A fast sampler for data simulation from spatial, and other, Markov random fields

Andee Kaplan

Iowa State University
ajkaplan@iastate.edu

June 22, 2017

Slides available at http://bit.ly/kaplan-phd

Joint work with M. Kaiser, S. Lahiri, and D. Nordman

Overview

Thesis: On advancing MCMC-based methods for Markovian data structures with applications to deep learning, simulation, and resampling

Goal: Develop statistical inference via Markov chain Monte Carlo (MCMC) techniques in complex data problems related to statistical learning, the analysis of network/graph data, and spatial resampling

Challenge: Develop model-based methodology, which is both *statistically rigorous* and *computationally scalable*, by exploiting conditional independence

1. Statistical properties of graph models used in deep machine learning and image classification
   (Ch. 2 & 3)
2. Fast methods for simulating spatial, network, and other data
   (Ch. 4 & 5)

This talk

- Markov random field models are popular for spatial or network data
- Rather than specifying a joint distribution directly, a model is specified through a set of full conditional distributions for each spatial location
- Conditional distributions are assumed to correspond to a valid joint (e.g., sufficient conditions in Kaiser and Cressie (2000))

Goal: A new, provably fast approach for simulating spatial/network data under a Markov model

Spatial Markov random field (MRF) models

Notation

- Variables $\{Y(\boldsymbol{s}_i) : i = 1, \ldots, n\}$ at locations $\{\boldsymbol{s}_i : i = 1, \ldots, n\}$
- Neighborhoods: $\mathcal{N}_i$ specified according to some configuration
- Neighboring Values: $\boldsymbol{y}(\mathcal{N}_i) = \{y(\boldsymbol{s}_j) : \boldsymbol{s}_j \in \mathcal{N}_i\}$
- Full Conditionals: $\{f_i(y(\boldsymbol{s}_i)|\boldsymbol{y}(\mathcal{N}_i), \boldsymbol{\theta}) : i = 1, \ldots, n\}$
  - $f_i(y(\boldsymbol{s}_i)|\boldsymbol{y}(\mathcal{N}_i), \boldsymbol{\theta})$ is conditional pmf/pdf of $Y(\boldsymbol{s}_i)$ given values for its neighbors $\boldsymbol{y}(\mathcal{N}_i)$
  - Often assume a common conditional cdf $F_i = F$ form $(f_i = f)$ for all i

Formulation adaptable to non-spatial data letting $\boldsymbol{s}_i$ be a marker for observation $Y(\boldsymbol{s}_i)$ (e.g., random graphs: $\boldsymbol{s}_i$ represents a potential edge and $Y(\boldsymbol{s}_i) \in \{0, 1\}$)

Common neighborhood structures

4-nearest neighborhood

Defined by locations in cardinal directions

$$\begin{array}{ccc} \cdot & * & \cdot \\ * & \boldsymbol{s}_i & * \\ \cdot & * & \cdot \end{array}$$

$$\mathcal{N}_i = \{\boldsymbol{s}_i \pm (0,1)\} \bigcup \{\boldsymbol{s}_i \pm (1,0)\}$$

8-nearest neighborhood

Also includes neighboring diagonals

$$\begin{array}{ccc} * & * & * \\ * & \boldsymbol{s}_i & * \\ * & * & * \end{array}$$

$$\mathcal{N}_i = \{\boldsymbol{s}_i \pm (0,1)\} \bigcup \{\boldsymbol{s}_i \pm (1,0)\} \bigcup \{\boldsymbol{s}_i \pm (1,-1)\} \bigcup \{\boldsymbol{s}_i \pm (1,1)\}$$

Exponential family examples

1. Conditional Gaussian (3 parameters):

$$f_i(y(\boldsymbol{s}_i)|\boldsymbol{y}(\mathcal{N}_i), \alpha, \eta, \tau) = \frac{1}{\sqrt{2\pi\tau}} \exp\left(-\frac{[y(\boldsymbol{s}_i) - \mu(\boldsymbol{s}_i)]^2}{2\tau^2}\right)$$

   $Y(\boldsymbol{s}_i)$ given neighbors $\boldsymbol{y}(\mathcal{N}_i)$ is normal with variance τ^2 and mean

$$\mu(\boldsymbol{s}_i) = \alpha + \eta \sum_{\boldsymbol{s}_j \in \mathcal{N}_i} [y(\boldsymbol{s}_j) - \alpha]$$

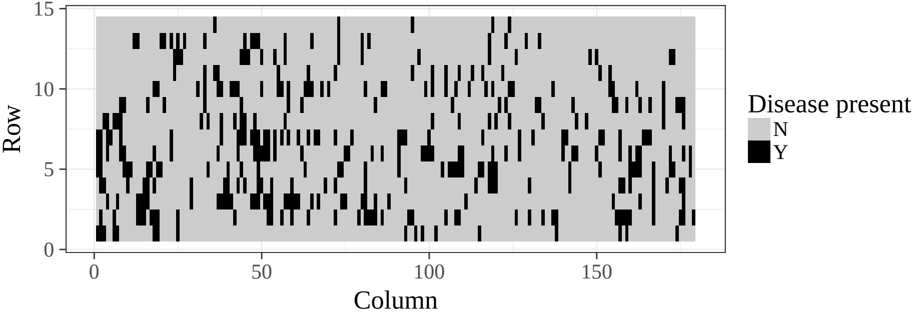
2. Conditional Binary (2 parameters):
   $Y(\boldsymbol{s}_i)$ given neighbors $\boldsymbol{y}(\mathcal{N}_i)$ is Bernoulli $p(\boldsymbol{s}_i, \kappa, \eta)$ where

$$\text{logit}[p(\boldsymbol{s}_i, \kappa, \eta)] = \text{logit}(\kappa) + \eta \sum_{\boldsymbol{s}_j \in \mathcal{N}_i} [y(\boldsymbol{s}_j) - \kappa]$$

In both examples, η represents a dependence parameter.

Illustrative Example

- For context, illustrate some common simulation demands arising in inference about spatial Markov models
- Spatial dataset from Besag (1977)
- Binary observations located on a 14 × 179 indicating the presence or absence of footrot in endive plants



Three spatial binary models

1. Isotropic centered autologistic model (Caragea and Kaiser 2009; Besag 1972; Besag 1977)

2. Centered autologistic model with two dependence parameters

3. Centered autologistic model as in (2) but having large scale structure determined by regression on the horizontal coordinate u_i of each spatial location $\boldsymbol{s}_i = (u_i, v_i)$.

Three models (Cont'd)

Conditional mass function of the form

$$f_i(y(\boldsymbol{s}_i)|\boldsymbol{y}(\mathcal{N}_i), \boldsymbol{\theta}) = \frac{\exp[y(\boldsymbol{s}_i)A_i\{\boldsymbol{y}(\mathcal{N}_i)\}]}{1 + \exp[y(\boldsymbol{s}_i)A_i\{\boldsymbol{y}(\mathcal{N}_i)\}}, \quad y(\boldsymbol{s}_i) = 0, 1,$$

with

Model	Natural parameter function
(1)	$A_i\{\boldsymbol{y}(\mathcal{N}_i)\} = \log\left(\frac{\kappa}{1-\kappa}\right) + \eta \sum\limits_{\boldsymbol{s}_j \in \mathcal{N}_i} \{y(\boldsymbol{s}_j) - \kappa\}$
(2)	$A_i\{\boldsymbol{y}(\mathcal{N}_i)\} = \log\left(\frac{\kappa}{1-\kappa}\right) + \eta_u \sum\limits_{\boldsymbol{s}_j \in N_{u,i}} \{y(\boldsymbol{s}_j) - \kappa\} + \eta_v \sum\limits_{\boldsymbol{s}_j \in N_{v,i}} \{y(\boldsymbol{s}_j) - \kappa\}$
(3)	$A_i\{\boldsymbol{y}(\mathcal{N}_i)\} = \log\left(\frac{\kappa_i}{1-\kappa_i}\right) + \eta_u \sum\limits_{\boldsymbol{s}_j \in N_{u,i}} \{y(\boldsymbol{s}_j) - \kappa_i\} + \eta_v \sum\limits_{\boldsymbol{s}_j \in N_{v,i}} \{y(\boldsymbol{s}_j) - \kappa_i\},$ $\log\left(\frac{\kappa_i}{1-\kappa_i}\right) = \beta_0 + \beta_1 \boldsymbol{u}_i$

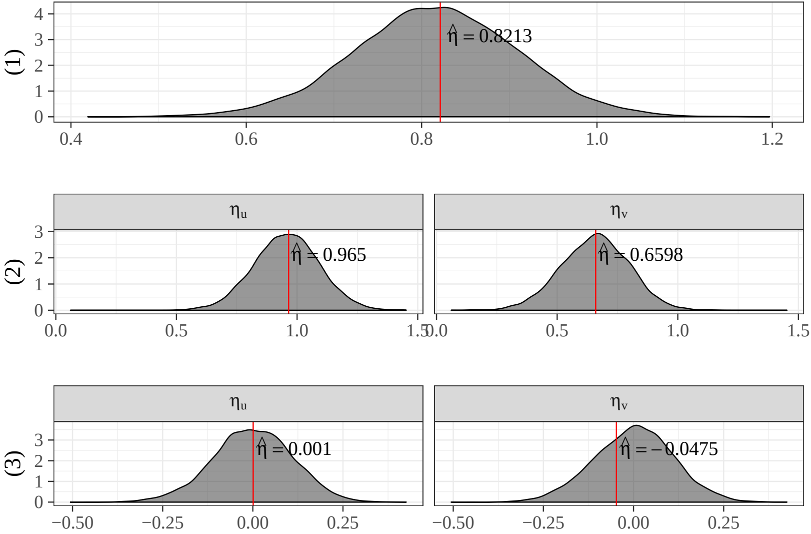
Bootstrap percentile confidence intervals

- Fit three models of increasing complexity to these data via pseudo-likelihood (Besag 1975)
- Apply simulation (parametric bootstrap) to obtain reference distributions for statistics based on the resulting estimators
- This involves the Gibbs sampler (due to the conditional model specification), where computational demands arise

	Model (1)		Model (2)			Model (3)			
	η	κ	η_u	η_v	κ	η_u	η_v	β_0	β_1
2.5%	0.628	0.107	0.691	0.378	0.106	-0.225	-0.221	-1.822	-0.003
50%	0.816	0.126	0.958	0.660	0.125	0.000	0.004	-1.600	-0.001
97.5%	1.001	0.145	1.220	0.921	0.145	0.209	0.214	-1.391	0.001

Bootstrap percentile confidence intervals in all three autologistic models

Sampling distributions via bootstrap simulation



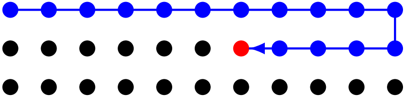
Common Spatial Simulation Approach

With common conditionally specified models for spatial lattice, standard MCMC simulation approach via Gibbs sampling is:

Starting from some initial $\boldsymbol{Y}_*^{(j)} \equiv \{Y_*^{(j)}(\boldsymbol{s}_1), \dots, Y_*^{(j)}(\boldsymbol{s}_n)\}$,

1. Moving row-wise, for $i = 1, \dots, n$, individually simulate/update $Y_*^{(j+1)}(\boldsymbol{s}_i)$ for each location $\boldsymbol{s}_i$ from conditional cdf F given

$$Y_*^{(j+1)}(\boldsymbol{s}_1), \dots, Y_*^{(j+1)}(\boldsymbol{s}_{i-1}), \quad Y_*^{(j)}(\boldsymbol{s}_{i+1}), \dots, Y_*^{(j)}(\boldsymbol{s}_n)$$



2. n individual updates provide 1 full Gibbs iteration.
3. Repeat 1-2 to obtain M resampled spatial data sets $\boldsymbol{Y}_*^{(j)}$, $j = 1, \dots, M$ (e.g., can burn-in, thin, etc.)

Endive data timing

- Endive example dataset simulations performed with the proposed (conclique-based) Gibbs sampler to follow
- Reported results would have been virtually identical with the same number of iterations to the standard sequential Gibbs sampler
- By model, generation of the reference distribution using the standard sampler would have taken approximately
    1. 25.31 minutes longer
    2. 31 minutes longer
    3. 40.7 minutes longer
- Conclique MRF sampler had running times
    1. 8.15 seconds
    2. 14.74 seconds
    3. 95.71 seconds

Concliques

Cliques – Hammersley and Clifford (1971)

Singletons and sets of locations such that each location in the set is a neighbor of all other locations in the set
Example: Four nearest neighbors gives cliques of sizes 1 and 2

The Converse of Cliques – Concliques (Kaiser, Lahiri, and Nordman 2012)

Sets of locations such that no location in the set is a neighbor of any other location in the set

4 Nearest Neighbors

·	*	·
*	***s***	*
·	*	·

Concliques 4 Nearest Neighbors

1	2	1	2
2	1	2	1
1	2	1	2
2	1	2	1

8 Nearest Neighbors

*	*	*
*	***s***	*
*	*	*

Concliques 8 Nearest Neighbors

1	2	1	2
3	4	3	4
1	2	1	2
3	4	3	4

Generalized spatial residuals (Kaiser, Lahiri, and Nordman 2012)

Definition

- $F(y|\boldsymbol{y}(\mathcal{N}_i), \boldsymbol{\theta})$ is the conditional cdf of $Y(\boldsymbol{s}_i)$ under the model
- Substitute random variables, $Y(\boldsymbol{s}_i)$ and neighbors $\{Y(\boldsymbol{s}_j) : \boldsymbol{s}_j \in \mathcal{N}_i\}$, into (continuous) conditional cdf to define residuals:

$$R(\boldsymbol{s}_i) = F(Y(\boldsymbol{s}_i)|\{Y(\boldsymbol{s}_j) : \boldsymbol{s}_j \in \mathcal{N}_i\}, \boldsymbol{\theta}).$$

Key Property

Let $\{\mathcal{C}_j : j = 1, \dots, q\}$ be a collection of concliques that partition the integer grid. Under the conditional model, **spatial residuals** *within* **a conclique are iid Uniform**$(0, 1)$**-distributed**:

$$\{R(\boldsymbol{s}_i) : \boldsymbol{s}_i \in \mathcal{C}_j\} \overset{iid}{\sim} \text{Uniform}(0, 1) \qquad \text{for } j = 1, \dots, q$$

Conclique-based Gibbs sampler

Using the conditional independence of random variables at locations within a conclique we propose a conclique-based Gibbs sampling algorithm for sampling from a MRF.

1. Split locations into Q disjoint concliques, $\mathcal{D} = \cup_{i=1}^{Q} \mathcal{C}_i$.
2. Initialize the values of $\{Y^{(0)}(\boldsymbol{s}) : \boldsymbol{s} \in \{\mathcal{C}_2, \dots, \mathcal{C}_Q\}\}$.
3. Starting from $\mathcal{C}_1$ for the i^{th} iteration, draw $\{Y^{(i)}(\boldsymbol{s}) : \boldsymbol{s} \in \mathcal{C}_1\}$ as random sample where $Y^{(i)}(\boldsymbol{s}) \overset{iid}{\sim} F(y(\boldsymbol{s}) | Y^{(i-1)}(\boldsymbol{t}), \boldsymbol{t} \in \mathcal{N}(\boldsymbol{s}))$
4. Update observations conclique-wise (using previous conclique updates).
    - For $j = 2, \dots, Q$, draw $\{Y^{(i)}(\boldsymbol{s}) : \boldsymbol{s} \in \mathcal{C}_j\}$ as random sample where $Y^{(i)}(\boldsymbol{s}) \overset{iid}{\sim} F(y(\boldsymbol{s}) | \{Y^{(i)}(\boldsymbol{t}), \boldsymbol{t} \in \mathcal{N}(\boldsymbol{s}) \cap \mathcal{C}_k \text{ where } k < j\}, \{Y^{(i-1)}(\boldsymbol{t}), \boldsymbol{t} \in \mathcal{N}(\boldsymbol{s}) \cap \mathcal{C}_k \text{ where } k > j\})$

This works by conditional independence & because neighbors for updating one conclique always belong to other concliques.

It's (computationally) fast!

- Because we are using batch updating vs. standard (i.e., single-location-wise) updating in a Gibbs sampler, this approach is **computationally fast**
- A flexible `R` package using `Rcpp` (called `conclique`, to appear on CRAN) that implements a conclique-based Gibbs sampler while allowing the user to specify an arbitrary model.
- More numerical comparisons to the standard Gibbs to follow

It's (provably) fast!

- While computationally fast, the MCMC sampler is also provably geometrically ergodic (i.e., the MCMC mixes at a fast rate) in a general sense, which is unusual for spatial data.
- State-of-the-art general theory for proving geometric ergodicity of Gibbs samplers exists only for two-state samplers (i.e., drift & minorization conditions) (Johnson and Burbank 2015).
    - For common 4-nearest neighbor spatial models, there are exactly 2 concliques (two stages in the conclique-based Gibbs sampler).
    - One can formally prove that the spatial sampler proposed is geometrically ergodic for many conditional spatial models (Gaussian, Gamma, Inverse-gamma, Beta, Binomial, etc.)

The conclique-based Gibbs sampler works

Conclique positivity condition

The full conditionals for the MRF model specify a valid joint distribution $\Pi(\cdot)$ for $(Y(\boldsymbol{s}_1), \ldots, Y(\boldsymbol{s}_n))$ with support $\mathcal{X} \subset \mathbb{R}^n$. It holds that $\mathcal{X} = \mathcal{X}_1 \times \cdots \times \mathcal{X}_Q$ where $\mathcal{X}_i$ denotes the marginal support of observations $\{Y(s_j) : s_j \in \mathcal{C}_i\}$ with locations in conclique $\mathcal{C}_i, i = 1, \ldots, Q$.

Theorem

Under the conclique positivity condition, the conclique-based Gibbs sampler is Harris ergodic with stationary distribution $\Pi(\cdot)$ *and, for any initialization* $x \in \mathcal{X}$, *the sampler converges monotonically to* $\Pi(\cdot)$ *in total variation as the number of iterations* $m \to \infty$, *i.e.,*
$\sup_{A \in \mathcal{F}} |P^{(m)}(x, A) - \Pi(A)| \downarrow 0$ *as* $m \to \infty$.

Geometric ergodicity

- $\Pi(\cdot)$ the joint distribution of observations $\{Y(\boldsymbol{s}_1), \dots, Y(\boldsymbol{s}_n)\}$ induced by a MRF specification
- $P^{(m)}(x, \cdot)$ the transition distribution at the mth iteration of the sampler with initialization $x \in \mathcal{X}$

Geometric ergodicity

The sampler is *geometrically ergodic* if there exists some real-valued function $G : \mathcal{X} \to \mathbb{R}$ and some constant $t \in (0, 1)$ which satisfy

$$\sup_{A \in \mathcal{F}} |P^{(m)}(x, A) - \Pi(A)| \leq G(x) t^m \text{ for any } x \in \mathcal{X},$$

where $\mathcal{F}$ denotes the σ-algebra associated with the joint support $\mathcal{X} \subset \mathbb{R}^n$.

Theoretical results

Theorem

Suppose a MRF model for $\{Y(\boldsymbol{s}_i) : i = 1, \dots, n\}$ admits two concliques and the conclique positivity condition holds with $\mathcal{X} = \mathcal{X}_1 \times \mathcal{X}_2 \subset \mathbb{R}^n$. If either $\mathcal{X}_1$ or $\mathcal{X}_2$ is compact and the full conditionals are continuous in conditioning variables $\boldsymbol{y}(\mathcal{N}_i)$, then, the conclique-based Gibbs sampler is geometrically ergodic with stationary distribution given by the joint, $\Pi(\cdot)$.

Ensures geometric ergodicity of the conclique-based Gibbs sampler for several four-nearest neighbor MRF models including (1) the autologistic binary, (2) the conditional binomial, (3) the conditional Beta, and (4) the Multinomial distributions as well as (5) the windsorized Poisson model of Kaiser and Cressie (1997).

Theoretical results (cont'd)

Geometric ergodicity of the conclique-based Gibbs sampling algorithm can also be established for four-nearest neighborhood MRF models with unbounded support

Theorem

Suppose $\{Y(\boldsymbol{s}_i) : i = 1, \ldots, n\}$, *having locations on a regular lattice in* $\mathbb{R}^2$, *follow a MRF model with a exponential family form and a four-nearest neighborhood structure. Then, the conclique-based Gibbs sampler is geometrically ergodic for the following cases.*

Theoretical results (cont'd)

(a) The conditional Gaussian model having conditional variance τ^2 and density

$$f_i(y(\boldsymbol{s}_i)|\boldsymbol{y}(\mathcal{N}_i)) = \frac{1}{\sqrt{2\pi\tau}} \exp\left\{-\frac{1}{2\tau^2}(y(\boldsymbol{s}_i) - \mu(\boldsymbol{s}_i))\right\}, \quad y(\boldsymbol{s}_i) \in \mathbb{R},$$

and conditional mean

$$\mu(\boldsymbol{s}_i) = \alpha + \eta \sum_{s_j \in \mathcal{N}_i} \{y(s_j) - \alpha\}$$

where $|\eta| < 0.25$ and $\alpha \in \mathbb{R}$.

Theoretical results (cont'd)

(b) The conditional (centered) Inverse Gaussian model with conditional density form

$$f_i(y_i|\boldsymbol{\theta}) = \exp\left\{\frac{A_{1i}(\boldsymbol{y}(\mathcal{N}_i))}{2} y(\boldsymbol{s}_i) - \frac{A_{2i}(\boldsymbol{y}(\mathcal{N}_i))}{2}\frac{1}{y(\boldsymbol{s}_i)} - B_i(\boldsymbol{y}(\mathcal{N}_i)) + C(y(\boldsymbol{s}_i))\right\},\ y(\boldsymbol{s}_i) \geq 1$$

where

$$A_{1i}(\boldsymbol{y}(\mathcal{N}_i)) = \frac{\lambda}{\mu^2} + \eta_1 \sum_{\boldsymbol{s}_j \in \mathcal{N}_i} \left(\frac{1}{y(\boldsymbol{s}_j)} - \frac{1}{\mu} - \frac{1}{\lambda}\right)$$

$$A_{2i}(\boldsymbol{y}(\mathcal{N}_i)) = \lambda + \eta_2 \sum_{\boldsymbol{s}_j \in \mathcal{N}_i} (y(\boldsymbol{s}_j) - \mu)$$

and $\mu, \lambda > 0$, $0 \leq \eta_1 \leq \lambda^2/4\mu(\lambda+\mu), 0 \leq \eta_2 \leq \lambda^2/4\mu$.

Theoretical results (cont'd)

(c) The conditional (centered) Truncated Gamma model with conditional density

$$f_i(y(\boldsymbol{s}_i)|\boldsymbol{\theta}) = \exp\left\{A_{1i}(\boldsymbol{y}(\mathcal{N}_i))\log(y_i) - A_{2i}(\boldsymbol{y}(\mathcal{N}_i))y_i - B_i(\boldsymbol{y}(\mathcal{N}_i))\right\}, \; y(\boldsymbol{s}_i) \geq 1$$

where

$$A_{1i}(\boldsymbol{y}(\mathcal{N}_i)) = \alpha_1 + \eta \sum_{\boldsymbol{s}_j \in \mathcal{N}_i} \log(y(\boldsymbol{s}_j)) \quad \text{and} \quad A_{2i}(\boldsymbol{y}(\mathcal{N}_i)) = \alpha_2$$

for $\eta > 0, \alpha_1 > -1, \alpha_2 > 0$.

Simulation comparisons

Quantitative framework from Turek et al. (2017) to compare conclique-based and sequential Gibbs sampler efficiency

1. Mixing effectiveness (algorithmic efficiency)
2. Computational demands of the algorithm (computational efficiency)

Algorithmic efficiency criterion:

$$A = \min_{1 \le i \le n} \left\{ \left(1 + 2 \sum_{j=1}^{\infty} \rho_i(j) \right)^{-1} \right\}$$

Computational efficiency criterion:

$$C = \begin{cases} \sum\limits_{k=1}^{Q} \text{samp}(\{Y(\boldsymbol{s}_i) : \boldsymbol{s}_i \in \mathcal{C}_k\} | \mathcal{C}_j, j \neq k) & \text{Conclique-based} \\ \sum\limits_{k=1}^{n} \text{samp}(Y(\boldsymbol{s}_k) | Y(\boldsymbol{s}_j), j \neq k) & \text{Sequential} \end{cases}$$

Simulation comparisons (Cont'd)

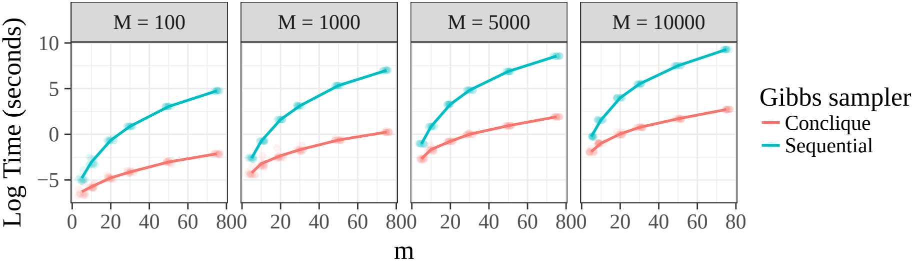
Going back to simulation from 3 binary models for spatial endive data

Gibbs	Model (1)		Model (2)		Model (3)	
	A	C	A	C	A	C
Conclique	0.807	2.9×10^{-4}	0.745	2.7×10^{-4}	0.72	3×10^{-4}
Standard	0.809	0.029	0.749	0.029	0.704	0.024

- Measures of algorithmic and computational efficiency, A and C, for three autologistic models on a 40×40 grid
- Estimates of A determined by average from 10 chains ($10,000$ iter)
- Estimates of C determined by average running times of
    - $20,000$ conclique-based samp($\{Y(\boldsymbol{s}) : \boldsymbol{s} \in \mathcal{C}_k\}$)
    - $16,000$ sequential samp($Y(\boldsymbol{s}_k)|Y(\boldsymbol{s}_j), j \neq k$)

Timing simulations

Comparisons of log time for simulation of $M = 100, 1000, 5000, 10000$ four-nearest neighbor Gaussian MRF datasets on a lattice of size $m \times m$ for various size grids, $m = 5, 10, 20, 30, 50, 75$, using sequential and conclique-based Gibbs samplers



For $10,000$ iterations/samples on 75×75 grid, conclique-based took 15.05 seconds and sequential took 1.076197×10^4 seconds $\approx$ 2.99 hours.

Another application example (Goodness of Fit)

An important question for Markov random field models with spatial data is

How to assess/diagnose fit?

Composite Hypothesis

$$H_0(C): \text{The conditional distributions of } \{Y(\boldsymbol{s}_i) : i = 1, \dots, n\} \\ \text{are } F(y(\boldsymbol{s}_i)|\boldsymbol{y}(\mathcal{N}_i), \boldsymbol{\theta})$$

where $\boldsymbol{\theta} \in \Theta$ is some *unknown* parameter value

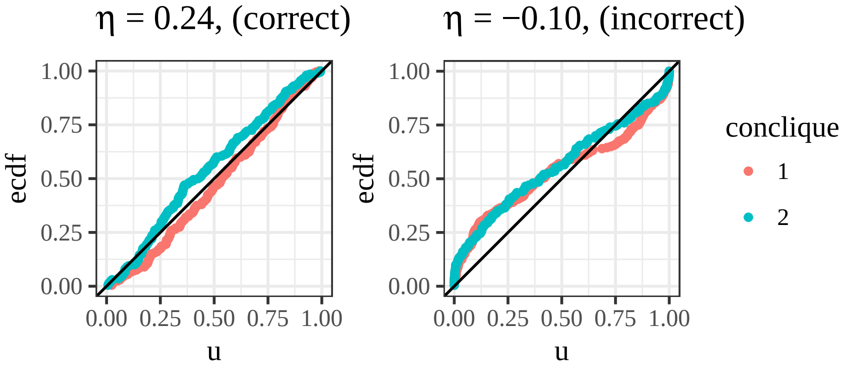
- Kaiser, Lahiri, and Nordman (2012) provide a methodology for performing GOF tests using concliques
- Conclique-based Gibbs sampling allows for fast approximation of the reference distribution for the GOF test statistics

Simple example

Gaussian Conditional Model - 20 × 20 Lattice, 4-nearest Neighbors

Let $Y(\boldsymbol{s}_i)|\boldsymbol{y}(\mathcal{N}_i) \sim N(\mu(\boldsymbol{s}_i), \tau^2)$, where $\mu(\boldsymbol{s}_i) = \alpha + \eta \sum_{\boldsymbol{s}_j \in \mathcal{N}_i} (y(\boldsymbol{s}_j) - \alpha)$.

Truth: $\alpha = 10, \tau^2 = 2, \eta = 0.24$.



From residuals to test statistics

Residual Empirical Distribution

Divide locations $\{\boldsymbol{s}_i\}_{i=1}^{n}$ into concliques: $\mathcal{C}_j$, $j = 1, \dots, q$
For j^{th} conclique, empirical cdf and and its difference to Uniform$(0, 1)$ cdf

$$G_{jn}(u) = \frac{1}{|\mathcal{C}_j|} \sum_{\boldsymbol{s}_i \in \mathcal{C}_j} I[R(\boldsymbol{s}_i) \le u]$$

$$W_{jn}(u) \equiv n^{1/2}\left[G_{jn}(u) - u\right]; \quad u \in [0, 1]$$

Test Statistics

$$\begin{aligned} T_{1n} &= \max_{j=1,\dots,q} \sup_{u \in [0,1]} |W_{jn}(u)| \\ T_{2n} &= \frac{1}{q} \sum_{j=1}^{q} \left(\int_0^1 |W_{jn}(u)|^2 du \right)^{1/2} \end{aligned}$$

Asymptotic behavior of test statistics T_{kn} is non-trivial (resampling is helpful to approximate distributions)

GOF methodology in practice

In application, a conditional distribution F model is formulated/specified.

1. Fit model $\hat{\theta}$ to original data $Y_1, \dots, Y_n$
2. Compute generalized residuals and test statistics: T_{kn}
3. Simulate spatial data $Y_1^*, \dots, Y_n^*$ from fitted cond. cdf: $F_{\hat{\theta}}$
4. Fit model to simulated data: $\hat{\theta}^*$
5. Compute generalized residuals and test statistics: T_{kn}^* from $Y_1^*, \dots, Y_n^*$ and $F_{\hat{\theta}^*}$
6. Do 3-5 many times
7. Result is reference distribution for test statistic T_{kn}

In simulating/resampling step 3 for spatial data, can use conclique-based Gibbs sampler due to the conditional specification F for each location.

Simulated example

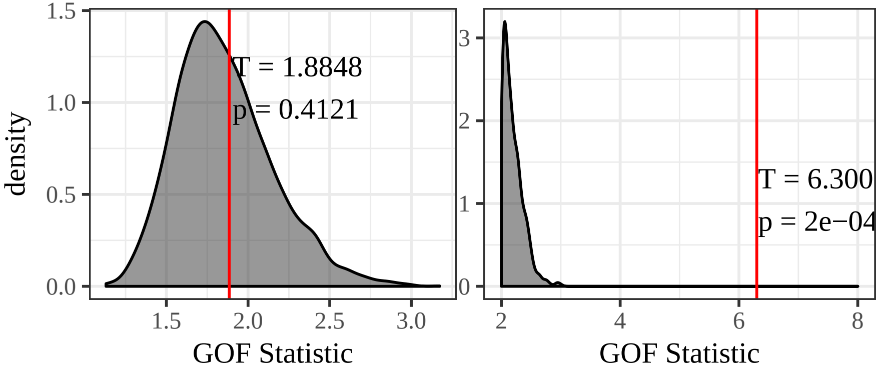
Proposed R package can do these tests with concliques

- Simulated one realization of lognormal conditionals on 20 × 20:

  $\log Y(\boldsymbol{s}_i)$ given neighbors $\{\boldsymbol{s}_i + (0, \pm 1), \boldsymbol{s}_i + (\pm 1, 0)\}$ is normal with variance τ^2 and mean $\mu(\boldsymbol{s}_i) = \alpha + \eta \sum_{\boldsymbol{s}_j \in \mathcal{N}_i} [\log y(\boldsymbol{s}_j) - \alpha]$
- Fit Gaussian MRF & fit lognormal MRF to data $Y(\boldsymbol{s}_i)$ using pseudo-likelihood

Model	Expected Value α	Conditional Variance τ^2	Dependence η	Model $p-$value
True	10	2	0.24	
Log-Gaussian	9.83	2.3	0.21	0.4121176
Gaussian	8.70362×10^4	3.5162355×10^{10}	0.17	0.00019996

Reference distributions



Bootstrapped reference distributions for the maximum across concliques of the Kolmogorov-Smirnov statistic from data generated from a four-nearest neighbor lognormal MRF with $\tau^2 = 2, \alpha = 10, \eta = 0.24$ and fit with a lognormal (left) and Gaussian (right) model.

conclique

R package (to appear on CRAN) can be installed via GitHub using the following R code.

```
devtools::install_github("andeek/conclique")
```

- Convenience functions lattice_4nn_torus and min_conclique_cover
- Gibbs samplers run_conclique_gibbs and run_sequential_gibbs
- GOF functions spatial_residuals and gof_statistics
- Bootstrap function bootstrap_gof

Extending `conclique`

One of the **key advantages** to using conclique-based approaches for simulation (and GOF tests) is the ability to consider non-Gaussian conditional models that go beyond a four-nearest neighbor structure.

`conclique` is generalizable in

- Dependence structure - beyond four-nearest neighbor
- Conditional distribution for each spatial location - beyond Gaussian and binary
- Generalized spatial residuals - for a user-supplied conditional distribution
- GOF statistics - aggregation beyond mean and max

Perks

Geometric Ergodicity

- Guaranteed convergence rate to the target joint data distribution for many (common) spatial MRF models
- With other established results, can obtain CLTs and Monte Carlo sample size assessments (Chan and Geyer 1994; Jones and others 2004; Hobert et al. 2002; Roberts, Rosenthal, and others 1997)

Speed & Flexibility

- Computationally more efficient alternative to the standard (sequential) Gibbs sampler
- Same general applicability in allowing accessible simulation for a wide variety of MRFs
  - Not limited to any one model or family or models
  - Can be applied to irregular lattices and non-standard neighborhoods

Future work and ideas

- Goodness-of-fit test for network data
  - The model-based method of resampling re-frames network into a collection of (Markovian) neighborhoods by using covariate information
  - Creates concliques on a graph structure
  - Use a conditionally specified network distribution (Casleton, Nordman, and Kaiser (2017)) to sample network data in a blockwise conclique-based Gibbs sampler.
- Bootstrap theory for approximating GOF statistics is ongoing work
- More user friendly API for `conclique` to appear on CRAN

Thank you

Questions?

- Slides – http://bit.ly/kaplan-phd
- Contact
  - Email – ajkaplan@iastate.edu
  - Twitter – http://twitter.com/andeekaplan
  - GitHub – http://github.com/andeek

References I

Besag, Julian. 1972. "Nearest-Neighbour Systems and the Auto-Logistic Model for Binary Data." *Journal of the Royal Statistical Society. Series B (Methodological)*. JSTOR, 75–83.

———. 1974. "Spatial Interaction and the Statistical Analysis of Lattice Systems." *Journal of the Royal Statistical Society. Series B (Methodological)*. JSTOR, 192–236.

———. 1975. "Statistical Analysis of Non-Lattice Data." *The Statistician*. JSTOR, 179–95.

———. 1977. "Some Methods of Statistical Analysis for Spatial Data." *Bulletin of the International Statistical Institute* 47 (2): 77–92.

Besag, Julian, and David Higdon. 1999. "Bayesian Analysis of Agricultural Field Experiments." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 61 (4). Wiley Online Library: 691–746.

Caragea, Petruța C, and Mark S Kaiser. 2009. "Autologistic Models with Interpretable Parameters." *Journal of Agricultural, Biological, and Environmental Statistics* 14 (3). Springer: 281.

Casleton, Emily, Daniel J Nordman, and Mark S Kaiser. 2017. "A Local Structure Model for Network Analysis." *Statistics and Its Interface* 10 (2). International Press of Boston, Inc.: 355–67.

Chan, Kung Sik, and Charles J Geyer. 1994. "Discussion: Markov Chains for Exploring Posterior Distributions." *The Annals of Statistics* 22 (4). JSTOR: 1747–58.

Hammersley, John M, and Peter Clifford. 1971. "Markov Fields on Finite Graphs and Lattices." *Unpublished*.

Hobert, James P, Galin L Jones, Brett Presnell, and Jeffrey S Rosenthal. 2002. "On the Applicability of Regenerative Simulation in Markov Chain Monte Carlo." *Biometrika*. JSTOR, 731–43.

Johnson, Alicia A, and Owen Burbank. 2015. "Geometric Ergodicity and Scanning Strategies for Two-Component Gibbs

References II

Samplers." *Communications in Statistics - Theory and Methods* 44 (15): 3125–45.

Jones, Galin L, and others. 2004. "On the Markov Chain Central Limit Theorem." *Probability Surveys* 1 (299-320): 5–1.

Kaiser, Mark S. 2007. "Statistical Dependence in Markov Random Field Models." *Preprint* 1. Citeseer.

Kaiser, Mark S, and Noel Cressie. 1997. "Modeling Poisson Variables with Positive Spatial Dependence." *Statistics & Probability Letters* 35 (4). Elsevier: 423–32.

———. 2000. "The Construction of Multivariate Distributions from Markov Random Fields." *Journal of Multivariate Analysis* 73 (2). Elsevier: 199–220.

Kaiser, Mark S, Soumendra N Lahiri, and Daniel J Nordman. 2012. "Goodness of Fit Tests for a Class of Markov Random Field Models." *The Annals of Statistics* 40 (1). Institute of Mathematical Statistics: 104–30.

Roberts, Gareth O, Jeffrey S Rosenthal, and others. 1997. "Geometric Ergodicity and Hybrid Markov Chains." *Electron. Comm. Probab* 2 (2): 13–25.

Turek, Daniel, Perry de Valpine, Christopher J Paciorek, Clifford Anderson-Bergman, and others. 2017. "Automated Parameter Blocking for Efficient Markov Chain Monte Carlo Sampling." *Bayesian Analysis* 12 (2). International Society for Bayesian Analysis: 465–90.