# Lecture 8
# Models for Non-Gaussian Data

Dennis Sun
Stanford University
Stats 253

July 10, 2015

1. So Far

2. AR Processes for Binary Data

3. (Tractable) Estimation Techniques

4. Data Example

1. So Far

2. AR Processes for Binary Data

3. (Tractable) Estimation Techniques

4. Data Example

# The Methods so Far

- So far, all the methods we've seen assume the model

  $$Y = \text{trend} + \text{noise},$$

  where the noise is correlated.
- The methods work best when the noise is Gaussian, but they make sense (e.g., best linear unbiased) as long as $Y$ is continuous (although it may be worth trying transformations to achieve Gaussianity).
- This model makes very little sense when $Y$ is Poisson or binary.

# A More General Model

- Instead, we can suppose that the trend models the mean of the random variable $Y$:

$$\mathrm{E}[Y] = \text{trend}.$$

- In the case where $Y$ is normal, this reduces to the model

$$Y = \text{trend} + \text{noise}.$$

- But when $Y$ is binary, the trend models $p = \mathrm{P}(Y = 1)$, e.g.,

$$p_i = \frac{\exp\left\{\mathbf{x}_i^T \boldsymbol{\beta}\right\}}{1 + \exp\left\{\mathbf{x}_i^T \boldsymbol{\beta}\right\}} \quad p(y_i) = \frac{\exp\left\{y_i \mathbf{x}_i^T \boldsymbol{\beta}\right\}}{1 + \exp\left\{\mathbf{x}_i^T \boldsymbol{\beta}\right\}} \propto \exp\left\{y_i \mathbf{x}_i^T \boldsymbol{\beta}\right\}$$

- When $\mathrm{E}[y_i] = \mu(\mathbf{x}_i^T \boldsymbol{\beta})$, where $y_i$ is from an exponential family, we call this a **generalized linear model**. Examples include logistic regression and Poisson regression.

# A Correlated Model for Binary Data?

- How would you model Bernoulli random variables $y_1, ..., y_n$ with means $p_1, ..., p_n$ so that they are correlated?
- For a given (positive-semidefinite) covariance matrix, there may not exist Bernoulli random variables with that covariance.
- Covariance modeling is hopeless.
- How about an autoregressive process?

$$p(y_i|\mathbf{y}_{-i}) \propto \exp\left\{y_i\Big(\mathbf{x}_i^T\boldsymbol{\beta} + \phi\sum_j w_{ij}(y_j - \mu(\mathbf{x}_j^T\boldsymbol{\beta}))\Big)\right\}$$

(Note: $\sum_j w_{ij} = 1$.) This is also called the **autologistic model**.

1. So Far

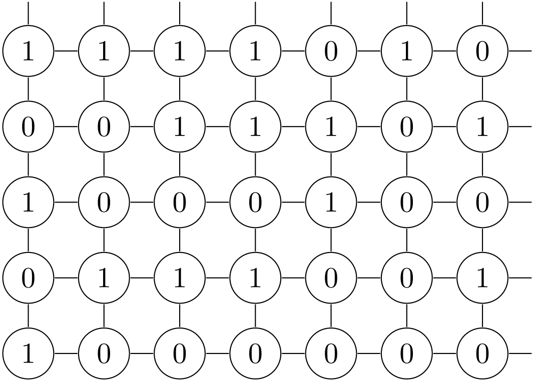
2. AR Processes for Binary Data

3. (Tractable) Estimation Techniques

4. Data Example

# Ising Model

The **Ising model** is a binary process on the lattice, used in statistical mechanics to model particle spins.



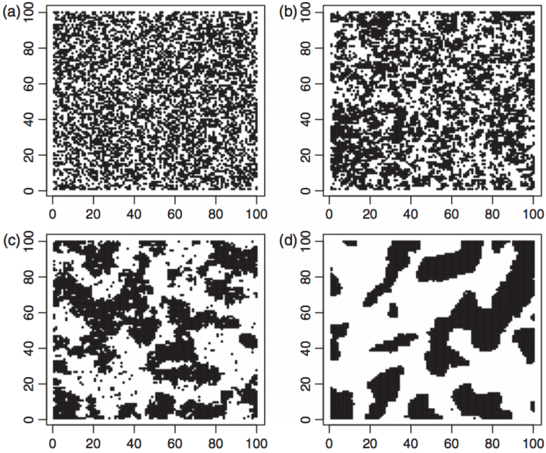
| | | | | | | |
|---|---|---|---|---|---|---|
| 1 | 1 | 1 | 1 | 0 | 1 | 0 |
| 0 | 0 | 1 | 1 | 1 | 0 | 1 |
| 1 | 0 | 0 | 0 | 1 | 0 | 0 |
| 0 | 1 | 1 | 1 | 0 | 0 | 1 |
| 1 | 0 | 0 | 0 | 0 | 0 | 0 |

It is a special case of the autologistic model for $\mathbf{x}_i^T \boldsymbol{\beta} = \log \frac{\pi}{1-\pi}$.

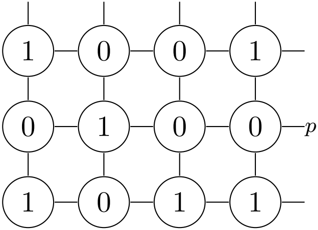
$$p(y_s | \mathbf{y}_{-s}) \propto \exp \left\{ \frac{\phi}{4} \sum_{s' \in N(s)} y_s y_{s'} + y_s (\log \frac{\pi}{1 - \pi} - \phi \pi) \right\}.$$

# Ising Model

Simulations of the Ising Model for $\phi = 1.6, 4.8, 8.0, 16.$



# What is its likelihood?



$$p(y_s|\mathbf{y}_{-s}) = \frac{\exp\left\{\frac{\phi}{4}\sum_{s'\in N(s)} y_s y_{s'} + y_s(\text{logit}(\pi) - \phi\pi)\right\}}{1+\exp\left\{\frac{\phi}{4}\sum_{s'\in N(s)} y_{s'} + (\text{logit}(\pi) - \phi\pi)\right\}}$$

By Brook's lemma, the distribution of $\mathbf{y}$ (up to a normalizing constant) is

$$\frac{p_\phi(\mathbf{y})}{p_\phi(\mathbf{0})} \propto \prod_{s=1}^{n} \frac{p_\phi(y_s|y_1,\ldots,y_{s-1},0,\ldots,0)}{p_\phi(0|y_1,\ldots,y_{s-1},0,\ldots,0)}$$

$$= \exp\left\{\frac{\phi}{4}\sum_{s=1}^{n}\sum_{s'\in N(s),s'<s} y_s y_{s'} + \sum_{s=1}^{n} y_s(\text{logit}(\pi) - \phi\pi)\right\}$$

$$= \exp\left\{\frac{\phi}{8}\sum_{s=1}^{n}\sum_{s'\in N(s)} y_s y_{s'} + (\text{logit}(\pi) - \phi\pi)\sum_{s=1}^{n} y_s\right\}$$

# What is its likelihood?

To get the likelihood, we need the normalizing constant. Because probabilities have to sum to 1, this is just the sum of the above over all possible $\mathbf{y}$:

$$p_\phi(\mathbf{y}) = \frac{\frac{p_\phi(\mathbf{y})}{p_\phi(\mathbf{0})}}{\sum_{\mathbf{y}} \frac{p_\phi(\mathbf{y})}{p_\phi(\mathbf{0})}}$$

$$= \frac{\exp\left\{\frac{\phi}{8}\sum_{s=1}^{n}\sum_{s'\in N(s)} y_s y_{s'} + (\text{logit}(\pi) - \phi\pi)\sum_s y_s\right\}}{\sum_{\mathbf{y}} \exp\left\{\frac{\phi}{8}\sum_{s=1}^{n}\sum_{s'\in N(s)} y_s y_{s'} + (\text{logit}(\pi) - \phi\pi)\sum_s y_s\right\}}.$$

There are $2^n$ terms in the denominator. For a $15 \times 15$ lattice, then there are $2^{225}$ terms, which is huge! Evaluating the likelihood is intractable, much less optimizing it over $\phi$.

1. So Far

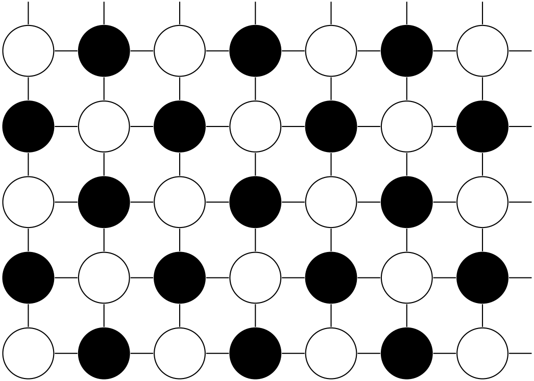
2. AR Processes for Binary Data

3. (Tractable) Estimation Techniques

4. Data Example

# Coding Estimator (Besag, 1974)

Suppose we color each site so that no site has a neighbor of the same color:



Then what is the likelihood of the observations at the black sites, conditional on the white sites?

$$p_\phi(\mathbf{y}_{black}|\mathbf{y}_{white}) = \prod_{i\ black} p_\phi(y_i|\mathbf{y}_{white}) = \prod_{i\ black} p_\phi(y_i|\mathbf{y}_{-i}).$$

# Coding Estimator

$$p_\phi(\mathbf{y}_{black}|\mathbf{y}_{white}) = \prod_{i\ black} p_\phi(y_i|\mathbf{y}_{white}) = \prod_{i\ black} p_\phi(y_i|\mathbf{y}_{-i}).$$

So one estimator is

$$\hat{\phi}_{black} = \underset{\phi}{\operatorname{argmax}} \prod_{i\ black} p_\phi(y_i|\mathbf{y}_{-i}).$$

This estimator is consistent and asymptotically normal as the number of sites increases.

But we could have reversed the roles of black and white and obtained another estimator

$$\hat{\phi}_{white} = \underset{\phi}{\operatorname{argmax}} \prod_{i\ white} p_\phi(y_i|\mathbf{y}_{-i}).$$

# Pseudolikelihood Estimator (Besag, 1975)

At the point where we have two estimators

$$\hat{\phi}_{black} = \operatorname*{argmax}_{\phi} \prod_{i\ black} p_{\phi}(y_i|\mathbf{y}_{-i})$$

$$\hat{\phi}_{white} = \operatorname*{argmax}_{\phi} \prod_{i\ white} p_{\phi}(y_i|\mathbf{y}_{-i}),$$

why not just maximize the product of the conditionals over all $i$?

$$\hat{\phi} = \operatorname*{argmax}_{\phi} \prod_{i} p_{\phi}(y_i|\mathbf{y}_{-i})$$

What the heck is $\prod_i p_{\phi}(y_i|\mathbf{y}_{-i})$? It's certainly not $p_{\phi}(\mathbf{y})$, nor a likelihood of any kind, so we call it the **pseudo-likelihood**.

In general, if you multiply all the conditionals together and pretend it's a likelihood, it's called a pseudo-likelihood.

# Results for the Pseudolikelihood

- Maximum Pseudolikelihood Estimators (MPLEs) are also consistent and asymptotically normal.

  (These essentially follow from the fact that the coding estimators are consistent and normal, plus lots of regularity conditions.)

- Standard errors are not available in closed form.

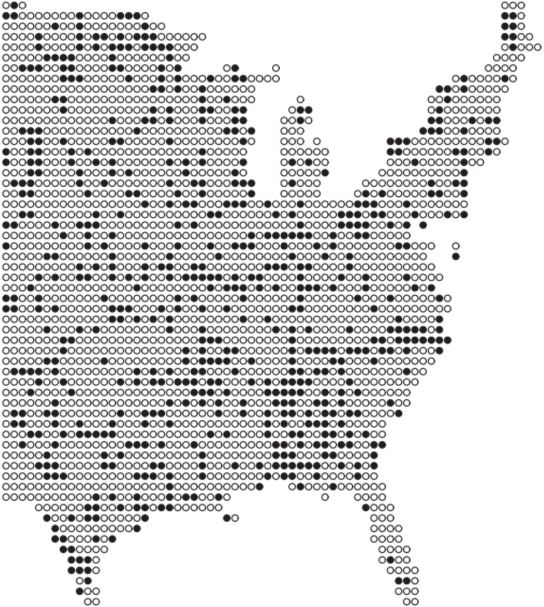
1. So Far

2. AR Processes for Binary Data

3. (Tractable) Estimation Techniques

4. Data Example

# Cancer Incidence Rates



# Maximum Pseudolikelihood Estimators

$$\prod_s p(y_s|\mathbf{y}_{-s}) = \prod_s \frac{\exp\left\{\frac{\phi}{4}\sum_{s'\in N(s)} y_s y_{s'} + y_s(\text{logit}(\pi) - \phi\pi)\right\}}{1+\exp\left\{\frac{\phi}{4}\sum_{s'\in N(s)} y_{s'} + (\text{logit}(\pi) - \phi\pi)\right\}}$$

$$= \prod_s \frac{\exp\left\{y_s(\alpha + \phi\frac{1}{4}\sum_{s'\in N(s)} y_{s'})\right\}}{1+\exp\left\{\alpha + \phi\frac{1}{4}\sum_{s'\in N(s)} y_{s'}\right\}}$$

where $\alpha = \text{logit}(\pi) - \phi\pi$. So we can obtain the MPLE by logistic regression of $y_s$ on an intercept and the average of its neighbors.

$$\hat{\alpha} = -1.49 \qquad \hat{\phi} = 1.70$$

$$\hat{\pi} = .259$$