, pp. 491–505.
- → John Moussouris (1974): Gibbs and Markov Random Systems with Constraints. Journal of Statistical Physics, 10, pp. 11-33.
- → Brian D. Ripley and Frank P. Kelly (1977): Markov point processes. Journal of the London Mathematical Society, 15, pp. 188–192.