Markov random fields II

Rasmus Waagepetersen

November 24, 2023

Outline:

1. Auto-logistic (Ising) and auto-Poisson models
2. Estimation for Ising model
3. Bayesian Image analysis
4. Gibbs sampler (MCMC algorithm)
5. Phase-transition for Ising model

Brooks vs. Hammersley-Clifford

Given a set of (alledgedly) full conditionals we can use either Brooks or H-C to identify candidate for a joint (unnormalized) density $p(\cdot)$. In both cases we need to check that $p(\cdot)$ can be normalized and that it is consistent with the given full conditionals.

On disadvantage of Brooks is that it in principle yields $n!$ solutions (possible non-uniqueness) and it does not inform on the form of $p(\cdot)$.

For H-C, we can construct the interaction functions using the full conditionals in a systematic way following the proof of 1. $\Rightarrow$ 2. For given y these interaction functions and hence $p(\cdot)$ are uniquely determined by the full conditionals. Moreover, we can easily check that the constructed interaction functions are consistent with the full conditionals since

$$p_i(x_i|x_{-i}) \propto \frac{p_i(x_i|x_{-i})}{p_i(y_i|x_{-i})} = \frac{p(x)}{p(x_{-i}, y_i)} = \prod_{C: i \in C} \phi(x_C)$$

Gaussian MRF

Consider graph $G = (V, E)$ and full conditionals

$$p_i(x_i | x_{-i}) \propto \exp(-\frac{1}{2\kappa_i}(x_i - \mu_i + \sum_{l \sim i} \beta_{il}(x_l - \mu_l))^2)$$

where $\beta_{ij}/\kappa_i = \beta_{ji}/\kappa_j$.

Letting $y = (\mu_l)_{l \in V}$, we have following H-C that

$$\phi_\emptyset = p(\mu) \text{ (to be determined)} \quad \phi_i(x_i) = \exp(-\frac{1}{2\kappa_i}(x_i - \mu_i)^2)$$

$$\phi_{\{i,j\}}(x_i, x_j) = \exp(-\frac{\beta_{ij}}{\kappa_i}(x_i - \mu_i)(x_j - \mu_j))$$

and $\phi_C(x_C) = 1$ for $\#C > 2$.

Note $\phi_{\{i,j\}}$ covers both pairs (i, j) and (j, i).

How ?

Start with $\phi_\emptyset = p(\mu)$ (which we do not know yet).

By construction in proof of H-C:

$$\phi_i(x_i) = \frac{p(x_i, \mu_{-i})}{p(\mu)} = \frac{p_i(x_i|\mu_{-i})}{p_i(\mu_i|\mu_{-i})} = \exp(-\frac{1}{2\kappa_i}(x_i - \mu_i)^2)$$

Note that this implies $p(x_i, \mu_{-i}) = \phi_i(x_i)p(\mu)$.

Next,

$$\phi_{\{i,j\}}(x_i, x_j) = \frac{p(x_i, x_j, \mu_{-\{i,j\}})}{p(\mu)\phi_i(x_i)\phi_j(x_j)} = \frac{p(x_i, x_j, \mu_{-\{i,j\}})}{p(x_i, \mu_{-i})\phi_j(x_j)} =$$

$$\frac{p_j(x_j|x_i, \mu_{-\{i,j\}})}{p_j(\mu_j|x_i, \mu_{-\{i,j\}})\phi_j(x_j)} = \exp(-\frac{\beta_{ji}}{\kappa_j}(x_i - \mu_i)(x_j - \mu_j))$$

Proceeding in the same way, we obtain $\phi_C(x_C) = 1$ for all C of cardinality $\#C > 2$ (of course we only need to consider C that are cliques with respect to G)

Considering the constructed $p(x)/p(\mu)$ and letting $Q_{ij} = \beta_{ij}/\kappa_i$ (with $\beta_{ii} = 1$) we see that this is the unnormalized density of $N(\mu, Q^{-1})$ provided Q is positive definite.

If Q is positive definite we can conclude

$$\phi_\emptyset = p(\mu) = (2\pi)^{-\#V/2}|Q|^{1/2}$$

For a Gaussian vector $X = (X_l)_{l \in V}$ we know that the full conditional of X_i only depends on those X_j for which $Q_{ij} \neq 0$.

Hence, X is a MRF with respect to $G \Leftrightarrow Q_{ij} = Q_{ji}$ differs from zero only if $\{i, j\} \in E$.

Note that the above is an example that ϕ_C can be equal to one also for C that is in fact a clique.

Auto-logistic model

Consider 2D rectangular $L \times K$ lattice V with horizontal/vertical neighbours. Only possible cliques are then singletons or pairs of horizontal or vertical neighbours.

Consider stochastic vector X on $\{0,1\}^V$ with

$$p_i(x_i|x_{-i}) = \frac{\exp(\alpha x_i + \beta \sum_{j \in N_i} x_i x_j)}{1 + \exp(\alpha + \beta \sum_{j \in N_i} x_j)}$$

Note $p_i(1|x_{-i})$ corresponds to logistic regression with covariate given by number of neighbouring 1's.

Following construction in proof of Hammersley-Clifford with $y = (0, \ldots, 0)$ we obtain

$$\phi_i(x_i) = \exp(\alpha x_i) \quad \phi_{\{i,j\}}(x_{\{i,j\}}) = \exp(\beta x_i x_j)$$

We do not need to consider C with $\#C > 2$ since such a C can not be a clique.

Unknown normalizing constant

Hence joint density is

$$p(x) = p(0)\exp(\alpha \sum_{l \in V} x_l + \beta \sum_{\{i,j\} \in E} x_i x_j)$$

Sum defining

$$p(0) = \left[\sum_{x \in \{0,1\}^V} \exp(\alpha \sum_{l \in V} x_l + \beta \sum_{\{i,j\} \in E} x_i x_j) \right]^{-1}$$

has 2^{LK} terms !

Finite but in general impossible to compute exactly.

Hence we only know $p(\cdot)$ up to proportionality.

Boundary conditions

- free boundary: pixels at edges have only 2 or 3 neighbours
- fixed boundary: we condition on fixed values of boundary pixels. Then all interior "random" pixels have 4 neighbours
- toroidal (similar to circulant): edge pixels neighbours of pixels on opposite edge. E.g. pixel $(1, j)$ becomes neighbour of pixel (L, j). Hence all pixels have 4 neighbours.

Symmetric case

Suppose all pixels have 4 neighbours (fixed or toroidal boundary).

If $\sum_{j \in N_i} = 2$ we may want $p_i(0|x_{-i}) = p_i(1|x_{-i})$.

This is achieved with $\alpha = -2\beta$.

Ising model

Autologistic model is another name for the very famous Ising model (from statistical physics). In statistical physics $0, 1$ are replaced by $-1, 1$ representing "spins" of elementary particles in piece of iron.

An equivalent form is

$$p(x) \propto \exp(\tilde{\alpha} \sum_{l \in V} x_l + \tilde{\beta} \sum_{\{i,j\} \in E} 1[x_i = x_j]) \tag{1}$$

That is, with $\tilde{\beta} > 0$, the model assigns large probabilities to x with many neighbours of equal value.

If $x_i \in \{0, 1\}$ and all pixels have four neighbours then (1) is equivalent to auto-logistic with $\alpha = \tilde{\alpha} - 4\tilde{\beta}$ and $\beta = 2\tilde{\beta}$.

Simulation of MRF

Gaussian MRF: use sparse matrix Cholesky decomposition of precision matrix.

General MRF: Markov chain Monte Carlo. Here we consider the so-called Gibbs sampler

Gibbs sampler

Idea: generate Markov chain $X^1, X^2, \dots$ so X^n converges to the distribution p of X.

Reasonable requirement: p is invariant distribution for Markov chain. That is, if $X^i \sim p$ then also $X^{i+1} \sim p$. This is implied by reversibility:

$$P(X^i \in A, X^{i+1} \in B) = P(X^i \in B, X^{i+1} \in A) \quad \text{when } X^i \sim p$$

(set B equal to sample space S of X. Then reversibility implies $P(X^i \in A) = P(X^{i+1} \in A)$)

Gibbs sampler update: given $X^i = x^i$ pick l in V and let $X^{i+1} = (X^i_{-l}, Y_l)$ where Y_l is sampled from conditional distribution of $X_l | X_{-l} = x^i_{-l}$.

l can be chosen at random in V or we can run through V in a systematic order.

Gibbs update is reversible

Let $S = \prod_{l \in V} S_l$ be sample space of X.

$$P(X^i \in A, X^{i+1} \in B) = \int_S \int_{S_l} 1[(x_{-l}, y_l) \in B, x \in A] p_l(y_l | x_{-l}) \mathrm{d}y_l p(x) \mathrm{d}x$$

Moreover, using a change of variable,

$$P(X^i \in B, X^{i+1} \in A) = \int_S \int_{S_l} 1[(x_{-l}, y_l) \in A, x \in B] p_l(y_l | x_{-l}) \mathrm{d}y_l p(x) \mathrm{d}x$$

$$= \int_S \int_{S_l} 1[x \in A, (x_{-l}, y_l) \in B] p_l(x_l | x_{-l}) \mathrm{d}x_l p(x_{-l}, y_l) \mathrm{d}x_{-l} \mathrm{d}y_l$$

These two integrals are equal since

$$p(x) p_l(y_l | x_{-l}) = p(y_l, x_{-l}) p_l(x_l | x_{-l})$$

Under weak regularity conditions one can show that the Gibbs sampler Markov chain converges to $p(\cdot)$.

I.e. $X^1, X^2, \ldots$ serves as a random sample of (dependent) observations from $p(\cdot)$.

Estimation

Suppose we have observed realization of auto-logistic model.

Likelihood is

$$p(x;\alpha,\beta) = p(0;\alpha;\beta)\exp(\alpha\sum_{i\in V} x_i + \beta\sum_{i\sim j} x_i x_j)$$

Problem: normalizing constant

$$c(\alpha,\beta) = [p(0;\alpha,\beta)]^{-1} = \sum_{x\in V^{\{0,1\}}} \exp(\alpha\sum_{i\in V} x_i + \beta\sum_{i\sim j} x_i x_j)$$

can not be evaluated exactly and is difficult to approximate numerically.

Besag's pseudo-likelihood

Likelihood function for auto-logistic is intractable due to unknown normalizing constant

Julian Besag suggested to maximize the pseudo-likelihood (product of full conditionals)

$$PL(\alpha, \beta) = \prod_{i \in V} p_i(x_i | x_{-i}; \alpha, \beta)$$

Not likelihood except if X_i's independent.

Score of log pseudo-likelihood is an unbiased estimating function

$$\mathbb{E} \frac{\mathrm{d}}{\mathrm{d}\alpha \mathrm{d}\beta} \log p_i(X_i | X_{-i}; \alpha, \beta) = 0$$

(Bartlett identity) and one can show that PL estimates are asymptotically normal.

Computationally straightforward - formally equivalent to logistic regression.

Bayesian image analysis

Consider a pixel image $X = (X_i)_{i \in V}$ where X_i represents the color/intensity for pixel i.

Suppose we observe "dirty" image Y where

$$Y_i = X_i + \epsilon_i$$

where ϵ_i represents independent zero-mean noise terms with some density $\epsilon_i \sim f$.

We want to reconstruct X given observation y of Y !

Idea behind Bayesian image analysis: represent prior beliefs about X using a probability distribution and infer X using posterior distribution $X|Y = y$.

Pixel values continuous

Suppose $X_i \in \mathbb{R}$. We believe neighbouring pixel values are similar. We might model this using Gaussian MRF introduced in previous lecture. I.e. with $\mu_i = \mu$ and $\kappa_i = \kappa$,

$$X_i | X_{-i} = x_{-i} \sim N(\mu + \frac{1}{\#N_i + \tau} \sum_{k \sim i} (x_k - \mu), \frac{\kappa}{\#N_i + \tau})$$

That is the conditional mean of X_i is essentially μ corrected with average deviations for neighbours. Joint density is of form

$$p(x) \propto \exp[-\frac{1}{2}(x-\mu)^{\mathsf{T}}(Q + \tau I)(x - \mu)]$$

Assume $\epsilon_i \sim N(0, \sigma^2)$. Posterior is

$$p(x|y) \propto p(y|x)p(x) \propto \exp[-\frac{1}{2\sigma^2} \sum_{i \in V} (y_i - x_i)^2] p(x) \qquad (2)$$

which is again a Gaussian MRF.

Posterior is known exactly (we can evaluate normalizing constant).

Note also: posterior Gaussian MRF is well-defined also with $\tau = 0$ in which case it does not depend on μ.

This is nice since we then do not need to specify μ.

Image segmentation

Image consists of two types (e.g. black or white) homogeneous regions. We may take $X_i \in \{0, 1\}$ with 0 for black and 1 for white.

Homogeneity: most neighbouring pixel values are of the same type $\Rightarrow$ use Ising model as prior !

Assume again Gaussian noise. Then posterior is

$$p(x|y) \propto \exp(-\frac{1}{2\sigma^2}\sum_{i\in V}(y_i - x_i)^2 + \tilde{\alpha}\sum_{i\in V} x_i + \tilde{\beta}\sum_{i\sim j} 1[x_i = x_j])$$

Again MRF distribution !

This time normalizing constant intractable but we can at least simulate posterior using Gibbs sampler.

We may want to use symmetric prior with $\tilde{\alpha} = 0$.

Contingency tables and graphical models

Consider a K-way contingency table given by combinations of K factors where the kth factor has values in set S_k.

For example 3 factors Smoker $S_1 = \{yes, no\}$, lung cancer $S_2 = \{yes, no\}$, Age $S_3 = \{young, middle, old\}$.

Consider an individual/object which is classified according to random values of these factors - leads to discrete random vector X that takes value $x = (x_1, \dots, x_K)$ if factor l takes the value x_l. E.g. outcome could be $(yes, no, middle)$ if person is middle-aged smoker without lung cancer.

Let

$$p(x) = P(X = x)$$

for x in sample space $S = \prod_{k=1}^{K} S_k$. E.g. $p(yes, no, middle)$ is probability of above outcome.

Suppose we have n individuals with vectors $X_1, \ldots, X_n$. Let N_x denote the number of individuals with $X_i = x$.

We can model vector of numbers $N = (N_x)_{x \in S}$ of individuals for each combination x of factor levels using a multinomial distribution $N \sim \text{multinomial}(n, (p(x))_{x \in S})$.

Imposing a MRF structure on probability $p(x)$ allows us to study conditional independence properties of various factors. E.g. is smoking conditionally independent of lung cancer given age ? (OK, not true :))

Conditional independence structure can be visualized via accompanying graph where vertices represent factors.

Phase transitition

Ernst Ising proposed his model as a model for ferromagnets. The spins represent orientations of iron-atoms. If majority of spins either + or - then the piece of iron is a magnet.

Consider the model with $\tilde{\alpha} = 0$ (no prefence for either + or -)

In one dimension, the Ising model is a Markov chain. According to the central limit theorem $M = \frac{1}{\sqrt{n}} \sum_{i \in V} x_i$ will converge to a zero mean normal distribution. I.e. distribution centered on configurations with roughly equal numbers of + and -.

In two or more dimensions the picture is completely different. There exists a critical value $\tilde{\beta}_c \approx 0.88$ so that for $\tilde{\beta} < \tilde{\beta}_c$, the distribution of M is unimodal, while for $\tilde{\beta} > \tilde{\beta}_c$, the distribution is bi-modal ! I.e. either majority of + or majority of - !

You can observe this by simulation: run a Gibbs sampler for large number of iterations starting from a random starting point (X_i^1 + or - with probability 0.5 each and initial spins independent).

For super critical $\tilde{\beta} > 0.88$ the Markov chain will end up in configurations dominated by either $+$ or -. And once in a configuration with majority of $+$ it takes a (very) long time to move to a configuration with a majority of - (and vice versa).

If β sub critical roughly equal amount of $+$ and -

Exercises

1. Identify the ϕ_C functions for the auto-logistic model (following proof of the Hammersley-Clifford theorem, use $y = (0, 0, \ldots, 0)$).
2. Use Brook's lemma to identify $p(\cdot)$ for the auto-logistic model. Does the result depend on the order of the factorization ?
3. Show that (1) is equivalent to the auto-logistic model in the case where all pixels have 4 neighbours

   Hint: $1[x_i = x_j] = x_i x_j + (1 - x_i)(1 - x_j)$ when $x_i, x_j \in \{0, 1\}$.
4. Auto-Poisson: suppose $X_i | X_{-i} = x_{-i}$ is Poisson with mean $\exp(\alpha + \beta \sum_{j \in N_i} x_j)$ with neighbourhood structure as for the auto-logistic. Find the joint distribution of X. Show that it is well-defined when $\beta \leq 0$ (meaning $\sum_{x \in S} h(x) < \infty$) but not ($\sum_{x \in S} h(x) = \infty$) when $\beta > 0$ and $h(\cdot)$ denotes the unnormalized simultaneous density.

Exercises continued

5. How can you simulate a Gaussian MRF when the Cholesky decomposition $Q = LL^{\mathsf{T}}$ has been obtained for the precision matrix ?
6. Show that the posterior distribution (2) is a Gaussian MRF. Also show that the posterior does not depend on μ when $\tau = 0$.

   Hint: if $Z \sim N_n(\xi, K^{-1})$, then

   $$p(z) \propto \exp(-\frac{1}{2}z^{\mathsf{T}}Kz + z^{\mathsf{T}}K\xi).$$

7. Implement and run Gibbs sampler for the Ising model (1) with $x_i \in \{0, 1\}$. Use fixed boundary with all boundary pixels equal to 1. Consider the symmetric case $\tilde{\alpha} = 0$ and values of $\tilde{\beta} = 0.4, 0.7, 0.9$. What do you observe ? (some code available on webpage).

Exercises continued

8.
   8.1 Show that the score function of pseudo-likelihood is unbiased.
   8.2 Implement pseudo-likelihood for auto-logistic model when a fixed boundary condition is used (use `R`-procedure `glm`) (some code available on webpage).
   8.3 Estimate α and β from the data set `isingdata.txt` using fixed boundary condition.

       The data was generated from (1) with $\tilde{\alpha} = 0$ and $\tilde{\beta} = 0.4$. Do your estimates of α and β seem reasonable compared to this ?

9. The data set `imageAnoisy.txt` contains a binary (black/white) image corrupted by iid normal noise with mean zero and standard deviation 0.25. You can read and view it using `temp=as.matrix(read.table("imageAnoisy.txt"))` and `image(temp)`. Adapt the previously constructed Gibbs sampler to sample from the posterior distribution when the Ising model is used as a prior. Use toroidal edge correction and try out different $\tilde{\beta}$ values.

Exercises continued

10. Consider the posterior distribution in exercise 6 with $\tau = 0$. Show that the posterior mean is

$$\hat{x} = \frac{1}{\sigma^2}(Q + \frac{1}{\sigma^2} I)^{-1} y$$

Compute the posterior mean based on the image data from previous exercise ($\sigma^2 = 0.25$). Try out varying values of κ.

Hint: use the sketch code `bayesian_GMRF.R`. Explain what is going on. Note moreover that $x = K^{-1}y \Leftrightarrow Kx = y$. If K is positive definite, $K = U^\mathsf{T} U$ for an upper triangular U. Thus we can solve $Kx = y \Leftrightarrow U^T Ux = y$ in two steps involving first U^T and next U. Each step is computationally efficient because U and U^T are triangular matrices.