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

Andrea Kaplan  
Iowa State University

Follow this and additional works at: https://lib.dr.iastate.edu/etd

Part of the Statistics and Probability Commons

Recommended Citation
Kaplan, Andrea, "On advancing MCMC-based methods for Markovian data structures with applications to deep learning, simulation, and resampling" (2017). Graduate Theses and Dissertations. 15546.
https://lib.dr.iastate.edu/etd/15546

This Dissertation is brought to you for free and open access by the Iowa State University Capstones, Theses and Dissertations at Iowa State University Digital Repository. It has been accepted for inclusion in Graduate Theses and Dissertations by an authorized administrator of Iowa State University Digital Repository. For more information, please contact digirep@iastate.edu.
On advancing MCMC-based methods for Markovian data structures with applications to deep learning, simulation, and resampling

by

Andrea Kaplan

A dissertation submitted to the graduate faculty
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Major: Statistics

Program of Study Committee:
Daniel Nordman, Co-major Professor
Stephen Vardeman, Co-major Professor
Kris De Brabanter
Mark Kaiser
Daniel Nettleton

Iowa State University
Ames, Iowa
2017
DEDICATION

For Carrie, I wouldn’t be here without your unwavering support and remarkable patience.
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>LIST OF TABLES</td>
<td>v</td>
</tr>
<tr>
<td>LIST OF FIGURES</td>
<td>viii</td>
</tr>
<tr>
<td>ACKNOWLEDGEMENTS</td>
<td>xii</td>
</tr>
<tr>
<td>ABSTRACT</td>
<td>xiii</td>
</tr>
<tr>
<td>CHAPTER 1. INTRODUCTION</td>
<td>1</td>
</tr>
<tr>
<td>1.1 Restricted Boltzmann machines</td>
<td>1</td>
</tr>
<tr>
<td>1.2 Conclique-based Gibbs sampler</td>
<td>3</td>
</tr>
<tr>
<td>CHAPTER 2. PROPERTIES AND BAYES FITTING OF RESTRICTED BOLTZMANN MACHINES</td>
<td>6</td>
</tr>
<tr>
<td>Abstract</td>
<td>6</td>
</tr>
<tr>
<td>2.1 Introduction</td>
<td>7</td>
</tr>
<tr>
<td>2.2 Background</td>
<td>11</td>
</tr>
<tr>
<td>2.3 Explorations of model properties through simulation</td>
<td>18</td>
</tr>
<tr>
<td>2.4 Model Fitting</td>
<td>28</td>
</tr>
<tr>
<td>2.5 Discussion</td>
<td>35</td>
</tr>
<tr>
<td>CHAPTER 3. A NOTE ON THE INSTABILITY AND DEGENERACY OF DEEP LEARNING MODELS</td>
<td>38</td>
</tr>
<tr>
<td>Abstract</td>
<td>38</td>
</tr>
<tr>
<td>3.1 Introduction</td>
<td>39</td>
</tr>
</tbody>
</table>
LIST OF TABLES

Table 2.1 Parameters used to fit a test case with \( V = H = 4 \). This parameter vector was chosen as a sampled value of \( \theta \) that was not near the convex hull of the sufficient statistics for a grid point in Figure 2.8 with < 5% near-degeneracy. .................................................. 30

Table 2.2 The values used for the hyperparameters for all three fitting methods. A rule of thumb is imposed which decreases prior variances for the model parameters as the size of the model increases and also shrinks \( \theta_{interaction} \) more than \( \theta_{main} \). The common \( C \) defining \( C_1 \) and \( C_2 \) in the BwTPLV method is chosen by tuning. .................................................. 32

Table 2.3 The effective sample sizes for a chain of length \( M = 1000 \) regarding all 16 probabilities for possible vector outcomes of visibles. BwTNLV would require at least 4.7 times as many MCMC iterations to achieve the same amount of effective information about the posterior distribution. ........ 34

Table 4.1 Full conditional distributions of three binary MRF models for the endive data. .................................................. 58

Table 4.2 Bootstrap percentile confidence intervals in all three autologistic models. 58

Table 4.3 Bootstrap p-values for a goodness-of-fit (GOF) assessment of the three centered autologistic models with four-nearest neighbor structure, (a) the isotropic model with one dependence parameter, (b) the model with two dependence parameters, and (c) the model with two dependence parameters and a marginal mean structure based on the regression on the horizontal location component, \( u_i \), fit to the endive data. ................. 59
Table 4.4 Measures of algorithmic and computational efficiency, A and C, for three autologistic models presented in Section 4.2.2 on a 40 × 40 grid. We compare the metrics for a conclique-based Gibbs sampler and a sequential sampler. .......................................................... 65

Table 5.1 The test statistic, maximum of KS statistics, calculated using residuals from both correct and incorrect models. .......................................................... 79

Table 5.2 Maximum pseudo-likelihood estimates for η, κ, and ρ² for data realized from a four-nearest neighbor Gaussian MRF with ρ² = 2, κ = 10, η = .24. 79

Table D.1 The transition densities for each of the three Gibbs sampling techniques, CGS, RQGS, and RSGS, where 1(·) is the indicator function and x−i = {xj, j ≠ i}. .......................................................... 95
## LIST OF FIGURES

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>Ten 4-pixel images simulated from a degenerate model (a) compared to ten 4-pixel images simulated from a non-degenerate model (b). The degenerate model places almost all probability on one outcome, causing the image to be generated repeatedly, whereas the non-degenerate model shows more realistic variation.</td>
<td>8</td>
</tr>
<tr>
<td>2.2</td>
<td>An example restricted Boltzmann machine (RBM), consisting of two layers, one hidden ($\mathcal{H}$) and one visible ($\mathcal{V}$), with no connections within a layer. Hidden nodes are indicated by gray filled circles and the visible nodes indicated by unfilled circles.</td>
<td>11</td>
</tr>
<tr>
<td>2.3</td>
<td>A small example restricted Boltzmann machine (RBM), with two nodes, one hidden and one visible.</td>
<td>19</td>
</tr>
<tr>
<td>2.4</td>
<td>Two perspectives of the convex hull of the “statistic space” in three dimensions for the toy RBM with one visible and one hidden node.</td>
<td>20</td>
</tr>
<tr>
<td>2.5</td>
<td>Contour plots for the three parametric mean functions of sufficient statistics for a RBM with one visible and one hidden node.</td>
<td>21</td>
</tr>
<tr>
<td>2.6</td>
<td>ELPR($\theta$)/$n_v$ for various values of $\theta$ for the tiny example model. Recall here $n_v$ is the number of visible nodes and here is 1. This quantity is large for large magnitudes of $\theta$.</td>
<td>22</td>
</tr>
</tbody>
</table>
Figure 2.7  The absolute difference between coordinates of model expectations, $E[\mathbf{X} | \boldsymbol{\theta}]$, and expectations given independence, $E[\mathbf{X} | \boldsymbol{\theta}^+]$ for a RBM with one visible and one hidden node. As an indicator of uninterpretability, note that differences in expectations increase as the dependence parameter $\theta_{11}$ deviates from zero. 

Figure 2.8  Results from the numerical experiment, here looking at the fraction of models that were near-degenerate for each combination of magnitude of $\boldsymbol{\theta}$ and model size, where $\boldsymbol{\theta} = (\theta_{\text{main}}, \theta_{\text{interaction}})$ is split into main and interaction parameters. Black lines show the contour levels for fraction of near-degeneracy, while the thick black line shows the level where the fraction of near-degenerate models is .05.

Figure 2.9  The sample mean value of $\text{ELPR}(\boldsymbol{\theta})/V$ at each grid point for each combination of magnitude of $\boldsymbol{\theta}$ and model size. As the magnitude of $\boldsymbol{\theta}$ grows, so does the value of this metric, indicating typical instability in the model.

Figure 2.10  The sample mean of the maximum component of the absolute difference between the model expectation vector, $E[\mathbf{X} | \boldsymbol{\theta}]$, and the expectation vector given independence, $E[\mathbf{X} | \boldsymbol{\theta}^+]$. Larger magnitudes of $\boldsymbol{\theta}$ correspond to larger differences, thus indicating reduced interpretability.

Figure 2.11  The fraction of near-degenerate models for each magnitude of $\boldsymbol{\theta}$. For each number $n_v$ of visibles in the model, the fraction near-degenerate moves away from zero at larger values of $||\boldsymbol{\theta}||$ as the number $n_h$ of hidden variables increases and the slope becomes steeper as $n_h$ increases as well.

Figure 2.12  The autocorrelation functions (ACF) for the posterior probabilities of all $2^4 = 16$ possible outcomes for the vector of four visibles assessed at multiple lags for each method with BwTNLV in black and BwTNML in red. As expected, ACF corresponding to the method that marginalizes out the hidden variables from the likelihood decreases to zero at a much faster rate, indicating better mixing for the chain.
Figure 2.13  Posterior probabilities of $16 = 2^4$ possible realizations of 4 visibles using two of the three Bayesian fitting techniques, BwTPLV and BwTNML. The black vertical lines show the true probabilities of each vector of visibles based on the parameters used to generate the training data while the red vertical lines show the empirical distribution. BwTNML produces the best fit for the data, however is also the most computationally intensive and least feasible with a real dataset.

Figure 4.1  The endive dataset, a $14 \times 179$ rectangular lattice with binary data encoding the presence or absence of footrot in endive plants from Besag (1977).

Figure 4.2  Sampling distribution of the dependence parameters ($\eta$, $\eta_u$, and $\eta_v$) for the three centered autologistic models with four-nearest neighbor structure, (a) the isotropic model with one dependence parameter, (b) the model with two dependence parameters, and (c) the model with two dependence parameters and a marginal mean structure based on the regression on the horizontal location component, $u_i$.

Figure 4.3  Comparisons of log time for simulation of $M = 100, 1000, 5000, 10000$ four-nearest neighbor 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.

Figure 5.1  Minimal conclique cover for a given lattice of size $20 \times 20$ using a four-nearest neighbors structure wrapped on a torus. This results in 2 concliques.

Figure 5.2  Comparisons of log time for simulation of four-nearest neighbor Gaussian MRF data 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.
Figure 5.3  By conclique, the empirical cdf of generalized spatial residuals based on data from a Gaussian MRF with $\rho^2 = 2, \kappa = 10, \eta = 0.24$ when computing residuals from the correct model (left) and an incorrect model with $\eta = -0.10$ (right). The black line represents the cdf of a Uniform(0,1).

Figure 5.4  A summary plot of the bootstrapped reference distribution for the maximum across concliques of the Kolmogorov-Smirnov statistic from data generated from a four-nearest neighbor Gaussian MRF with $\rho^2 = 2, \kappa = 10, \eta = 0.24$. We fail to reject the null hypothesis, meaning this is an adequate model for the data.
ACKNOWLEDGEMENTS

I would like to take a moment and thank everybody who has helped me get to this point. There are those that have helped me through the long night that is graduate school and those that helped convince me that graduate school was the right choice for me. I could write a thesis-length ‘thank you’ section, but I will try to keep this brief.

To my advisors, Dr. Dan Nordman and Dr. Steve Vardeman, I want to thank you for being such wonderful mentors. I couldn’t ask for more patient and understanding people to work with. Working with you has been an inspiration to me. I want to thank you for giving me such positive role models and showing me how to be a researcher, educator, and a statistician.

I would like to thank my committee members for their time and patience: Dr. Mark Kaiser, Dr. Kris DeBrabanter, and Dr. Dan Nettleton. I would also like to thank the Department of Statistics for encouraging a collaborative environment and giving me an opportunity to thrive. I have been very lucky to have mentors in the department in many capacities who should also be acknowledged. Thank you to Dr. Heike Hofmann, Dr. Ulrike Genschel, Dr. Dan Nettleton, and Dr. Di Cook for the guidance that you have given me in my time at Iowa State. I hope you know that I appreciated your time and mentorship greatly.

Finally, I would like to thank my friends and family for always believing in my success and providing me with the support to realize it. Thank you to my wife, Carrie Chennault. I simply don’t know where I would be without you. You have helped make the last five years fly by with adventures and laughs. For my Dad, thank you for always lending an ear even when you have no idea what I’m talking about. I know you’re proud of me, thank you for always telling me. To my friends, thank you for keeping me sane during this time. If we’ve shared a belay, bike ride, or beer, I’m talking to you.
Markov chain Monte Carlo (MCMC) is a computational statistical approach for numerically approximating distributional quantities useful for inference that might otherwise be intractable to directly calculate. A challenge with MCMC methods is developing implementations which are both statistically rigorous and computationally scalable to large data sets. This work generally aims to bridge these aspects by exploiting conditional independence, or Markov structures, in data models. Chapter 2 investigates the model properties and Bayesian fitting of a graph model with Markovian dependence used in deep machine learning and image classification, called a restricted Bolzmann machine (RBM), and Chapter 3 presents a framework for describing inherent instability in a general class of models which includes RBMs. Chapters 4 and 5 introduce a fast method for simulating data from a Markov Random Field (MRF) by exploiting conditional independence specified in the model and a flexible R package that implements the approach in C++.
CHAPTER 1. INTRODUCTION

This thesis is most generally focused on the development of 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.

In particular, MCMC is a computational statistical approach for numerically approximating distributional quantities useful for inference that might otherwise be intractable to directly calculate. However, a challenge with MCMC methods is developing implementations which are both statistically rigorous and computationally scalable to large data sets. This work generally aims to bridge these aspects by exploiting conditional independence, or Markov structures, in data models. We investigate several problems in this context, such as (1) a statistical quantification of graph models used in deep machine learning and image classification and (2) the development of new, fast methods for simulating spatial, network, and other data with complex dependencies.

1.1 Restricted Boltzmann machines

In recent years, restricted Boltzmann machines (RBMs) have risen to prominence in the data mining and machine learning due to their connection to deep learning, specifically in stacked RBMs (see R. Salakhutdinov and Hinton 2009; R. Salakhutdinov and Hinton 2012; Srivastava, Salakhutdinov, and Hinton 2013; Le Roux and Bengio 2008 for examples). A RBM is an undirected graphical model (for discrete or continuous random variables) with two layers, one hidden and one visible, used for describing data generation (Smolensky 1986). By
incorporating a hidden layer, RBMs are thought to have the ability to encode very complex and rich structures in data, making them attractive for supervised learning. However, the statistical properties of this model for conceptualizing data are largely unexplored in the literature, and the commonly cited fitting methodology remains heuristic-based (Hinton, Osindero, and Teh 2006). In Chapter 2, we provide steps toward a thorough understanding of the model and its behavior from the perspective of statistical theory and then explore the possibility of a rigorous fitting methodology via MCMC. We have found the RBM model class to be concerning in two fundamental ways, which suggests that such models should be used with caution for inference.

First, these models are often unsatisfactory as a conceptualization of how data are generated. Recalling Fisher (1922), the aim of a statistical model is to represent data in a compact way. However, RBMs often fail to generate outcomes that resemble realistic data by only returning the data configuration with the highest probability probability under the model. In other words, when sampling data from a degenerate RBM, only a small number of output possibilities have probability greater than zero, and thus a random sample of images may consist of several copies of the same one image (or small number of images).

In addition to degeneracy, RBMs also easily exhibit a type of instability, whereby small differences in the data values lead to dramatically different probabilities for data. In practice, this is seen when a single pixel change in an image results in a wildly different classification. Such model properties have recently been observed in RBMs (J. Li 2014), as well as other deep architectures (Szegedy et al. 2013; Nguyen, Yosinski, and Clune 2014). In Chapter 3, we consider the problem of quantifying instability for general probability models defined on sequences of observations, where each sequence of length \( N \) has a finite number of possible outcomes. The results presented apply to large classes of models commonly used in random graphs, network analysis, and machine learning contexts (including the RBM model class).

Beyond model properties of the RBM class, the fitting of RBM models is problematic. As the size of these models grows, both maximum likelihood and Bayesian methods of fitting quickly become intractable. But even when fitting is feasible, there are unique challenges due
to the overparameterization of RBMs. In Chapter 2, three MCMC-based Bayesian modeling techniques are proposed and compared, each with the goal of avoiding parts of the RBM parameter space that yield improper models. With increased computational intensity comes an a potential improvement in fitting accuracy, but at the cost of feasibility. Additionally, because these RBM models are highly overparametrized, any principled fitting method will seek to reproduce the empirical distribution of most training sets. Hence, there is very little smoothing to be gained in fitting RBM models with sophisticated statistical methods. We conclude from this work that any model built using these structures, like a deep Boltzmann machine, is unlikely to achieve prediction or inference in a fully principled way without limiting the flexibility of the model class. This concern with overfitting is addressed via specification of a Bayesian prior to varying levels of success.

1.2 Conclique-based Gibbs sampler

For spatial data (including graph, network, and other data structures), conditionally specified models can be usefully formulated on the basis of an underlying Markov random field (MRF; Besag 1974). This approach often provides an attractive alternative to direct specification of a full joint data distribution, which may be difficult for large, correlated data structures (e.g. spatial data). Hence, the model is defined by prescribing a full conditional distribution for each spatial location which functionally depends on a set of observations with neighboring locations in the conditional model statement. Such Markov random field models have a natural and well-known connection to the Gibbs sampler through the conditionally specified distributions; see Besag (1994) and Kaiser and Cressie (2000). The current Gibbs strategy for simulating from such MRF models involves single-site, or sequential, updating, whereby each observation in the field is simulated or updated individually (Besag, York, and Mollie 1991).

Chapter 4 develops a new and fast way to simulate data, which has provable MCMC accuracy (or convergence rate) properties, such as geometric ergodicity\(^1\). As explained in Chapter

\(^1\)Geometric ergodicity in a Markov chain refers to the speed at which the chain converges to target distributions.
4, the simulation approach involves a type of block, or group-wise, updating Gibbs sampler based on the notion of concliques, where the latter are sets of locations which are mutually non-neighboring (cf. (Kaiser, Lahiri, and Nordman 2012)). The simulation approach is based on using the Markov random field model structure to split data into groups of non-neighboring spatial observations, called concliques. Under a hypothesized Markov model structure, spatial residuals within each conclique are independent and identically distributed as uniform variables. This means that data can be generated with a MCMC technique (i.e., a Gibbs sampler) whereby observations within each conclique are simultaneously and independently updated. A real data example is first provided to initially highlight the flexibility and speed of the method. In addition, numerical studies show that the proposed sampler exhibits algorithmic, or mixing, performance resembling that of the current standard for simulating spatial data via the sequential Gibbs sampler. While mixing behavior may be similar, the proposed sampler is illustrated to have substantial advantages in computational speed or efficiency compared to the standard Gibbs approach. These findings are useful in that model-based simulation from MRF specifications plays an important role in statistical inference, particularly in generating or approximating reference distributions for spatial statistics which might otherwise be intractable to obtain. Since the latter can require generation of large collections of spatial datasets, the proposed conclique-based Gibbs sampler can be helpful to simulating and performing inference in a computationally manageable timeframe.

Chapter 4 further formally established that the proposed conclique-based Gibbs sampling method is geometrically ergodic for many four-nearest neighbor MRF models, including models for both continuous and discrete spatial data, under minimal assumptions. Not only does this guarantee a particularly fast mixing rate for the Markov chain, but geometric ergodicity can be used with the results of Chan and Geyer (1994) to establish central limit theorems (Jones and others 2004; Hobert et al. 2002; Roberts, Rosenthal, and others 1997). This result holds, because in the four-nearest neighbor structure, it is possible to partition the spatial locations into two concliques, where observations associated with one conclique are conditionally independent given the observations in the other conclique. Currently, the sequential Gibbs
sampler for spatial datasets cannot be shown to be geometrically ergodic (for a data set of realistic size) because the state-of-the-art MCMC theory for proving geometric ergodicity is limited to less than 4-components in the Gibbs sampler; see, Johnson and Burbank (2015); Hobert and Geyer (1998); Tan and Hobert (2009); Doss and Hobert (2010); Jones and Hobert (2004); and Johnson and Jones (2015).

Chapter 5 then presents a flexible R (R Core Team 2017) package (called **conclique**, to appear on CRAN) that implements a conclique-based Gibbs sampler while allowing the user to specify an arbitrary model. **conclique** is implemented in **Rcpp** (Eddelbuettel and Francois 2011), which improves the speed of the software but also remains flexible enough to allow for the user to specify all aspects of the model, including an arbitrary Markov dependence structure. By implementing the conclique-based Gibbs sampler in R, the methodology has greater potential impact for the analysis of spatial and/or network data.
CHAPTER 2. PROPERTIES AND BAYES FITTING OF RESTRICTED BOLTZMANN MACHINES

A paper to be submitted to the *Journal of the American Statistical Association*

Andee Kaplan, Daniel J. Nordman, and Stephen B. Vardeman

**Abstract**

A restricted Boltzmann machine (RBM) is an undirected graphical model constructed for discrete or continuous random variables, with two layers, one hidden and one visible, and no conditional dependency within a layer. In recent years, RBMs have risen to prominence due to their connection to deep learning. By treating a hidden layer of one RBM as the visible layer in a second RBM, a deep architecture can be created. RBMs are thought to thereby have the ability to encode very complex and rich structures in data, making them attractive for supervised learning. However, the generative behavior of RBMs is largely unexplored. In this paper, we discuss the relationship between RBM parameter specification in the binary case and model properties such as degeneracy, instability and uninterpretability. We also describe the difficulties that arise in likelihood-based and Bayes fitting of such (highly flexible) models, especially as Gibbs sampling (quasi-Bayes) methods are often advocated for the RBM model structure.
2.1 Introduction

The data mining and machine learning communities have recently shown great interest in deep learning, specifically in stacked restricted Boltzmann machines (RBMs) (see R. Salakhutdinov and Hinton 2009; R. Salakhutdinov and Hinton 2012; Srivastava, Salakhutdinov, and Hinton 2013; Le Roux and Bengio 2008 for examples). A RBM is a probabilistic undirected graphical model (for discrete or continuous random variables) with two layers, one hidden and one visible, with no conditional dependency within a layer (Smolensky 1986). These models have reportedly been used with success in classification of images (Larochelle and Bengio 2008; Srivastava and Salakhutdinov 2012). However, the model properties are largely unexplored in the literature and the commonly cited fitting methodology remains heuristic-based and relies on rough approximation (Hinton, Osindero, and Teh 2006). In this paper, we provide steps toward a fuller understanding of the model class and its properties from the perspective of statistical model theory, and we then explore the possibility of a rigorous fitting methodology. We find the RBM model class to be concerning in two fundamental ways.

First, the models can be unsatisfactory as conceptualizations of how data are generated. That is, recalling Fisher (1922), the aim of a statistical model is to represent data in a compact way. Neyman and Box further state that a model should “provide an explanation of the mechanism underlying the observed phenomena” (Lehmann 1990; G. E. P. Box 1967). At issue, simulation from RBMs can often produce data lacking realistic variability so that such models may thereby fail to satisfactorily reflect a true data generation process. Such behavior relates to model degeneracy (or near degeneracy), which is a statistical concern in that most data processes being modeled are realistically not degenerate in their spectrum of potential outcomes. For example, when sampling data from a nearly degenerate RBM, only a small number of output possibilities have probability greater than zero, and thus a sample of images will all be copies of the same one image (or small number of images). An example of ten 4-pixel images simulated from a nearly degenerate RBM model is compared to ten 4-pixel images simulated from a non-degenerate RBM model in Figure 2.1. The degenerate model places
Figure 2.1: Ten 4-pixel images simulated from a degenerate model (a) compared to ten 4-pixel images simulated from a non-degenerate model (b). The degenerate model places almost all probability on one outcome, causing the image to be generated repeatedly, whereas the non-degenerate model shows more realistic variation.

almost all probability on one outcome, causing the image to be generated repeatedly, whereas the non-degenerate model shows more potentially realistic variation.

In addition to such degeneracy, we find that RBMs can easily exhibit types of instability, related to how sensitive the model probability structure can be to small changes in data outcomes. In practice, this may be seen when a single pixel change in an image results in a wildly different classification in an image classification problem. Occurrences of such behavior have recently been documented in RBMs (J. Li 2014), as well as other deep architectures (Szegedy et al. 2013; Nguyen, Yosinski, and Clune 2014). We describe potential issues of model instability, degeneracy and uninterpretability for the RBM class, which are related properties, and examine the presence of these in Section 2.3 through simulations of small, manageable examples.

In addition to model properties, we also explore the quality of estimation in fitting these models. The fitting can be problematic for two reasons, the first being computational and the
second concerning flexibility. As the size of these models grows, both maximum likelihood and Bayesian methods of fitting quickly become intractable. The literature often suggests Markov chain Monte Carlo (MCMC) tools for approximate maximization of likelihoods to fit these models (e.g., Gibbs sampling to exploit conditional structure in hidden and visible variables), but little is said about the attributes of realized estimates (Hinton 2010; Hinton, Osindero, and Teh 2006). Related to this, these MCMC algorithms require updating potentially many latent variables (hiddens) which can critically influence convergence in MCMC-based likelihood methods. Applying basic statistical principles in fitting RBM models of tractable sizes, we compare three fully Bayesian techniques involving MCMC, which are computationally more accessible than direct maximum likelihood and also aim to avoid parts of a RBM parameter space that yield unattractive models. As might be expected, with greater computational complexity comes an increase in fitting accuracy, but at the cost of practical feasibility.

While the computational concerns are inconvenient in fitting, issues due to model flexibility are potentially more concerning. For a RBM model with enough hidden variables, it has been shown that any distribution for the visibles can be approximated arbitrarily well (Le Roux and Bengio 2008; Montufar and Ay 2011; and Montúfar, Rauh, and Ay 2011). However, for cell data, the empirical distribution of an observed training set of visible variables provides a highest likelihood benchmark, before any parametric model class is even introduced and applied to obtain a refinement of model fit. As a consequence, we find that any fully principled fitting method based on the likelihood for a RBM with enough hidden variables will seek to reproduce the (discrete) empirical distribution of a training set. This aspect can be undesirable, and perhaps even unexpected compared to most modeling scenarios, in that no “smoothed distribution” may result when fitting a RBM model of sufficient size with a rigorous likelihood-based method. We are therefore led to be skeptical that models that involve these structures (like a deep Boltzmann machine) can achieve useful prediction or inference in a principled way without intentionally limiting the flexibility of the fitted model.

Notions of weight-decay (penalization) and sparsity (regularization) have been suggested in the RBM literature as practical remedies for both over-fitting and poor mixing in the Markov
chain during fitting (Hinton 2010; Tieleman 2008; Cho, Ilin, and Raiko 2012). Both $L_1$ and $L_2$
type penalties are mentioned to achieve weight-decay, though the benefits of these particular
forms are unknown in different situations. However, the degree to which regularization and
penalization are used by practitioners is not clear because these concepts are not (by default)
a part of the accepted fitted methodology (Hinton 2002; Carreira-Perpinan and Hinton 2005).
In this paper, we attempt to address the concerns with overfitting and poor mixing in an
alternative, and perhaps more transparent manner, via specification of a Bayesian prior in
Section 2.4.1.

This paper is structured as follows. Section 2.2 formally defines the RBM including the
joint distribution of hidden and visible variables and explains the model's connection to deep
learning. Additionally, measures of model impropriety and methods of quantifying/detecting it
are defined. Section 2.3 details our explorations into the model behavior and potential propriety
issues with the RBM class. We examine three Bayesian fitting techniques intended to avoid
model impropriety in Section 2.4 and conclude with a discussion in Section 2.5. Appendix A
provides proofs for results on RBM parameterizations and data codings described in Section
2.3.1.2.

While applications of the RBM have claimed some successes in classification problems,
it is unclear if the model class allows one to go beyond fitting to other statistical matters of
importance in using the models, such as quantification of the uncertainty of estimation and
the formulation of predictive distributions. These are closely tied to the probability properties
of the RBM model for explaining data generation and it is hoped that the current exposition
contributes to a better understanding in this direction.
2.2 Background

2.2.1 Restricted Boltzmann machines

A restricted Boltzmann machine (RBM) is an undirected graphical model specified for discrete or continuous random variables, binary variables being most commonly considered. In this paper, we consider the binary case for concreteness. A RBM architecture has two layers, hidden (\(\mathcal{H}\)) and visible (\(\mathcal{V}\)), with no dependency connections within a layer. An example of this structure is in Figure 2.2 with the hidden nodes indicated by gray circles and the visible nodes indicated by white circles. A common use for RBMs is to create features for use in classification. For example, binary images can be classified through a process that treats the pixel values as the visible variables \(v_i\) in a RBM model (Hinton, Osindero, and Teh 2006).

2.2.1.1 Joint distribution

Let \(\mathbf{x} = (h_1, \ldots, h_{n_H}, v_1, \ldots, v_{n_V})\) represent the states of the visible and hidden nodes in a RBM for some integers \(n_v, n_H \geq 1\). Each single binary random variable, visible or hidden, will take its values in a common coding set \(\mathcal{C}\), where we allow one of two possibilities for the coding.
set, $\mathcal{C} = \{0, 1\}$ or $\mathcal{C} = \{-1, 1\}$, with “1” always indicating the “high” value of the variable. While $\mathcal{C} = \{0, 1\}$ may be a natural starting point, we argue in Section 2.3 that the coding $\mathcal{C} = \{-1, 1\}$ induces more interpretable model properties for the RBM. A standard parametric form for probabilities corresponding to a potential vector of states, $X = (H_1, \ldots, H_{n_H}, V_1, \ldots, V_{n_V})$, for the nodes is

$$f_\theta(x) \equiv p_\theta(X = x) = \frac{\exp \left( \sum_{i=1}^{n_V} \sum_{j=1}^{n_H} \theta_{ij} v_i h_j + \sum_{i=1}^{n_V} \theta_{iv} v_i + \sum_{j=1}^{n_H} \theta_{hj} h_j \right)}{\gamma(\theta)}, \quad x \in \mathcal{C}^{n_H+n_V} \tag{2.1}$$

where $\theta \equiv (\theta_{11}, \ldots, \theta_{1n_H}, \ldots, \theta_{n_V1}, \ldots, \theta_{n_Vn_H}, \theta_{1v}, \ldots, \theta_{v1}, \ldots, \theta_{vn_H}) \in \mathbb{R}^{n_V+n_H+n_V \times n_H}$ denotes the vector of model parameters and the denominator

$$\gamma(\theta) = \sum_{x \in \mathcal{C}^{n_H+n_V}} \exp \left( \sum_{i=1}^{n_V} \sum_{j=1}^{n_H} \theta_{ij} v_i h_j + \sum_{i=1}^{n_V} \theta_{iv} v_i + \sum_{j=1}^{n_H} \theta_{hj} h_j \right)$$

is the normalizing function that ensures the probabilities ((2.1)) sum to one. For $x = (h_1, \ldots, h_{n_H}, v_1, \ldots, v_{n_V}) \in \mathcal{C}^{n_H+n_H}$ and

$$t(x) = (h_1, \ldots, h_{n_H}, v_1, \ldots, v_{n_V}, v_1h_1, \ldots, v_{n_V}h_{n_H}) \in \mathcal{C}^{n_H+n_V+n_H+n_V}, \tag{2.2}$$

let $\mathcal{T} = \{t(x) : x \in \mathcal{C}^{n_H+n_V}\} \subset \mathbb{R}^{n_V+n_H+n_H+n_V}$ be the set of possible values for the vector of variables needed to compute probabilities ((2.1)) in the model, and write $Q_\theta(x) = \sum_{i=1}^{n_H} \sum_{j=1}^{n_V} \theta_{ij} h_i v_j + \sum_{i=1}^{n_H} \theta_{hi} h_i + \sum_{j=1}^{n_V} \theta_{v_j} v_j$ for the “neg-potential” function. The RBM model is parameterized by $\theta$ containing two types of parameters, main effects and interaction effects. The main effects parameters ($\{\theta_{vi}, \theta_{hi}\}_{i=1, \ldots, n_V}$) weight the values of the visible $v_i$ and hidden $h_j$ nodes in probabilities ((2.1)) and the interaction effect parameters ($\theta_{ij}$) weight the values of the connections $v_i h_j$, or dependencies, between hidden and visible layers.

Due to the potential size of the model, the normalizing constant $\gamma(\theta)$ can be practically impossible to calculate, making simple estimation of the model parameter vector problematic. In model fitting, a kind of Gibbs sampling can be tried due to the conditional architecture of the RBM (i.e. visibles given hiddens or vice versa). Specifically, the conditional independence of nodes in each layer (given those nodes in the other layer) allows for stepwise simulation of both hidden layers and model parameters (e.g., see the contrastive divergence of Hinton (2002) or Bayes methods in Section 2.4).
2.2.1.2 Connection to Deep Learning

RBM have risen to prominence in recent years due to their connection to deep learning (see Hinton, Osindero, and Teh 2006; R. Salakhutdinov and Hinton 2012; Srivastava, Salakhutdinov, and Hinton 2013 for examples). By stacking multiple layers of RBMs in a deep architecture, proponents of the models claim to produce the ability to learn “internal representations that become increasingly complex, which is considered to be a promising way of solving object and speech recognition problems” (R. Salakhutdinov and Hinton 2009, 450). The stacking is achieved by treating a hidden layer of one RBM as the visible layer in a second RBM, and so on, until the desired multi-layer architecture is created.

2.2.2 Degeneracy, instability, and uninterpretability

The highly flexible nature of a RBM (having as it does $n_H + n_V + n_H * n_V$ parameters) creates at least three kinds of potential issues in model impropriety that we will call degeneracy, instability, and uninterpretability. In this section we define these characteristics, consider how to quantify these in a RBM, and point out relationships among them.

2.2.2.1 Near-degeneracy

In Random Graph Model theory, model degeneracy means there is a disproportionate amount of probability placed on only a few elements of the sample space, $\mathcal{X}$, by the model (Handcock 2003). For random graph models, $\mathcal{X}$ denotes all possible graphs that can be constructed from a set of nodes and an exponentially parameterized random graph model has a distribution of the form

$$ f_\theta(x) = \frac{\exp(\theta^T t(x))}{\gamma(\theta)}, x \in \mathcal{X}, $$

where $\theta \in \Theta \subset \mathbb{R}^q$ is the model parameter, and $t : \mathcal{X} \rightarrow \mathbb{R}^q$ is a vector of statistics based on the adjacency matrix of a graph. Here, as earlier, $\gamma(\theta) = \sum_{x \in \mathcal{X}} \exp(\theta^T t(x))$ is the normalizing
function. Let $C$ denote the convex hull of the potential outcomes of sufficient statistics, $\{t(x) : x \in X\}$, under the model above. Handcock (2003) classifies an exponentially parametrized random graph model at $\theta$ as near-degenerate if the mean value of the vector of sufficient statistics under $\theta$, $\mu(\theta) = \mathbb{E}_\theta t(X)$, is close to the boundary of $C$. Intuitively, if a model is near-degenerate in the sense that only a small number of elements of the sample space $X$ have positive probability, the expected value $\mathbb{E}_\theta t(X)$ is an average of that same small number of values of $t(x)$ (defining the boundary of the hull $C$) and can be expected to not be pulled deep into the interior of $C$.

A RBM model can be thought to exhibit an analogous form of near-degeneracy when there is a disproportionate amount of probability placed on a small number of elements in the sample space of visibles and hiddens, $C^{nv+n_h}$. Using the idea of Handcock (2003), when the random vector $t(x) = (v_1, \ldots, v_{nv}, h_1, \ldots, h_{n_h}, v_1h_1, \ldots, v_{nv}h_{n_h}) \in \mathcal{T} \equiv \{t(x) : x \in C^{nv+n_h}\}$ from ((2.2)), appearing in the neg-potential function $Q_{\theta}(-)$, has a mean vector $\mu(\theta) \in \mathbb{R}^{nv+n_h+nv+n_h}$ close to the boundary of the convex hull of $\mathcal{T}$, then the RBM model can be said to exhibit near-degeneracy at $\theta \in \mathbb{R}^{nv+n_h+nv+n_h}$. Here the mean of $t(x)$ is

$$\mu(\theta) = \mathbb{E}_\theta t(X) = \sum_{x \in C^{nv+n_h}} \{t(x)f_\theta(x)\} = \sum_{x \in C^{nv+n_h}} \left\{ \frac{\exp\left( \sum_{i=1}^{nv} \sum_{j=1}^{n_h} \theta_{ij}v_ih_j + \sum_{i=1}^{nv} \theta_{vi}v_i + \sum_{j=1}^{n_h} \theta_{hj}h_j \right)}{\sum_{x' \in C^{nv+n_h}} \exp\left( \sum_{i=1}^{nv} \sum_{j=1}^{n_h} \theta_{ij}v_ih_j + \sum_{i=1}^{nv} \theta_{vi}v_i + \sum_{j=1}^{n_h} \theta_{hj}h_j \right)} \right\}.$$ 

### 2.2.2.2 Instability

Considering exponential families of distributions, Schweinberger (2011) introduced a concept of model deficiency related to instability. Instability can be roughly thought of as excessive sensitivity in the model, where small changes in the components of potential data outcomes, $x$, may lead to substantial changes in the probability function $f_\theta(x)$. Furthermore, model instability can be viewed on a spectrum of potential sensitivity in probability structure, with model degeneracy included as a extreme or limiting case of instability. To quantify "instabil-
ity” more rigorously (particularly beyond the definition given by Schweinberger (2011)) it is useful to consider how RBM models might be expanded to incorporate more and more visibles. When increasing the size of RBM models, it becomes necessary to grow the number of model parameters (and in this process one may also arbitrarily expand the number of hidden variables used). To this end, let $\theta_n \equiv (\theta_{v1}, \ldots, \theta_{vn}, \theta_{h1}, \ldots, \theta_{hn}, \theta_{11}, \ldots, \theta_{nn}), n \geq 1$, denote an element of a sequence of RBM parameters indexed by the number $n$ of visibles ($V_1, \ldots, V_n$) and define a (scaled) extremal log-probability ratio of the RBM model at $\theta_n$ as

$$
\frac{1}{n} \log \left[ \frac{\max_{(v_1, \ldots, v_n) \in \mathcal{E}^{n_v}} P_{\theta_n}(v_1, \ldots, v_n)}{\min_{(v_1, \ldots, v_n) \in \mathcal{E}^{n_v}} P_{\theta_n}(v_1, \ldots, v_n)} \right] \equiv \frac{1}{n_v} \text{ELPR}(\theta_n)
$$

where $P_{\theta_n}(v_1, \ldots, v_n) \propto \sum_{(h_1, \ldots, h_n) \in \mathcal{E}^{n_H}} \exp \left( \sum_{i=1}^{n_v} \theta_{vi} v_i + \sum_{j=1}^{n_H} \theta_{hi} h_i + \sum_{i=1}^{n_v} \sum_{j=1}^{n_H} \theta_{ij} v_i h_j \right)$ is the RBM probability of observing outcome $(v_1, \ldots, v_n)$ for the visible variables $(V_1, \ldots, V_n)$ under parameter vector $\theta_n$, after marginalization of hidden variables.

In formulating a RBM model for a potentially large number of visibles (i.e., as $n_v \to \infty$), we will say that the ratio (\ref{eq:ELPR}) needs to stay bounded for the sequence of RBM models to be stable. That is, we make the following convention.

**Definition 1 (S-unstable RBM).** Let $\theta_n \in \mathbb{R}^{n_v + n_H + n_H \times n_v}, n \geq 1$, be an element of a sequence of RBM parameters where the number of hidden, $n_H \equiv n_H(n_v) \geq 1$, can be an arbitrary function of the number $n$ of visibles. A RBM model formulation is **Schweinberger-unstable** or **S-unstable** if

$$
\lim_{n_v \to \infty} \frac{1}{n_v} \text{ELPR}(\theta_n) = \infty.
$$

In other words, the RBM model sequence is unstable if, given any $c > 0$, there exists an integer $n_c > 0$ so that $\frac{1}{n_V} \text{ELPR}(\theta_{n_v}) > c$ for all $n_v \geq n_c$. This definition of **S-unstable** is a generalization or re-interpretation of the “unstable” concept of Schweinberger (2011) in that here RBM models for visibles $(v_1, \ldots, v_n)$ do not form an exponential family and the dimensionality of $\theta_n$ is not fixed, but rather grows with $n_v$.

S-unstable RBM model sequences are undesirable for several reasons. One is that, as mentioned above, small changes in data can lead to overly-sensitive changes in probability
under the data model. Consider, for example,

$$\Delta(\theta_{n_v}) \equiv \max \left\{ \log \frac{P_{\theta_{n_v}}(v)}{P_{\theta_{n_v}}(v^*)} : v \& v^* \in \mathcal{C}^{n_v} \text{ differ in exactly one component} \right\},$$
denoting the biggest log-probability ratio for a one component change in data outcomes (visibles) at a RBM parameter $\theta_{n_v}$. We then have the following result.

**Proposition 1.** Let $c > 0$ and let $\text{ELPR}(\theta_{n_v})$ be as in ((3.1)) for an integer $n_v \geq 1$. If $\frac{1}{n_v} \text{ELPR}(\theta_{n_v}) > c$, then $\Delta(\theta_{n_v}) > c$.

In other words, if the probability ratio ((3.1)) is too large, then a RBM model sequence will exhibit large probability shifts for very small changes in data configurations (i.e., will exhibit instability). Such instability can be a concern in a model for the similar reasons to degeneracy: as outcomes vary over the sample space, the geography of probabilities is extremely rugged, with deep pits following sharp mountains. Recall the applied example of RBM models as a means to classify images. For data as pixels in an image, the instability result in Proposition 1 manifests itself as a one pixel change in an image (one component of the visible vector) resulting in a large shift in the probability, which in turn could result in a vastly different classification of the image. Examples of this behavior have been presented in Szegedy et al. (2013) for other deep learning models, in which a one pixel change in a test image results in a wildly different classification.

Additionally, S-unstable RBM model sequences may be formally connected to the near-degeneracy of Section 2.2.2.1 (in which model sequences place all probability on a small portion of their sample spaces). To see this, define an arbitrary modal set of possible outcomes (i.e. set of highest probability outcomes) in RBM models with parameters $\theta_{n_v}, n_v \geq 1$ as

$$M_{e, \theta_{n_v}} \equiv \left\{ v \in \mathcal{C}^{n_v} : \log P_{\theta_{n_v}}(v) > (1 - \varepsilon) \max_{v'} P_{\theta_{n_v}}(v') + \varepsilon \min_{v'} P_{\theta_{n_v}}(v') \right\}$$

for a given $0 < \varepsilon < 1$. Then S-unstable model sequences are guaranteed to be degenerate, as the following result shows.

**Proposition 2.** For an S-unstable RBM model sequence and any $0 < \varepsilon < 1$,

$$P_{\theta_{n_v}} \left( (v_1, \ldots, v_{n_v}) \in M_{e, \theta_{n_v}} \right) \to 1 \text{ as } n_v \to \infty.$$
In other words, S-unstable RBM model sequences are guaranteed to stack up all probability on a specific set of outcomes for visibles, which could potentially be arbitrarily narrow. Proofs of Propositions 1 and 2 appear in Kaplan, Nordman, and Vardeman (2017). These findings also have counterparts in results in Schweinberger (2011), but unlike results there, we do not limit consideration to exponential family forms with a fixed number of parameters.

2.2.2.3 Uninterpretability

For spatial Markov models, Kaiser (2007) defines a measure of model impropriety he calls *uninterpretability*, which is characterized by dependence parameters in a model being so extreme that marginal mean-structures fail to hold as anticipated by a model statement. We adapt this notion to RBM models as follows. Note that in a RBM, the parameters $\theta_{v_1}, \ldots, \theta_{v_{nv}}$ and $\theta_{h_1}, \ldots, \theta_{h_{nh}}$ are naturally associated with main effects of visible and hidden variables and can be interpreted as (logit functions of) means for variables $V_1, \ldots, V_{nv}, H_1, \ldots, H_{nh}$ in a model with no interaction parameters, $\theta_{ij} = 0, i = 1, \ldots, n_v, j = 1, \ldots, n_h$. That is, with no interaction parameters, we have from (2.1) that

$$P_\theta(V_i = 1) \propto e^{\theta_{vi}} \quad \text{and} \quad P_\theta(H_j = 1) \propto e^{\theta_{hj}}, \quad i = 1, \ldots, n_v, j = 1, \ldots, n_h$$

so that, for example, $\text{logit}(P_\theta(V_i = 1)) = \theta_{vi}$ (or $2\theta_{vi}$) under the coding $\mathcal{C} = \{0,1\}$ (or $\{-1,1\}$). Hence, these main effect parameters have a clear interpretation under an independence model (one with $\theta_{ij} = 0$) but this interpretation can break down as interaction parameters increase in magnitude relative to the size of the main effects. In such cases, the main effect parameters $\theta_{vi}$ and $\theta_{hj}$ are no longer interpretable in the models (statements of marginal means) and the dependence parameters are so large as to dominate the entire model probability structure (also destroying the model interpretation of dependence as local conditional modifications of an overall marginal mean structure). Whether or not parameter interpretation is itself a goal in a given application of RBM models, this concept of interpretation can provide an additional device for examining other aspects of model propriety related to instability and degeneracy. As explained by Kaiser (2007), models with interpretable dependence parameters typically
correspond to non-degenerate models, while degradation in interpretability is often associated with model drift into degeneracy. To assess which parameter values $\mathbf{\theta}$ may cause difficulties in interpretation, we use the difference $E[X|\mathbf{\theta}] - E[X|\mathbf{\theta}^*]$ between two model expectations: $E[X|\mathbf{\theta}]$ at $\mathbf{\theta}$ and expectations $E[X|\mathbf{\theta}^*]$ where $\mathbf{\theta}^*$ matches $\mathbf{\theta}$ for all main effects but otherwise has $\theta_{ij} = 0$ for $i = 1, \ldots, n_v, j = 1, \ldots, n_h$. Hence, $\mathbf{\theta}^*$ and $\mathbf{\theta}$ have the same main effects but $\mathbf{\theta}^*$ has 0 dependence parameters. Uninterpretability is then avoided at a parametric specification $\mathbf{\theta}$ if the model expected value at $\mathbf{\theta}$ is not very different from the corresponding model expectation under independence. Using this, it is possible to investigate what parametric conditions lead to uninterpretability in a model versus those that guarantee interpretable models. If $E[X|\mathbf{\theta}] - E[X|\mathbf{\theta}^*]$ is large, then the RBM model with parameter vector $\mathbf{\theta}$ is said to be uninterpretable.

The quantities to compare in the RBM case are

$$E[X|\mathbf{\theta}] = \sum_{x \in \mathbb{Q}^{n_v+n_h}} x f_{\mathbf{\theta}}(x) = \sum_{x \in \mathbb{Q}^{n_v+n_h}} x \frac{\exp \left( \sum_{i=1}^{n_v} \sum_{j=1}^{n_h} \theta_{ij} v_i h_j + \sum_{i=1}^{n_v} \theta_{vi} v_i + \sum_{j=1}^{n_h} \theta_{hj} h_j \right)}{\sum_{x \in \mathbb{Q}^{n_v+n_h}} \exp \left( \sum_{i=1}^{n_v} \sum_{j=1}^{n_h} \theta_{ij} v_i h_j + \sum_{i=1}^{n_v} \theta_{vi} v_i + \sum_{j=1}^{n_h} \theta_{hj} h_j \right)}$$

and

$$E[X|\mathbf{\theta}^*] = \sum_{x \in \mathbb{Q}^{n_v+n_h}} x \frac{\exp \left( \sum_{i=1}^{n_v} \theta_{vi} v_i + \sum_{j=1}^{n_h} \theta_{hj} h_j \right)}{\sum_{x \in \mathbb{Q}^{n_v+n_h}} \exp \left( \sum_{i=1}^{n_v} \theta_{vi} v_i + \sum_{j=1}^{n_h} \theta_{hj} h_j \right)}$$

### 2.3 Explorations of model properties through simulation

We next explore and numerically explain the relationship between values of $\mathbf{\theta}$ and the three notions of model impropriety, near-degeneracy, instability, and uninterpretability, for RBM models of varying sizes.
2.3.1 Tiny example

To illustrate the ideas of model near-degeneracy, instability, and uninterpretability in a RBM, we consider first the smallest possible (toy) example that consists of one visible node \(v_1\) and one hidden node \(h_1\) that are both binary. A schematic of this model can be found in Figure 2.3. Because it seems most common, we shall begin by employing 0/1 encoding of binary variables (both \(h_1\) and \(v_1\) taking values in \(\mathcal{C} = \{0, 1\}\)). (Eventually we shall argue in Section 2.3.1.2 that \(-1/1\) coding has advantages.)

![Figure 2.3: A small example restricted Boltzmann machine (RBM), with two nodes, one hidden and one visible.](image)

2.3.1.1 Impropriety three ways

For this small model, we are able to investigate the symptoms of model impropriety, beginning with near-degeneracy. To this end, recall from Section 2.2.2.1 that one characterization requires consideration of the convex hull of possible values of statistics \(t(x)\),

\[
\mathcal{T} = \{t(x) : x = (v_1, h_1) \in \{0, 1\}^2 \} = \{(v_1, h_1, v_1 h_1) : v_1, h_1 \in \{0, 1\}\}
\]

appearing in the RBM probabilities for this model. As this set is in three dimensions, we are able to explicitly illustrate the shape of boundary of the convex hull of \(\mathcal{T}\) and explore the behavior of the mean vector \(\mu(\theta) = E_{f(x)} t(x)\) as a function of the parameter vector \(\theta\). Figure 2.4 shows the convex hull of our “statistic space,” \(\mathcal{T} \subset \{0, 1\}^3\), for this toy problem from two perspectives (enclosed by the unit cube \([0, 1]^3\), the convex hull of \([0, 1]^3\)). In this small model, note that the convex hull of \(\mathcal{T}\) does not fill the unrestricted hull of \([0, 1]^3\) because of the relationship between the elements of \(\mathcal{T} = \{(v_1, h_1, v_1 h_1 : v_1, h_1 \in \{0, 1\}\} \) (i.e. \(v_1 h_1 = 1\) only if \(v_1 = h_1 = 1\)).
Figure 2.4: Two perspectives of the convex hull of the “statistic space” in three dimensions for the toy RBM with one visible and one hidden node.

We can compute the mean vector for $t(x)$ as a function of the model parameters as

$$\mu(\theta) = E_\theta[t(X)] = \sum_{x=(v_1,h_1)\in\{0,1\}^2} \left\{ t(x) \frac{\exp(\theta_{11}v_1 + \theta_{h1}h_1 + \theta_{v1}v_1)}{\gamma(\theta)} \right\} = \begin{bmatrix} \frac{\exp(\theta_{11})+\exp(\theta_{11}+\theta_{v1}+\theta_{h1})}{\gamma(\theta)} \\ \frac{\exp(\theta_{h1})+\exp(\theta_{11}+\theta_{v1}+\theta_{h1})}{\gamma(\theta)} \\ \frac{\exp(\theta_{v1})+\exp(\theta_{11}+\theta_{v1}+\theta_{h1})}{\gamma(\theta)} \end{bmatrix}$$

where $\gamma(\theta) = \sum_{h_1=0}^{1} \sum_{v_1=0}^{1} \exp(\theta_{11}v_1 + \theta_{h1}h_1 + \theta_{v1}v_1)$. The three parametric coordinate functions of $\mu(\theta)$ can be represented as in Figure 2.5. (Contour plots for three coordinate functions are shown in columns for various values of $\theta_{11}$, which can be interpreted here as an absolute log-odds ratio as the visible changes between 0 and 1.) In examining these, we see that as coordinates of $\theta$ grow larger in magnitude, at least one mean function for the entries of $t(x)$ approaches a value 0 or 1, forcing $\mu(\theta) = E_\theta t(x)$ to be near to the boundary of the convex hull of $\mathcal{T}$, as a sign of model near-degeneracy. Thus, for a very small example we can see the relationship between values of $\theta$ and model degeneracy.

Secondly, we can look at ELPR$(\theta)$ from ((3.1)) for this tiny model in order to consider model instability as a function of RBM parameters. Recall that large values of ELPR$(\theta)$ are associated with an extreme sensitivity of the model probabilities $f_\theta(x)$ to small changes in $x$.
Figure 2.5: Contour plots for the three parametric mean functions of sufficient statistics for a RBM with one visible and one hidden node.

(see Proposition 1). The quantity $\text{ELPR}(\theta)$ for this small RBM is

$$\text{ELPR}(\theta) = \log \left[ \frac{\max_{(v_1,\ldots,v_{n_v})\in\mathcal{C}^{n_v}} P_{\theta v_1}(v_1,\ldots,v_{n_v})}{\min_{(v_1,\ldots,v_{n_v})\in\mathcal{C}^{n_v}} P_{\theta v_1}(v_1,\ldots,v_{n_v})} \right] = \log \left[ \frac{\max_{v_1, h_1\in\mathcal{C}} \sum \exp \{ \theta_{11} v_1 + \theta_{h_1} h_1 + \theta_{v_1} v_1 \}}{\min_{v_1, h_1\in\mathcal{C}} \sum \exp \{ \theta_{11} v_1 + \theta_{h_1} h_1 + \theta_{v_1} v_1 \}} \right].$$

Figure 2.6 shows contour plots of $\text{ELPR}(\theta)/n_v$ for various values of $\theta$ in this model with $n_v = 1$. We can see that this quantity is large for large magnitudes of $\theta$, especially for large values of the dependence/interaction parameter $\theta_{11}$. This suggests instability as $|\theta|$ becomes large, agreeing also with the concerns about near-degeneracy produced by consideration of $\mu(\theta)$.

Finally to consider the effect of $\theta$ on potential model uninterpretability, we can look at the difference between model expectations, $E[X|\theta]$, and expectations given independence, $E[X|\theta^*]$ for the tiny toy RBM model where $X = (V_1,H_1,V_1 H_1)$. This difference is given by

$$E[X|\theta] - E[X|\theta^*] = \left[ \frac{\exp(\theta_{11} + \theta_{2} + 2\theta_{h_1}) - \exp(\theta_{11} + 2\theta_{h_1})}{\exp(\theta_{11} + \theta_{2}) + \exp(\theta_{11} + \theta_{2} + \theta_{h_1})} \right] \left[ \frac{\exp(\theta_{11} + 2\theta_{h_1} + \theta_{1}) - \exp(2\theta_{h_1} + \theta_{1})}{\exp(\theta_{11} + 2\theta_{h_1} + \theta_{1}) + \exp(2\theta_{h_1} + \theta_{1})} \right].$$
Figure 2.6: ELPR($\theta$)/$n_v$ for various values of $\theta$ for the tiny example model. Recall here $n_v$ is the number of visible nodes and here is 1. This quantity is large for large magnitudes of $\theta$.

Figure 2.7: The absolute difference between coordinates of model expectations, $E[X|\theta]$, and expectations given independence, $E[X|\theta^*]$ for a RBM with one visible and one hidden node. As an indicator of uninterpretability, note that differences in expectations increase as the dependence parameter $\theta_{11}$ deviates from zero.

Again, we can inspect these coordinate functions of this vector difference to look for a relationship between parameter values and large values of $E[X|\theta] - E[X|\theta^*]$ as a signal of uninterpretability for the toy RBM.

Figure 2.7 shows that the absolute difference between coordinates of the vector of model expectations, $E[X|\theta]$ and corresponding expectations $E[X|\theta^*]$ given independence grow for the toy RBM as the values of $\theta$ are farther from zero, especially for large magnitudes of the dependence parameter $\theta_{11}$. This is a third indication that parameter vectors of large magnitude lead to model impropriety in a RBM.
2.3.1.2 Data encoding

Multiple encodings of the binary variables are possible. For example, we could allow hiddens \((H_1, \ldots, H_{n_H}) \in \{0,1\}^{n_H}\) and visibles \((V_1, \ldots, V_{n_V}) \in \{0,1\}^{n_V}\), as in the previous sections or we could instead encode the state of the variables as \([-1,1]^{n_H}\) and \([-1,1]^{n_V}\). This will result in variables \(t(X)\) from \((2.2)\) satisfying \(t(x) \in \{0,1\}^{n_H+n_V+n_H+n_V}\) or \(t(x) \in \{-1,1\}^{n_H+n_V+n_H+n_V}\) depending on how we encode “on” and “off” states in the nodes.

The \(-1/1\) data encoding has the benefit of providing a guaranteed-to-be non-degenerate model at \(\theta = 0 \in \mathbb{R}^{n_H+n_V+n_H+n_V}\), where the zero vector then serves as the natural center of the parameter space and induces the simplest possible model properties for the RBM (i.e., at \(\theta = 0\), all variables are independent uniformly distributed on \([-1,1]^{n_V}\)). The proof of this and further exploration of the equivalence of the \(\theta\) parameterization of the RBM model class and parameterization by \(\mu(\theta)\) is in the on-line supplementary materials. Hence, while from some computing perspectives \(0/1\) coding might seem most natural, the \(-1/1\) coding is far more convenient and interpretable from the point of view of statistical modeling, where it makes parameters simply interpreted in terms of symmetrically defined main effects and interactions. Under the data encoding \(-1/1\), the parameter space centered at 0 is also helpful for framing parameter configurations that are undesirably large (i.e., these are naturally parameters that have moved too far away from \(0\) where the RBM model is anchored to be trivially describable and completely problem-free). In light of all of these matters we will henceforth employ the \(-1/1\) coding.

2.3.2 Exploring manageable examples

To explore the impact of RBM parameter vector \(\theta\) magnitude on near-degeneracy, instability, and uninterpretability, we consider models of small size. For \(n_H, n_V \in \{1, \ldots, 4\}\), we sample 100 values of \(\theta\) with various magnitudes (details to follow). For each set of parameters we then calculate metrics of model impropriety introduced in Section 2.2.2 based on \(\mu(\theta)\),
ELPR(\(\theta\))/n_v, and the absolute coordinates of \(E[X|\theta] - E[X|\theta']\). In the case of near-degeneracy, we classify each model as near-degenerate based on the distance of \(\mu(\theta)\) from the boundary of the convex hull of \(\mathcal{F}\) and look at the fraction of models that are “near-degenerate,” meaning they are within a small distance \(\varepsilon > 0\) of the boundary of the convex hull. We define “small” through a rough estimation of the volume of the hull for each model size. We pick \(\varepsilon_0 = 0.05\) for \(n_H = n_V = 1\) and then, for every other \(n_H\) and \(n_V\), set \(m = n_H + n_V + n_V n_H\) and pick \(\varepsilon\) so that \(1 - (1 - 2\varepsilon_0)^3 = 1 - (1 - 2\varepsilon)^m\). In this way, we roughly scale the volume of the “small distance” to the boundary of the convex hull to be equivalent across model dimensions.

In our numerical experiment, we split \(\theta = (\theta_{\text{main}}, \theta_{\text{interaction}})\) into \(\theta_{\text{main}}\) and \(\theta_{\text{interaction}}\), in reference to which variables in the probability function the parameters correspond (whether they multiply a \(v_i\) or a \(h_j\) or they multiply a \(v_i h_j\)), and allow the two types of terms to have varying average magnitudes, \(|\theta_{\text{main}}|/(n_H + n_V)\) and \(|\theta_{\text{interaction}}|/(n_H n_V)\). These average magnitudes vary on a grid between 0.001 and 3 with 24 breaks, yielding 576 grid points. (By looking at the average magnitudes, we are able to later consider the potential benefit of shrinking each parameter value \(\theta_i\) towards zero in a Bayesian fitting technique.) At each point in the grid, 100 vectors \((\theta_{\text{main}})\) are sampled uniformly on a sphere with radius corresponding to the first coordinate in the grid and 100 vectors \((\theta_{\text{interaction}})\) are sampled uniformly on a sphere with radius corresponding to the second coordinate in the grid via sums of squared and scaled iid Normal(0,1) variables. These vectors are then paired to create 100 values of \(\theta\) with magnitudes at each point in the grid.

The results of this numerical study are summarized in Figures 2.8, 2.9, and 2.10. From these three figures, it is clear that all three measures of model impropriety show higher values for larger magnitudes of the parameter vectors, supporting the RBM model properties developed in Section 2.2. As a compounding issue, these figures show that, as model grow in size, it becomes easier for more parameter configurations to push RBM models into near-degeneracy, instability and uninterpretability. Additionally, since there are \(n_H n_V\) interaction terms in \(\theta\) versus only \(n_H + n_V\) main effect terms, for large models there are many more interaction
Figure 2.8: Results from the numerical experiment, here looking at the fraction of models that were near-degenerate for each combination of magnitude of $\theta$ and model size, where $\theta = (\theta_{\text{main}}, \theta_{\text{interaction}})$ is split into main and interaction parameters. Black lines show the contour levels for fraction of near-degeneracy, while the thick black line shows the level where the fraction of near-degenerate models is .05.

parameters than main effects in the models. And so, severely limiting the magnitude of the individual interactions may well help prevent model impropriety.

Figure 2.11 shows the fraction of near-degenerate models for each magnitude of $\theta$ for each model architecture. For each number $n_v$ of visibles in the model, as the number $n_h$ of hiddens increase, the fraction near-degenerate diverges from zero at increasing rates for larger values of $||\theta||$. This shows that, as model size gets larger, the risk of degeneracy starts at a slightly larger magnitude of parameters, but very quickly increases until reaching close to 1.

These manageable examples indicate that RBMs are near-degenerate, unstable, and un-interpretable for large portions of the parameter space with large $||\theta||$. These problematic aspects require serious consideration when using RBM models, on top of the additional matter of principled/rigorous fitting of RBM models.
Figure 2.9: The sample mean value of ELPR($\theta$)/$V$ at each grid point for each combination of magnitude of $\theta$ and model size. As the magnitude of $\theta$ grows, so does the value of this metric, indicating typical instability in the model.

Figure 2.10: The sample mean of the maximum component of the absolute difference between the model expectation vector, $E[X|\theta]$, and the expectation vector given independence, $E[X|\theta^*]$. Larger magnitudes of $\theta$ correspond to larger differences, thus indicating reduced interpretability.
Figure 2.11: The fraction of near-degenerate models for each magnitude of $\theta$. For each number $n_v$ of visibles in the model, the fraction near-degenerate moves away from zero at larger values of $||\theta||$ as the number $n_h$ of hidden variables increases and the slope becomes steeper as $n_h$ increases as well.
2.4 Model Fitting

Typically, fitting a RBM via maximum likelihood (ML) methods will be infeasible due mainly to the intractability of the normalizing term $\gamma(\theta)$ in a model ((2.1)) of any realistic size. Ad hoc methods are used instead, which aim to avoid this problem by using stochastic ML approximations that employ a small number of MCMC draws (i.e., contrastive divergence, (Hinton 2002)).

However, computational concerns are not the only issues with fitting a RBM using ML. In addition, a RBM model, with the appropriate choice of parameters and number of hiddens, has the potential to re-create any distribution for the data (i.e., reproduce any specification of cell probabilities for the binary data outcomes). For example, Montufar and Ay (2011) show that any distribution on $\{0,1\}^n$ can be approximated arbitrarily well by a RBM with $2^{n-1} - 1$ hidden units. We provide a small example that illustrates that in fact there can be many such approximations. For simplicity, consider a model with two visible variables $(V_1, V_2)$ and one hidden $H_1$ so that there are four possible data realizations for $(V_1, V_2)$ given by $(\pm 1, \pm 1)$ and we may express the model probabilities as

$$ P(V_1 = v_1, V_2 = v_2|\theta_{v_1}, \theta_{v_2}, \theta_{h_1}, \theta_{11}, \theta_{21}) \propto \exp(v_1 \theta_{v_1} + v_2 \theta_{v_2}) \sum_{h \in \{\pm 1\}} \exp(h[\theta_{h_1} + \theta_{11}v_1 + \theta_{21}v_2]), $$

for $(v_1, v_2) \in \{-1, 1\}^2$, in terms of real-valued parameters $\theta_{v_1}, \theta_{v_2}, \theta_{h_1}, \theta_{11}, \theta_{21}$. Given some specified cell probabilities, say

$$ 0 \leq p(-1,-1), p(1,-1), p(-1,1), p(1,1), $$

for the outcomes $(\pm 1, \pm 1)$, the appendix provides technical details to show parameter values $(\theta_{v_1}, \theta_{v_2}, \theta_{h_1}, \theta_{11}, \theta_{21})$ may be chosen to match any given cell probabilities with arbitrary closeness. In fact, when the cell probabilities ((A.2)) are all strictly positive, parameters in the RBM model can be specified to reflect these probabilities exactly. And, when one or more of the cell probabilities ((A.2)) are zero, the corresponding RBM probabilities $P(V_1 = v_1, V_2 = v_2|\theta_{v_1}, \theta_{v_2}, \theta_{h_1}, \theta_{11}, \theta_{21})$ may never be identically zero (due to exponential terms in the model) but parameters can be still selected to make the appropriate RBM cell probabilities arbitrarily small. Furthermore,
as development in the appendix shows, only it is possible to approximate any distribution on
the visibles arbitrarily well (cf. Montufar and Ay 2011), but dramatically different parameter
settings can induce the same RBM model (beyond mere symmetries in parameterization). A
further, and somewhat odd consequence of the RBM parameterization is then that, when fitting
the RBM model by likelihood-based methods, we already know the nature of the answer before
we begin: namely, such fitting will simply reproduce the empirical distribution of the training
data if sufficiently many hidden are in the model and there can be no model refinements or
smoothing from the RBM. That is, based on a random sample of vectors of visible variables,
the model for the cell probabilities with the highest likelihood over all possible model classes
(i.e., RBM-based or not) is the empirical distribution, and the over-parameterization of the
RBM model itself ensures that this empirical distribution can be arbitrarily well-approximated.
For illustration, considering the simple example from above with \( n \) iid observations, each con-
sisting of two realized visibles \((V_1, V_2)\) the empirical cell frequencies from the sample are the
highest likelihood choices for the cell probabilities \(p_{(-1,-1)}, p_{(1,-1)}, p_{(-1,1)}, p_{(1,1)}\) in (A.2) before
any parametric model is introduced for improved analysis, and the discussion above indicates
that RBM model parameters can be chosen to re-create this empirical distribution to an ar-ritrarily close degree. Not only does RBM model fitting based on ML seek to reproduce the
empirical distribution, whenever this empirical distribution contains empty cells, fitting steps
for the RBM model will further aim to choose parameters that necessarily diverge to infinity
in magnitude in order to zero-out the corresponding RBM cell probabilities. In data applic-
ations with a large sample space, it is unlikely that the training set will include at least one of
each possible vector outcome (unlike this small example). This implies that some RBM model
parameters must diverge to +\( \infty \) to mimic the empirical distribution with empty cells and, as
we have already discussed in Section 2.3, large magnitudes of \( \theta \) lead to model impropriety in
the RBM.

Here we consider what might be done in a principled manner to prevent both overfitting
and model impropriety, testing on a \( n_v = n_H = 4 \) case that already stretches the limits of what
is computable - in particular we consider Bayes methods.
2.4.1 Bayesian model fitting

To avoid model impropriety for a fitted RBM, we wish to avoid parts of the parameter space \( \mathbb{R}^{n_V+n_H+n_V\times n_H} \) that lead to near-degeneracy, instability, and uninterpretability. Motivated by the insights in Section 2.3.2, one idea is to shrink \( \theta = (\theta_{\text{main}}, \theta_{\text{interaction}}) \) toward 0 by specifying priors that place low probability on large values of \( ||\theta|| \), specifically focusing on shrinking \( \theta_{\text{interaction}} \) more than \( \theta_{\text{main}} \). This is similar to an idea advocated by Hinton (2010) called weight decay, in which a penalty is added to the interaction terms in the model, \( \theta_{\text{interaction}} \), shrinking their magnitudes.

Table 2.1: Parameters used to fit a test case with \( V = H = 4 \). This parameter vector was chosen as a sampled value of \( \theta \) that was not near the convex hull of the sufficient statistics for a grid point in Figure 2.8 with \(< 5\% \) near-degeneracy.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
<th>Parameter</th>
<th>Value</th>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \theta_1 )</td>
<td>-1.104376</td>
<td>( \theta_{11} )</td>
<td>-0.0006334</td>
<td>( \theta_{31} )</td>
<td>-0.0038301</td>
</tr>
<tr>
<td>( \theta_2 )</td>
<td>-0.2630044</td>
<td>( \theta_{12} )</td>
<td>-0.0021401</td>
<td>( \theta_{32} )</td>
<td>0.0032237</td>
</tr>
<tr>
<td>( \theta_3 )</td>
<td>0.3411915</td>
<td>( \theta_{13} )</td>
<td>0.0047799</td>
<td>( \theta_{33} )</td>
<td>0.0020681</td>
</tr>
<tr>
<td>( \theta_4 )</td>
<td>-0.2583769</td>
<td>( \theta_{14} )</td>
<td>0.0025282</td>
<td>( \theta_{34} )</td>
<td>0.0041429</td>
</tr>
<tr>
<td>( \theta_{h1} )</td>
<td>-0.1939302</td>
<td>( \theta_{21} )</td>
<td>0.0012975</td>
<td>( \theta_{41} )</td>
<td>0.0089533</td>
</tr>
<tr>
<td>( \theta_{h2} )</td>
<td>-0.0572858</td>
<td>( \theta_{22} )</td>
<td>0.0000253</td>
<td>( \theta_{42} )</td>
<td>-0.0042403</td>
</tr>
<tr>
<td>( \theta_{h3} )</td>
<td>-0.2101802</td>
<td>( \theta_{23} )</td>
<td>-0.0004352</td>
<td>( \theta_{43} )</td>
<td>-0.000048</td>
</tr>
<tr>
<td>( \theta_{h4} )</td>
<td>0.2402456</td>
<td>( \theta_{24} )</td>
<td>-0.0086621</td>
<td>( \theta_{44} )</td>
<td>0.0004767</td>
</tr>
</tbody>
</table>

We considered a test case with \( n_V = n_H = 4 \) and parameters given in Table 2.1. This parameter vector was chosen as a sampled value of \( \theta \) at which the resulting RBM model would not be clearly degenerate. We simulated \( n = 5,000 \) realizations of visibles as a training set and fit the RBM using three Bayes methodologies. These involved the following set-ups with choice of prior distribution \( \pi(\theta) \) for parameters \( \theta \).

1. A “trick” prior. Here we cancel out normalizing term in the likelihood (from \( \gamma(\theta) \) in \((2.1)) \) so that resulting full conditionals of \( \theta \) are multivariate Normal. Namely, this involves a prior of the form

\[
\pi(\theta) \propto \gamma(\theta)^\nu \exp \left( -\frac{1}{2C_1} \theta_{\text{main}}^\prime \theta_{\text{main}} - \frac{1}{2C_2} \theta_{\text{interaction}}^\prime \theta_{\text{interaction}} \right),
\]
where
\[ \gamma(\theta) = \sum_{x \in \mathcal{X}^{m+n}} \exp \left( \sum_{i=1}^{n_v} \sum_{j=1}^{n_H} \theta_{ij} v_i h_j + \sum_{i=1}^{n_v} \theta_{v_i} v_i + \sum_{j=1}^{n_H} \theta_{h_j} h_j \right) \]
for hyperparameter choices \(0 < C_2 < C_1\). The unknown hidden variables \(h_j\) are also directly treated as latent variables and are sampled in each MCMC iterative draw from the posterior distribution. This is the method of J. Li (2014). We will refer to this method as Bayes with Trick Prior and Latent Variables (BwTPLV).

2. A truncated Normal prior. Here we use independent spherical normal distributions as priors for \(\theta_{\text{main}}\) and \(\theta_{\text{interaction}}\), which are truncated at \(3\sigma_{\text{main}}\) and \(3\sigma_{\text{interaction}}\), respectively, based on standard deviation hyperparameters \(0 < \sigma_{\text{interaction}} < \sigma_{\text{main}}\). Full conditional distributions are not conjugate, and simulation from the posterior was accomplished using a geometric adaptive Metropolis Hastings step (Zhou 2014) and calculation of likelihood normalizing constant. (This computation is barely feasible for a problem of this size and would be infeasible for larger problems.) Here the hidden variables \(h_j\) are again carried along in the MCMC implementation as latent variables. We will refer to this method as Bayes with Truncated Normal prior and Latent Variables (BwTnL).

3. A truncated Normal prior and marginalized likelihood. Here we marginalize out the hidden variables \(h = (h_1, \ldots, h_{n_H})\) in \(f_{\theta}(x)\), and use the truncated Normal priors applied to the marginal probabilities for visible variables given by
\[ g_{\theta}(v) \propto \sum_{h \in \mathcal{H}^{m+n}} \exp \left( \sum_{i=1}^{n_v} \sum_{j=1}^{n_H} \theta_{ij} v_i h_j + \sum_{i=1}^{n_v} \theta_{v_i} v_i + \sum_{j=1}^{n_H} \theta_{h_j} h_j \right), \quad v \in \mathcal{Y}^m. \]
We will refer to this method as Bayes with Truncated Normal prior and Marginalized Likelihood (BwTNML).

The three fitting methods are ordered above according to computational feasibility in a real-data situation, with BwTPLV being the most computationally feasible due to conjugacy and BwTNML the least feasible due to the marginalization and need for an adaptive Metropolis Hastings step. All three methods require choosing the values of hyperparameters. In each case, we have chosen these values based on a rule of thumb that shrinks \(\theta_{\text{interaction}}\) more than \(\theta_{\text{main}}\). Additionally, BwTPLV requires additional tuning (i.e. a tuning parameter \(C > 0\) in
Table 2.2) to choose $C_1$ and $C_2$, reducing its appeal. The forms used for the hyperparameters in our simulation are presented in Table 2.2. It should be noted that, due to the common

Table 2.2: The values used for the hyperparameters for all three fitting methods. A rule of thumb is imposed which decreases prior variances for the model parameters as the size of the model increases and also shrinks $\theta_{\text{interaction}}$ more than $\theta_{\text{main}}$. The common $C$ defining $C_1$ and $C_2$ in the BwTPLV method is chosen by tuning.

<table>
<thead>
<tr>
<th>Method</th>
<th>Hyperparameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>BwTPLV</td>
<td>$C_1$</td>
<td>$\frac{C - 1}{n_H + n_V}$</td>
</tr>
<tr>
<td></td>
<td>$C_2$</td>
<td>$\frac{C}{n_H + n_V}$</td>
</tr>
<tr>
<td>BwTNLV</td>
<td>$\sigma^2_{\text{main}}$</td>
<td>$\frac{1}{n_H + n_Y}$</td>
</tr>
<tr>
<td></td>
<td>$\sigma^2_{\text{interaction}}$</td>
<td>$\frac{1}{n_Y n_V}$</td>
</tr>
<tr>
<td>BwTNML</td>
<td>$\sigma^2_{\text{main}}$</td>
<td>$\frac{1}{n_H + n_Y}$</td>
</tr>
<tr>
<td></td>
<td>$\sigma^2_{\text{interaction}}$</td>
<td>$\frac{1}{n_Y n_V}$</td>
</tr>
</tbody>
</table>

prior distributions for $\theta$, both BwTNLV (method 2 above) and BwTNML (method 3) are drawing from the same stationary posterior distribution for vectors of visibles. A fundamental difference between these two methods lies in how well these two chains mix and how quickly they arrive at the target posterior distribution. After a burn-in period of 50 iterations selected by inspecting the trace plots, we assess the issue of mixing in two ways. First, the autocorrelation functions (ACF) from each posterior sample corresponding to a model probability for a visible vector outcome $\mathbf{v} = (v_1, v_2, v_3, v_4) \in \{\pm 1\}^4$ (i.e., computed from $\theta$ under ((2.1))) are determined and plotted in Figure 2.12 with BwTNLV in black and BwTNML in red. As expected, ACF corresponding to the method (BwTNML) that marginalizes out the hidden variables from the likelihood decreases to zero at a much faster rate, indicating better mixing for the chain.

Secondly, we can assess the mixing of the BwTNLV/BwTNML chains using the notion of effective sample size. If the MCMC chain were truly iid draws from the target distribution, then for the parameter $p^{(i)}$ denoting the probability of the $i$th vector outcome for the four visibles $\mathbf{v} = (v_1, v_2, v_3, v_4) \in \{\pm 1\}^4$, $i = 1, \ldots, 16$, its estimate as the average $\bar{p}^{(i)}$ of posterior sample versions would be approximately Normal with mean given by the posterior marginal mean of $p^{(i)}$, and variance given by $\sigma_i^2 / M$, where $\sigma_i^2$ is the true posterior variance of $p^{(i)}$ and $M$ is the length of the chain. However, with the presence of correlation in our chain, the asymptotic variance of $\bar{p}^{(i)}$ is instead approximately some $C_i / M$, where $C_i$ is some positive constant such
Figure 2.12: The autocorrelation functions (ACF) for the posterior probabilities of all $2^4 = 16$ possible outcomes for the vector of four visibles assessed at multiple lags for each method with BwTNLV in black and BwTNML in red. As expected, ACF corresponding to the method that marginalizes out the hidden variables from the likelihood decreases to zero at a much faster rate, indicating better mixing for the chain.
that \( C_i > \sigma_i^2 \). We can use an overlapping block-means approach (Gelman, Shirley, and others 2011) to get a crude estimate for \( C_i \) as \( \hat{C}_i = bS_i^2 \), where \( S_i^2 \) denotes the sample variance of overlapping block means \( \{ p_{j}^{(i)} = \frac{\sum_{k=j}^{j+b-1} p_k^{(i)}}{b} \}_{j=1}^{M-b+1} \) of length \( b \) computed from the posterior samples \( \{ p_k^{(i)} \}_{k=1}^{M} \). We compare it to an estimate of \( \sigma_i^2 \) using sample variance \( \hat{\sigma}_i^2 \) of the raw chain, \( \{ p_k^{(i)} \}_{k=1}^{M} \). Formally, we approximate the effective sample size of the length \( M \) MCMC chain as

\[
M_{eff}^{(i)} = M \frac{\hat{\sigma}_i^2}{\hat{C}_i}.
\]

For both BwTNL and BwTNML methods, effective sample sizes for a chain of length \( M = 1000 \) for inference about each of the \( 2^4 = 16 \) model probabilities are presented in Table 2.3. These range from 304.57 to 1229.39 for BwTNML, while BwTNLV only yields between 65.05 and 132.61 effective draws. Thus, BwTNLV would require at least 4.7 times as many iterations to be run of the MCMC chain in order to achieve the same amount of effective information about the posterior distribution. For this reason, consistent with the ACF results in Figure 2.12, BwTNLV does not seem to be an effective method for fitting the RBM, though computing resources can hinder use of the alternative BwTNML involving marginalization of hidden variables.

Table 2.3: The effective sample sizes for a chain of length \( M = 1000 \) regarding all 16 probabilities for possible vector outcomes of visibles. BwTNLV would require at least 4.7 times as many MCMC iterations to achieve the same amount of effective information about the posterior distribution.

<table>
<thead>
<tr>
<th>Outcome</th>
<th>BwTNLV</th>
<th>BwTNML</th>
<th>Outcome</th>
<th>BwTNLV</th>
<th>BwTNML</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>73.00</td>
<td>509.43</td>
<td>9</td>
<td>83.47</td>
<td>394.90</td>
</tr>
<tr>
<td>2</td>
<td>65.05</td>
<td>472.51</td>
<td>10</td>
<td>95.39</td>
<td>327.35</td>
</tr>
<tr>
<td>3</td>
<td>87.10</td>
<td>1229.39</td>
<td>11</td>
<td>70.74</td>
<td>356.56</td>
</tr>
<tr>
<td>4</td>
<td>72.64</td>
<td>577.73</td>
<td>12</td>
<td>81.40</td>
<td>338.30</td>
</tr>
<tr>
<td>5</td>
<td>71.67</td>
<td>452.01</td>
<td>13</td>
<td>105.98</td>
<td>373.59</td>
</tr>
<tr>
<td>6</td>
<td>66.49</td>
<td>389.78</td>
<td>14</td>
<td>132.61</td>
<td>306.91</td>
</tr>
<tr>
<td>7</td>
<td>84.30</td>
<td>660.37</td>
<td>15</td>
<td>82.15</td>
<td>365.30</td>
</tr>
<tr>
<td>8</td>
<td>75.46</td>
<td>515.09</td>
<td>16</td>
<td>98.05</td>
<td>304.57</td>
</tr>
</tbody>
</table>

\( M = 1000 \) for inference about each of the \( 2^4 = 16 \) model probabilities are presented in Table 2.3. These range from 304.57 to 1229.39 for BwTNML, while BwTNLV only yields between 65.05 and 132.61 effective draws. Thus, BwTNLV would require at least 4.7 times as many iterations to be run of the MCMC chain in order to achieve the same amount of effective information about the posterior distribution. For this reason, consistent with the ACF results in Figure 2.12, BwTNLV does not seem to be an effective method for fitting the RBM, though computing resources can hinder use of the alternative BwTNML involving marginalization of hidden variables.

Figure 2.13 shows the posterior probability of each possible \( v \in \{-1,1\}^4 \) after fitting the RBM model according to method 1 (BwTPLV using trick prior) and method 3 (BwTNML)
(excluding method 2 (BwTNL) that seeks the same posterior as method 3). The black vertical lines show the true probabilities of each image based on the parameters used to generate the training set while the red vertical lines show the empirical distribution for the training set of 5,000 vectors. From these posterior predictive checks, it is evident that BwTNML produces the best fit to the data. Furthermore, along with the discussion of Section 2.4, Figure 2.13 also shows that it can be undesirable to seek to perfectly re-create an empirical distribution in fitting RBM models (i.e., true model probabilities may differ substantially). The priors in the BwTNML method constrain the RBM model fit to avoid replication of the empirical distribution and better estimate the underlying true data generating probabilities. However, this method requires a marginalization step to obtain the probability function of visible observations alone, which is infeasible for a model with $n_H$ of any real size.

2.5 Discussion

RBM models constitute an interesting class of undirected graphical models that are thought to be useful for supervised learning tasks. However, when viewed as generative statistical models, RBMs are prone to forms of model impropriety such as near-degeneracy, S-instability, and uninterpretability. Additionally, these models are difficult to fit using a rigorous methodology, due to the dimension of the parameter space coupled with the size of the latent variable space.

In this paper, we have presented three fully Bayes-principled MCMC-based methods for fitting RBMs. Common practice is to use a kind of MCMC to overcome fitting complexities. However because of the size of the space to be filled with MCMC iterates, convergence and mixing of these methods will be slow. Marginalization over the latent variables in the model is shown to improve mixing, but is numerically intractable due to the necessity of repeated calculation of the normalizing constant. Due to the extreme flexibility in this model class, rigorous likelihood-based fitting for a RBM will typically aim to merely return the
Figure 2.13: Posterior probabilities of $16 = 2^4$ possible realizations of 4 visibles using two of the three Bayesian fitting techniques, BwTPLV and BwTNML. The black vertical lines show the true probabilities of each vector of visibles based on the parameters used to generate the training data while the red vertical lines show the empirical distribution. BwTNML produces the best fit for the data, however is also the most computationally intensive and least feasible with a real dataset.
(discrete) empirical distribution for visibles, meaning any practitioner should be aware of this and employ some form of regularization or penalization in the model.

Ultimately, it is not clear that RBM models are useful as generative models. Furthermore, without the appropriate generative behavior in the RBM model, the ability to quantify uncertainty in the estimated model parameters becomes impossible and the model loses useful application to prediction problems (e.g., realistic predictive distributions). In the case of classification, predictive distributions not be the ultimate goal, but when S-instability is present, small (imperceptible) differences in the data may lead to greatly different probabilities and thus greatly different classifications. For these reasons, we are skeptical about RBMs as data generative tools.
CHAPTER 3. A NOTE ON THE INSTABILITY AND DEGENERACY OF DEEP LEARNING MODELS

A paper to be submitted to the *Journal of the American Statistical Association*

Andee Kaplan, Daniel J. Nordman, and Stephen B. Vardeman

Abstract

A probability model exhibits instability if small changes in a data outcome result in large, and often unanticipated, changes in probability. For correlated data structures found in several application areas, there is increasing interest in predicting/identifying such sensitivity in model probability structure. We consider the problem of quantifying instability for general probability models defined on sequences of observations, where each sequence of length $N$ has a finite number of possible outcomes. A sequence of probability models results, indexed by $N$, that accommodates data of expanding dimension. Model instability is formally shown to occur when a certain log-probability ratio under such models grows faster than $N$. In this case, a one component change in the data sequence can shift probability by orders of magnitude. Also, as instability becomes more extreme, the resulting probability models are shown to tend to degeneracy, placing all their probability on potentially small portions of the sample space. These results on instability apply to large classes of models commonly used in random graphs, network analysis, and machine learning contexts.
3.1 Introduction

We consider the behavior, and the potential impropriety, of probability models built to incorporate a sequence of discrete observations with length $N$. Let $(X_1,\ldots,X_N)$ denote a set of discrete random variables with a finite sample space, $\mathcal{X}^N$. That is, $\mathcal{X}$ with $|\mathcal{X}| < \infty$ represents a finite set of potential outcomes for each single variable $X_i$, and the data sequence $(X_1,\ldots,X_N)$ takes values in the $N$-fold product space $\mathcal{X}^N$. For each $N$, let $P_{\theta_N}$ denote a probability model on $\mathcal{X}^N$, under which $P_{\theta_N}(x_1,\ldots,x_N) > 0$ is the probability of the data outcome $(x_1,\ldots,x_N) \in \mathcal{X}^N$. In this, we assume that the model support of $P_{\theta_N}$ is the sample space $\mathcal{X}^N$. This framework produces a series $P_{\theta_N}$ of probability models, indexed by a generic sequence of parameters $\theta_N$, to describe data of each length $N \geq 1$. The size and structure of such parameters are without restriction, and natural cases include those where $\theta_N \in \mathbb{R}^{q(N)}$ for some arbitrary integer-valued function $q(\cdot) \geq 1$. We will refer to this model class as *Finitely Supported Finite Sequence (FSFS) models*.

Section 3.2 provides several examples of FSFS models commonly used in graph/network analysis and machine learning (i.e., deep learning models). Section 3.3 establishes formal results regarding the propriety of FSFS models with regard to stability. A FSFS probability model sequence exhibits instability if small changes in the components of a data outcome $(x_1,\ldots,x_N)$ can result in large changes in probability $P_{\theta_N}(x_1,\ldots,x_N)$. The concept of instability, introduced in the field of statistical physics by Ruelle (1999), was extended to include a notion of detection and quantification for certain exponential family models by Schweinberger (2011). For similar exponential models, particularly in connection to random graphs/networks, Handcock (2003) considered (mean-based) characterizations for so-called model degeneracy, whereby a probability model places all mass on a small subset of the sample space and produces undesirably low variability in model outcomes. As described by Schweinberger (2011), model instability and model degeneracy are related by viewing degeneracy as an extreme or limiting form of instability. The instability results of Schweinberger (2011) were developed for the case of discrete exponential family models. The main results here concern a general measure of model...
instability, appropriate across the whole FSFS model class. This can be used to identify when certain maximal probabilities in FSFS models are too extreme relative to the length $N$ and may thereby induce a potentially undesirable probability structure. In this case, a one component change in the data sequence may shift probability by orders of magnitude, and FSFS models are rigorously shown to become degenerate as the measure of instability increases. Lastly, Section 3.4 emphasizes the implications of our model propriety results and proofs of the main results appear in Appendix C.

### 3.2 Examples

Many model families fall under the umbrella of FSFS models. For illustration, this section presents three specific examples of FSFS models, including models with deep architectures.

#### 3.2.1 Discrete exponential family models

For discrete random variables $X = (X_1, \ldots, X_N)$ with sample space $\mathcal{X}^N$, $|\mathcal{X}| < \infty$, consider an exponential family model for $X$ with probability mass function of the form

$$p_{\lambda}(x) = \exp \left[ \eta^T(\lambda) g_N(x) - \psi(\lambda) \right], \quad x \in \mathcal{X}^N,$$

depending on parameter vector $\lambda \in \Lambda \subset \mathbb{R}^k$ and natural parameter function $\eta : \mathbb{R}^k \mapsto \mathbb{R}^L$ with fixed positive integers $k$ and $L$ denoting their dimensions. Above, $g_N : \mathcal{X}^N \mapsto \mathbb{R}^L$ is a vector of sufficient statistics, while

$$\psi(\lambda) = \log \sum_{x \in \mathcal{X}^N} \exp \left[ \eta^T(\lambda) g_N(x) \right], \quad \lambda \in \Lambda,$$

denotes the normalizing function, and $\Lambda = \{ \lambda \in \mathbb{R}^k : \psi(\lambda) < \infty, k \leq q(N) \}$ is the parameter space.

Defining $P_{\theta_N}(x) \equiv p_{\lambda_N}(x)$ with $\theta_N = \lambda_N$ to be a sequence of elements of $\Lambda \subset \mathbb{R}^k$ and noting that $P_{\theta_N}(x) > 0$ for all $x \in \mathcal{X}^N$, these discrete exponential family models are special cases of the FSFS models. Such exponential models arise with spatial data on a lattice (Besag 1974),
network data (Wasserman and Faust 1994; Handcock 2003), and even standard independence models for discrete data, such as with \( N \) iid Bernoulli random variables. (Note that for random graphs or networks with, say, \( m \) nodes, one may wish to consider \( N = \binom{m}{2} \) edges as binary (presence/absence) variables \( X_i \). In this case, the length \( N \) of data sequence may naturally increase as a function of \( m \).) For these exponential models, the dimension of the parameter \( \theta_N \) remains constant over each \( N \), as \( \theta_N \) lies in a parameter space of fixed Euclidean dimension \( k \). This need not be true for other types of FSFS models considered in Sections 1.1-3.2.3. Schweinberger (2011) considered instability in such exponential models (e.g., for random graphs) for sequences of fixed parameters \( \theta_N = \lambda \in \mathbb{R}^k, N \geq 1 \).

3.2.2 Restricted Boltzmann machines

A restricted Boltzmann machine (RBM) is an undirected graphical model specified for discrete or continuous random variables, with binary variables being most common (cf. Smolensky 1986). A RBM architecture has two layers, hidden (\( \mathcal{H} \)) and visible (\( \mathcal{V} \)), with conditional independence within each layer. Let \( \mathbf{X} = (X_1, \ldots, X_N) \) denote the \( N \) random variables for visibles with support \( \mathcal{X}^N \) and \( \mathbf{H} = (H_1, \ldots, H_{N_H}) \) denote the \( N_H \) random variables for hiddens with support \( \mathcal{X}^{N_H} \) where \( \mathcal{X} = \{-1, 1\} \). For parameters \( \alpha \in \mathbb{R}^{N_H}, \beta \in \mathbb{R}^N, \) and \( \Gamma \) as a matrix with dimension \( N_H \times N \), the RBM model for \( \mathbf{X} = (\mathbf{X}, \mathbf{H}) \) then has the joint probability mass function

\[
P_{\theta_N}(\mathbf{x}) = \exp \left[ \alpha^T \mathbf{h} + \beta^T \mathbf{x} + \mathbf{h}^T \Gamma \mathbf{x} - \psi(\theta_N) \right], \quad \mathbf{x} = (\mathbf{h}, \mathbf{x}) \in \mathcal{X}^{N+N_H},
\]

where

\[
\psi(\theta_N) = \log \sum_{\mathbf{x} \in \mathcal{X}^{N+N_H}} \exp \left[ \alpha^T \mathbf{h} + \beta^T \mathbf{x} + \mathbf{h}^T \Gamma \mathbf{x} \right], \quad \theta_N \in \Theta_N,
\]

is the normalizing function. Let \( \theta_N = (\alpha, \beta, \Gamma) \in \Theta_N \subset \mathbb{R}^{q(N)} \) with \( q(N) = N + N_H + N \times N_H \) denote the vector of parameters for the RBM. The probability mass function for the visible variables \( X_1, \ldots, X_N \) follows from marginalizing this joint specification:

\[
P_{\theta_N}(\mathbf{x}) = \sum_{\mathbf{h} \in \mathcal{X}^{N_H}} P_{\theta_N}(\mathbf{x}, \mathbf{h}), \quad \mathbf{x} \in \mathcal{X}^N.
\]
Note that the vector of model parameters $\theta_N$, of size $q(N)$, grows in size as a function of sample dimension $N$ to accommodate the dimension of visible variables $X_1, \ldots, X_N$, and one may further choose the number $N_H$ of hidden variables to change with $N$ as well. In particular, the number $N_H$ of hiddens may also potentially and arbitrarily increase with $N$. Additionally, as $|\mathcal{X}| = 2$ and $P_{\theta_N}(x) > 0$ for all $x \in \mathcal{X}_N$, the RBM model specification for visibles is a FSFS model. This example also indicates that models formed by marginalizing a base FSFS model (e.g., a type of exponential family model) is again a FSFS model class.

### 3.2.3 Deep learning

Consider two models with “deep architecture” that contain multiple hidden (or latent) layers in addition to a visible layer of data, namely a deep Boltzmann machine (R. Salakhutdinov and Hinton 2009) and a deep belief network (Hinton, Osindero, and Teh 2006). Let $M$ denote the number of hidden layers included in the model and let $N_{(H,1)}, \ldots, N_{(H,M)}$ denote the numbers of hidden variables within each hidden layer. Then the random vector $\tilde{X} = \{H_1^{(1)}, \ldots, H_{N_{(H,1)}}^{(1)}, \ldots, H_1^{(M)}, \ldots, H_{N_{(H,M)}}^{(M)}; X\}$ collects both the hidden variables $\{H_i^{(j)} : i = 1, \ldots, N_{(H,j)}; j = 1, \ldots, M\}$ and visible variables $X = (X_1, \ldots, X_N)$ in a deep probabilistic model. Each variable outcome will again lie in $\mathcal{X} = \{-1, 1\}$.

**Deep Boltzmann machine (DBM).** The DBM class of models maintains conditional independence within all layers in the model by stacking RBM models and only allowing conditional dependence between neighboring layers. The joint probability mass function for a DBM is

$$P_{\theta_N}(\tilde{x}) = \exp \left[ \sum_{i=1}^{M} \alpha^{(i)} x^{(i)} + \beta^{(i)} x + h^{(1)} \Gamma(0) x + \sum_{i=1}^{M-1} h^{(i)} \Gamma(i) h^{(i+1)} - \psi(\theta_N) \right],$$

for $\tilde{x} = (h^{(1)}, \ldots, h^{(M)}, x) \in \mathcal{X}^{N_{(H,1)} + \cdots + N_{(H,M)} + N}$ where

$$\psi(\theta_N) = \log \sum_{\tilde{x} \in \mathcal{X}^{N_{(H,1)} + \cdots + N_{(H,M)} + N}} \exp \left[ \sum_{i=1}^{M} \alpha^{(i)} x^{(i)} + \beta^{(i)} x + h^{(1)} \Gamma(0) x + \sum_{i=1}^{M-1} h^{(i)} \Gamma(i) h^{(i+1)} \right],$$

for $\theta_N \in \Theta_N$ is the normalizing function and parameters in the model are $\beta \in \mathbb{R}^V$, $\alpha^{(i)} \in \mathbb{R}^{N_{(H,i)}}$ for $i = 1, \ldots, M$, along with a matrix $\Gamma(0)$ of dimension $N_{(H,1)} \times N$, and matrices $\Gamma^{(i)}$ of dimension
$N_{(H,i)} \times N_{(H,i+1)}$ for $i = 1, \ldots, M - 1$. Let $\theta_N = (\alpha^{(1)}, \ldots, \alpha^{(M)}, \beta, \Gamma^{(0)}, \ldots, \Gamma^{(M-1)}) \in \Theta_N \subset \mathbb{R}^{q(N)}$ denote the combined vector of parameters with total length $q(N) = N_{(H,1)} + \cdots N_{(H,M)} + N + N_{(H,1)} * N + N_{(H,2)} * H_{(H,1)} + \cdots + N_{(H,M)} * H_{(H,M)-1}$.

The probability mass function for the visible random variables $X_1, \ldots, X_N$ follows from this joint specification as

$$P_{\theta_N}(x) = \sum_{(h^{(1)}, \ldots, h^{(M)}) \in \mathcal{X}_{N_{(H,1)}}^{\cdots N_{(H,M)}}} P_{\theta_N}(\tilde{x}), \quad x \in \mathcal{X}^N$$

Again like the RBM case, the DBM model specification examples a FSFS model.

**Deep belief network (DBN).** A DBN resembles a DBM in that there are multiple layers of latent random variables stacked in a deep architecture with no conditional dependence between layers. The difference between the DBM and DBN models is that all but the last stacked layer in a DBN are Bayesian networks (see Pearl 1985), rather than RBMs. Thus for visibles $X_1, \ldots, X_N$ with support $\mathcal{X}^N$, a DBN is also a FSFS model if the number $q(N)$ of components in the parameter vector is dependent on the dimension of the visibles. Commonly, as in logistic belief nets (Neal 1992), a “weight” parameter is placed on each interaction between visibles, $X_1, \ldots, X_N$, and the first layer of latent variables, $H^{(1)}_1, \ldots, H^{(1)}_{N_{(H,1)}}$, in the definition of a FSFS model.

### 3.3 Instability results

To define or measure instability in FSFS models, it is useful to consider the behavior of a data model sequence $P_{\theta_N}$. A relevant quantity to this end is a (scaled) extremal log-probability ratio (ELPR)

$$\frac{1}{N} \log \left[ \frac{\max_{(x_1, \ldots, x_N) \in \mathcal{X}^N} P_{\theta_N}(x_1, \ldots, x_N)}{\min_{(x_1, \ldots, x_N) \in \mathcal{X}^N} P_{\theta_N}(x_1, \ldots, x_N)} \right] \equiv \frac{1}{N} \text{ELPR}_N(\theta_N).$$

The main idea is that, in formulating FSFS models for potentially increasing numbers of variables (i.e., for $N \to \infty$), the ratio ((3.1)) should remain bounded to better ensure model stability,
requiring that the largest probability possible under $P_{\theta_N}$ maintain a fixed order of magnitude relative to the smallest probability allowed under the same model. Specifically, the log of the ratio should grow at mostly linearly with the sample size $N$. This leads to the following definition.

**Definition 2** (S-unstable FSFS). Let $\theta_N \in \mathbb{R}^{q(N)}$ be a sequence of FSFS model parameters where the size of the model $q(N)$ is a function of the number of random variables $N$. A FSFS model formulation is Schweinberger-unstable or S-unstable if, as the number of variables increase ($N \to \infty$),

$$
\lim_{N \to \infty} \frac{1}{N} \text{ELPR}(\theta_N) \equiv \lim_{N \to \infty} \frac{1}{N} \log \left[ \frac{\max_{(x_1, \ldots, x_N) \in \mathcal{X}^N} P_{\theta_N}(x_1, \ldots, x_N)}{\min_{(x_1, \ldots, x_N) \in \mathcal{X}^N} P_{\theta_N}(x_1, \ldots, x_N)} \right] = \infty.
$$

In other words, a model is S-unstable if, given any $C > 0$, there exists an integer $N_C > 0$ so that $\frac{1}{N} \text{ELPR}_N(\theta_N) > C$ for all $N \geq N_C$. A FSFS model formulation may be termed S-stable if it fails to be S-unstable.

This definition of S-unstable is a generalization or reinterpretation of “unstable” used in Schweinberger (2011) by allowing non-exponential family models (e.g. RBM and DBM models in Sections 1.1-3.2.3) and an increasing number of parameters. While this definition differs in form and scope from the original, it does match that in Schweinberger (2011) for the special case of exponential models (cf. Section 3.2.1) considered there.

S-unstable FSFS model sequences may be undesirable for several reasons. One is that small changes in data can lead to overly-sensitive changes in probability. Consider, for example, the quantity given by

$$
\Delta(\theta_N) \equiv \max \left\{ \log \frac{P_{\theta_N}(x)}{P_{\theta_N}(x')} : x \neq x' \in \mathcal{X}^N \text{ differ in exactly one component} \right\},
$$

which represents the biggest log-probability ratio for a one component change in data outcomes at a FSFS parameter $\theta_N$. We then have the following (non-asymptotic) result.

**Proposition 3.** Let $\text{ELPR}(\theta_N)$ be as in (3.1) for an integer $N \geq 1$. For a given $C > 0$, if

$$
\frac{1}{N} \text{ELPR}_N(\theta_N) > C,
$$

In other words, a model is S-unstable if, given any $C > 0$, there exists an integer $N_C > 0$ so that $\frac{1}{N} \text{ELPR}_N(\theta_N) > C$ for all $N \geq N_C$. A FSFS model formulation may be termed S-stable if it fails to be S-unstable.
then

$$\Delta N(\theta_N) > C.$$ 

Again, if the probability ratio ((3.1)) is too large, then the FSFS model will exhibit large changes in probability for very small differences in the data configuration, which exemplifies the intuitive notion of instability.

Additionally, S-unstable FSFS model sequences are connected to degenerate models, where model *degeneracy* typically entails placing all probability on a small portion of the sample space. For perspective, note that differing sizes of $1/N \cdot \text{ELPR}(\theta_N)$ in ((3.1)) may induce a spectrum of levels of “stability” and Proposition 3 indicates increasing sensitivity of model probabilities (i.e., for one component changes in outcomes) as ((3.1)) increases. Furthermore, as the instability measure ((3.1)) grows, FSFS model sequences are guaranteed to slide into full degeneracy as Proposition 4 shows. Define a $\varepsilon$-modal set

$$M_{\varepsilon, \theta_N} = \left\{ x \in \mathcal{X}^N : \log P_{\theta_N}(x) > (1 - \varepsilon) \max_{x^* \in \mathcal{X}^N} P_{\theta_N}(x^*) + \varepsilon \min_{x^* \in \mathcal{X}^N} P_{\theta_N}(x^*) \right\}$$

of possible outcomes, for a given $0 < \varepsilon < 1$.

**Proposition 4.** For an unstable FSFS model in Definition 2, and for any given $0 < \varepsilon < 1$,

$$P_{\theta_N}(\{(x_1, \ldots, x_N) \in M_{\varepsilon, \theta_N}\}) \rightarrow 1 \text{ as } N \rightarrow \infty.$$ 

In other words, in S-unstable FSFS models, all probability in the model formulation with a large number of random variables will concentrate mass on an $\varepsilon$-modal set, where $\varepsilon$ can be made arbitrarily small. The associated mode set could potentially be quite small, in which case Proposition 4 would suggest that the unstable model asymptotically stacks all probability on a few outcomes.

**Remark.** There is a further generalization the notion of instability in Definition 2 meant to address independent replications of data sequences. That is, one might consider data as $n$ independent and identically distributed replications $X_1, \ldots, X_n$, where each $X_i = (X_{i,1}, \ldots, X_{i,N}) \in \mathcal{X}^N$ follows a common FSFS model with probabilities $P_{\theta_i}(x) > 0$, $x \in \mathcal{X}^N$ and $|\mathcal{X}| < \infty$, for
\(i = 1, \ldots, n\). This leads to a total of \(n + N\) random variables in the joint model. However, the
definition of S-unstable and Propositions 3-4 still hold for such iid replications. This is because
\[
\left( \max_{x \in \mathcal{X}^N} P_{\theta_N}(x) \right)^n
\]
is the largest probability possible under the joint model for the \(n\) replications
while
\[
\left( \min_{x \in \mathcal{X}^N} P_{\theta_N}(x) \right)^n
\]
is the smallest probability. Thus, for the combined replications \(X_1, \ldots, X_n\),
the analog definition of the extremal log-probability becomes

\[
\frac{\text{extremal log-probability ratio}}{\# \text{ random variables in the model}} = \frac{1}{n \cdot N} \log \left[ \frac{\left( \max_{x \in \mathcal{X}^N} P_{\theta_N}(x) \right)^n}{\left( \min_{x \in \mathcal{X}^N} P_{\theta_N}(x) \right)^n} \right] = \frac{1}{N} \log \frac{\max_{x \in \mathcal{X}^N} P_{\theta_N}(x)}{\min_{x \in \mathcal{X}^N} P_{\theta_N}(x)}
\]

\[
= \frac{1}{N} \text{ELPR}(\theta_N),
\]
implying that the definition of an S-unstable FSFS model sequence is invariant to the level \((n)\)
of independent replication. Consequently, overall model instabilities may be characterized by
those of one observation from the common FSFS model.

### 3.4 Implications

For a large class of models that covers a broad range of applications (including “deep
learning”), we have developed a formal definition of instability in model probability structure
and elucidated multiple consequences of instability. We have shown for FSFS models that
instability manifests through small changes in data leading to potentially large changes in
probability as well as the potential to place all probability on certain modal subsections of the
sample space, which could be potentially small. The FSFS model class is quite broad and,
particularly in developing FSFS models for large data sets, some caution should be used in
parameter specification to control effects of model instability.
CHAPTER 4. A FAST SAMPLER FOR DATA SIMULATION FROM SPATIAL, AND OTHER, MARKOV RANDOM FIELDS

A paper to be submitted to the *Journal of Graphical and Computational Statistics*

Andee Kaplan, Mark S. Kaiser, Soumendra N. Lahiri, and Daniel J. Nordman

Abstract

For spatial and network data, a model may be formulated on the basis of a Markov random field (MRF) structure and the specification of a conditional distribution for each observation. This piece-wise conditional approach often provides an attractive alternative to directly specifying a full joint data distribution, which may be difficult for large correlated data. At issue, fast simulation of data from such MRF models is often an important consideration, particularly when repeated generation of large numbers of data sets is required (e.g., for approximating reference distributions for statistics). However, the standard Gibbs strategy for simulating data from a spatial MRF models involves individual-site updates from conditional distributions, which is often challenging and computationally slow even for one complete iteration of relatively small sample size. As a remedy, we describe a fast way to simulate from MRF models, based on the concept of “concliques”, (i.e., groups of non-neighboring observations). The proposed simulation scheme is computationally fast due to its ability to lower the number of steps necessary to complete one iteration of a Gibbs sampler. We motivate the simulation method, formally establish its validity, and assess its computational performance through numerical studies, where speed advantages are shown. In addition to numerical evidence, we also formally prove that the proposed Gibbs sampler for simulating MRF data is geometrically
ergodic (i.e., exhibits fast convergence rates) for simulating data from many commonly used spatial MRF models. Such general convergence results are typically unusual for spatial data generation but made possible here through the proposed sampling scheme.

4.1 Introduction

For modeling large-scale correlated data sets, conditionally specified models can often be usefully formulated on the basis of an underlying Markov random field structure (MRF; Besag 1974). This approach involves specifying a full conditional distribution for each observation, which often depends functionally on other (neighboring) observations in the conditional model statement. Model formulation in this conditional, component-wise fashion can provide an attractive alternative to direct specification of a full joint data distribution, which may be difficult to approach for correlated data structures (e.g., spatial data). Such MRF models have become popular for modeling temporally- or spatially-dependent areal data (Cressie 1993), image segmentation (Zhang, Brady, and Smith 2001), computer vision (S. Z. Li 2012), and positron emission tomography (Higdon 1998), among other challenging applications including the analysis of networks (cf. Strauss and Ikeda 1990; Hoff, Raftery, and Handcock 2002; Casleton, Nordman, and Kaiser 2017). In addition to providing a route for model formulation, another reason for the popularity of MRF specifications is that observation-wise conditional distributions fit naturally with the Gibbs sampler (S. Geman and Geman 1984) for generating data realizations via Markov chain Monte Carlo (MCMC) methods (Gelfand and Smith 1990). That is, a well-known close connection exists between MRF formulations and the Gibbs sampler through conditionally specified distributions; see Besag (1994) and Kaiser and Cressie (2000).

Accordingly, a dominant current strategy for sampling from a MRF model involves a single-site, or sequential, update strategy with a Gibbs sampler (Besag, York, and Mollie 1991) whereby each observation in the field is simulated or updated individually, in turn, from its conditional distribution given all other current observational values. However, while simple to set up, sequential Gibbs updating can be inherently slow computationally as each complete Gibbs
A single iteration requires the same number of updates as data points in the MRF model. Consequently, even for relatively small data sets (e.g., hundreds of spatial points), there can be substantial time investments in just one run of the standard Gibbs sampler. The computational burdens are then further compounded by multiple iterations of this sampler in order to create a large collection of simulated data sets, as potentially required for ensuring appropriate mixing of the sampler (i.e., burn-in) and for adequately establishing some Monte Carlo approximation of interest (e.g., numerically approximating the distribution of a statistic). Some block updating methods have been developed to speed up simulation from the Gaussian MRF model in particular (see Rue 2001; also Rue and Held 2005 for an overview of relevant work) but these require manipulation of potentially large covariance matrices and, unlike the sequential Gibbs approach, usually have no clear extension to other MRF models.

In this paper, we introduce a simple, fast scheme for sampling from general MRF models in a manner that exploits conditional independence under the MRF model among subcollections of non-neighboring observations defined by “concliques.” For goodness-of-fit testing of a spatial MRF specification, Kaiser, Lahiri, and Nordman (2012) (hereafter [KLN]) introduced concliques as a type of converse to “cliques,” where the latter are commonly encountered with MRFs as sets of locations involving shared neighbors (e.g., Hammersley and Clifford 1971). [KLN] used concliques to develop spatial residuals and large-sample tests for the fit of MRF models. However, as a separate issue from assessing model formulation, the notion of concliques can also have alternative implications for the application of Gibbs sampling to MRFs. We show that this is indeed the case, using concliques to establish a formal approach to fast simulation from MRF models. The resulting simulation method is a block updating Gibbs sampler that applies to any valid conditionally specified MRF model under mild conditions. The proposed conclique-based approach has the advantage of being computationally more efficient than the sequential Gibbs update strategy, particularly for generating large collections of data sets, while maintaining similar rates of mixing. Similarly to the standard Gibbs sampler, the conclique-based strategy is also generally applicable compared to the simulation approaches for MRF models (cf. the end of this section).
In Section 4.2 we present some background of MRF formulations and motivate simulation from such models with an example tied to spatial binary data. We describe concliques and the conclique-based Gibbs sampler for MRFs in Section 4.3, where the proposed simulation method is also formally validated under mild model conditions. Section 4.4 then provides a numerical comparison of the mixing and computational efficiency of the proposed sampling scheme to the standard (sequential) Gibbs sampler for MRFs, where substantial speed advantages are shown.

Section 4.5 then provides a theoretical development which indicates that the proposed MCMC sampler has guaranteed fast convergence rates to the target joint data distribution for many commonly used MRF models. That is, the conclique-based Gibbs sampler is proven to be geometrically ergodic for a large class of MRF models, which includes general four-nearest neighborhood structures. Not only does this guarantee a mixing rate for the involved Markov chain, but geometric ergodicity can be used with other established results to obtain central limit theorems and Monte Carlo sample size assessments (Chan and Geyer 1994; Jones and others 2004; Hobert et al. 2002; Roberts, Rosenthal, and others 1997). The traditional sequential Gibbs sampling method for such spatial MRF models (or for other realistic MRFs for spatial data) cannot similarly be proven to be geometrically ergodic because current technology for generally establishing geometric ergodicity of a Gibbs sampler is limited to less than 4-components in the sampler, see, among others Johnson and Burbank (2015); Hobert and Geyer (1998); Tan and Hobert (2009); Doss and Hobert (2010); Jones and Hobert (2004); Johnson and Jones (2015). Hence, in addition to substantial computational speed-ups, the proposed conclique-based Gibbs sampler is shown to have useful theoretical properties for simulating spatial and other data.

To frame the simulation method to follow, we end this section with a brief overview of other simulation approaches for MRF models. While a joint data distribution, at least in theory, may be constructable from conditional distributions in a MRF specification, the normalizing terms involved are often intractable to determine in practice (cf. Kaiser and Cressie 2000). For example, this holds true for the binary or autologistic models considered in Section 4.2.2 for illustration. Hence, direct simulation from the joint distribution induced by a MRF model also
often becomes intractable, which motivates the traditional use of a sequential Gibbs sampler based on the observation-wise conditional distributions. To be clear, the proposed conclique-based Gibbs sampler to follow for simulating MRF data is meant to be computationally more efficient alternative to the standard Gibbs sampler. There are, however, simulation alternatives to Gibbs sampling altogether.

For example, through the use of coupling from the past (Propp and Wilson 1996), perfect sampling may also be employed to sample from a MRF specification (cf. Møller 1999). For certain autologistic models, in particular, perfect sampling has received much consideration for simulating spatial binary data on a lattice (cf. Hughes, Haran, and Caragea 2011; Hughes 2014; Friel and Pettitt 2004). But, due to the method’s intricacies, perfect sampling does generally require more effort to set up than the Gibbs sampling considered here, as there is no exact rule for chain coupling. Additionally, justification of perfect sampling also imposes certain monotonicity requirements in arguments of conditional distributions (cf. Møller 1999) not required in Gibbs sampling. Furthermore, when considering Gaussian MRF models, we also note that several possibilities exist for data simulation, even perfectly, including versions of direct sampling and circulant embedding (Rue 2001; Rue and Held 2005; Møller and Waagepetersen 2003; Davies, Bryant, and others 2013). However, even for Gaussian MRF models, the simplicity of the Gibbs sampler is difficult to beat. Ultimately, for MRF specifications, Gibbs sampling plays as a natural and flexible role in simulation from a broad variety of (discrete or continuous) data structures on both regular (e.g. gridded spatial) and irregular (e.g. network data) lattices. To this end, the proposed conclique-based Gibbs sampler can offer substantial advantages in computational speed and efficiency over the traditional sequential Gibbs sampler, while maintaining the same general applicability in allowing accessible simulation for a wide variety of MRFs. This helps to expand the practical possibilities of simulation from large and complex MRF data structures.
4.2 MRF formulation and illustration

4.2.1 MRF formulation

We introduce some notation for MRF models, using a description typical in an applied spatial context for concreteness. Let \( \{ s_i : i = 1, \ldots, n \} \) denote a set of locations, generically indexed in some Euclidean space (e.g. \( \mathbb{R}^2 \)), and \( \{ Y(s_i) : i = 1, \ldots, n \} \) denote a corresponding collection of univariate random variables, where \( Y(s_i) \) represents an observation associated with location \( s_i \). A MRF formulation commonly involves specifying a neighborhood for each location \( s_i \), which consists of locations whereby the full conditional distribution of \( Y(s_i) \) is functionally dependent on observations at these locations. Let \( f_i \) denote the conditional density (or mass) function of \( Y(s_i)\) given all other observations \( \{ Y(s_j) = y(s_j) : j \neq i \} \). Additionally, let \( \mathcal{N}_i \equiv \{ s_j : i \neq j \text{ and } f_i \text{ depends functionally on } y(s_j) \} \) represent the neighborhood for location \( s_i \) and state a corresponding set of neighborhood observations as \( y(\mathcal{N}_i) \equiv \{ y(s_j) : s_j \in \mathcal{N}_i \} \). Under a defining MRF assumption, it holds that

\[
f_i(y(s_i)|\{y(s_j) : j \neq i\}) = f_i(y(s_i)|y(\mathcal{N}_i)),
\]

or the full conditional distribution for \( Y(s_i) \), given all other data values, depends only on those observations \( y(\mathcal{N}_i) \) given by the neighborhood \( \mathcal{N}_i \). The data model follows by prescribing a conditional density (4.1) for each observation \( i = 1, \ldots, n \), which allows for a wide variety of models in construction. Random variables may be discrete or continuous, neighborhoods may or may not vary in size across locations, and the conditional distribution \( f_i \) can include parameters \( \theta \) or spatial covariates with a form that can potentially depend on the location. One common example of conditional densities in a MRF specification (4.1) involves an exponential family of the form given by

\[
f_i(y(s_i)|y(\mathcal{N}_i), \theta) = \exp [A_i(y(\mathcal{N}_i))y(s_i) - B_i(y(\mathcal{N}_i)) + C(y(s_i))] \tag{4.2}
\]

where \( A_i(\cdot) \) are the natural parameter functions, \( B_i(\cdot) \) is a function of \( y((\mathcal{N}_i)) \) only through \( A_i(\cdot) \), and \( C(\cdot) \) is a known function. Under an assumption of pairwise dependence (or cliques of at
most size two), Besag (1974) showed a necessary form for $A_i(\cdot)$ is

$$A_i(y(\mathcal{N}_i)) = \alpha_i + \sum_{s_j \in \mathcal{N}_i} \eta_{i,j} y(s_j)$$

with a parameter $\alpha_i \in \mathbb{R}$ and dependence parameters $\eta_{i,j} = \eta_{j,i}$. Lee, Kaiser, and Cressie (2001) generalized this parameterization result for including cliques of any size in potential neighborhood formulations, while

$$A_i(y(\mathcal{N}_i)) = \tau^{-1}(\kappa_i) + \sum_{s_j \in \mathcal{N}_i} \eta_{i,j} \{y(s_j) - \kappa_j\}$$

where $\eta_{i,j} = \eta_{j,i}$ again represents the dependence parameters in the model while $\kappa_i$ denotes a large scale parameter associated with a function $\tau^{-1}(\cdot)$ that maps expected values to natural parameters. The centered version is intended to clarify interpretation of dependence effects in the model as well as to facilitate the interpretation of $\kappa_i$ as an unconditional mean, with appropriate dependence parameters; see also Kaiser (2007). Hardouin and Yao (2008), and Kaiser and Caragea (2009).

A common conditional density form for all observation ($f_i = f$ for all $i$) is often assumed with such models. One simple example of such an exponential family MRF model is a conditional Gaussian model with density

$$f_i(y(s_i)|y(\mathcal{N}_i), \alpha, \eta, \tau) = \frac{1}{\sqrt{2\pi \tau}} \exp \left( -\frac{[y(s_i) - \mu(s_i)]^2}{2\tau^2} \right), \quad y(s_i) \in \mathbb{R} \quad \text{(4.3)}$$

determined by variance $\tau^2 > 0$ and conditional mean

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

where $\eta$ is a single dependence parameter (e.g. $|\eta| < 0.25$) and $\alpha \in \mathbb{R}$ represents an unconditional mean. Another example is a centered version of an autologistic model where $Y(s_i)$ given neighbors $y(\mathcal{N}_i)$ is Bernoulli($p(s_i, \kappa, \eta)$) distributed with

$$\text{logit}(p(s_i, \kappa, \eta)) = \text{logit}(\kappa) + \eta \sum_{s_j \in \mathcal{N}_i} \{y(s_j) - \kappa\}.$$ 

depending on a large scale (i.e. unconditional mean) parameter $\kappa \in (0,1)$ and dependence parameter $\eta \in \mathbb{R}$. 

Hence, dependence enters such models through scaling of sums of (mean centered) neighboring values, whereby dependence parameters of zero induce a model form which would standardly be used under an independence assumption. For spatial data modeling, two standard neighboring structures are four- and eight-nearest neighbors. More specifically, for observations on a lattice, a four-nearest neighborhood is defined by locations in cardinal directions as \( \mathcal{N}_i = \{ s_i \pm (1,1) \} \cup \{ s_i \pm (1,0) \} \cup \{ s_i \pm (0,1) \} \cup \{ s_i \pm (0,0) \} \), while the eight-nearest neighbor neighborhood \( \mathcal{N}_i = \{ s_i \pm (0,1) \} \cup \{ s_i \pm (1,0) \} \cup \{ s_i \pm (1,1) \} \cup \{ s_i \pm (-1,1) \} \cup \{ s_i \pm (-1,0) \} \cup \{ s_i \pm (-1,-1) \} \), further includes neighboring diagonal A visual representation of these two structures is presented below, where * represent neighbors of \( s_i \) and \( \cdot \) represent non-neighbors in the lattice.

\[
\begin{array}{ccc}
\cdot & * & \cdot \\
* & s_i & * \\
\cdot & * & \cdot \\
\end{array}
\quad \quad \quad \quad \\
\begin{array}{ccc}
* & * & * \\
* & s_i & * \\
* & * & * \\
\end{array}
\]

We will often assume that a valid joint distribution for \( \{ Y(s_1), \ldots, Y(s_n) \} \) exists from the conditionals specified (4.1). Arnold et al. (2001) provide conditions necessary for this, while Kaiser and Cressie (2000) describe the construction of conditionals under conditions which guarantee a valid joint distribution. While we again use a model formulation framed toward spatial data, note that the same MRF elements translate to other non-spatial data settings (e.g., for network or random graph data, “locations” \( s_i \) may mark or denote the position of a potential edge while \( Y(s_i) \) may represent the presence/absence of the edge).

4.2.2 Illustrative example

To motivate the simulation approach to follow, we first present a small example of how simulation from MRF models may arise and be required in practice. For this, we consider a spatial dataset from Besag (1977) consisting of binary observations located on a \( 14 \times 179 \) grid indicating the presence or absence of footrot in endive plants. Figure 4.1 shows the endive data where black pixels indicate a value of 1, or presence of the disease. As illustration, we consider fitting three models of increasing complexity to these data via pseudo-likelihood (Be-
Figure 4.1: The endive dataset, a $14 \times 179$ rectangular lattice with binary data encoding the presence or absence of footrot in endive plants from Besag (1977).

and apply simulation to obtain reference distributions for statistics based on the resulting estimators. As this involves a type of parametric bootstrap approximation for sampling distributions, the speed of the proposed sampler becomes an important consideration in swiftly rendering a large number of spatial data sets from differing models. The three models we will consider for the spatial binary data are (a) an isotropic centered autologistic model (Caragea and Kaiser 2009; Besag 1972; Besag 1977), (b) a centered autologistic model with two dependence parameters, and (c) a centered autologistic model as in (b) but having large scale structure determined by regression on the horizontal coordinate $u_i$ of each spatial location $s_i = (u_i, v_i)$. For each model, a four-nearest neighborhood is used (with natural adjustments for border observations) and the resulting conditional mass function has the form

$$f_i(y(s_i)|y(A_i), \theta) = \frac{\exp[y(s_i)A_i(y(A_i))]}{1 + \exp[y(s_i)A_i(y(A_i))]}, \quad y(s_i) = 0, 1,$$

with natural parameter functions, $A_i(y(A_i)) \equiv A_i(y(A_i))(\theta)$ given in Table 4.1, which depend on a vector of model parameters that we denote generically as $\theta$ for notational purposes; Table 4.1 further describes the various model-wise parameters which, upon collection, would represent $\theta$ here for a model.

Pseudo-likelihood again yields parameter estimates $\hat{\theta}$ for each of the three models. To calibrate confidence intervals based on $\hat{\theta}$ for a model, normal approximations are difficult to use because standard errors for such pseudo-likelihood estimators depend intricately on the underlying dependence among spatial observations, with no tractable form (cf. Guyon 1982). Instead, simulation in the form of a model-based bootstrap may be applied to estimate the
sampling distribution of \( \hat{\theta} \). Using the full conditional distributions given by the estimates \( \hat{\theta} \), as a proxy for the unknown parameters \( \theta \), we generated 10,000 spatial samples for each binary MRF model based on the proposed Gibbs sampler (after a burn-in of 1,000 and thinning by a factor of 5 which were conservative selections by trace plots). Each simulated spatial dataset was of the same size as the endive data, so that resulting collection of simulated samples provides a set of re-creations of the original data under each model. A (bootstrap) parameter estimate, say \( \hat{\theta}^* \), is obtained from each generated spatial sample for a given model, and the resulting empirical distribution of bootstrap estimates across simulated data sets approximates the sampling distribution of \( \hat{\theta} \) (where, technically, the relationship between \( \hat{\theta}^* \) and \( \hat{\theta} \) in the bootstrap world aims to mimic that of \( \theta \) and \( \hat{\theta} \)). As illustration, Figure 4.2 displays the sampling distributions for the pseudo-likelihood estimators of dependence parameters (e.g., \( \eta, \eta_u, \eta_v \)) in the three models, as estimated through bootstrap simulation. From these distributional approximations, Table 4.2 shows 95% (percentile bootstrap) confidence intervals for all model parameters. The intervals suggest that spatial dependence is a significant aspect of Models (a) and (b), but that most of the explanatory power of Model (c) lies in the model’s large scale structure as opposed to dependence. Hence, as in this data example, simulation from MRF models can be helpful for quantifying the uncertainty in parameter estimation, provided that the simulation can be conducted in a computationally fast and practical way. The latter is of interest in the simulation development presented in the next section.

As another example of MRF model-based simulation in inference, we also consider reference distributions for goodness-of-fit (GOF) model assessments for all three models. Besag (2001) assessed the fit of the isotropic autologistic model (Model (a)) using a Monte Carlo test and concluded that the model is a poor fit for these data. We repeat the assessment using a GOF test statistic \( T(\hat{\theta}) \) from [KLN], which is based on pseudo-estimates and suggests an inadequate model fit for large test statistic values according to large-sample theory developed in [KLN]. In order to approximate a reference distribution for testing, we use the collection of the same bootstrap simulated data sets to evaluate test statistic analogs and compute p-values as reported in Table 4.3. Our results support the lack-of-fit conclusion of Besag (2001)
Figure 4.2: Sampling distribution of the dependence parameters ($\eta$, $\eta_u$, and $\eta_v$) for the three centered autologistic models with four-nearest neighbor structure, (a) the isotropic model with one dependence parameter, (b) the model with two dependence parameters, and (c) the model with two dependence parameters and a marginal mean structure based on the regression on the horizontal location component, $u_i$. 
Table 4.1: Full conditional distributions of three binary MRF models for the endive data.

<table>
<thead>
<tr>
<th>Model</th>
<th>Natural parameter function</th>
</tr>
</thead>
<tbody>
<tr>
<td>(a) Isotropic centered autologistic model with $\kappa \in (0,1)$, $\eta \in \mathbb{R}$, and $\mathcal{M}_i = {s_i \pm (1,0), s_i \pm (0,1)}$</td>
<td>$A_i{y(\mathcal{M}<em>i)} = \log \left( \frac{1}{1 - \kappa_i} \right) + \eta \sum</em>{s_j \in \mathcal{M}_i} {y(s_j) - \kappa}$</td>
</tr>
<tr>
<td>(b) Centered autologistic model with $\kappa \in (0,1)$ and dependence parameters $\eta_u, \eta_v \in \mathbb{R}$ in horizontal/vertical directions with $\mathcal{M}<em>{u,i} = {s_i \pm (1,0)}$, $\mathcal{M}</em>{v,i} = {s_i \pm (0,1)}$</td>
<td>$A_i{y(\mathcal{M}<em>i)} = \log \left( \frac{1}{1 - \kappa_i} \right) + \eta_u \sum</em>{s_j \in \mathcal{M}<em>{u,i}} {y(s_j) - \kappa} + \eta_v \sum</em>{s_j \in \mathcal{M}_{v,i}} {y(s_j) - \kappa}$</td>
</tr>
<tr>
<td>(c) Centered autologistic model as in (b) with scale parameter $\kappa_i$ determined by logistic regression ($\beta_0, \beta_1 \in \mathbb{R}$) on the horizontal coordinate $u_i$ of location $s_i = (u_i, v_i)$</td>
<td>$A_i{y(\mathcal{M}<em>i)} = \log \left( \frac{\kappa_i}{1 - \kappa_i} \right) + \eta_u \sum</em>{s_j \in \mathcal{M}<em>{u,i}} {y(s_j) - \kappa_i} + \eta_v \sum</em>{s_j \in \mathcal{M}_{v,i}} {y(s_j) - \kappa_i}$, $\log \left( \frac{\kappa_i}{1 - \kappa_i} \right) = \beta_0 + \beta_1 u_i$</td>
</tr>
</tbody>
</table>

Table 4.2: Bootstrap percentile confidence intervals in all three autologistic models.

<table>
<thead>
<tr>
<th></th>
<th>Model (a)</th>
<th>Model (b)</th>
<th>Model (c)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\eta$</td>
<td>$\kappa$</td>
<td>$\eta_u$</td>
</tr>
<tr>
<td>2.5%</td>
<td>0.628</td>
<td>0.107</td>
<td>0.691</td>
</tr>
<tr>
<td>50%</td>
<td>0.816</td>
<td>0.126</td>
<td>0.958</td>
</tr>
<tr>
<td>97.5%</td>
<td>1.001</td>
<td>0.145</td>
<td>1.220</td>
</tr>
</tbody>
</table>

cconcerning the isotropic autologistic model ($p$-value=0.037), though Models (b) and (c) are more compatible with these data ($p$-values=0.879 and 0.361 respectively) by adding directional model structure (e.g., in neighborhood or large scale parameters). The simulations above were performed with the proposed (conclique-based) Gibbs sampler. Had we used the standard sequential Gibbs, the reported results would have been virtually identical with the same number of iterations. However, the generation of the reference distribution using the standard sampler would have taken approximately 29.04 minutes longer for Model (a), approximately 34.95 minutes longer for Model (b), and approximately 34.4 minutes longer for Model (c) compared to the proposed MRF sampler (which had running times of 5.34 seconds, 5.96 seconds, and 71.95 seconds, respectively).
Table 4.3: Bootstrap p-values for a goodness-of-fit (GOF) assessment of the three centered autologistic models with four-nearest neighbor structure, (a) the isotropic model with one dependence parameter, (b) the model with two dependence parameters, and (c) the model with two dependence parameters and a marginal mean structure based on the regression on the horizontal location component, \( u_i \), fit to the endive data.

<table>
<thead>
<tr>
<th>Model</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>(a)</td>
<td>0.037</td>
</tr>
<tr>
<td>(b)</td>
<td>0.879</td>
</tr>
<tr>
<td>(c)</td>
<td>0.361</td>
</tr>
</tbody>
</table>

These timings are based on C++ implementations of both samplers in R (R Core Team (2017)); available in an associated R package, `conclique`, to appear on CRAN) using a 1.7 GHz processor. Furthermore, the simulation studies above were based on 10,000 data generations as a pragmatic choice for a number of sampling iterations that might be employed in practice. However, if, say, one wished to use enough Monte Carlo samples so that reported bootstrap confidence intervals would not change (or be guaranteed with some level of confidence) to second or third decimal place accuracy, then a data-based approach could further be used to estimate number of iterations required (cf. Raftery, Lewis, and others 1992). For example, with the endive data and using the 10,000 simulated data sets for each model as a pilot collection, the sample size method of Liu, Nordman, and Meeker (2016) estimates that 5,641,429 sampling iterations would be needed for either the standard or proposed samplers to determine bootstrap confidence intervals for dependence parameters in Model (a) to the second decimal place with 90% confidence (3,341,725 and 31,508,527 sizes in Models (b) and (c) respectively). For perspective, the time savings of the proposed sampler now becomes on the order of months compared to the standard Gibbs approach.

### 4.3 Conclique-based Gibbs sampling algorithms

Recall that the MRF model involves specifying a conditional distribution \( f_i \) for each observation \( Y(s_i) \) as in (4.1), which depends on observations \( y(\mathcal{N}_i) \) in an associated neighborhood \( \mathcal{N}_i \) for location \( s_i \). From this model formulation, a `conclique` is defined by [KLN] as a singleton set or a set of locations such that no location in the set is a neighbor of any other location in the set. For any MRF specification, a collection of \( Q \) concliques given by \( \mathcal{C}_1, \ldots, \mathcal{C}_Q \) can be
found which partition the available spatial locations as \( \bigcup_{i=1}^{Q} \mathcal{C}_i = \{s_1, \ldots, s_n\} \) with \( \mathcal{C}_i \cap \mathcal{C}_j = \emptyset \) for \( i \neq j \).

For example, if we consider spatial data on a regular grid and the common four-nearest neighborhood structure as in (4.4), then it is possible to partition locations into two concliques, as labeled on the lattice below.

```
four-nearest
1 2 1 2
2 1 2 1
1 2 1 2
2 1 2 1
```

```
eight-nearest
1 2 1 2
3 4 3 4
1 2 1 2
3 4 3 4
```

On the other hand, for an eight-nearest neighbor scheme (4.4), locations can be partitioned into four concliques illustrated above. Note that if a single conclique in the conclique example above with four-nearest neighbors is subdivided, the resulting subsets also constitute concliques (e.g., in the conclique example above with four-nearest neighbors, any nontrivial subset of “2”'s could be replaced by “3”'s to create a collection of three concliques). Hence, ideally, one wishes to have minimal collection of concliques, or a so-called minimal conclique cover [cf. KLN], whereby \( Q \) is as small as possible. For example, minimal conclique covers have sizes \( Q = 2 \) and \( Q = 4 \), respectively, for the four- and eight-nearest neighbor schemes above.

For goodness-of-fitting testing with MRF models, [KLN] used concliques to define generalized spatial residuals based on a conditional probability integral transform. That is, let \( F_i \) denote the cumulative distribution function (cdf) for the conditional density \( f_i \) of observation \( Y(s_i) \) in (4.1). If we assume such cdfs are continuous for simplicity, then the residual for location \( s_i \) is defined as \( R(s_i) = F_i(Y(s_i) : \{Y(s_j) : s_j \in N_i\}) \), by substituting observations into the conditional cdf form. A main result of [KLN] for assessing MRF specifications is that, under the MRF model, these residuals are iid Uniform(0,1) distributed within each conclique: that is,

\[
\{R(s) : s \in \mathcal{C}_j\} \text{ are a Uniform}(0,1) \text{ random sample,}
\] (4.5)
for each $j = 1, \ldots, Q$. While dependence can exist between residuals from differing concliques, [KLN] developed goodness-of-fit test statistics by comparing residuals per conclique to a uniform reference distribution and pooling discrepancies across concliques. However, apart from model assessment, one could alternatively interpret the result in (4.5) as a means of simulating or generating observations for an entire conclique: draw a random sample of Uniform$(0,1)$ variables, say $\{U(s) : s \in \mathcal{C}_j\}$, and compute $Y(s) \equiv R^{-1}(U(s)), s \in \mathcal{C}_j$. For generating data from a MRF model, this suggests a way to formulate a Gibbs sampler in an alternative fashion to the standard sequential Gibbs approach, whereby updates are conducted independently and simultaneously per conclique, which we describe next.

The algorithm for a conclique-based composition Gibbs sampler (CGS) is presented below. Additionally, Appendix D describes two further conclique-based samplers in the form of a random sequence scan (RQGS) and random scan (RSGS) Gibbs sampler. The CGS updates each conclique in a fixed order for each iteration, whereas the RQGS updates all concliques in a randomly selected order according to a fixed permutation probability while the RSGS randomly updates one conclique in each iteration while maintaining the other conclique according to a fixed component selection probability. While CGS is the most commonly used Gibbs sampling scheme, we present RSGS and RQGS for completeness as these possess theoretical properties of potential interest in some cases (e.g. reversibility; cf. Johnson and Burbank 2015). In the following, let $Y^{(i)}(s)$ denote the value of an observation at location $s$ at the $i$th iteration of the Gibbs sampler for $i = 0, 1, \ldots$.

**Conclique-based CGS Algorithm.** Let $M \geq 1$ denote the number of complete Gibbs iterations:

1. Split locations into $Q \geq 2$ disjoint concliques, $\mathcal{C}_1, \ldots, \mathcal{C}_Q$.
2. Initialize the values of $\{Y^{(0)}(s) : s \in \{\mathcal{C}_2, \ldots, \mathcal{C}_Q\}\}$.
3. For iteration $i = 1, \ldots, M$,
   1. Considering all locations $s_j \in \mathcal{C}_1$, sample $\{Y^{(i)}(s_j) : s_j \in \mathcal{C}_1\}$ by independently drawing $Y^{(i)}(s_j) \sim f_j(\cdot|\{Y^{(i-1)}(s), s \in \mathcal{N}_j\})$ from conditionals in (4.1).
2. Set $\ell = 2$. Considering all locations $s_j \in C_\ell$, sample $\{Y(i)(s_j) : s_j \in C_\ell\}$ by independently drawing $Y(i)(s_j) \sim f_j(i|Y(i)(N_j))$ with conditioning observations $y(i)(N_j) \equiv \cup_{k=1}^{\ell-1}\{Y(i)(s) : s \in N_j \cap C_k\} \cup \cup_{k=\ell+1}^{Q}\{Y(i-1)(s) : s \in N_j \cap C_k\}$, where the second set union is defined as empty if $\ell = Q$.

3. For $Q > 2$, repeat step 2 for each $\ell = 3, \ldots, Q$.

Note that, at the $i$th sampling iteration, observations with locations the $\ell$th conclique $C_\ell$ ($1 < \ell < Q$) are updated conditional on the observations associated with other concliques, where observations with locations in concliques $C_1, \ldots, C_{\ell-1}$ are updated before conclique $C_\ell$ at the $i$th stage of iteration while observations associated with concliques $C_{\ell+1}, \ldots, C_Q$ are updated after conclique $C_\ell$ at the $i$th stage. This sampling plan is operational because, at each conclique update, the neighboring observations needed for conditioning in target conditional distributions, by design, never belong to the same conclique being updated. Essentially, the sampler exploits a group type of conditional independence that induced by the MRF model, specified at the individual observation level $f_i$.

Under a mild condition on the MRF model, presented next, the conclique-based Gibbs sampler can be shown to be valid.

**Condition 1** (Conclique positivity condition). The full conditionals (4.1) for the MRF model specify a valid joint distribution $\Pi(\cdot)$ for $(Y(s_1), \ldots, Y(s_n))$ with density/mass function $\pi(\cdot)$ having support $\mathcal{X} \subset \mathbb{R}^n$. For the collected $C_1, \ldots, C_Q$ of concliques under the MRF model, it holds that $\mathcal{X} = \mathcal{X}_1 \times \cdots \times \mathcal{X}_Q$ where $\mathcal{X}_i$ denotes the support of the marginal density of observations $\{Y(s_j) : s_j \in C_i\}$ with locations in conclique $C_i, i = 1, \ldots, Q$.

We again assume the conditional specification of the MRF yields a valid joint distribution $\Pi(\cdot)$ for observations $\{Y(s_1), \ldots, Y(s_n)\}$; see Kaiser and Cressie (2000) and references therein for details on joint construction. Condition 1 also involves a minimal assumption regarding the support of observations among concliques in order to guarantee appropriate transition properties of the proposed CGS. This conclique-wise positivity condition is implied by
Besag (1974)’s original positivity condition, stating that the joint support may be expressed as the cross product of marginal supports across individual observations. (The latter was used by Besag (1974) to determine a joint distribution from conditional specifications (1) in some MRF models.) Under Condition 1, Theorem 1 next shows that the conclique-based CGS is guaranteed to capture the target joint data distribution $\Pi$ as the number of Gibbs iterations increase. That is, the sampler is provably Harris ergodic (i.e., $\phi$-irreducible, aperiodic and Harris recurrent with invariant distribution $\Pi(\cdot)$ for a measure $\phi$); see Harris (1956). To state the result, let $P^{(m)}(x,A), A \in \mathcal{F}$, denote the transition distribution of the conclique-based CGS after $m \geq 1$ complete iterations from an initializing point $x \in \mathcal{X}$, where $\mathcal{F}$ denotes the appropriate $\sigma$-algebra associated with $\mathcal{X} \subset \mathbb{R}^n$.

**Theorem 1.** Suppose Condition 1 holds. Then the conclique-based CGS Gibbs sampler (presented above) is Harris ergodic with stationary distribution given by the joint data distribution, $\Pi(\cdot)$ and, for any initialization $x \in \mathcal{X}$, the sampler will monotonically converge to $\Pi(\cdot)$ in total variation distance as the number of iterations $m \to \infty$, i.e.,

$$\sup_{A \in \mathcal{F}} |P^{(m)}(x,A) - \Pi(A)| \downarrow 0 \quad \text{as } m \to \infty.$$ 

The conclusion of Theorem 1 also holds for the two additional conclique-wise Gibbs samplers (RQGS and RSGS) described in Appendix D. See Appendix E for details on the proof.

### 4.4 Simulation comparisons

By exploiting a systematic type of group-wise conditional independence induced a MRF formulation (4.1), the intent of the conclique Gibbs sampler is to provide a more efficient method for simulating data from MRF models than the sequential Gibbs sampler that is directly implied by the observation-wise conditional distributions in the original MRF specification (4.1). To numerically investigate performance, we employ a quantitative framework from Turek et al. (2017) and explore two contributors to MCMC efficiency in terms of both mixing effectiveness (or algorithmic capability to produce approximately independent samples from the target joint
data distribution) as well as computational demands of the algorithm (related to computing speed). Because the conclique-based Gibbs sampler is a block updating Gibbs sampler, this may be compared to the traditional sequential (scalar) sampler in both senses of efficiency with metrics of Turek et al. (2017). We compare the methods for simulating data $Y(s_i)$ from a spatial MRF model at a set of $n$ locations $s_i, i = 1, \ldots, n$ on a regular grid, as explained in the following.

To assess mixing or algorithmic efficiency, we consider the location-wise minimum number of effective samples per actual sample, which is defined in terms of a quantity

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

where $\rho_i(j), j \geq 1$, denotes the process autocorrelation function for the chain generations of observation $Y(s_i)$ over MCMC/Gibbs iterations. The quantity $A$ corresponds to the minimum inverse of integrated autocorrelations among the chains output by location (Roberts, Rosenthal, and others 2001; Turek et al. 2017). For $M$ full iterations of a Gibbs sampler, the number of approximately independent full data sets is then approximated as $A/M$, after adjusting for the largest autocorrelation among MCMC iterations incurred a sampling location. Hence the intent of this measure is to capture the “worst case scenario” in terms of chain mixing for a Gibbs sampler type, where small values of $A$ indicate poorer mixing properties for a sampler. For a given MRF model and sample size $n$, the quantity $A$ is estimated from a realized chain produced by a Gibbs sampler and we use a kernel estimator based on sample autocorrelations as provided in the package LaplacesDemon (Statisticat and LLC. 2016).

As a different consideration, computational efficiency (computing speed) is to be measured in units of algorithmic run-time per MCMC iteration. For the purpose of generating data from a spatial MRF model, we will compare the total computational cost of simulating an observation at each spatial location with respect to both conclique and sequential Gibbs samplers. For this, let $\text{samp}(\cdot |)$ generally denote the time needed to randomly sample once from a (scalar or vector-valued) conditional distribution. Then the computational cost $C$ for each of the two
sampling schemes is given respectively as

\[
C = \begin{cases} 
\sum_{k=1}^{Q} \text{samp}(\{Y(s_i) : s_i \in C_k\} | \mathcal{E}_j, j \neq k) & \text{Conclique-based} \\
\sum_{k=1}^{n} \text{samp}(Y(s_k)|Y(s_j), j \neq k) & \text{Sequential},
\end{cases}
\]

which does not include any initial computational overhead, such as memory allocation or assigning location to concliques. If the average cost \( \text{samp}(\{Y(s) : s \in C_k\} | \mathcal{E}_j, j \neq k)/n_k \) to jointly sample all \( n_k \) (say) values of a conclique is less than the average cost \( C/n \) to individually sample all \( n \) data values, then the conclique Gibbs sampler will exhibit computational improvements over the sequential update Gibbs sampler. Reported values of \( A \) are determined by the average of estimation from 10 chains with different starting values and values of \( C \) are determined by recording running times of 20,000 conclique-based \( \text{samp}(\{Y(s) : s \in C_k\} | \mathcal{E}_j, j \neq k) \) and 16,000 sequential \( \text{samp}(Y(s_k)|Y(s_j), j \neq k) \) in \( \mathbb{R} \) and averaging the results.

For illustration, we first consider the two metrics \( A \) and \( C \) for the three spatial binary models of varying complexity introduced in Section 4.2.2 for a 40 \( \times \) 40 grid and 10,000 iterations. The results for all three models are in Table 4.4. Table 4.4 indicates that, as the models become more complex (more parameters), their algorithmic or mixing efficiency is slightly decreased and the computational complexity remains similar. In other words, to achieve the same number of effective draws from the target joint data distribution, more iterations become necessary as the underlying model becomes slightly more complex, though the computational demands per Gibbs iteration are relatively unchanged for these models. However, comparing the conclique-based to the sequential-based samplers, the conclique-based is at least 81 times faster than the sequential-based sampler in any model case here. This example suggests that the algorithmic/mixing efficiencies of both conclique-based and sequential Gibbs samplers are

<table>
<thead>
<tr>
<th>Gibbs algorithm</th>
<th>Model (a)</th>
<th>Model (b)</th>
<th>Model (c)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>A</td>
<td>C</td>
<td>A</td>
</tr>
<tr>
<td>Conclique</td>
<td>0.809</td>
<td>3.6 \times 10^{-4}</td>
<td>0.747</td>
</tr>
<tr>
<td>Sequential</td>
<td>0.809</td>
<td>0.029</td>
<td>0.756</td>
</tr>
</tbody>
</table>
Figure 4.3: Comparisons of log time for simulation of $M = 100, 1000, 5000, 10000$ four-nearest neighbor 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.

often similar, which is true for the models and sample sizes considered in Table 4.4 as well as for all other MRF models and grid sizes encountered in our investigations. Nevertheless, the actual time savings of the conclique-based sampler over the sequential Gibbs can be quite substantial, where the computational benefits can greatly increase as the desired number of MCMC iterations grows. The latter is relevant to enhancing the effective MCMC sample size when simulating data from MRF models of increasing complexity.

To illustrate and compare computational speeds in a larger context, we also evaluated timing results for simulation of data from the conditional Gaussian specification in (4.3), considering various spatial grid sizes and numbers of iterations for each Gibbs sampler. In (4.3), the exact parameter values are immaterial to the timing study, but we chose $\alpha = 0$, $\tau^2 = 1$ and $\eta = 0.2$ there. For timing reference, we implemented the standard sequential and conclique-based Gibbs samplers using C++ implementations in an available R package conclique on a 1.7 GHz processor. Figure 5.2 summarizes log running times for simulating $M$ data sets from the Gaussian MRF model on a grid of size $n = m \times m$, for $m = 5, 10, 20, 30, 50, 75$ and $M = 100, 1000, 5000, 10000$. 
For small grid sizes (e.g. 5 × 5 or 10 × 10) the time savings are minimal, but as grid sizes become larger the time saving for the conclique verses sequential Gibbs sampler is substantial. For example, the conclique-based Gibbs sampler took 15.05 seconds and the sequential-based Gibbs sampler took $1.076 \times 10^4$ seconds ($\approx 2.99$ hours) to simulate 10,000 spatial data sets of size $75 \times 75$. Essentially, as the number $M$ of iterations increases, the computational time is linear in $M$ with both samplers for a given spatial sample size $n = m \times m$. However, from Figure 5.2, computational time increases exponentially faster for the sequential Gibbs sampler compared to the conclique approach as number of observations increases through $m$. Consequently, the conclique-based Gibbs sampler can be dramatically more time efficient for simulating large collections of even moderately sized samples.

### 4.5 Ergodicity results

In addition to potential benefits for computational speeds, the structure of the conclique-based Gibbs sampler also allows some important theoretical properties to be generally and formally shown. Our goal here is to establish that the sampling approach is guaranteed to be geometrically ergodic or, equivalently, to exhibit a geometrically fast mixing rate as a function of the number of iterations. Recalling notation from Section 3, let $\Pi(\cdot)$ again denote the joint distribution of observations $\{Y(s_1), \ldots, Y(s_n)\}$ induced by a MRF specification (4.1) and let $P^{(m)}(x, \cdot)$ represent the transition distribution at the $m$th iteration of the sampler with initialization $x \in \mathcal{X}$. Then, 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

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

where again $\mathcal{F}$ denotes the $\sigma$-algebra associated with the joint support $\mathcal{X} \subset \mathbb{R}^n$ of $\{Y(s_1), \ldots, Y(s_n)\}$ (cf. Condition 1). Hence, if the conclique-based sampler is geometrically ergodic, then the distribution $P^{(m)}(x, \cdot)$ induced by the sampler converges geometrically fast to the target (joint) distribution in total variation norm, regardless of initialization $x \in \mathcal{X}$. It turns out that the conclique-based Gibbs sampler (CGS) can, in fact, be proven to be
geometrically ergodic for a general type of MRF model, namely, a MRF specification exhibiting two concliques. While the result is specialized to two conclique MRF models, this model class is surprisingly large in spatial applications where four-nearest neighborhood structures are commonly used, which creates two concliques (cf. Section 4.3). In contrast, geometric ergodicity of the standard sequential Gibbs sampler is not possible to similarly establish for this collection of MRF models or more generally. This is because current theory for the geometric ergodicity of a Gibbs sampler is restricted to less than 4-components in the Gibbs sampler (See, among others Johnson and Burbank 2015; Hobert and Geyer 1998; Tan and Hobert 2009; Doss and Hobert 2010; Jones and Hobert 2004; Johnson and Jones 2015).

In other words, the conclique-based approach allows fast theoretical mixing properties to be established in sampling from important collections of spatial MRF models, which has been previously intractable to do with Gibbs sampling. This finding is summarized in Theorem 2.

**Theorem 2.** Suppose a MRF model for \( \{Y(s_i) : i = 1, \ldots, n\} \) admits two concliques and assume Condition 1 holds with \( \mathcal{X} = \mathcal{X}_1 \times \mathcal{X}_2 \subset \mathbb{R}^n \) (where \( \mathcal{X}_i \) denotes the support of observations associated with conclique \( i = 1, 2 \)). Additionally, suppose that either \( \mathcal{X}_1 \) or \( \mathcal{X}_2 \) is compact and that the full conditionals (4.1) are continuous in conditioning variables \( y(s_i), i = 1, \ldots, n \). Then, the conclique-based Gibbs sampler (CGS) is geometrically ergodic with stationary distribution given by the full joint, \( \Pi(\cdot) \).

**Remark.** Theorem 1 also holds for two additional Gibbs samplers (RQGS and RSGS) described in Appendix D in addition to the CGS. See Appendix E for proof.

Theorem 2 automatically ensures geometric ergodicity of the conclique-based sampler for several four-nearest neighbor MRF models having compact support \( \mathcal{X} \) for \( \{Y(s_1), \ldots, Y(s_n)\} \), such as the autologistic binary, the conditional binomial, the conditional Beta, and the Multinomial distributions as well as the windsorized Poisson model of Kaiser and Cressie (1997).

Furthermore, the geometric ergodicity of the conclique-based Gibbs sampling algorithm can also be established for four-nearest neighborhood MRF models with unbounded support. Theorem 3 treats three such cases of conditional distributions in the form of centered versions
of the Gaussian, Inverse Gaussian, and Truncated Gamma MRF models. These models belong to exponential families, with conditional densities of the form
\[
f_i(y(s_i)|\{y(\mathcal{N}_i)\}) = \exp \left[ \sum_{k=1}^{K} A_{ki}(y(\mathcal{N}_i)) T_k(y(s_i)) - B_i(y(\mathcal{N}_i)) + C(y(s_i)) \right]
\]
which is a generalization of (4.2) involving further possible statistics \(T_k(y(s_i))\) of observation \(y(s_i)\) in the conditional density along with associated natural parameter functions \(A_{ki}(y(\mathcal{N}_i))\) based on neighboring observations \(y(\mathcal{N}_i)\) (cf. Kaiser 2007). Additionally, \(f_i\) in (4.6) may depend on further model parameters, as indicated in models considered in Theorem 3 next.

**Theorem 3.** Suppose \(\{Y(s_i) : i = 1, \ldots, n\}\) with positions on a regular lattice in \(\mathbb{R}^2\) follow a MRF model with a common conditional distribution form (4.1) belonging to one of the following exponential families with a neighborhood \(\mathcal{N}_i \subset \{s_i, \pm (0, 1), s_i, \pm (1, 0)\}\), \(i = 1, \ldots, n\) (i.e. four-nearest neighborhood structure). Then, the conclique-based Gibbs sampler (CGS) is geometrically ergodic for each of the following

(a) The conditional Gaussian model from (4.3) having conditional variance \(\tau^2\) and density
\[
f_i(y(s_i)|y(\mathcal{N}_i)) = \frac{1}{\sqrt{2\pi}\tau} \exp \left\{ -\frac{1}{2\tau^2} (y(s_i) - \mu(s_i)) \right\}, \quad y(s_i) \in \mathbb{R},
\]
and conditional mean
\[
\mu(s_i) = \alpha + \eta \sum_{s_j \in \mathcal{N}_i} \{y(s_j) - \alpha\}
\]
where \(|\eta| < 0.25\) and \(\alpha \in \mathbb{R}\).

(b) The conditional (centered) Inverse Gaussian model with conditional expectations
\[
E(y(s_i)|y(\mathcal{N}_i)) = \sqrt{A_{2i}(y(\mathcal{N}_i))/A_{1i}(y(\mathcal{N}_i))} \quad \text{and} \quad E(1/y(s_i)|y(\mathcal{N}_i)) = \sqrt{A_{1i}(y(\mathcal{N}_i))/A_{2i}(y(\mathcal{N}_i))} + 1/A_{1i}(y(\mathcal{N}_i))
\]
and conditional density form
\[
f_i(y_i|\theta) = \exp \left\{ \frac{A_{1i}(y(\mathcal{N}_i))}{2} y(s_i) - \frac{A_{2i}(y(\mathcal{N}_i))}{2} \frac{1}{y(s_i)} - B_i(y(\mathcal{N}_i)) + C(y(s_i)) \right\}, \quad y(s_i) \geq 1
\]
where
\[
A_{1i}(y(\mathcal{N}_i)) = \frac{\lambda}{\mu^2} + \eta_1 \sum_{s_j \in \mathcal{N}_i} \left( \frac{1}{y(s_j)} - \frac{1}{\mu} - \frac{1}{\lambda} \right)
\]
\[
A_{2i}(y(\mathcal{N}_i)) = \lambda + \eta_2 \sum_{s_j \in \mathcal{N}_i} (y(s_j) - \mu)
\]
and $\mu, \lambda > 0$, $0 \leq \eta_1 \leq \lambda^2/4\mu(\lambda + \mu), 0 \leq \eta_2 \leq \lambda^2/4\mu$. In this model, the parameters $\mu$ and $\lambda$ control the large scale mean, while $\eta_1$ and $\eta_2$ control the dependence. Under an independence model ($\eta_1 = \eta_2 = 0$), the mean of $Y(s_i)$ is $\mu$ while the mean of $1/Y(s_i)$ is $1/\mu + 1/\lambda$.

The conditional (centered) Truncated Gamma model where $Y(s_i|y(M_i))$ is gamma (supported on $[1, \infty)$) with scale parameter $A_{1i}(y(M_i)) + 1$ and shape parameter $1/A_{2i}(y(M_i))$ with conditional density

$$f_i(y(s_i)|\theta) = \exp \{A_{1i}(y(M_i))\log(y_i) - A_{2i}(y(M_i))y_i - B_i(y(M_i))\}, \quad y(s_i) \geq 1$$

where

$$A_{1i}(y(M_i)) = \alpha_1 + \eta \sum_{s_j \in M_i} \log(y(s_j)) \quad \text{and} \quad A_{2i}(y(M_i)) = \alpha_2$$

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

The centered parameterization of Inverse Gaussian and Truncated Gamma models in Theorem 3 provides an analog of the centering formulation developed in Caragea and Kaiser (2009) for spatial binary models. Hence, despite the theoretical limitation to conditional specifications with two concliques, the geometric ergodicity of conclique-based Gibbs sampler extends to simulating data from many MRF models for spatial data.

### 4.6 Discussion

Fast simulation of data from MRF models is often an important task in statistical inference, which may be difficult to approach for correlated data structures (e.g., spatial data). We have presented a fast approach to simulating data from MRF models that employs conclique-based Gibbs sampling as an alternative to the standard single-site (sequential) Gibbs strategy while maintaining general applicability compared to other simulation approaches for MRF models. In fact, this strategy is applicable to data structures that do not fall on a regular lattice, like network data.
In this paper, we have motivated the simulation method, formally established its validity, and assessed its computational performance through numerical studies, where speed advantages are shown. In addition to numerical evidence, we also presented formal proofs that the proposed Gibbs sampler for simulating MRF data is geometrically ergodic for simulating data from many commonly used spatial MRF models. Such general convergence results are typically unusual for spatial data generation but made possible here through the proposed sampling scheme.

In additional to the spatial MRF models explored in this paper, a MRF model can be used for random graphs and networks, modeling the incidence of edges between two graph nodes with a conditional distribution for each edge. For even a graph with a small number of nodes, there can be many edges, meaning that the single-site (sequential) Gibbs sampler would be computationally time demanding but the conclique-based approach could be applied to reduce the number of computations in each iteration. For example, in the graph models of (Casleton, Nordman, and Kaiser 2017), there is a geographic notion (a radius of influence) whereby nodes in the graph which are far apart in distance (or other covariates) will not share an edge. It follows that collections of nodes which are separated by a distance under the MRF model can be used to define concliques for edges in the graph (observations as edges which do not neighbor other edges in the graph). In this case, the proposed conclique-based Gibbs sampler could be used to allow inference on model parameters or GOF tests through parametric bootstrap samples in a computationally feasible amount of time.
CHAPTER 5. conclique: SIMULATION AND GOODNESS OF FIT FOR SPATIAL AND OTHER MARKOVIAN DATA STRUCTURES

A paper to be submitted to The R Journal

Andee Kaplan, Mark S. Kaiser, Soumendra N. Lahiri, and Daniel J. Nordman

Abstract

For spatial and network data, conditionally specified models can be formulated on the basis of an underlying Markov random field (MRF). This approach often provides an attractive alternative to direct specification of a full joint data distribution, which may be difficult for large correlated data. However, simulation from such MRF models can be challenging, even with relatively small sample sizes. We describe a new Gibbs algorithm for simulating data from MRF models, where the proposed simulation scheme is computationally fast due to its ability to lower the number of steps necessary to run a single Gibbs iteration. We demonstrate use of a flexible R package (called conclique, to appear on CRAN) that implements the proposed (conclique-based) Gibbs sampler and also performs related goodness-of-fit tests for MRF models.

5.1 Introduction

Markov random field (MRF) models are common for spatial data (as well as graph, network and other data structures). Rather than specifying a joint distribution directly, a model is specified through a set of full conditional distributions for the spatial locations. The R package conclique provides a fast way to simulate from a MRF model using a modified Gibbs sampler
and implements a formal goodness-of-fit (GOF) test for model assessment (Kaiser, Lahiri, and Nordman 2012).

Throughout this paper, we will employ the following notational conventions: see also Kaiser and Cressie (2000) for spatial MRF specifications. Spatial locations are denoted \( s_i \), where the set of all locations \( \{s_i : i = 1, \ldots, n\} \) is assumed to lie on a regular lattice, for simplicity. The values of random variables at these locations will be denoted \( \{Y(s_i) : i = 1, \ldots, n\} \) and spatial neighborhoods, \( \mathcal{N}_i \) will be pre-specified for each location \( s_i \), where the conditional distribution of \( Y(s_i) \) shall depend on observations and locations in \( \mathcal{N}_i \subset \{s_j : j \neq i\} \). We then denote the neighboring values as \( y(\mathcal{N}_i) = \{y(s_j) : s_j \in \mathcal{N}_i\} \) and define the full conditional distributions for each spatial location as \( \{f_i(y(s_i)|y(\mathcal{N}_i), \Theta) : i = 1, \ldots, n\} \), where each \( f_i(y(s_i)|y(\mathcal{N}_i), \Theta) \) is conditional a probability mass or probability density function (pmf/pdf) of \( Y(s_i) \) given values for its neighbors \( y(\mathcal{N}_i) \). Lastly, we let \( F_i(y(s_i)|y(\mathcal{N}_i), \Theta) \) be the conditional distribution function (cdf) of \( Y(s_i) \) given values for its neighbors \( y(\mathcal{N}_i) \) and a potential set \( \Theta \) of parameter values. For simplicity, we will assume a common conditional form \( (f_i = f \text{ and } F_i = F) \) for all locations, which occurs often in practice.

In the following sections, we introduce the \texttt{conclique} package through demonstration of simulation from a particular MRF model as well as an associated simulation-based test for assessing GOF of spatial MRF models from Kaiser, Lahiri, and Nordman (2012). Currently \texttt{conclique} can be installed via GitHub using the following \texttt{R} code.

### 5.2 Concliques

Concliques are defined as sets of locations such that no location in the set is a neighbor of any other location in the set (Kaiser, Lahiri, and Nordman 2012) For example, the singleton sets from all locations provide a trivial set of concliques for any neighborhood structure. We seek to create a set of maximally sized concliques, \( \{C_i : i = 1, \ldots, Q\} \), for a given neighborhood specification (a minimal number of concliques needed to cover or partition the locations \( \{s_1, \ldots, s_n\} \) of observations) using the \texttt{conclique} package and assign conclique labels to positions on a grid.
Figure 5.1: Minimal conclique cover for a given lattice of size $20 \times 20$ using a four-nearest neighbors structure wrapped on a torus. This results in 2 concliques.

for easy plotting. Figure 5.1 shows an example of concliques for spatial data on a $20 \times 20$ lattice with a four-nearest neighbor structure, where the four-nearest neighborhood $\mathcal{N}_i$ of a location $s_i$ is illustrated as

```
* 
* $s_i$ *
* 
```

with *’s above as the positions of neighbors. For a four-nearest neighbor model, the set of maximal concliques has $Q = 2$, as indicated in Figure 5.1. For comparison, an eight-nearest neighbor model will yield a set of maximal concliques containing $Q = 4$ concliques.

The conclique package comes with the convenience functions `lattice_4nn_torus`, `min_conclique_cover`, and `assign_concliques`, which create a four-nearest neighbor lattice structure of a given dimension $n \times n$ wrapped on a torus (in this case, $20 \times 20$), generate a minimal conclique cover, and assign the conclique labels to a grid, respectively. With these helper functions, one may set up a framework for our example for generating spatial data on a lattice.
5.2.1 Simulation

By exploiting the conditional independence of concliques, a batch updating Gibbs sampler can be used to simulate spatial values from this model. The algorithm for this sampler is as follows, where $Y^{(j)}(s)$ denotes the value of the observation at location $s$ after iteration $j \geq 0$ of the algorithm:

1. Split locations into $Q$ disjoint concliques, $\bigcup_{i=1}^{Q} \mathcal{C}_i = \{s_1, \ldots, s_n\}$.
2. Initialize the values of $\{Y^{(0)}(s) : s \in \{\mathcal{C}_2, \ldots, \mathcal{C}_Q\}\}$.
3. For $i = 1, \ldots, M$,
   1. Considering all $s_j \in \mathcal{C}_1$, sample $\{Y^{(i)}(s_j) : s_j \in \mathcal{C}_1\}$ by independently drawing $Y^{(i)}(s_j) \sim F(\cdot | Y^{(i-1)}(s), s \in N_j)$
   2. Set $\ell = 2$. Considering all $s_j \in \mathcal{C}_\ell$, sample $\{Y^{(i)}(s_j) : s_j \in \mathcal{C}_\ell\}$ by independently drawing $Y^{(i-1)}(s_j) \sim F(\cdot | Y^{(i)}(s), s \in N_j \cap \mathcal{C}_k$ where $\ell < k$, $\{Y^{(i)}(s), s \in N_j \cap \mathcal{C}_k$ where $\ell > k\}$)
   3. For $Q > 2$, repeat step 2 for each $\ell = 3, \ldots, Q$.

This is a departure from the traditional algorithm for sampling spatial data, which samples each location sequentially. In the sequential Gibbs sampler, one iteration will consist of $n$ steps, whereas with the conclique based Gibbs sampler, one iteration will consist of $Q$ steps (where a block update is performed with simultaneous, independent draws). The goal is to speed up computation by incorporating this batch updating. For demonstration purposes, we simulate data from a four-nearest neighbor MRF model with Gaussian full conditional:

$$f_i(y(s_i)|y(-\mathcal{N}_i), \kappa, \eta, \rho) = \frac{1}{\sqrt{2\pi\rho}} \exp\left(-\frac{[y(s_i) - \mu(s_i)]^2}{2\rho^2}\right), \; y(s_i) \in \mathbb{R}$$

with conditional variance $\rho^2$ and conditional mean

$$\mu(s_i) = \kappa + \eta \sum_{s_j \in \mathcal{N}_i} [y(s_j) - \kappa],$$

where $\eta$ represents a dependence parameter and $\kappa$ represents a location parameter specifying the (unconditional) mean of each observation (Kaiser, Caragea, and Furukawa 2012). As an example, we will simulate from the Gaussian MRF with $\rho^2 = 2$, $\kappa = 20$, $\eta = 0.24$ using both a
conclique-based Gibbs sampler and a sequential Gibbs sampler to compare run time for 10,000 iterations of the sampler generating data on a $20 \times 20$ lattice.

In this example, the conclique-based Gibbs sampler took 2.34 seconds and the sequential-based Gibbs sampler took 43.95 seconds $\approx 0.73$ minutes to simulate 10,000 spatial data sets of size $20 \times 20$ on a 1.7 GHz processor. As the grid size increases, the time savings become more significant. For 10,000 iterations/samples on $75 \times 75$ grid, conclique-based took 15.05 seconds and sequential took $1.076 \times 10^4$ seconds $\approx 2.99$ hours. This is shown in Figure 5.2, which summarizes log times for simulating 10,000 Gaussian MRF data sets on a grid of size $n = m \times m$, for various values of $m$.

The conclique package comes equipped with all the functions necessary to simulate from this Gaussian MRF with a single dependence parameter using both the conclique and sequential approach (as shown above). However, the user can also supply his own model specifications for simulation, see Section 5.4 for more details.
5.3 Goodness of fit test

Beyond simulation of data with Markovian dependence structure, **conclique** also includes functionality for GOF tests of MRF models. The theory for the GOF methodology is presented in Kaiser, Lahiri, and Nordman (2012).

5.3.1 Generalized spatial residuals

We can define a *generalized spatial residual* through substitution of a random variable, $Y(s_i)$ and its neighbors $\{Y(s_j) : s_j \in \mathcal{N}_i\}$, into a corresponding (continuous) conditional cdf:

$$
R(s_i) = F(Y(s_i)|\{Y(s_j) : s_j \in \mathcal{N}_i\}, \theta), \quad i = 1, \ldots, n,
$$

where $F(y|y(\mathcal{N}_i), \theta) \equiv F(y|y(\mathcal{N}_i), \theta)$ denotes the conditional cdf of observation $Y(s_i)$, assumed to have a common form $F$ for all locations. Similar residuals, $R(s_i)$, can be defined for non-continuous $F$ too. It then holds that, within a clique, the generalized spatial residuals are iid Uniform$(0,1)$-distributed if $F(|\cdot|)$ corresponds to the true form of the underlying conditional distribution in a MRF model specification (cf. Kaiser, Lahiri, and Nordman 2012).

The spatial residuals are obtainable using the **conclique** package. For demonstration, we continue with our example from Section 5.2 using spatial data on a $20 \times 20$ grid from a four-nearest neighbor Gaussian MRF with $\rho^2 = 2$, $\kappa = 20$, $\eta = 0.24$. To illustrate the use of generalized spatial residuals as a fit assessment tool, we generate data from the true model and compute residuals using the true model parameters as well as misspecified parameters, say one with $\eta = -0.10$. We compare the empirical cdf of the generalized spatial residuals within each clique, $\{R(s_i) : s_i \in \mathcal{C}_j\}_{j=1}^Q$ to the cdf of a Uniform$(0,1)$, where $Q = 2$ in our example with a four-nearest neighborhood. The results are in Figure 5.3. Under the incorrect model, we can see that, for both for both cliques, all of the residuals are below the Uniform$(0,1)$ cdf, where the departure of uniformity suggests a poor model fit (which, in fact, is by design here to illustrate the behavior of residuals under a misspecified model).
Figure 5.3: By conclique, the empirical cdf of generalized spatial residuals based on data from a Gaussian MRF with $\rho^2 = 2$, $\kappa = 10$, $\eta = 0.24$ when computing residuals from the correct model (left) and an incorrect model with $\eta = -0.10$ (right). The black line represents the cdf of a Uniform(0,1).

### 5.3.2 Test statistics

In order to combine generalized spatial residuals across concliques to form a single test statistic, we look at the empirical cdf of the residuals and its difference to the Uniform(0,1) cdf using some discrepancy measure (like the Kolmogorov-Smirnov statistic) and combine using an aggregation function (like maximum); see also Kaiser, Lahiri, and Nordman (2012).

Both the Kolmogorov-Smirnov statistic and the Cramèr-von-Mises criterion are implemented in conclique, and any user specified aggregation can be supplied; mean and max are reasonable choices. In this example, we consider the generalized spatial residuals as in Figure 5.3, with residuals computed for $Q = 2$ concliques (under both the correct and incorrect conditional Gaussian model with four-nearest neighbors). We use the Kolmogorov-Smirnov (KS) statistic to compare residuals within each conclique to a Uniform(0,1) distribution and create a single test statistic by taking the maximum of KS statistics over two concliques. For illustration, the test statistic is calculated using residuals from both correct and incorrect models, as reported in Table 5.1. We can see that the incorrectly specified model has a higher value
Table 5.1: The test statistic, maximum of KS statistics, calculated using residuals from both correct and incorrect models.

<table>
<thead>
<tr>
<th></th>
<th>True</th>
<th>Incorrect</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1.69</td>
<td>4.29</td>
</tr>
</tbody>
</table>

of the statistic, which points to potentially rejecting this model. However, without a reference distribution, the meaning behind the size of a test statistic becomes impossible to judge. To this end, we have the ability to generate spatial data from a model using our conclique-based Gibbs simulation functions in conclique and thereby simulate (or parametrically bootstrap) a reference distribution for conclique-based test statistics.

5.3.3 Getting a reference distribution

In order to answer the question, “Could the data have plausibly arisen from the fitted model?”, we will employ a GOF test using the framework laid out in Sections 5.3.1 and 5.3.2 and use a parametric bootstrap (implemented via the conclique-based Gibbs sampler from Section 5.2.1) to obtain the reference distribution for our statistic. Prior to using the conclique-based Gibbs sampler, we fit the model using a pseudo-likelihood approach (Besag 1974) and obtained estimates for \( \hat{\eta}, \hat{\kappa}, \) and \( \hat{\rho}^2 \) to the previous data generated from a four-nearest neighbor Gaussian MRF with \( \rho^2 = 2, \kappa = 10, \eta = 0.24 \). These estimates can be found in Table 5.2. From the fitted model, we can then use a convenience function in conclique, called bootstrap_gof to obtain the reference distribution of the GOF statistics that we have chosen (the maximum across concliques of the Kolmogorov-Smirnov statistic). This function will return an approximated p-value associated with the GOB test statistic (i.e., the probability of a larger statistic value), a vector of quantile values from the reference distribution (if desired), and a summary plot of the distribution (if desired). See Figure 5.4 for details. From the reference distribution generated
Figure 5.4: A summary plot of the bootstrapped reference distribution for the maximum across concliques of the Kolmogorov-Smirnov statistic from data generated from a four-nearest neighbor Gaussian MRF with $\rho^2 = 2, \kappa = 10, \eta = 0.24$. We fail to reject the null hypothesis, meaning this is an adequate model for the data.

by the parametric bootstrap within conclique, the test statistic is 1.337 with a p-value of 0.892. In testing the fit of the conditional Gaussian model to the data, we fail to reject the null hypothesis, meaning the four-nearest neighbor Gaussian model is not an obviously inadequate one for these data. We should expect this result because this exercise knowingly fit the correct, or true data-generating, MRF model (Gaussian MRF with one dependence parameter and a four-nearest neighbor structure). For contrast, testing the GOF of this same Gaussian MRF model with four-nearest neighbors to synthetic data generated from a spatial log-normal model gives a p-value of $10^{-4}$ from the same conclique-based GOF procedure, indicating the GOF method detects this departure in distributional class. While we focused on a Gaussian MRF for demonstration, through the use of conclique, many MRF distributional types and forms, including non-Gaussian ones, can potentially be considered in GOF model tests.

5.4 Extending conclique

For simulating from a Gaussian MRF (e.g., with four-nearest neighbor structure), other potential simulation options also exist, such as determining the joint multivariate normal dis-
tribution for the data from specified conditional Gaussian distributions (i.e. finding the spatial
covariance matrix for the observations) and considering direct simulation or by using circulant
embedding; see Cressie (1993) and Rue and Held (2005)]. One of the key advantages to us-
ing conclique-based approaches for simulation, as well as GOF tests, is the ability to consider
non-Gaussian conditional models that go beyond a four-nearest neighbor structure.

5.4.1 Dependence structure

The conclique package comes with a function to create a lattice of specified dimension
with four-nearest neighbor structure wrapped on a torus (lattice_4nn_torus). However, the
user can create his or her own dependence structure by creating an igraph (Csardi and Nepusz
2006) object with each location as a node and the dependence as the edges. The dimension of
the lattice must be kept as an attribute for the object, this is accomplished using

\texttt{igraph::set.graph.attribute(lattice, "dimvector", dimvec)}

where \texttt{dimvec} is a vector storing the dimensions of the lattice, for example \texttt{c(N, N)}.

5.4.2 Conditional distribution

Additionally, the user can specify a different conditional distribution for each spatial lo-
cation (cf. Kaiser and Cressie 2000) for a description of such constructions). In order to
accomplish this, the user must provide a sampler function which takes as parameters

- \textbf{data} A list containing two elements, \texttt{sums} and \texttt{nums}, which contain the sum of the data
  in each neighborhood as well as the number of locations in the neighborhood for each
  point in the conclique. This list also can contain the \texttt{x} and \texttt{y} coordinates of each spatial
  location if desired.
- \textbf{params} A named list of parameter values, that parameterize the conditional distribution.
This function should return a sampled data set of the same size as the original data. For example, the sampler function for the single dependency parameter Gaussian model is reproduced below.

```r
gaussian_single_param_sampler <- function(data, params) {
  rho <- params$rho
  kappa <- params$kappa
  eta <- params$eta
  sums <- data$sums[[1]]
  nums <- data$nums[[1]]

  mean_structure <- kappa + eta*(sums - nums*kappa)
  rnorm(length(mean_structure))*rho + mean_structure
}
```

### 5.4.3 Spatial residuals

In order to extend `conclique` to use an arbitrary conditional distribution for defining the generalized spatial residuals, the user must specify a cdf function which takes, as parameters, the same parameters as the sampler function from Section 5.4.2, `data` and `params`.

For example, the cdf function for the single dependency parameter Gaussian model is reproduced below.

```r
gaussian_single_param_cdf <- function(data, params) {
  rho <- params$rho
  kappa <- params$kappa
  eta <- params$eta
  sums <- data$sums[[1]]
  nums <- data$nums[[1]]
```
mean_structure <- kappa + eta*(sums - nums*kappa)

pnorm(data$data, mean = mean_structure, sd = rho)
}

5.5 Conclusion

conclique is a fast and flexible implementation of a conclique-based method for simulating MRF data and performing associated GOF tests. By employing a conclique-based sampling method, many spatial models become feasible (like Bernoulli, Poisson, truncated Gamma with one or multiple dependence parameters). Through the use of conditional independence within a conclique (Kaiser, Lahiri, and Nordman 2012) and an Rcpp back end (Eddelbuettel and François 2011), this method offers benefits for computational speed in simulating MRF data compared to traditional spatial Gibbs sampling methods, allowing for a broad range of applications and inference using model-based simulation.
APPENDIX A. ON THE ONE-TO-ONE CORRESPONDENCE BETWEEN PARAMETERS AND MOMENTS IN THE RBM DISTRIBUTION

For integers $m,n \geq 1$, consider random vectors $V_m \equiv (V_1, \ldots, V_m)$ and $H_n \equiv (H_1, \ldots, H_n)$, where the random variables $V_i$ and $H_j$ assume values in $\{1, -1\}$, $i = 1, \ldots, m$, $j = 1, \ldots, n$. We suppose the vector $(V_m, H_n)$ follows a restricted Boltzmann machine (RBM) distribution, i.e. has probability mass function

$$P(V_1 = v_1, \ldots, V_m = v_m, H_1 = h_1, \ldots, H_n = h_n | \theta) \propto \exp \left[ \sum_{i=1}^{m} v_i \theta_{v_i} + \sum_{j=1}^{n} h_j \theta_{h_j} + \sum_{i=1}^{m} \sum_{j=1}^{n} v_i h_j \theta_{ij} \right],$$

with $m + n + mn$ real parameters $\theta_{v_i}, \ldots, \theta_{v_m}, \theta_{h_1}, \ldots, \theta_{h_n}$ and $\theta_{ij}$, $i = 1, \ldots, m$, $j = 1, \ldots, n$ collected in a parameter vector denoted as $\theta \in \mathbb{R}^{m+n+mn}$. Denote the distribution of $(V_m, H_n)$ as $\text{RBM}(m, n, \theta)$. In the following, for convenience, we say a symmetric random variable $X$ has a Bernoulli$(1/2)$ distribution if $P(X = 1) = 1/2 = P(X = -1)$.

Use the notation $V_m * H_n = (V_1 H_1, \ldots, V_m H_n, V_1 H_1, \ldots, V_m H_n)$. For the $\text{RBM}(m, n, \theta)$ model, consider in the function $g : \mathbb{R}^{m+n+mn} \to \mathcal{I} \subset \mathbb{R}^{m+n+mn}$ given by

$$g(\theta) = E_{\theta}(V_m, H_n, V_m * H_n)$$

$$= \sum_{v_m \in \{\pm 1\}^m, h_n \in \{\pm 1\}^n} (v_m, h_n, v_m * h_n) P(V_m = v_m, H_n = h_n | \theta)$$

for $\theta \in \mathbb{R}^{m+n+mn}$; above $\mathcal{I} \equiv \{g(\theta) : \theta \in \mathbb{R}^{m+n+mn}\}$ denotes the image of the mapping.

We have the following properties:
The function \( g : \mathbb{R}^{m+n+mn} \to \mathcal{S} \) is one-to-one; see Theorem 4 below. Hence, these moments uniquely characterize the parameters in the RBM\((m,n,\theta)\) model.

3. The image \( \mathcal{S} \) is a symmetric set (i.e., \( g(\theta) = -g(-\theta) \)) and must be a connected (not necessarily convex) region in \( \mathbb{R}^{m+n+mn} \), without any voids in it (by the continuity of \( g(\theta) \)).

4. The center of the image \( \mathcal{S} \) (the zero vector in \( \mathbb{R}^{m+n+mn} \)) is given only by \( \theta = 0 \) (the zero vector in the parameter space). This corresponds to the case where all the random vectors \( V_1, \ldots, V_m, H_1, \ldots, H_n \) are iid with a symmetric Bernoulli(\(1/2\)) distribution; see Theorem 5.

**Theorem 4.** Let \( m,n \geq 1 \) and suppose the random vector \((V_m,H_n) = (V_1, \ldots, V_m, H_1, \ldots, H_n)\) follows a RBM\((m,n,\theta)\) distribution. Let \( \theta_1, \theta_2 \in \mathbb{R}^{m+n+mn} \) and let \( P_1 \) and \( E \) denote probability and expectation under \( \theta_i, i = 1,2 \). Then, the following are equivalent

1. \( E_1 V_i = E_2 V_i, E_1 H_j = E_2 H_j \) and \( E_1 V_i H_j = E_2 V_i H_j \) for any \( i = 1, \ldots, m, j = 1, \ldots, n \).
2. \( \theta_1 = \theta_2 \).
3. \( P_1 \) and \( P_2 \) are the same distribution for \((V_m,H_n)\).

**Proof.** We first establish the equivalence of claims 2 and 3. If claim 2 holds, then claim 3 holds trivially. Suppose claim 3 holds and pick \( \ell \in \{1, \ldots, m\} \). Let \( \theta_{v_1}, \ldots, \theta_{v_m}, \theta_{h_1}, \ldots, \theta_{h_n}, \theta_{i_j}, i = 1, \ldots, m, j = 1, \ldots, n \), denote the components of \( \theta_1 \) and let \( \hat{\theta}_{v_1}, \ldots, \hat{\theta}_{v_m}, \hat{\theta}_{h_1}, \ldots, \hat{\theta}_{h_n}, \hat{\theta}_{i_j}, i = 1, \ldots, m, j = 1, \ldots, n \), denote the analogous components of \( \theta_2 \). Then, by assumption,

\[
\exp \left[ 2\theta_{v_\ell} + 2 \sum_{j=1}^{n} h_j \theta_{i_j} \right] = \frac{P_1 \left( \{ V_\ell = 1 \} \cap \bigcap_{i \neq \ell}^{m} \{ V_i = v_i \} \cap \bigcap_{j=1}^{n} \{ H_j = h_j \} \right)}{P_1 \left( \{ V_\ell = -1 \} \cap \bigcap_{i \neq \ell}^{m} \{ V_i = v_i \} \cap \bigcap_{j=1}^{n} \{ H_j = h_j \} \right)}
\]

\[
= \frac{P_2 \left( \{ V_\ell = 1 \} \cap \bigcap_{i \neq \ell}^{m} \{ V_i = v_i \} \cap \bigcap_{j=1}^{n} \{ H_j = h_j \} \right)}{P_2 \left( \{ V_\ell = -1 \} \cap \bigcap_{i \neq \ell}^{m} \{ V_i = v_i \} \cap \bigcap_{j=1}^{n} \{ H_j = h_j \} \right)} \exp \left[ 2\hat{\theta}_{v_\ell} + 2 \sum_{j=1}^{n} h_j \hat{\theta}_{i_j} \right]
\]

for any \( \{v_i\}_{i=1,i \neq \ell}^{m} \cap \{h_j\}_{j=1}^{n} \subseteq \{\pm 1\} \). Equating exponents, we have \( \theta_{v_\ell} - \hat{\theta}_{v_\ell} = \sum_{j=1}^{n} h_j (\hat{\theta}_{i_j} - \theta_{i_j}) \) for any \( \{h_j\}_{j=1}^{n} \subseteq \{\pm 1\} \). Taking all \( h_j = 1 \) or all \( h_j = -1 \) for \( 1 \leq j \leq n \) yields \( \theta_{v_\ell} = \hat{\theta}_{v_\ell} \). This result further implies \( \hat{\theta}_{i_1} - \theta_{i_1} = -\sum_{j=2}^{n} h_j (\hat{\theta}_{i_j} - \theta_{i_j}) \) for any given \( \{h_j\}_{j=2}^{n} \subseteq \{\pm 1\} \), so that \( \theta_{i_1} = \hat{\theta}_{i_1} \) must hold. Iterating this argument, sequentially for each \( j = 2, \ldots, m \), shows \( \theta_{i_j} = \hat{\theta}_{i_j} \) for any
$1 \leq j \leq n$ in addition to $\theta_{v_i} = \hat{\theta}_{v_i}$. As $\ell \in \{1,\ldots,m\}$ was arbitrary, we have $\theta_{v_i} = \hat{\theta}_{v_i}$ and $\theta_{ij} = \hat{\theta}_{ij}$ for any $1 \leq i \leq m$, $1 \leq j \leq n$.

Now pick $k \in \{1,\ldots,n\}$ so that, by the RBM$(m,n,\theta)$ probability structure as above, we have analogously that
\[
\theta_{hk} + \sum_{i=1}^{m} v_i \theta_{ik} = \hat{\theta}_{hk} + \sum_{i=1}^{m} v_i \hat{\theta}_{ik}
\]
for any $\{v_i\}_{i=1}^{m} \subset \{\pm 1\}$, so that $\theta_{hk} = \hat{\theta}_{hk}$. As $k \in \{1,\ldots,n\}$ was arbitrary, we have now established claim 2 from claim 3.

To establish the equivalence of claims 1 and 3 in Theorem 4 under the RBM$(m,n,\theta)$ model, we note that claim 3 easily implies claim 1. Hence, it suffices now to establish claim 3 from claim 1. This follows immediately from Lemma 1 below and the fact that the variables $V_1,\ldots,V_m$ are conditionally independent given $H_1,\ldots,H_n$ in the RBM$(m,n,\theta)$ model (and likewise $H_1,\ldots,H_n$ are independent given $V_1,\ldots,V_m$).

\begin{lemma}
Let $m,n \geq 1$. Let $P_1,P_2$ denote two probability distributions for a random vector $(V_m,H_n) = (V_1,\ldots,V_m,H_1,\ldots,H_n)$ supported on $\{a,b\}^{m+n}$, for some real constants $a \neq b \in \mathbb{R}$, such that $P_i(V_m = v_m,H_n = h_n) > 0$, $i = 1,2$, for any $v_m \in \{a,b\}^m$, $h_n \in \{a,b\}^n$. Suppose additionally that, under $P_1$ or $P_2$, the variables $V_1,\ldots,V_m$ are conditionally independent given $H_n = h_n \in \{a,b\}^n$ and that the variables $H_1,\ldots,H_n$ are conditionally independent given $V_m = v_m \in \{a,b\}^m$.

Let $E_{i\ell}$ denote expectation under the distribution $P_i$, $i = 1,2$. Then, for any $v_m \in \{a,b\}^m,h_n \in \{a,b\}^n$,
\[
P_1(V_m = v_m,H_n = h_n) = P_2(V_m = v_m,H_n = h_n),
\]
follows if $E_1V_i = E_2V_i$, $E_1H_j = E_2H_j$, $E_1V_iH_j = E_2V_iH_j$ holds for all $i \in \{1,\ldots,m\}$, $j \in \{1,\ldots,n\}$.

\begin{proof}
The case $m = n = 1$ follows from Lemma 2. We now use an induction argument, assuming that the result of Lemma 1 holds for some given order of $(m,n)$ with $m,n \geq 1$ and show the the result continues to hold for $(m+1,n)$ or $(m,n+1)$. We shall treat the case $(m+1,n)$ (where the other case $(m,n+1)$ follows by symmetrical arguments).
Considering random variables \((V_{m+1}, H_n) = (V_1, \ldots, V_m, V_{m+1}, H_1, \ldots, H_n)\), the induction hypothesis applies to \((V_m, H_n)\) where \(V_m = (V_1, \ldots, V_m)\). Hence, for any \(v_m \in \{a, b\}^m, h_n \in \{a, b\}^n\), we have \(P_1(V_m = v_m, H_n = h_n) = P_2(V_m = v_m, H_n = h_n) > 0\), \(P_1(H_n = h_n) = P_2(H_n = h_n) > 0\) and consequently

\[
P_1(V_m = v_m|H_n = h_n) = P_2(V_m = v_m|H_n = h_n).
\] (A.1)

Likewise, by the induction hypothesis applied to \((V_{m-1}, H_n)\) for \(V_{m-1} = (V_2, \ldots, V_{m+1})\), we have

\[
1 - P_1(V_{m+1} = b|H_n = h_n) = P_1(V_{m+1} = a|H_n = h_n)
= P_2(V_{m+1} = a|H_n = h_n)
= 1 - P_2(V_{m+1} = b|H_n = h_n)
\]

for any \(h_n \in \{a, b\}^n\). Now fix \(v_m \in \{a, b\}^m, h_n \in \{a, b\}^n\) and let \(v_{m+1} = (v_{m+1}, v_m)\) for \(v_{m+1} \in \{a, b\}\). Then, by the conditional independence assumption,

\[
P_i(V_{m+1} = v_{m+1}|H_n = h_n) = P_i(V_{m+1} = v_{m+1}|H_n = h_n)P_i(V_m = v_m|H_n = h_n)
\]

for \(i = 1, 2\), so that we have \(P_1(V_{m+1} = v_{m+1}|H_n = h_n) = P_2(V_{m+1} = v_{m+1}|H_n = h_n)\) by ((B.1)). Consequently, it follows that

\[
P_1(V_{m+1} = v_{m+1}, H_n = h_n) = P_1(V_{m+1} = v_{m+1}|H_n = h_n)P_1(H_n = h_n)
= P_2(V_{m+1} = v_{m+1}|H_n = h_n)P_2(H_n = h_n)
= P_2(V_{m+1} = v_{m+1}, H_n = h_n)
\]

by \(P_1(H_n = h_n) = P_2(H_n = h_n)\). As \(v_{m+1} \in \{a, b\}^{m+1}, h_n \in \{a, b\}^n\) were arbitrary, this completes the proof of Lemma 1.

\[\square\]

**Lemma 2.** Suppose the discrete random vector \((X, Y)\) has support \(\{a, b\} \times \{a, b\}\), for some \(a \neq b \in \mathbb{R}\), under two probability distributions \(P_1\) and \(P_2\). Let \(E_i\) denote expectation under the distribution \(P_i, i = 1, 2\). Then, the following are equivalent

1. \(E_1X = E_2X, E_1Y = E_2Y, E_1XY = E_2XY\), where \(E_i\) denotes expectation under \(P_i, i = 1, 2\).
2. \(P_1(X = x, Y = y) = P_2(X = x, Y = y)\) for \(x, y \in \{a, b\}\), i.e., \((X, Y)\) has the same distribution under \(P_1\) and \(P_2\).
Proof. Denote the four probabilities by which the random vector \((X, Y)\) assume pairs \((b, b), (b, a), (a, b), (a, a)\), respectively, as \(c_1, c_2, c_3, 1 - c_1 - c_2 - c_3\) under \(P_1\) and \(d_1, d_2, d_3, 1 - d_1 - d_2 - d_3\) under \(P_2\). The conditions \(E_1X = E_2X, E_1Y = E_2Y, E_1XY = E_2XY\) with \(b \neq a\) imply that

\[
(c_1 + c_2) = (d_1 + d_2), \quad (c_1 + c_3) = (d_1 + d_3) \quad (b - a)c_1 + a(2c_1 + c_2 + c_3) = (b - a)d_1 + a(2d_1 + d_2 + d_3).
\]

As \((2c_1 + c_2 + c_3) = (2d_1 + d_2 + d_3)\) and \(b \neq a\), we conclude that \(c_1 = d_1\), from which it follows that \(c_2 = d_2\) and \(c_3 = d_3\). Hence, claim 1 implies claim 2 in Lemma 2. Claim 2 also trivially implies claim 1.

\[\Box\]

**Theorem 5.** Let \(m, n \geq 1\) and suppose the random vector \((V_m, H_n) = (V_1, \ldots, V_m, H_1, \ldots, H_n)\) follows a \(\text{RBM}(m, n, \theta)\) distribution. Then, the following are equivalent

1. \(EV_i = 0, EH_j = 0\) and \(EV_iH_j = 0\) for any \(i = 1, \ldots, m, j = 1, \ldots, n\).
2. \(\theta = 0 \in \mathbb{R}^{m+n+mn}\).
3. \(V_1, \ldots, V_m, H_1, \ldots, H_n\) are iid Bernoulli(1/2) random variables.

Proof. We first establish the equivalence of claims 2 and 3. If claim 2 holds, then claim 3 follows easily from the resulting uniform cell probabilities: \(P(V_1 = v_1, \ldots, V_m = v_m, H_1 = h_1, \ldots, H_n = h_n|\theta) = 2^{-mn}\) for any \(v_i, h_j \in \{\pm 1\}\), \(i = 1, \ldots, m, j = 1, \ldots, n\).

Now suppose claim 3 holds. Pick \(\ell \in \{1, \ldots, m\}\). Then, by the \(\text{RBM}(m, n, \theta)\) probability structure and the iid Bernoulli assumption, we have the conditional probability

\[
\frac{1}{2} = P \left( V_\ell = v_\ell \left| \bigcap_{i=1, i \neq \ell}^m \{V_i = v_i\} \cap \bigcap_{j=1}^n \{H_j = h_j\} \right. \right)
\]

\[
\equiv \frac{\exp(v_\ell \theta_\ell + v_\ell \sum_{j=1}^n h_j \theta_{ij})}{\exp(\theta_\ell + \sum_{j=1}^n h_j \theta_{ij}) + \exp(-\theta_\ell - \sum_{j=1}^n h_j \theta_{ij})}
\]

for any subsets \(\{v_i\}_{i=1}^m, \{h_j\}_{j=1}^n \subseteq \{\pm 1\}\). Equating conditional probabilities when \(v_\ell = 1\) or when \(v_\ell = -1\), we have \(\theta_\ell = -\sum_{j=1}^m h_j \theta_{ij}\) for any given \(\{h_j\}_{j=1}^n \subseteq \{\pm 1\}\). Taking all \(h_j = 1\) or all \(h_j = -1\) for \(1 \leq j \leq n\) yields \(\theta_\ell = 0\). This result further implies \(\theta_{\ell 1} = -\sum_{j=2}^n h_j \theta_{\ell j}\) for any given \(\{h_j\}_{j=2}^n \subseteq \{\pm 1\}\), so that \(\theta_{\ell 1} = 0\) must hold. Iterating this argument, sequentially for each
\( j = 2, \ldots, m \), shows \( \theta_{kj} = 0 \) for any \( 1 \leq j \leq n \) in addition to \( \theta_{v_l} = 0 \). As \( \ell \in \{1, \ldots, m\} \) was arbitrary, we have \( \theta_{v_i} = 0 \) and \( \theta_{ij} = 0 \) for any \( 1 \leq i \leq m, 1 \leq j \leq n \).

Now pick \( k \in \{1, \ldots, n\} \) so that, by the RBM \((m, n, \theta)\) probability structure and the iid assumption, we have the conditional probability

\[
\frac{1}{2} = P \left( H_k = h_k \left| \bigcap_{j \neq k} \{ H_j = h_j \} \right. \right) = \frac{\exp(h_k \theta_{h_k})}{\exp(\theta_{h_k}) + \exp(-\theta_{h_k})}
\]

for any \( \{h_j\}_{j=1}^n \subset \{\pm 1\} \). Considering \( h_k = 1 \) or \( h_k = -1 \), we conclude that \( \theta_{h_k} = -\theta_{h_k} \) or \( \theta_{h_k} = 0 \). As \( k \in \{1, \ldots, n\} \) was arbitrary, we have now established claim 2 from claim 3.

To establish the equivalence of claims 1 and 2 in Theorem 5 under the RBM \((m, n, \theta)\) model, we note that claim 2 again implies claim 3, which then easily implies claim 1. By Theorem 4, if claim 1 holds then the only possibility for this is \( \theta = 0 \in \mathbb{R}^{m+n+mn} \). \( \square \)
APPENDIX B. FLEXIBILITY OF THE RBM MODEL

In an RBM model with enough hidden variables, parameter values may be chosen to match any given cell probabilities with arbitrary closeness. Additionally, when one or more of the cell probabilities ((A.2)) are zero, the corresponding RBM probabilities may never be identically zero (due to exponential terms in the model) but parameters can be still selected to make the appropriate RBM cell probabilities arbitrarily small. We show this for a model with two visible variables \((V_1, V_2)\) and one hidden \(H_1\).

To demonstrate, we assume \(p_{(-1,-1)} > 0\) (without loss of generality) in the specified cell probabilities ((A.2)) and replace parameters \(\theta_{11}, \theta_{21}\) with \(\Delta_1 \equiv \theta_{11} + \theta_{21}\) and \(\Delta_2 \equiv \theta_{11} - \theta_{21}\). We may then prescribe values of \(\theta_{v_1}, \theta_{v_2}, \theta_{h_1}, \Delta_1, \Delta_2\) so that the model probability ratio

\[
P(V_1 = v_1, V_2 = v_2 | \theta_{v_1}, \theta_{v_2}, \theta_{h_1}, \Delta_1, \Delta_2) / P(V_1 = -1, V_2 = -1 | \theta_{v_1}, \theta_{v_2}, \theta_{h_1}, \Delta_1, \Delta_2)
\]

matches the corresponding ratio \(p(v_1,v_2)/p_{(-1,-1)}\) over three values of \((v_1,v_2) = (-1,-1), (1,1), (1,-1), (-1,1)\).

For instance, assuming the cell probabilities from ((A.2)) are all positive, these probabilities can be exactly reproduced by choosing

\[
\theta_{v_1} = \frac{1}{2} \log \left( \frac{p_{(-1,-1)} \exp(\theta_{h_1} - \Delta_1) + \exp(-\theta_{h_1} + \Delta_1)}{p_{(-1,-1)} \exp(\theta_{h_1} + \Delta_2) + \exp(-\theta_{h_1} - \Delta_2)} \right),
\]

\[
\theta_{v_2} = \frac{1}{2} \log \left( \frac{p_{(-1,-1)} \exp(\theta_{h_1} - \Delta_1) + \exp(-\theta_{h_1} + \Delta_1)}{p_{(-1,-1)} \exp(\theta_{h_1} - \Delta_2) + \exp(-\theta_{h_1} + \Delta_2)} \right)
\]

and selecting \(\theta_{h_1}, \Delta_1, \Delta_2\) to solve

\[
\frac{p_{(1,1)p_{(-1,-1)}}}{p_{(-1,1)p_{(1,-1)}}} = \frac{\ell(|\theta_{h_1}|) + \ell(|\Delta_1|)}{\ell(|\theta_{h_1}|) + \ell(|\Delta_2|)},
\]

(B.1)

based on a monotonically increasing function \(\ell(x) \equiv \exp(-2x) + \exp(2x), \ x \geq 0\). If \([p_{(1,1)}p_{(-1,-1)}]/[p_{(-1,1)}p_{(1,-1)}] \geq 1\), one can pick any values for \(\theta_{h_1}, \Delta_2 \in \mathbb{R}\) and solve ((B.1)) for
\[ |\Delta_1|; \text{ likewise, when } \frac{[p_{1,1}p_{(-1,-1)}]}{[p_{(-1,1)}p_{(1,-1)}]} < 1 \text{ in } ((B.1))) \text{, one may solve for } |\Delta_2| \text{ upon choosing any values for } \theta_{h_1}, \Delta_1 \in \mathbb{R}. \]

Alternatively, if exactly one specified cell probability in \(((A.2))\) is zero, say \(p_{(1,1)}\) (without loss of generality), we can select parameters \(\theta_{h_1}, \theta_{v_2}\) as above based on a sequence \((\theta_{h_1}, \Delta_1, \Delta_2) \equiv (\theta_{h_1}^{(m)}, \Delta_1^{(m)}, \Delta_2^{(m)}), m \in \{1, 2, \ldots, \} \) of the remaining parameter values such that \(\lim_{m \to \infty} |\Delta_1^{(m)}| = \infty\)

and \(\lim_{m \to \infty}(|\theta_{h_1}^{(m)}| + |\Delta_2^{(m)}|)/|\Delta_1^{(m)}| = 0\) hold. This guarantees that the resulting RBM model matches the given cell probabilities \(((A.2))\) in the limit:

\[ \lim_{m \to \infty} P(V_1 = v_1, V_2 = v_2 | \theta_{v_1}, \theta_{v_2}, \theta_{h_1}, \Delta_1, \Delta_2) = p_{(v_1,v_2)}, (v_1,v_2) \in \{(\pm 1, \pm 1)\}. \quad (B.2) \]

If exactly two specified probabilities in \(((A.2))\) are zero, say \(p_{(1,1)}\) and \(p_{(-1,1)}\) (without loss of generality), then a limit approximation as in \(((B.2))\) follows by picking \(\theta_{v_1}\) as above based on any choices of \((\theta_{h_1}, \Delta_1, \Delta_2)\) and choosing a sequence of \(\theta_{v_2} \equiv \theta_{v_2}^{(m)}\) values for which \(\theta_{v_2}^{(m)} \to -\infty.\)
APPENDIX C. PROOFS OF INSTABILITY RESULTS

Proof of Proposition 3. We prove the contrapositive, supposing that $\Delta(\theta_N) \leq C$ holds for some $C > 0$ and show $\text{ELPR}(\theta_N) \leq NC$. Let $x_{\min} \equiv \arg\min_{x \in \mathcal{X}^N} P_{\theta_N}(x)$ and $x_{\max} \equiv \arg\max_{x \in \mathcal{X}^N} P_{\theta_N}(x)$. Note there exists a sequence $x_{\min} \equiv x_0, x_1, \ldots, x_k \equiv x_{\max}$ in $\mathcal{X}^N$ of component-wise switches to move from $x_{\min}$ to $x_{\max}$ in the sample space (i.e. $x_i, x_{i+1} \in \mathcal{X}^N$ differ in exactly 1 component for $i = 0, \ldots, k$) for some integer $k \in \{0, 1, \ldots, N\}$. Under the FSFS model, recall $P_{\theta_N}(x) > 0$ holds so that $\log P_{\theta_N}(x)$ is well-defined for each outcome $x \in \mathcal{X}^N$. Then, if $k > 0$, it follows that

$$\text{ELPR}(\theta_N) = \log \left[ \frac{P_{\theta_N}(x_{\max})}{P_{\theta_N}(x_{\min})} \right] = \sum_{i=1}^{k} \log \left( \frac{P_{\theta_N}(x_i)}{P_{\theta_N}(x_{i-1})} \right) \leq \sum_{i=1}^{k} \left| \log \left( \frac{P_{\theta_N}(x_i)}{P_{\theta_N}(x_{i-1})} \right) \right| \leq k \Delta_N(\theta_N) \leq NC,$$

using $k \leq N$ and $\Delta(\theta_N) \leq C$. If $k = 0$, then $x_{\max} = x_{\min}$ and the same bound above holds.

Proof of Proposition 4. where $|\mathcal{X}| < \infty$ holds in the FSFS model. We may suppose $|\mathcal{X}| > 1$ (i.e., $\mathcal{X}^N$ has more than one outcome) because otherwise the model is trivially degenerate for all $N \geq 1$. Fix $0 < \varepsilon < 1$. Then, $x_{\max} \in M_{\varepsilon, \theta_N}$, so $P_{\theta_N}(M_{\varepsilon, \theta_N}) \geq P_{\theta_N}(x_{\max}) > 0$. If $x \in \mathcal{X}^N \setminus M_{\varepsilon, \theta_N}$, then by definition $P_{\theta_N}(x) \leq [P_{\theta_N}(x_{\max})]^{1-\varepsilon}[P_{\theta_N}(x_{\min})]^\varepsilon$ holds so that

$$1 - P_{\theta_N}(M_{\varepsilon, \theta_N}) = \sum_{x \in \mathcal{X}^N \setminus M_{\varepsilon, \theta_N}} P_{\theta_N}(x) \leq (|\mathcal{X}|^N)[P_{\theta_N}(x_{\max})]^{1-\varepsilon}[P_{\theta_N}(x_{\min})]^{\varepsilon}.$$

From the lower bound on $P_{\theta_N}(M_{\varepsilon, \theta_N})$ and the upper bound on $1 - P_{\theta_N}(M_{\varepsilon, \theta_N})$, it follows that

$$\frac{1}{N} \log \left[ \frac{P_{\theta_N}(M_{\varepsilon, \theta_N})}{1 - P_{\theta_N}(M_{\varepsilon, \theta_N})} \right] \geq \frac{1}{N} \log \left[ \frac{P_{\theta_N}(x_{\max})}{(|\mathcal{X}|^N)[P_{\theta_N}(x_{\max})]^{1-\varepsilon}[P_{\theta_N}(x_{\min})]^{\varepsilon}} \right] = \frac{\varepsilon}{N} \log \left[ \frac{P_{\theta_N}(x_{\max})}{P_{\theta_N}(x_{\min})} \right] - \log |\mathcal{X}| \to \infty$$
as \( N \to \infty \) by the definition of an unstable FSFS model (cf. Definition 2). Consequently, 
\[
P_{\mathbf{g}_N}(M_{\mathbf{g}_N}) \to 1 \quad \text{as} \quad N \to \infty
\]
as claimed. \( \square \)
APPENDIX D. ADDITIONAL CONCLIQUE-BASED GIBBS SAMPLING STRATEGIES

The random sequence scan (RQGS) updates all concliques in a randomly selected order according to a fixed permutation probability. There are $Q!$ possible update orders. Let the $i^{th}$ update order be denoted $i(1),\ldots,i(Q)$ and $q = \{q_1,\ldots,q_Q\}$ be the permutation probabilities such that $\sum_{i=1}^{Q} q_i = 1$. The random scan (RSGS) randomly updates one conclique in each iteration while fixing the others according to a fixed component selection probability $p = \{p_1,\ldots,p_Q\}$ where $p_i > 0$ and $\sum_{i=1}^{Q} p_i = 1$ (Johnson and Burbank 2015). The algorithms are as follows.

Conclique-based RQGS Algorithm, Let $M \geq 1$ denote the number of complete Gibbs iterations:

1. Split locations into $Q \geq 2$ disjoint concliques, $\mathcal{C}_1,\ldots,\mathcal{C}_Q$.

2. Initialize the values of $\{Y^{(0)}(s) : s \in \mathcal{C}_1,\ldots,\mathcal{C}_Q\}$.

3. Draw a permutation of indices $\alpha$ according to the fixed permutation probabilities, $q = \{q_1,\ldots,q_Q\}$.

4. For iteration $i = 1,\ldots,M$,

   (a) Considering all locations $s_j \in \mathcal{C}_{\alpha(1)}$, sample $\{Y^{(i)}(s_j) : s_j \in \mathcal{C}_1\}$ by independently drawing $Y^{(i)}(s_j) \sim f_j(\cdot|\{Y^{(i-1)}(s) : s \in \mathcal{N}_j\})$ from conditionals in (4.1).

   (b) Set $\ell = 2$. Considering all locations $s_j \in \mathcal{C}_{\ell}$, sample $\{Y^{(i)}(s_j) : s_j \in \mathcal{C}_\ell\}$ by independently drawing $Y^{(i)}(s_j) \sim f_j(\cdot|Y^{(i)}_{\alpha(\ell)}(\mathcal{N}_j))$ with conditioning observations $Y^{(i)}_{\alpha(\ell)}(\mathcal{N}_j) \equiv$
\[ \bigcup_{k=1}^{\alpha(l)-1} \{ Y^{(i)}(s) : s \in \mathcal{N}_j \cap \mathcal{C}_k \} \cup \bigcup_{k=\alpha(l)+1}^{Q} \{ Y^{(i-1)}(s) : s \in \mathcal{N}_j \cap \mathcal{C}_k \}, \]
where the second set union is defined as empty if \( \alpha(l) = Q \).

(c) For \( Q > 2 \), repeat step 2 for each \( \ell = 3, \ldots, Q \).

**Conclique-based RSGS Algorithm.** Let \( M \geq 1 \) denote the number of complete Gibbs iterations:

1. Split locations into \( Q \geq 2 \) disjoint concliques, \( \mathcal{C}_1, \ldots, \mathcal{C}_Q \).
2. Initialize the values of \( \{ Y^{(0)}(s) : s \in \{ \mathcal{C}_1, \ldots, \mathcal{C}_Q \} \} \).
3. Draw an index \( h \) according to the fixed component selection probability \( p = \{ p_1, \ldots, p_Q \} \).
4. For iteration \( i = 1, \ldots, M \),
   
   (a) Considering all locations \( s_j \in \mathcal{C}_h \), sample \( \{ Y^{(i)}(s_j) : s_j \in \mathcal{C}_h \} \) by independently drawing
       
       \[ Y^{(i)}(s_j) \sim f_j(\cdot | \{ Y^{(i-1)}(s) : s \in \mathcal{N}_j \}) \text{ from conditionals in (4.1).} \]
   
   (b) For each \( s \in \mathcal{C}_j \) where \( j \neq h \), \( Y^{(i)}(s) = Y^{(i-1)}(s) \).

Each of the three Gibbs sampling techniques (including CGS) has a corresponding transition density \( k(x, y) \), and one-step transition distribution \( P(x, A) = \int_A k(x, y) \mu(y) \) based on an initialization \( x \in \mathcal{X} \). These densities are given in Table D.1.

Table D.1: The transition densities for each of the three Gibbs sampling techniques, CGS, RQGS, and RSGS, where \( I(\cdot) \) is the indicator function and \( x_{-i} = \{ x_j, j \neq i \} \).

<table>
<thead>
<tr>
<th>Sampling Technique</th>
<th>Transition Density</th>
</tr>
</thead>
<tbody>
<tr>
<td>CGS</td>
<td>( k_{CGS}(x, y) = f(y_1</td>
</tr>
<tr>
<td>RQGS</td>
<td>( k_{RQGS}(x, y) = \sum_{i=1}^{Q} \sum_{i=1}^{Q} q_i f(y_{i(1)}</td>
</tr>
<tr>
<td>RSGS</td>
<td>( k_{RSGS}(x, y) = \sum_{i=1}^{Q} p_i f(y_j</td>
</tr>
</tbody>
</table>
Lemma 3. Assuming that the full conditionals for the MRF model specify a valid joint distribution, all three conclique-based Gibbs samplers (CGS, RQGS, and RSGS) yield the joint distribution $\Pi(\cdot)$ of $(Y(s_1), \ldots, Y(s_n))$ as the invariant distribution.

Proof. Let $y = (y_1, \ldots, y_Q)$ and $x = (x_1, \ldots, x_Q)$ with $x_i, y_i \in \mathbb{R}^{n_i}$ for integers $l_1, \ldots, l_Q$. Let $l(j) = \sum_{i=1}^{Q} l_i$. Let $\pi(x)$ denote the joint density of $\Pi(\cdot)$ with respect to a dominating measure $\mu$. Then, the one-step transition kernel in the Gibbs sampler has a density $k_\sigma(y, x), \sigma \in \{\text{CGS, RQGS, RSGS}\}$ as specified in Table D.1 with respect to the dominating measure. Pick or fix $x \in \mathcal{X} \subset \mathbb{R}^{n}$ in the joint support of $\pi(\cdot)$ (i.e. $\pi(x) > 0$). Let $S$ be a nonempty subset of $\{1, 2, \ldots, Q\}$ and, for $x = (x_1, \ldots, x_Q)$ as above, let $x_S = \{x_i : 1 \leq i \leq Q, i \in S\}$ and $x_{-S} = \{x_i : 1 \leq i \leq Q, i \notin S\}$. For $i = 1, \ldots, Q$, write $\pi(x_i | \cdot)$ to denote the conditional density for conclique $i$ values given values (·) of the other conclique observations and write $\pi(x_S)$ to denote the marginal density of observations belonging to a conclique indexed by $S \subset \{1, \ldots, Q\}$.

Then, for the CGS strategy, recall the transition density $k_{\text{CGS}}(y, x) = \prod_{i=1}^{Q} \pi(x_i | x_{(1, \ldots, i-1), y_{(1, \ldots, i)}})$ (where for $i = 1$, we have $\pi(x_1 | y_{-1})$ notationally as well as $\pi(x_Q | x_{(1, \ldots, Q-1)} = \pi(x_Q | x_{-Q})$ for $i = Q$).
We have

\[ \int \pi(y) k_{CGS}(y, x) d\mu(y) \]

\[ = \int_{\mathbb{R}^{(1)}} \pi(y) \pi(x_1|y_{-1}) \pi(x_2|x_1, y_{-1(2)}) \cdots \pi(x_Q|x_{Q-1}) d\mu(y) \]

\[ = \pi(x_Q|x_{-Q}) \int_{\mathbb{R}^{(1)}} \pi(y_{-1}) \pi(x_1|y_{-1}) \pi(x_2|x_1, y_{-1(2)}) \cdots \pi(x_{Q-1}|x_{-(Q-1, Q)}, y_Q) d\mu(y_{-1}) \]

\[ = \pi(x_Q|x_{-Q}) \int_{\mathbb{R}^{(2)}} \pi(y_{-1}, x_1) \pi(x_2|x_1, y_{-1(2)}) \cdots \pi(x_{Q-1}|x_{-(Q-1, Q)}, y_Q) d\mu(y_{-1}) \]

\[ = \pi(x_Q|x_{-Q}) \int_{\mathbb{R}^{(3)}} \pi(y_{-(1, 2)}, x_1) \pi(x_2|x_1, y_{-(1, 2)}) \cdots \pi(x_{Q-1}|x_{-(Q-1, Q)}, y_Q) d\mu(y_{-(1, 2)}) \]

\[ \vdots \]

\[ = \pi(x_Q|x_{-Q}) \int_{\mathbb{R}^{(Q)}} \pi(y_Q, x_{-Q}) d\mu(y_Q) \]

\[ = \pi(x_Q|x_{-Q}) \pi(x_{-Q}) \]

\[ = \pi(x), \]

establishing the result. Note that we technically assumed \( \mathbb{R}^n = \mathcal{X} \), i.e. \( \pi(y) > 0 \) for \( y \in \mathbb{R}^n \) above for simplicity, but without loss of generality (as the same follows by partitioning \( \mathbb{R}^n = \mathcal{X} \cup \mathbb{R}^n \setminus \mathcal{X} \) if necessary and partitioning \( \mathbb{R}^{(j)} \) along \( \{ x_{-(1, \ldots, j-1)} : \pi(x_{-(1, \ldots, j-1)}) > 0 \} \) for \( j = 1, \ldots, Q \) generally).
Likewise, for the RQGS strategy, it holds that

\[
\int \pi(y) k_{\text{RQGS}}(y, x) d\mu(y)
\]

\[
= \int \pi(y) \sum_{i=1}^{Q} q_i \pi(x_{i(1)} | y_{-i(1)}) \pi(x_{i(2)} | x_{i(1)}, y_{-(i(1), i(2))}) \cdots \pi(x_{i(Q)} | x_{-i(Q)}) d\mu(y)
\]

\[
= \sum_{i=1}^{Q} q_i \pi(x_{i(Q)} | x_{-i(Q)}) \int \pi(y) \pi(x_{i(1)} | y_{-i(1)}) \pi(x_{i(2)} | x_{i(1)}, y_{-(i(1), i(2))}) \cdots \pi(x_{i(Q-1)} | x_{-(i(Q-1), i(Q))}, y_{i(Q)}) d\mu(y)
\]

\[
= \sum_{i=1}^{Q} q_i \pi(x_{i(Q)} | x_{-i(Q)}) \int \pi(y_{i(Q)}, x_{-i(Q)}) d\mu(y_{i(Q)})
\]

\[
= \sum_{i=1}^{Q} q_i \pi(x_{i(Q)} | x_{-i(Q)}) \pi(x_{-i(Q)})
\]

\[
= \pi(x) \sum_{i=1}^{Q} q_i
\]

\[
= \pi(x),
\]

while for the RSGS strategy, by marginalizing over \( y_i \) for each \( i = 1, \ldots, Q \),

\[
\int \pi(y) k_{\text{RSGS}}(y, x) d\mu(y) = \int \pi(y) \sum_{i=1}^{Q} p_i \pi(x_i | y_{-i}) \mathbb{I}(x_{-i} = y_{-i}) d\mu(y)
\]

\[
= \sum_{i=1}^{Q} p_i \int \pi(y) \pi(x_i | y_{-i}) \mathbb{I}(x_{-i} = y_{-i}) d\mu(y)
\]

\[
= \sum_{i=1}^{Q} p_i \int \pi(y_{-i}) \pi(x_i | y_{-i}) \mathbb{I}(x_{-i} = y_{-i}) d\mu(y_{-i})
\]

\[
= \sum_{i=1}^{Q} p_i \int \pi(x_i, y_{-i}) \mathbb{I}(x_{-i} = y_{-i}) d\mu(y_{-i})
\]

\[
= \sum_{i=1}^{Q} p_i \pi(x_i, x_{-i})
\]

\[
= \pi(x)
\]

Thus all three sampling strategies yield the joint density, \( \pi(\cdot) \) as their invariant distribution. \( \square \)
Recall that a Markov chain is *Harris ergodic* if it is \( \phi \)-irreducible, aperiodic, Harris recurrent, and possesses invariant distribution \( \Pi \) for some measures \( \phi \) and \( \Pi \).

To prove Theorem 1, we employ the following Lemma based on Johnson (2009).

**Lemma 4** (Johnson, 2009). Let \( P \) denote the one-step transition kernel of a \( d \)-component Gibbs sampler. Assume \( P(x, \cdot) \) is absolutely continuous with respect to invariant distribution \( \pi \). Also, for CGS and RQGS, suppose \( P(x, A) > 0 \) for any \( x \in \mathcal{X} \) and \( A \in \mathcal{F} \) for which \( \Pi(A) > 0 \). On the other hand, suppose the \( d \)-step RSGS transition kernel \( P^d(x, A) > 0 \). Then the Gibbs sampler is Harris ergodic. The \( m \)-step transition kernel \( P^m(x, \cdot) \) converges to \( \Pi(\cdot) \) in total variation, i.e. as \( m \to \infty \),

\[
\sup_{A \in \mathcal{F}} |P^m(x, A) - \Pi(A)| \downarrow 0 \text{ as } m \to \infty.
\]

**Proof of Theorem 1.** By Lemma 3, it suffices to show that the three Gibbs samplers (CGS, RQGS, and RSGS) are Harris ergodic by applying Lemma 4.

Note that all three transition densities, \( k_{CGS}(\cdot, \cdot) \), \( k_{RQGS}(\cdot, \cdot) \), and \( k_{RSGS}(\cdot, \cdot) \) are positive on the support \( \mathcal{X} \subset \mathbb{R}^n \) of \( \pi(\cdot) \) by Condition 1.

Let \( A \in \mathcal{F} \) (where \( \mathcal{F} \) is the \( \sigma \)-algebra associated with \( \mathcal{X} \)) be such that \( \Pi(A) = 0 \). Then, by definition \( \Pi(A) = \int_A \pi(x) d\mu(x) = 0 \) where \( \mu \) is the dominating measure on \( \mathcal{X} \). Now, since the invariant distribution \( \pi \) is positive on \( A \subset \mathcal{X} \), this implies \( \mu(A) = 0 \). Therefore \( P_\sigma(x, A) = \int_A k_\sigma(y, x) d\mu(y) = 0 \) where \( \sigma \in \{\text{CGS, RQGS, RSGS}\} \). Thus, \( P(x, \cdot) \) is absolutely continuous with respect to invariant distribution \( \Pi \) for each of the sampling strategies.

Now, let \( A \subset \mathcal{F} \) (i.e. \( A \subset \mathcal{X} \)) be such that \( \Pi(A) > 0 \). Then \( \Pi(A) = \int_A \pi(x) d\mu(x) > 0 \), implying \( \mu(A) > 0 \) must hold by the positivity of \( \pi(\cdot) \) on \( A \). Since \( k_{CGS}(\cdot, \cdot) \), and \( k_{RQGS}(\cdot, \cdot) \) are positive on \( \mathcal{X} \), this implies \( P_{CGS}(x, A) > 0 \) and \( P_{RQGS}(x, A) > 0 \) hold for any \( x \in \mathcal{X} \). Finally, for \( d = Q \), the \( d \)-step transition kernel for RSGS is defined as follows:

\[
P_{RSGS}^Q(x, A) = P(X^{(i+Q)} \in A | X^{(i)} = x) = \int_A k_{RSGS}^Q(x, y) d\mu(y), A \in \mathcal{F},
\]
where
\[ k^{Q}_{\text{RSGS}}(x,y) = \int_{X^{(i)}} k_{\text{RSGS}}(x,z)k^{(Q-1)}_{\text{RSGS}}(z,y)d\mu(z). \]

We will proceed by induction to show \( P^{k}_{\text{RSGS}}(x,A) > 0 \) holds for any \( A \in \mathcal{F} \) (\( A \subset \mathcal{X} \)), \( x \in \mathcal{X} \), and \( k \geq 1 \). For \( Q = 1 \), pick and fix \( A \in \mathcal{F} \), \( x \in \mathcal{X} \). Then for \( k = 1 \), it holds that
\[ P^{1}_{\text{RSGS}}(x,A) = P(X^{(i+1)} \in A | X^{(i)} = x) = \int_{A} k_{\text{RSGS}}(x,y)d\mu(y) > 0 \]
due to the fact that \( k_{\text{RSGS}}(\cdot,\cdot) \) is positive on \( \mathcal{X} \). Now assume \( P^{k-1}_{\text{RSGS}}(x,A) > 0 \) holds for some integer \( k-1 \) and any \( A \in \mathcal{F} \) and \( x \in \mathcal{X} \). Then by the Fubini-Tonelli Theorem and the definition of \( k^{k}_{\text{RSGS}}(x,y) \) in terms of \( k^{k-1}_{\text{RSGS}}(x,y) \), we have
\[ P^{k}_{\text{RSGS}}(x,A) = \int_{A} k^{k}_{\text{RSGS}}(x,y)d\mu(y) \]
\[ = \int_{A} \int_{\mathcal{X}} k_{\text{RSGS}}(x,z)k^{(k-1)}_{\text{RSGS}}(z,y)d\mu(z)d\mu(y) \]
\[ = \int_{\mathcal{X}} k_{\text{RSGS}}(x,z) \int_{A} k^{(k-1)}_{\text{RSGS}}(z,y)d\mu(y)d\mu(z) \]
\[ = \int_{\mathcal{X}} k_{\text{RSGS}}(x,z)P^{k-1}_{\text{RSGS}}(z,A)d\mu(z) > 0. \]

Again, due to the fact that \( k_{\text{RSGS}}(\cdot,\cdot) \) is positive on \( \mathcal{X} \), as well as the induction assumption that \( P^{k-1}_{\text{RSGS}}(x,A) > 0 \) for some integer any \( A \in \mathcal{F} \), \( x \in \mathcal{X} \).

Thus, by Lemma 4 all three Gibbs sampling strategies are Harris ergodic. \( \square \)

**Proof of Theorem 2.** Recall that a Markov chain on a space \( \mathcal{X} \) is *geometrically ergodic* if there exists some function \( G : \mathcal{X} \to \mathbb{R} \) and some constant \( t \in (0,1) \) that satisfy
\[ \| P^n(x, \cdot) - \pi(\cdot) \| \leq G(x)t^n \quad \text{for any } x \in \mathcal{X}. \]

We shall use the following Lemma 5 to establish Theorem 2.

**Lemma 5** (Johnson and Burbank, 2015). Suppose a 2 component Gibbs sampler is Harris ergodic and for all \((y_1,y_2),(y_{1n},y_{2n})\) \( \in \mathcal{Y}_1 \times \mathcal{Y}_2 \) such that \((y_{1n},y_{2n}) \to (y_1,y_2)\),
\[ \pi(y_2 | \liminf_{n \to \infty} y_{1n}) \leq \liminf_{n \to \infty} \pi(y_2 | y_{1n}) \quad \text{and} \quad \pi(y_1 | \liminf_{n \to \infty} y_{2n}) \leq \liminf_{n \to \infty} \pi(y_1 | y_{2n}) \]
holds, where \( \pi(\cdot|\cdot) \) denotes the conditional densities for each of the two components of the sampler. Also suppose that there exist functions \( f : Y_1 \rightarrow [1, \infty) \) and \( g : Y_2 \rightarrow [1, \infty) \) and constants \( j, k, u, v > 0 \) such that \( ju < 1 \) and

\[
E[f(Y_1)|y_2] \leq jg(y_2) + k \quad \text{and} \quad E[g(Y_2)|y_1] \leq uf(y_1) + v. \tag{E.1}
\]

If \( C_d = \{y_2 : g(y_2) \leq d\} \) is compact for all \( d > 0 \), then the two component Gibbs sampler is geometrically ergodic.

Under the assumptions, Theorem 1 yields that all three Gibbs samplers are Harris ergodic with stationary distribution given by the full joint. We next apply Lemma 5 noting that we are assuming that \( Q = 2 \) concliques are available under the four-nearest neighborhood structure for the MRF model with lattice data. That is, in the notation of Lemma 5, we have a two component Gibbs sampler for \( (Y_1, Y_2) \) with components \( Y_1 = \{Y(s_i) : s_i \in \mathcal{C}_1\} \) and \( Y_2 = \{Y(s_i) : s_i \in \mathcal{C}_2\} \) defined by dividing observations \( (Y(s_1), \ldots, Y(s_n)) \) into \( Q = 2 \) concliques. By Theorem 2 assumptions, the full conditionals \( f_i(y(s_i)|y(\mathcal{N}_i)) \) from (4.1) are continuous in \( y(\mathcal{N}_i) \). Suppose the locations in conclique 1 may be written as \( s_{i_1}, \ldots, s_{i_\ell} \) for \( \{i_1, \ldots, i_\ell\} \subset \{1, \ldots, n\} \) and \( 1 \leq \ell \leq n \). Then the transition density \( \pi(y_1|y_2) \) of \( Y_1 \) (conclique 1 values) given \( Y_2 \) (conclique 2 values) may be written as \( \pi(y_1|y_2) = \prod_{j=1}^{\ell} f_{i_j}(y(s_{i_j})|y(\mathcal{N}_{i_j})) \) where by the Markov property, \( f_{i_j}(y(s_{i_j})|y(\mathcal{N}_{i_j})) = f_{i_j}(y(s_{i_j})|y(\mathcal{N}_{i_j})) \) holds as \( y(\mathcal{N}_{i_j}) \subset Y_2 \) for \( j = 1, \ldots, \ell \). Since each full conditional density \( f_{i_j}(y(s_{i_j})|y(\mathcal{N}_{i_j})) \) is continuous in \( y_2 \), the transition density \( \pi(y_1|y_2) \) is continuous in \( y_2 \) so that if \( (y_{1n}, y_{2n}) \rightarrow (y_1, y_2) \), then \( \pi\left(y_1|\lim_{n \rightarrow \infty}y_{2n}\right) = \pi(y_1|y_2) = \lim_{n \rightarrow \infty} \pi(y_1|y_{2n}) \) holds. The same argument holds upon switching the conditioning roles of \( Y_1 \) and \( Y_2 \) (conclique 1 and conclique 2). Thus, by Lemma 5, Theorem 2 will follow by establishing (E.1) holds with observations \( Y_1 \) and \( Y_2 \) from conclique 1 and 2, respectively.

To this end, define \( f(y_1) = 1 \) and \( g(y_2) = 1 \) for \( y_1 \in \mathcal{X}_1 \) and \( y_2 \in \mathcal{X}_2 \) where \( \mathcal{X}_i \) is the support of observations in \( \mathcal{C}_i \) for \( i = 1, 2 \). Without loss of generality, suppose \( \mathcal{X}_2 \) is compact under Theorem 2 assumptions. Then it holds that

\[
E(f(Y_1)|y_2) = 1 \leq j + 1 = jg(y_2) + k \quad \text{and} \quad E(g(Y_2)|y_1) = 1 \leq u + 1 = uf(y_1) + v
\]
for any constants \(j, u > 0\) such that \(j u < 1\) with \(k = 1\) and \(v = 1\). This verifies (E.1). Finally, let \(d > 0\). Then \(C_d = \{y_2 \in \mathcal{Y}_2 : g(y_2) \leq d\} \subset \mathcal{Y}_2\), where the latter set is compact. Thus, \(C_d\) is compact and Lemma 5 holds for any two component conclique based Gibbs sampler among CGS, RQGS, and RSGS. Therefore, CGS, RQGS, and RSGS samplers are geometrically ergodic. \(\square\)

**Proof of Theorem 3.**

**Gaussian:**

Let \(Y(s_i)\) be conditionally Gaussian distributed given \(Y(\mathcal{M}_i)\) using a four-nearest neighbor structure and having expected values \(\{\mu(s_i) : i = 1, \ldots, n\}\) and constant conditional variance \(\tau^2\) where

\[
\mu(s_i) = \alpha + \eta \sum_{s_j \in \mathcal{M}_i} \{y(s_j) - \alpha\}.
\]

Then, this model is of the form specified in (4.6) with

\[
A_{1i}(y(\mathcal{M}_i)) = \frac{1}{2\tau^2} \mu(s_i), \quad T_i(y(s_i)) = y(s_i), \quad B_i(y(\mathcal{M}_i)) = -2\tau^2 A_{1i}(y(\mathcal{M}_i))^2.
\]

This model specifies a valid joint distribution for \(|\eta| < 0.25\) (Cressie 1993) and from Theorem 1 we have that the conclique-based Gibbs strategies yield a Harris ergodic sampler for this model (as there are 2 concliques). Since the support for \(Y(s_i) | Y(\mathcal{M}_i)\) is not compact, we cannot use Theorem 2 to show geometric ergodicity. However by the proof of Theorem 2, it suffices to establish (E.1) and the compactness of \(C_d = \{y_2 \in g(y_2) \leq d\}\) for \(d > 0\) where, as in the proof of Theorem 2, \(Y_1 \in \mathcal{X}_1\) and \(Y_2 \in \mathcal{X}_2\) denote the observations in conclique \(\mathcal{C}_1\) and \(\mathcal{C}_2\), respectively, and \(f(\cdot)\) and \(g(\cdot)\) denote functions of \(Y_1\) and \(Y_2\) in (E.1). Here \(\mathcal{X}_i = \mathbb{R}^{m_i}\) holds where \(m_i\) denotes the number of observations (locations) in \(\mathcal{C}_i, i = 1, 2\).

Define

\[
f(y_1) = (y_1 - a, y_1 - a) + 1, y_1 \in \mathbb{R}^{m_1}\quad\text{and}\quad g(y_2) = (y_2 - b, y_2 - b) + 1, y_2 \in \mathbb{R}^{m_2},
\]

where \(\langle \cdot, \cdot \rangle\) denotes the vector dot product and

\[
a = \frac{1}{4} \frac{(1 + 4\eta)\alpha \eta}{\alpha - \eta^2} \quad\text{and}\quad b = \frac{1}{4} \frac{(1 + 4\eta)\alpha \eta}{\alpha - \eta^2}.
\]

Now, let \(c\) denote a generic constant that does not depend on \(y_2\) and define \(\mu_1 = E(Y_1 | Y_2)\) as a vector of conditional means under the Gaussian model corresponding to the locations in \(\mathcal{C}_1\).
In what follows, without loss of generality, we assume the \( m_1 \) locations in conclique \( C_1 \) are \( \mathcal{C}_1 = \{s_1, \ldots, s_{m_1}\} \) for simplicity. Then for \( y_2 \in \mathbb{R}^{m_2} \), we have

\[
E(f(Y_1)|y_2) = E((Y_1 - a, Y_1 - a) + 1|y_2)
\]

\[
= E((Y_1, Y_1) - 2(Y_1, a) + m_1 a^2 + 1|y_2)
\]

\[
= \langle \mu_1, \mu_1 \rangle + m_1 \tau^2 - 2\langle \mu_1, a \rangle + m_1 a^2 + 1
\]

\[
= \sum_{i=1}^{m_1} \left( \alpha + \eta \sum_{s_j \in \mathcal{K}_i} (y(s_j) - \alpha) \right)^2 + m_1 \tau^2 - 2\alpha \sum_{i=1}^{m_1} \left( \alpha + \eta \sum_{s_j \in \mathcal{K}_i} (y(s_j) - \alpha) \right) + m_1 a^2 + 1
\]

\[
= m_1 ((1 + 4\eta)^2 \alpha^2 + \alpha^2 + \tau^2) + 2(1 + 4\eta)\alpha(\eta - a)4\sum_{i=1}^{m_2} y_{2i} + \eta^2 \sum_{i=1}^{m_1} \left( \sum_{s_j \in \mathcal{K}_i} y(s_j) \right)^2
\]

\[
\leq m_1 ((1 + 4\eta)^2 \alpha^2 + \alpha^2 + \tau^2) + 2(1 + 4\eta)\alpha(\eta - a)4\sum_{i=1}^{m_2} y_{2i} + 4\eta^2 \sum_{i=1}^{m_1} \sum_{j=1}^{4} y(s_{ij})^2
\]

\[
= m_1 ((1 + 4\eta)^2 \alpha^2 + \alpha^2 + \tau^2) + 2(1 + 4\eta)\alpha(\eta - a)4\sum_{i=1}^{m_2} y_{2i} + 16\eta^2 \sum_{i=1}^{m_2} y_{2i}^2
\]

\[
= j \sum_{i=1}^{m_2} y_{2i}^2 + 2j (1 + 4\eta)\alpha(\eta - a) \sum_{i=1}^{m_2} y_{2i} + c
\]

\[
= j \sum_{i=1}^{m_2} y_{2i}^2 - 2j b \sum_{i=1}^{m_2} y_{2i} + c
\]

\[
= jg(y_2) + k.
\]

for some \( k > 0 \), where the inequality above follows from Jensen’s inequality and \( j = 16\eta^2 \) above. Similarly,

\[
E(g(Y_2)|y_1) \leq uf(Y_1) + v
\]

holds with \( u = 12\eta^2 \) and some \( v > 0 \). Thus, (E.1) of Lemma 5 is satisfied where \( ju = 16^2\eta^4 < 1 \) by the model assumption of \( |\eta| < 0.25 \).

Now let \( d > 0 \) and note that

\[
C_d = \{y_2 \in \mathbb{R}^{m_2} : g(y_2) \leq d\} = \{\langle y_2 - b, y_2 - b \rangle + 1 \leq d\} = \{||y_2 - b||^2 \leq d - 1\}
\]

is compact because \( C_d \) is empty for \( 0 < d < 1 \) and a closed ball of radius \( d - 1 \) when \( d \geq 1 \). Thus, Lemma 5 applies and the CGS, RQGS, and RSGS samplers are geometrically ergodic for the Gaussian case.
(Centered) Inverse Gaussian:

Let \( Y(s_i) \) be conditionally Inverse Gaussian (IG) distributed given \( Y(N_i) \) using a four-nearest neighbor structure with a density

\[
f_i(y(s_i)|y(N_i), \mu, \lambda, \eta_1, \eta_2) = \exp \left\{ \frac{A_{1i}(y(N_i))}{2} y_i - \frac{A_{2i}(y(N_i))}{2} \frac{1}{y_i} - B_i(y(N_i)) + C(y(s_i)) \right\}
\]

where

\[
A_{1i}(y(N_i)) = \alpha_1 + \eta_1 \sum_{j} \left( \frac{1}{y(s_j)} - \frac{1}{\mu} - \frac{1}{\lambda} \right) \quad \text{and} \quad A_{2i}(y(N_i)) = \alpha_2 + \eta_2 \sum_{j} (y(s_j) - \mu),
\]

for \( \mu, \lambda > 0, \eta_1, \eta_2 \geq 0, \) and \( \alpha_1 = \lambda / \mu^2 > 0, \alpha_2 = \lambda > 0. \)

For this model, the conditional mean of \( Y_i \) is \( \sqrt{A_{2i}(y(N_i))/A_{1i}(y(N_i))} \) and the conditional mean of \( 1/Y_i \) is \( \sqrt{A_{1i}(y(N_i))/A_{2i}(y(N_i)) + 1/A_{2i}(y(N_i))} \). In order for this model to be valid (i.e. \( A_{1i}(y(N_i)), A_{2i}(y(N_i)) \geq 0 \)), we need \( \lambda, \mu > 0 \) with \( \eta_1, \eta_2 \geq 0, \) or equivalently

\[
\alpha_1 - 4\eta_1 \left( \frac{1}{\mu} + \frac{1}{\lambda} \right) > 0 \quad \text{and} \quad \alpha_2 - 4\eta_2 \mu > 0,
\]

or \( 0 \leq \eta_1 \leq \frac{\lambda^2}{4\mu(\lambda + \mu)} \quad \text{and} \quad 0 \leq \eta_2 \leq \frac{\lambda}{4\mu}. \)

in the four-nearest neighborhood structure.

For technical reasons related to geometric ergodicity, we extend and close the IG model support from \( (0, \infty) \) to \( [0, \infty) \) without changing the joint distribution of \( (Y(s_1), \ldots, Y(s_n)) \). To accomplish this,

1. We declare the conditional distribution for \( Y(s_i)|\{Y(s_j): s_j \in N_i \} \) to be IG(1,1) if any conditioning variables are zero among \( \{Y(s_j): s_j \in N_i \} \), and

2. extend the density of any IG distribution to be \( \infty \) when the argument is zero, i.e. \( f_i(y|\cdot) = \infty \) at \( y = 0. \)

Let \( m_i \) be the number of locations in \( \mathcal{C}_i \) for \( i = 1, 2 \). Now, we can write

\[
A_{1i}(y(N_i)) = \tilde{\alpha}_1 + \eta_1 \sum_{s_j \in N_i, y(s_j)} \frac{1}{y(s_j)}, \quad \tilde{\alpha}_1 = \alpha_1 - 4\eta_1 \left( \frac{1}{\mu} + \frac{1}{\lambda} \right) > 0,
\]

\[
A_{2i}(y(N_i)) = \tilde{\alpha}_2 + \eta_2 \sum_{s_j \in N_i} y(s_j), \quad \tilde{\alpha}_2 = \alpha_2 - 4\eta_2 \mu > 0,
\]

where

\[
\tilde{\alpha}_1 = \alpha_1 - 4\eta_1 \left( \frac{1}{\mu} + \frac{1}{\lambda} \right), \quad \tilde{\alpha}_2 = \alpha_2 - 4\eta_2 \mu.
\]
assuming that none of \(y(s_j), s_j \in \mathcal{N}_i\) are zero.

By the structure of the IG conditional densities, it suffices to establish geometric ergodicity of the CGS, RSGS, and RQGS samplers by verifying (E.1) in an application of Lemma 5 with \(Y_1 = \{Y(s_i) : s_i \in \mathcal{G}_1\} \equiv (Y_{11}, \ldots, Y_{1m_1})\) and \(Y_2 = \{Y(s_i) : s_i \in \mathcal{G}_2\} \equiv (Y_{12}, \ldots, Y_{1m_2})\).

Now define

\[
f(y_1) = \sum_{i=1}^{m_1} y_{1i} + 1, y_1 \in [0, \infty)^{m_1} \quad \text{and} \quad g(y_2) = \sum_{i=1}^{m_2} y_{2i} + 1, y_2 \in [0, \infty)^{m_2}.
\]

In the following, write \(Y(s_i) = Y_i, i = 1, \ldots, n\) for simplicity. Then for \(y_2 \in [0, \infty)^{m_2}\), letting \(1(\cdot)\) denote the indicator function and defining \(\theta_1 = 4(\eta_2/\tilde{\alpha}_1)^{1/2}\), we have

\[
E(f(y_1)|y_2) = \sum_{\{1 \leq i \leq m_1: \text{some } Y_i = 0 \text{ for } s_j \in \mathcal{N}_i\}} E(IG(1, 1)) + \sum_{\{1 \leq i \leq m_1: \text{some } Y_i \neq 0 \text{ for } s_j \in \mathcal{N}_i\}} E(Y_1|Y_2) + 1
\]

\[
= \sum_{\{1 \leq i \leq m_1: \text{some } Y_i = 0 \text{ for } s_j \in \mathcal{N}_i\}} 1 + \sum_{\{1 \leq i \leq m_1: \text{some } Y_i \neq 0 \text{ for } s_j \in \mathcal{N}_i\}} \left(\frac{A_2(y(s_i.N))}{A_1(y(s_i.N))}\right)^{1/2} + 1
\]

\[
\leq m_1 + \sum_{\{1 \leq i \leq m_1: \text{some } Y_i \neq 0 \text{ for } s_j \in \mathcal{N}_i\}} \left(\frac{(\tilde{\alpha}_2)^{1/2} + \left(\frac{\eta_2 \sum_{j \in \mathcal{N}_i} y_j}{\tilde{\alpha}_1 + \eta_2 \sum_{j \in \mathcal{N}_i} 1/y_j}\right)^{1/2}}{(\tilde{\alpha}_1)^{1/2}} + 1\right)
\]

\[
\leq m_1 \left(1 + \left(\frac{\tilde{\alpha}_2}{\tilde{\alpha}_1}\right)^{1/2}\right) + \frac{\eta_2 m_1}{4} \sum_{i=1}^{m_1} \left(\frac{\sum_{j \in \mathcal{N}_i} y_j}{\sum_{j \in \mathcal{N}_i} 1/y_j}\right)^{1/2} + 1
\]

\[
\leq m_1 \left(1 + \left(\frac{\tilde{\alpha}_2}{\tilde{\alpha}_1}\right)^{1/2}\right) + 1 + \frac{1}{4} \theta_1 \sum_{i=1}^{m_1} \left(\sum_{j \in \mathcal{N}_i} y_j\right)^{1/2} \left(\sum_{j \in \mathcal{N}_i} y_j\right)^{1/2} + \theta_1 \left(\sum_{j \in \mathcal{N}_i} y_j\right)^{1/2} > 2\theta_1
\]

\[
\leq m_1 \left(1 + \left(\frac{\tilde{\alpha}_2}{\tilde{\alpha}_1}\right)^{1/2} + \frac{1}{2} \theta_1^2\right) + 1 + \frac{11}{4} \sum_{i=1}^{m_1} \sum_{j \in \mathcal{N}_i} y_j
\]

\[
\leq c + \frac{1}{2} g(y_2) + k.
\]

for \(j = \frac{1}{2}\) and some \(k > 0\). Similarly, it holds that

\[
E(g(y_2)|y_1) \leq uf(y_1) + v
\]
for $u = \frac{1}{4}$ and some $v > 0$. Thus, (E.1) of Lemma 5 is satisfied with $ju = \frac{1}{4} < 1$.

And for $d > 0$, we have

$$C_d = \{y_2 \in [0, \infty)^{m_2} : g(y_2) \leq d\} = \left\{y_2 \in [0, \infty)^{m_2} : \sum_{i=1}^{m_2} y_{2i} + 1 \leq d\right\}$$

which is compact as $C_d = \emptyset$ if $0 < d \leq 1$ and for $d > 1$, $C_d$ is a closed ball of radius $d - 1$ under the $L_1$ norm. Thus, Lemma 5 holds. So it follows that the CGS, RQGS, and RSGS are geometrically ergodic for the IG case.

(Centered) Truncated Gamma:

Let $Y(s_i)$ be conditionally Gamma distributed (truncated such that $Y(s_i) \geq 1$), given $Y(\mathcal{N}_i)$ using a four-nearest neighbor structure where

$$f_i(y(s_i) | \theta) = \exp\{A_{1i}(y(\mathcal{N}_i)) \log(y_i) - A_{2i}(y(\mathcal{N}_i))y_i - B_i(y(\mathcal{N}_i)))\}, \quad y(s_i) \geq 1,$$

where

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

for $\eta > 0, \alpha_1 > -1, \alpha_2 > 0$. That is $Y(s_i) | y(\mathcal{N}_i)$ is a gamma with scale and shape parameters $A_{1i}(y(\mathcal{N}_i)) + 1$ and $1/A_{2i}(y(\mathcal{N}_i))$.

This model specifies a valid joint distribution, so from Theorem 1 we have that the clique-based Gibbs strategies yield a Harris ergodic sampler for this model. Again, by the structure of the truncated Gamma conditionals, it suffices to establish geometric ergodicity of the CGS (and RSGS/RQGS) by verifying (E.1) in an application of Lemma 5 with $Y_1 = \{Y(s_i) : s_i \in \mathcal{C}_1 \} \equiv (Y_{11}, \ldots, Y_{1m_1})$ and $Y_2 = \{Y(s_i) : s_i \in \mathcal{C}_2 \} \equiv (Y_{12}, \ldots, Y_{1m_2})$ where $m_k$ denotes the number of observations/locations in clique $\mathcal{C}_k, k = 1, 2$.

Define functions of clique observations as

$$f(y_1) = \sum_{i=1}^{m_1} y_{1i} \quad \text{and} \quad g(y_2) = \sum_{i=1}^{m_2} y_{2i}.$$ 

Let $c$ and $\tilde{c}$ denote generic constants that do not depend on $y_2$ and let $I(\cdot)$ denote the indicator function. Then, for $y_2 \in [1, \infty)^{m_2}$, noting that the conditional truncated gamma distribution
of \( Y(s_i) \) has a mean bounded by that of a gamma variable with scale \( A_{1i}(y_{i\cdot}) \) and shape \( 1/A_{2i}(y_{i\cdot}) \) parameters, we have that

\[
E(f(Y) | y_2) = E \left( \sum_{i=1}^{m_1} Y_{1i} | y_2 \right) \\
= \sum_{i=1}^{m_1} E(Y_{1i} | y_2) \\
\leq \sum_{i=1}^{m_1} \frac{1}{\alpha_2} \left( \alpha_1 + \eta \sum_{j \in j_i} \log(y(s_j)) + 1 \right) \\
= c + \frac{4\eta}{\alpha_2} \sum_{j=1}^{m_2} \log(y_{2j}) \\
\leq c + \frac{4\eta}{\alpha_2} \sum_{j=1}^{m_2} \sqrt{y_{2j}} \\
= c + \frac{4\eta}{\alpha_2} \sum_{j=1}^{m_2} \left[ \sqrt{y_{2j}} \left( \sqrt{y_{2j}} \leq 2 \frac{4\eta}{\alpha_2} \right) + \sqrt{y_{2j}} \left( \sqrt{y_{2j}} > 2 \frac{4\eta}{\alpha_2} \right) \right] \\
\leq c + \frac{4\eta}{\alpha_2} \sum_{j=1}^{m_2} \left[ 2 \frac{4\eta}{\alpha_2} + \frac{y_{2j}}{2 \times 4\eta/\alpha_2} \right] \\
= \tilde{c} + \frac{1}{2} \sum_{j=1}^{m_2} y_{2j} \\
= jg(y_2) + k,
\]

for \( j = \frac{1}{2} \) and some \( k > 0 \), using above that \( \log(y) \leq \sqrt{y} \) for \( y \geq 1 \). Similarly,

\[
E(g(Y) | y_1) \leq uf(y_1) + v
\]

holds for \( u = \frac{1}{2} \) and some \( v > 0 \). Thus, (E.1) of Lemma 5 is satisfied where \( ju = \frac{1}{4} < 1 \).

Finally, for \( d > 0 \),

\[
C_d = \left\{ y_2 \in [1, \infty)^{m_2} : g(y_2) \leq d \right\} = \left\{ y_2 \in [1, \infty)^{m_2} : \sum_{i=1}^{m_2} y_{2i} \leq d \right\}
\]

is compact (\( C_d = \emptyset \) if \( 0 < d < m_2 \) and \( C_d \) is closed and bounded for \( d \geq m_2 \)). Thus, Lemma 5 holds. Ergo, the CGS, RQGS, and RSGS are geometrically ergodic for the Truncated Gamma case. \( \square \)


