The Pennsylvania State University
The Graduate School

ADVANCES IN SPATIAL STATISTICS AND INFERENCE METHODS FOR MARKOV POPULATION MODELS

A Dissertation in
Statistics
by
Adam Walder

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Doctor of Philosophy

August 2021

The dissertation of Adam Walder was reviewed and approved by the following:

Ephraim M. Hanks
Associate Professor, Department of Statistics
Chair of Graduate Studies
Dissertation Advisor
Chair of Committee

Murali Haran
Professor and Head, Department of Statistics

Aleksandra B. Slavković
Professor, Department of Statistics
Associate Dean for Graduate Education

Matthew Ferrari
Associate Professor, Department of Biology
Director of the Center for Infectious Disease Dynamics

Abstract

Spatial generalized linear mixed models (SGLMMs) commonly rely on Gaussian random fields (GRFs) to capture spatially correlated error. We investigate the results of replacing Gaussian processes with Laplace moving averages (LMAs) in SGLMMs. We demonstrate that LMAs offer improved predictive power when the data exhibits localized spikes in the response. SGLMMs with LMAs are shown to maintain analogous parameter inference and similar computing to Gaussian SGLMMs. We propose a novel discrete space LMA model for irregular lattices and construct conjugate samplers for LMAs with georeferenced and areal support. We provide a Bayesian analysis of SGLMMs with LMAs and GRFs over multiple data support and response types

We develop methods for privatizing spatial location data, such as spatial locations of individual disease cases. We propose two novel Bayesian methods for generating synthetic location data based on log-Gaussian Cox processes (LGCPs). We show that conditional predictive ordinate (CPO) estimates can easily be obtained for point process data. We construct a novel risk metric that utilizes CPO estimates to evaluate individual disclosure risks. We adapt the propensity mean square error (pMSE) data utility metric for LGCPs. We demonstrate that our synthesis methods offer an improved risk vs. utility balance in comparison to radial synthesis with a case study of Dr. John Snow's cholera outbreak data.

We demonstrate how to perform inference on Markov population processes with Laplace approximations. We derive a sparse covariance structure for the linear noise approximation (LNA) which offers a joint Gaussian likelihood for Markov population models based solely on the solution to a set of deterministic equations. We show that Laplace approximations allow inference with LNAs to be parallelized and require no stochastic infill. We also demonstrate that our method offers comparable accuracy to MCMC on a simulated Susceptible-Infected-Susceptible data set. We use Laplace approximations to fit a stochastic susceptible-exposed-infected-recovered system to the Princess Diamond COVID-19 cruise ship data set.

Table of Contents

List of Figures vii

List of Tables ix

Acknowledgments x

Chapter 1
Introduction 1

Chapter 2
Bayesian Analysis of Spatial Generalized Linear Mixed Models with Laplace Moving Average Random Fields 3

2.1 Introduction . . . 3
2.2 Spatial Models with Matérn Random Fields . . . 5
2.2.1 SGLMMs . . . 5
2.2.2 Random Fields with Matérn Covariance . . . 6
2.3 LMA Models for Point Referenced Data . . . 6
2.3.1 Laplace Moving Average Models as SPDEs . . . 7
2.3.2 Finite Element Approximations to Matérn GRFs . . . 8
2.3.3 Finite Element Approximations to Matérn LMAs . . . 9
2.3.4 Model Fitting: Continuous Space . . . 10
2.4 LMA Models in Discrete Space . . . 12
2.4.1 Graph Trend Filtering . . . 13
2.4.2 LMAs in Discrete Space . . . 13
2.4.3 L2 vs. L1 Penalization . . . 15
2.4.4 Model Fitting: Discrete Space . . . 16
2.5 Example Analyses . . . 17
2.5.1 Discrete Space Examples: Slovenia Stomach Cancer . . . 17
2.5.2 Discrete Space Examples: Columbus Crime Data . . . 19
2.5.3 Continuous Space Examples: Malaria in the Gambia, Africa . . . 21
2.5.4 Continuous Space Example: LAGOS . . . 23
2.6 Discussion . . . 25

Chapter 3
Privacy for Spatial Point Process Data **27**
3.1 Introduction . . . 27
3.2 Data Synthesis and Dissemination . . . 31
3.3 LGCPs . . . 32
3.3.1 Computational Details for Fitting Log-Gaussian Cox Processes . . . 33
3.4 Conditional Predictive Ordinate Estimates for Cox Processes . . . 34
3.5 Generating Fully Synthetic Location Data . . . 36
3.5.1 Radial Synthesis . . . 36
3.5.2 Additive Noise Synthesis . . . 36
3.5.3 Posterior Resampling Synthesis . . . 38
3.6 Evaluating Utility . . . 39
3.7 Evaluating Disclosure Risks . . . 40
3.7.1 Disclosure Risk Metric . . . 41
3.7.2 Radial Synthesis Disclosure Risks . . . 41
3.7.3 Additive Noise Synthesis Disclosure Risks . . . 42
3.7.4 Posterior Resampling Synthesis Disclosure Risks . . . 43
3.7.5 Differential Privacy . . . 43
3.8 Case Study: Dr. John Snow's Cholera Outbreak . . . 45
3.9 Discussion . . . 49

Chapter 4
Laplace Approximations for Fast Bayesian Inference on Markov Population Models **52**
4.1 Introduction . . . 52
4.2 Markov Population Models . . . 54
4.2.1 Example: The Stochastic SIS Model . . . 55
4.2.2 Diffusion Approximations . . . 56
4.3 Linear Noise Approximations to Stochastic Population Models . . . 57
4.3.1 Limiting Deterministic Systems . . . 57
4.3.2 Approximate Gaussian Processes . . . 58
4.3.3 Sparse Precision Matrices . . . 59
4.4 Laplace Approximations for Inference with LNAs . . . 60
4.4.1 Approximate Marginal Distributions for $\pi(\boldsymbol{\theta}|\boldsymbol{y})$ . . . 60
4.4.2 Gaussian Approximations of $\pi\left(\boldsymbol{X}_N|\boldsymbol{\theta},\boldsymbol{y}\right)$ . . . 61
4.4.3 Estimating $\pi\left(\theta_i|\boldsymbol{y}\right)$ . . . 62
4.5 Simulation Study . . . 63
4.6 Data Analysis: COVID Cruise Ship . . . 66
4.7 Discussion . . . 72

Chapter 5
Conclusion **74**
5.1 Concluding Remarks for Chapter 2 . . . 74
5.2 Concluding Remarks for Chapter 3 . . . 75

5.3 Concluding Remarks for Chapter 4 76

Appendix A
Chapter 2 Appendices **77**
A.1 CAR Models 77
A.2 SAR Models 77
A.3 Conditionally Independent Block Proposals 78
A.4 Details of Slovenia Data Analysis 78
A.5 Details of the Columbus Crime Dataset Analysis 80
A.6 Details of Malaria Data Analyses 81
A.7 Details of LAGOS Analysis 83

Appendix B
Chapter 3 Appendices **85**
B.1 Finite Element Approximations for Matérn GRFs 85
B.2 Approximate Inference for LGCPs 86
B.3 Circular Synthesis Disclosure Risk Details 87
B.4 Quadrature Scheme For Circular Domains 88
B.5 Prior Choice for (κ, ξ) 89

Appendix C
Chapter 4 Appendices **91**
C.1 FCLT for Poisson Processes 91
C.2 Solving for V(t) 91
C.3 SIS LNA 92
C.4 SEIR LNA 93

Bibliography **94**

List of Figures

2.1 (a) Standard normal, $\mathcal{N}(0,1)$, and scale one Laplace density plots. (b) Tails of the respective distributions. 7

2.2 Plot of observed incidence ratio of stomach cancer (SIR), reported as the ratio of observed occurrences divided by the expected count in municipality i. 18

2.3 Plot of crime rate in thousands in the 49 counties of Columbus, Ohio. . . 19

2.4 Plot of triangular mesh with $n = 288$ nodes and malaria frequency at 65 unique village locations. 22

2.5 Plot of median log total phosphorus (TP) recorded at 5526 unique lake locations. 24

3.1 A map of the 578 observed cholera deaths, streets, and water pumps in Soho, London. The Broad St. water pump (pump 7) was the source of the outbreak. 28

3.2 (A): An intensity surface plot with the posterior mean estimate for the spatial random field $\hat{\eta}(\boldsymbol{s})$ from the John Snow cholera outbreak dataset. (B): An ANS intensity surface plot for an additive noise spatial random field, $\psi(\boldsymbol{s})$, with spatial scale $\hat{\kappa}^2$ and noise-level $\sigma^2 = 10$. 37

3.3 (A): An intensity surface plot with the posterior mean estimate for the spatial random field $\hat{\eta}(\boldsymbol{s})$ from the John Snow cholera outbreak dataset. (B): An intensity surface plot for a resampled spatial random field, $\nu(\boldsymbol{s})$, with spatial scale $\hat{\kappa}^2$ and variance parameter $\hat{\xi}^2$. 38

3.4 A kernel density estimate of the Soho, London population in 1854. Cholera death locations are plotted in blue as well as the water pumps numbered 1-13, with pump 7 being the Broad St. water pump. The estimated total population is 21,345. 45

3.5 Plot of max disclosure risk vs. *pMSE* for *ANS*, radial synthesis, and *PRS* datasets. Note that only the min and max utility are plotted for *PRS*. . . 46

3.6 Plot of max disclosure risk vs. *pMSE* for *ANS* and *PRS* datasets with max disclosure risks all less than 0.005. The noise level σ^2 is displayed above each *ANS* point. All 15 *PRS* data sets are plotted. 47

3.7 A plot of the fitted intensity surface for the optimal *PRS* data set with max disclosure risk of 1.205e-17 and *pMSE* score of 0.0016. The true deaths (blue), *PRS* synthetic locations (red), and the Broad St. pump (black) are plotted as well. 48

4.1 A plot of the stochastic trajectory of $P(t)$ (red) generated from an SIS model with $N = 1000$, $P(0) = 0.05$, $\beta = 0.50$, and $\gamma = 0.33$. 56

4.2 A plot of the stochastic trajectory of $P(t)$ (red), the ODE fit $P^{\dagger}_{\hat{\boldsymbol{\theta}}}(t)$ (black) for the posterior mean estimates of the Laplace approximation $\hat{\boldsymbol{\theta}} = (0.4383, 0.2798)$, and the observed sample proportions $y_t/50$ (blue). 63

4.3 Marginal density plots of $\pi(\beta|\boldsymbol{y})$ and $\pi(\gamma|\boldsymbol{y})$ fit by Laplace approximations (black) and MCMC (blue). The true values of $\beta = 0.50$ and $\gamma = 0.33$ are plotted in red. 65

4.4 A plot of the observed proportions of seropositive individuals aboard the Princess Diamond cruise ship (black), the ODE solution $P^{\dagger}_{\boldsymbol{\theta}}(t)$ (blue) for posterior mean estimates of $\boldsymbol{\theta}$ from fitting a Laplace approximation, and 95% credible intervals of $P^{\dagger}_{\boldsymbol{\theta}}(t)$ generated by simulating from the LNA (red). 67

4.5 Marginal density plots ($\pi(\theta_i|\boldsymbol{y})$). 71

List of Tables

2.1 BCVS and ESS for ten-fold cross validation on the Slovenia stomach cancer outbreak dataset. 19

2.2 BCVS from ten-fold cross validation and ESS for Columbus Ohio Crime dataset. 20

2.3 Parameter estimates for discrete space data analysis examples of Sections (2.5.1–2.5.2) 21

2.4 BCVS from ten-fold cross validation and ESS for malaria in the Gambia, Africa. 23

2.5 BCVS for ten fold cross validation and ESS for the LAGOS dataset. . . . 24

2.6 Parameter estimates for continuous space data analysis examples of Sections (2.5.3–2.5.4) 25

3.1 Posterior mean estimates and corresponding 95% credible intervals (CI) for the effective range and marginal variance of the spatial random field for the confidential and *PRS* data sets. 48

4.1 Posterior mean estimates and 95% highest posterior density (HPD) intervals for the Laplace approximation (Laplace) and MCMC model fits on the simulated SIS data set. True parameter values are $\beta = 0.50$ and $\gamma = 0.33$. 65

4.2 Covid data 70

4.3 Posterior mean estimates and 95% HPD intervals for the SEIR model fit to the Princess Diamond cruise ship data set. 71

Acknowledgments

I would first like thank my thesis advisor Dr. Ephraim Hanks for always finding a way to make time for even the most meaningless of questions. Though there were many occurrences in which I felt utterly clueless, lost, and ashamed in my shortcomings, Dr. Hanks always made the extra effort to keep my spirits high and guide me through this long journey.

I would also like thank my friend group here in State College. To my closest friends Ilias and Emily who always managed to provide a different outlook on life during my graduate school years. To my informal roommates Matt, Mike, Tobia, and Umberto for always keeping the door open and the countless memories shared.

And lastly, to my closest friend Obadha, for his unwavering support over the last 16 years of friendship.

Chapter 1
Introduction

This work consists of three projects on Bayesian inference for spatial generalized linear mixed models, privacy for spatial point process data, and infectious disease dynamics. The range of topics encompassed in this work includes; penalization methods for SGLMMs, Gaussian process regression, spatial modeling, data privacy, statistical computing, stochastic differential equation modeling, and inference for mechanistic models. Each work discussed in Chapters 2–4 offer novel advances in statistical modeling and computational methods. We illustrate the usefulness of each proposed methodology with applications on several real-world data sets.

Chapter 2 of this work focuses on penalization methods for spatial generalized linear mixed models (SGLMMs). SGLMMs commonly rely on Gaussian random fields (GRFs) to capture spatially correlated error. We investigate the results of replacing Gaussian processes with Laplace moving averages (LMAs) in SGLMMs. We demonstrate that LMAs offer improved predictive power when the data exhibits localized spikes in the response. SGLMMs with LMAs are shown to maintain analogous parameter inference and similar computing to Gaussian SGLMMs. We propose a novel discrete space LMA model for irregular lattices and construct conjugate samplers for LMAs with georeferenced and areal support. We provide a Bayesian analysis of SGLMMs with LMAs and GRFs over multiple data support and response types. This work, co-authored by Ephraim Hanks, has been published in *Computational Statistics and Data Analysis (CSDA)*.

In Chapter 3, we develop methods for privatizing spatial location data, such as spatial locations of individual disease cases. We propose two novel Bayesian methods for generating synthetic location data based on log-Gaussian Cox processes (LGCPs). We show that conditional predictive ordinate (CPO) estimates can easily be obtained for point process data. We construct a novel risk metric that utilizes CPO estimates to evaluate individual disclosure risks. We adapt the propensity mean square error (pMSE)

data utility metric for LGCPs. We demonstrate that our synthesis methods offer an improved risk vs. utility balance in comparison to radial synthesis with a case study of Dr. John Snow's cholera outbreak data. This is a joint work with Ephraim Hanks and Aleksandra Slavković currently in review at *Journal of Computational and Graphical Statistics (JCGS)*.

In Chapter 4, we demonstrate how to perform inference on Markov population processes with Laplace approximations. We derive a sparse covariance structure for the linear noise approximation (LNA) which offers a joint Gaussian likelihood for Markov population models based solely on the solution to a set of deterministic equations. We show that Laplace approximations allow inference with LNAs to be parallelized and require no stochastic infill. We also demonstrate that our method offers comparable accuracy to MCMC on a simulated Susceptible-Infected-Susceptible data set. We use Laplace approximations to fit a stochastic susceptible-exposed-infected-recovered system to the Princess Diamond COVID-19 cruise ship data set. This is a joint work with Ephraim Hanks, currently being prepared for submission.

In Chapter 5, we highlight the key results of Chapters 2–4. We share our thoughts on the future directions of the three projects discussed in this work. We also discuss our thoughts on the current needs for statistical developments in the fields of inference methods for infectious disease models and spatial statistics.

Chapter 2
Bayesian Analysis of Spatial Generalized Linear Mixed Models with Laplace Moving Average Random Fields

[1]

2.1 Introduction

The Gaussian random field (GRF) possesses an intuitive dependence structure which offers a flexible fit for describing spatially and/or temporally correlated errors. These desirable features have popularized the use of Gaussian processes in spatial and temporal statistics, as well as design of experiments and other fields. Despite the GRF's flexible nature, Gaussian processes can over-smooth in the presence of local spikes (Paciorek and Schervish, 2004). In this work we consider the use of Laplace moving average models (LMAs) in place of traditional Gaussian processes in spatial generalized linear mixed models (SGLMMs)

LMAs have received sporadic attention over the past decade as alternatives to GRFs (Åberg and Podgórski, 2011; Wallin and Bolin, 2015; Opitz, 2016). However, there has been no systematic comparison of LMAs and GRFs for spatial generalized linear mixed models (SGLMMs) in the Bayesian framework. Our contributions in this work include

[1]This is a published work. Walder, A., & Hanks, E. M. (2020). Bayesian analysis of spatial generalized linear mixed models with Laplace moving average random fields. Computational Statistics & Data Analysis, 144, 106861

1. The development of a novel discrete space (areal) LMA model for irregular lattices.

2. The construction of conjugate samplers for both continuous (point-referenced) and discrete (areal) SGLMMs with LMAs.

3. A Bayesian analysis comparing the predictive power and computational efficiency of LMAs and GRFs over a range of scenarios, including continuous, binary, and count data collected both in discrete (areal) and point-referenced (geostatistical) spatial support.

Whittle (1954) demonstrated that continuous space GRFs with Matérn covariance arise as solutions to a stochastic partial differential equation (SPDE). Lindgren et al. (2011) constructed a sparse finite element representation of this Gaussian Matérn SPDE. As a result, a sparse form of the multivariate normal distribution can be used to fit Matérn GRF models in a computationally efficient manner (Rue, 2001). Bolin (2014) extended the finite element approximation of Lindgren et al. (2011) to the case of Type-G Matérn random fields of which the LMA is a special case. In Section 2.3, we provide a summary of this extension for the symmetric LMA. Following Bolin (2014), the LMA can similarly be expressed as a conditionally sparse Gaussian random field through the introduction of auxiliary data. We also provide insights for handling the computational issues associated with MCMC implementation.

Wallin and Bolin (2015) explored the LMA for geostatistical data with Gaussian responses. Though the discrete space model was claimed to be analogous, no further exploration was considered. Faulkner and Minin (2018) provided a Bayesian implementation of the graph trend filtering (GTF) estimates of Wang et al. (2016) for temporal data. Both works, found that replacing traditional Gaussian priors for Laplace priors provided a model with better adaptivity in the presence of local “spikes" in the response. We develop a novel discrete space analog to the continuous space LMA model that is an extension of Wang et al. (2016) for SGLMMs. Our model can easily be implemented in place of any Gaussian conditionally autoregressive (CAR) or simultaneously autoregressive (SAR) model. In Section 2.4, we propose a novel MCMC implementation of our Bayesian hierarchical model for discrete space SGLMMs.

The LMA models of this paper offer an intuitive alternative to traditional GRF SGLMMs. The discrete space and continuous space LMA models are constructed based on sparse matrix operations making for fast and efficient fitting. We provide Bayesian analyses based on our novel MCMC implementation of the LMA models over four separate data sets. Our MCMC implementation and model construction allows for

Bayesian inference with LMAs that is just as interpretable as with GRF models. In some cases, the LMA is shown to provide a better fit than the Gaussian model. Given the ease of implementation, and familiarity of inference, LMA models can be useful alternatives to GRF models.

The paper is organized as follows: In Section 2 we provide background material needed to develop our models. In Section 2.3 we discuss finite element approximations for continuous space LMAs. We also provide details related to the numerical issues involved with fitting LMAs via MCMC. In Section 2.4 we detail our discrete space LMA model and its relation to the GTF estimates of Wang et al. (2016). In Section 2.5 we consider four datasets on which the LMA model is compared to its GRF counterpart. We conclude with a discussion in Section 2.6.

2.2 Spatial Models with Matérn Random Fields

In this section we provide background information to assist in developing the framework of the hierarchical spatial models considered in this work. We begin by describing the SGLMM, and follow with a discussion of the Matérn random field as a solution to a stochastic partial differential equation (SPDE).

2.2.1 SGLMMs

Generalized linear models (GLMs) model the mean $\boldsymbol{\mu}$ of a distribution f with linear predictors through an invertible link function $g(\cdot)$. Generalized linear mixed models (GLMMs) are GLMs that allow for the inclusion of random effects in the linear predictor. The spatial GLMM (SGLMM) attempts to capture an unobserved spatially varying trend by imposing a dependence structure in the random effect.

Let $\{\boldsymbol{u}_i\}_{i=1}^N$ be a collection of locations observed in some spatial domain $\Omega \subset \mathbb{R}^d$. Consider $Y(\boldsymbol{u}_i) \sim f(y_i)$ such that the mean, $\mu_i = E\left(Y(\boldsymbol{u}_i) \mid \boldsymbol{\beta}, \eta(\boldsymbol{u}_i), \epsilon(\boldsymbol{u}_i)\right)$, is modeled through the link function

$$g(\mu_i) = \boldsymbol{x}_i'\boldsymbol{\beta} + \eta(\boldsymbol{u}_i) + \epsilon(\boldsymbol{u}_i). \tag{2.1}$$

In some cases an uncorrelated random effect, $\epsilon(\boldsymbol{u})|\sigma^2 \sim \mathcal{N}(0, \sigma^2)$, is included. In the Gaussian response case, σ^2, is thought of as homogeneous measurement error. The spatially varying random effect $\eta(\boldsymbol{u})|\boldsymbol{\theta}$ is included in (2.1) to capture spatial dependence. The hyper-parameters ($\boldsymbol{\theta}$) govern the mean and covariance structure of the spatial

random effect. The most common distribution assumed for $\eta(\boldsymbol{u})|\boldsymbol{\theta}$ is Gaussian. This work considers replacing the traditional Gaussian prior for a less common Laplace moving average (LMA). We provide a thorough comparison of the two prior choices over four datasets in Section 2.5.

2.2.2 Random Fields with Matérn Covariance

Stationary Matérn random fields arise as stationary solutions to the SPDE

$$(\kappa^2 - \triangle)^{\alpha/2}\eta(\boldsymbol{u}) = \xi\mathcal{W}(\boldsymbol{u}), \quad \boldsymbol{u} \in \mathbb{R}^d, \quad \alpha = \nu + d/2, \tag{2.2}$$

where $\triangle = \sum_{i=1}^{d} \frac{\partial^2}{\partial u_i^2}$ is the Laplacian operator, κ is a spatial scale parameter, α controls the smoothness of the realized fields, and ξ is a variance parameter. The SPDE in (2.2) is driven by Gaussian white noise, $\mathcal{W}(\boldsymbol{u})$. Whittle (1954, 1963) showed that stationary solutions to (2.2) are GRFs with Matérn covariance

$$C(\boldsymbol{u}, \boldsymbol{v}) = \frac{\phi^2}{2^{\nu-1}\Gamma(\nu)}(\kappa||\boldsymbol{v} - \boldsymbol{u}||)^{\nu} K_{\nu}(\kappa||\boldsymbol{v} - \boldsymbol{u}||), \quad \boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^d, \tag{2.3}$$

where $||\cdot||$ denotes the Euclidean norm in $\mathbb{R}^d$. The marginal variance is $\phi^2 = \frac{\xi^2\Gamma(\nu)}{\Gamma(\alpha)(4\pi)^{d/2}\kappa^{2\nu}}$ and the approximation of the effective range is given by $\rho = \sqrt{8\nu}/\kappa$ (Lindgren et al., 2011). Lindgren et al. (2011) provided a finite element representation of the SPDE in (2.2). In Section 2.3.2 we detail the approximation of Lindgren et al. (2011). The result is a sparse Gaussian Markov random field (GMRF) approximation to a GRF with Matérn covariance given in (2.3).

2.3 LMA Models for Point Referenced Data

GRF modeling remains appealing to scientists due to familiarity of implementation and the intuitive dependence structure of Gaussian processes. The Laplace distribution offers a wider tailed, sharper peaked, alternative to the Gaussian distribution (see Figure 2.1). In this Section we provide an overview of the results of Bolin (2014), which demonstrate that the LMA can be expressed as the solution to an SPDE similar to its Gaussian counterpart. We then lay out the finite element approximations of Lindgren et al. (2011) and Bolin (2014) which provide sparse representations of the analytic solutions of SPDEs driven by Gaussian and Laplace noise respectively. We conclude the Section by providing a novel exploration of model fitting via MCMC.

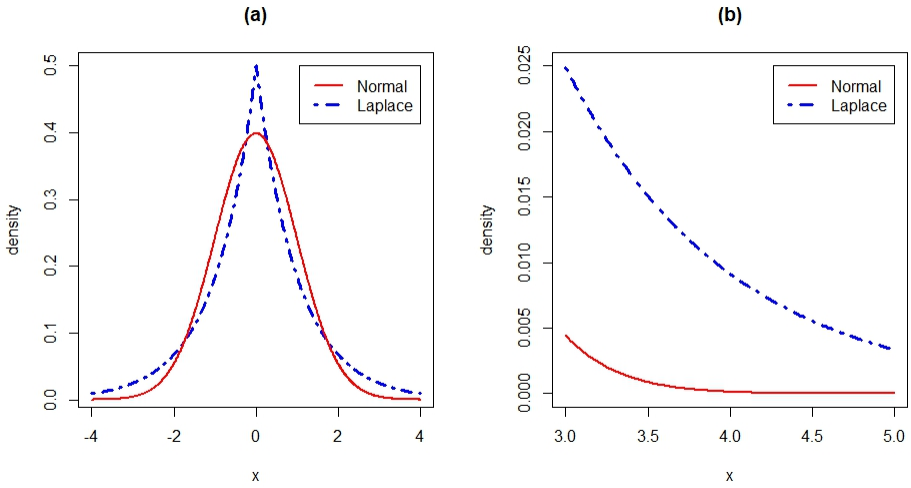

Figure 2.1: (a) Standard normal, $\mathcal{N}(0,1)$, and scale one Laplace density plots. (b) Tails of the respective distributions.

2.3.1 Laplace Moving Average Models as SPDEs

Gaussian priors often produce marginal distributions with light tails. Åberg and Podgórski (2011) suggested the use of LMAs to obtain asymmetric and heavier tailed marginals. Åberg and Podgórski (2011) showed that the LMA can be expressed as a convolution of a kernal with Laplace noise. Bolin (2014) showed that LMA models with Matérn covariance can equivalently be expressed as the solution to an SPDE by replacing the Gaussian white noise, $\mathcal{W}(\boldsymbol{u})$, in (2.2) with Laplace noise. Laplace noise on a compact set $\Omega \subset \mathbb{R}^d$ can be expressed as

$$\dot{\Lambda} = \sum_{k=1}^{\infty} \left(\Gamma_k + \sqrt{\Gamma_k} G_k \right) \delta_{\boldsymbol{u}_k}, \tag{2.4}$$

where $G_k \overset{iid}{\sim} \mathcal{N}(0,1)$, $\Gamma_k \sim e^{-\nu\gamma_k} W_k$, W_k are *iid* standard exponentials, γ_k are arrival times of a Poisson process with intensity one, and $\delta_{\boldsymbol{u}_k}$ is a Dirac distribution centered at $\boldsymbol{u}_k$ with each $\boldsymbol{u}_k$ uniformly distributed in Ω (see Bolin (2014) for details).

The SPDE defining an LMA process, $\eta(\boldsymbol{u})$, with Matérn covariance is given by

$$(\kappa^2 - \triangle)^{\alpha/2} \eta(\boldsymbol{u}) = \dot{\Lambda}(\boldsymbol{u}), \quad \alpha = \nu + d/2. \tag{2.5}$$

Both the SPDEs driven by Gaussian and Laplace noise produce random fields with Matérn covariance. However, the SPDE driven by Laplace noise provides a random

field with the ability to capture “spikey" spatial behavior better than GRFs. Unlike its Gaussian counterpart, there does not exist a closed form solution for the Laplace driven SPDE in (2.5). Bolin (2014) used a finite element approximation of the SPDE in (2.5) to fit LMA models using an EM algorithm. In Sections (2.3.2–2.3.3) we detail the finite element approximations to the Gaussian and LMA SPDEs proposed by Lindgren et al. (2011) and Bolin (2014) respectively.

2.3.2 Finite Element Approximations to Matérn GRFs

The finite element representation of the LMA model proposed by Bolin (2014) is an extension of the results in the Gaussian case. For the sake of comparison, we first detail the finite element approximation of Lindgren et al. (2011).

In Section 2.2.2 we expressed GRFs with stationary Matérn covariances as analytic solutions to SPDEs. Though the analytic solution provides useful insights, model fitting and parameter estimation are often facilitated by considering a numerical approximation. Lindgren et al. (2011) proposed the use of a finite element approximation to the stochastic weak formulation of the SPDE

$$(\kappa^2 - \triangle)^{\alpha/2}\eta(\psi) = \xi\mathcal{W}(\psi), \quad \alpha = \nu + d/2, \tag{2.6}$$

where $\{\psi\}$ is a set of test functions. The finite element method (FEM) solution begins by expressing the solution, $\eta(\boldsymbol{u})$, as a basis expansion

$$\eta(\boldsymbol{u}) = \sum_{i=1}^{n} \phi_i(\boldsymbol{u}) w_i, \quad \boldsymbol{u} \in \Omega, \tag{2.7}$$

where $\{\phi_i(\boldsymbol{u})\}_{i=1}^n$ is a set of basis functions on Ω. The solution in (2.6) is only required to hold for a finite collection of ψ_i. The Galerkin method approximate solution is obtained by setting $\{\psi_i\}_{i=1}^n = \{\phi_i\}_{i=1}^n$.

Lindgren et al. (2011) formulated an FEM approximation by considering $\{\phi_i(\boldsymbol{u})\}_{i=1}^n$ to be piecewise triangular basis functions. The basis functions are constructed by partitioning the spatial region of interest, $\Omega \subset \mathbb{R}^d$, into non-overlapping triangular regions. The corners of the triangles, referred to as vertices, are assigned n Gaussian weights, denoted w_i. Each ϕ_i is defined to be 1 at vertex i and 0 at all other vertices. Lindgren et al. (2011) derived the distribution of the weights

$$\boldsymbol{w}|\xi, \kappa \sim \mathcal{N}\left(\boldsymbol{0}, \xi^2 \boldsymbol{Q}_\alpha^{-2}\right), \tag{2.8}$$

where $\boldsymbol{Q}_\alpha$ is determined by the choice of α in (2.6).

Following Lindgren et al. (2011), the precision matrix in (2.8) is defined as

$$\boldsymbol{Q}_\alpha = \begin{cases} \boldsymbol{Q}_1 = \boldsymbol{L}, & \alpha = 1 \\ \boldsymbol{Q}_2 = \boldsymbol{L}\boldsymbol{C}^{-1}\boldsymbol{L}, & \alpha = 2 \\ \boldsymbol{Q}_\alpha = \boldsymbol{L}\boldsymbol{C}^{-1}\boldsymbol{Q}_{(\alpha-2)}\boldsymbol{C}^{-1}\boldsymbol{L}, & \alpha \geq 3 \end{cases}, \tag{2.9}$$

where $\boldsymbol{L} = \kappa^2\boldsymbol{C} + \boldsymbol{G}$. The matrices used to define $\mathbf{L}$, are given by

$$C_{ij} = \int_\Omega \phi_i(\boldsymbol{u})\phi_j(\boldsymbol{u})d\boldsymbol{u}, \tag{2.10}$$

$$G_{ij} = \int_\Omega \nabla\phi_i(\boldsymbol{u})\nabla\phi_j(\boldsymbol{u})d\boldsymbol{u}. \tag{2.11}$$

Note that $\mathbf{C}$ as defined in (2.10) is sparse, but its inverse, $\boldsymbol{C}^{-1}$, which is required in $\boldsymbol{Q}_\alpha$ for $\alpha \geq 2$, may not be. In turn, the resulting precision matrix, $\boldsymbol{Q}_\alpha$, may not be sparse. To ensure sparsity in $\boldsymbol{Q}_\alpha$, Lindgren et al. (2011) provided a GMRF approximation to the GRF representing the numerical solution to the SPDE in (2.6) by replacing $\mathbf{C}$ in (2.10) with the diagonal matrix,

$$C_{ii} = \int_\Omega \phi_i(\boldsymbol{u})d\boldsymbol{u}, \tag{2.12}$$

Under lattice refinement, the sparse representation of $\mathbf{C}$ as defined in (2.12) converges to the same solution given by $\mathbf{C}$ in (2.10) (see Lindgren et al. (2011) Appendix C.5). $\mathbf{C}$ in (2.12) is now a diagonal matrix relating to the volume of the regions produced by the mesh. $\mathbf{G}$ is a sparse matrix describing the connectivity of the mesh nodes.

2.3.3 Finite Element Approximations to Matérn LMAs

Bolin (2014) extended the results of Lindgren et al. (2011) to the case of Type-G Matérn random fields. We consider the FEM approximation for the symmetric LMA model with Matérn covariance. Similar to the Gaussian case, the FEM approximation begins with a stochastic weak formulation of the SPDE in (2.5) given by

$$(\kappa^2 - \triangle)^{\alpha/2}\eta(\psi) = \lambda\dot{\Lambda}(\psi). \tag{2.13}$$

Following Section 2.3.2 we construct piecewise linear basis functions $\{\phi_i\}_{i=1}^n$. Let $\Gamma_i \sim Gamma\,(\tau C_{ii}, \lambda^2)$ where C_{ii} is the i^{th} element of $\mathbf{C}$ in (2.12). Define $\boldsymbol{\Gamma} = diag(\Gamma_1, ..., \Gamma_n)$.

Bolin (2014) showed that the distribution of the basis expansion weights given by the Galerkin method is

$$\boldsymbol{K}_\alpha \boldsymbol{w}|\boldsymbol{\Gamma} \;\sim\; \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Gamma}), \tag{2.14}$$

where

$$\boldsymbol{K}_\alpha = \begin{cases} \boldsymbol{K}_2 = \boldsymbol{L}, & \alpha = 2 \\ \boldsymbol{K}_\alpha = \boldsymbol{L}\boldsymbol{C}^{-1}\boldsymbol{K}_{\alpha-2}, & \alpha = 4, 6, 8, ... \end{cases}, \tag{2.15}$$

with $\boldsymbol{L} = \kappa^2\boldsymbol{C} + \boldsymbol{G}$ defined as in the Gaussian case. We note that this Section contains all the information needed to construct and fit an FEM solution to LMA models. We refer the reader to Bolin (2014) for a mathematically rigorous construction of the FEM solution.

There currently exists no extension of the LMA approximation for odd valued α's. We also point out that (2.14) results in a sparse precision matrix for small values of α. For $\alpha = 2$, the precision matrix, $\boldsymbol{K}_2\boldsymbol{\Gamma}^{-1}\boldsymbol{K}_2$, defines a sparse representation for the roughest covariance function offered by the finite element approximation to the LMA. This corresponds to Matérn covariance with $\nu = 1$ for spatial models in $\mathbb{R}^2$ and $\nu = 1.5$ for $\mathbb{R}$.

In summary, we see that the LMA model can be expressed as a sparse conditionally Gaussian distribution conditioned on Gamma-distributed auxiliary variables. We discuss the numerical issues and computational considerations implied by this approximation in Section 2.3.4.

2.3.4 Model Fitting: Continuous Space

Parameter estimation of LMAs is difficult since no closed form of the likelihood exists. Podgórski and Wegener (2011) proposed a method of moments-based estimation for LMA models. In Section 2.3.3 we showed that expressing the LMA as an SPDE allows for inference in the likelihood framework. Bolin (2014) constructed an EM algorithm for parameter estimation following the FEM approximation described in Section 2.3.3. Wallin and Bolin (2015) considered the use of an MCEM algorithm in order to provide a computationally efficient extension to SGLMMs. Persistent numerical issues contributed to difficult parameter estimation, in both the works of Bolin (2014) and Wallin and Bolin (2015). The issues stem from the fact that $\mathrm{E}\left[\Gamma_i^{-1}|\cdot\right]$ is unbounded for small $\min_{1\leq i\leq n} |\tau C_{ii} - 1/2|$. Bolin (2014) suggested truncating the expected value at 1000 to avoid

numerical instabilities.

To our knowledge, there has been no systematic Bayesian comparison of SGLMMs with LMAs and GRFs fit via MCMC. Samplers for GRFs require samples from the full-conditionals of the n-dimensional basis expansion weights $\boldsymbol{w}$, variance parameter ξ, spatial scale parameter κ^2, fixed effects $\boldsymbol{\beta}$, and a homogeneous random effect variance σ^2 (if applicable). Samplers for the continuous space LMA require n additional auxiliary variables ($\Gamma_i's$) and a shape parameter τ.

For the sake of illustration, consider a continuous response point referenced model with N unique observed locations denoted $\{\boldsymbol{u}_i\}_{i=1}^N$. Define the N by n projection matrix ($\boldsymbol{A}$) with entries $A_{ij} = \phi_j(\boldsymbol{u}_i)$, where $\{\phi_l(\boldsymbol{u})\}_{l=1}^n$ are triangular basis functions formed as described in Section 2.3.2. Using the basis expansion of $\eta(\boldsymbol{u})$ (see (2.7) Section 2.3.2) and assuming homogeneous error measurement $\epsilon(\boldsymbol{u}_i) \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$, we can write the discretized likelihood as $[\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{w}, \sigma^2] \sim \mathcal{N}(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{A}\boldsymbol{w}, \sigma^2\boldsymbol{I})$.

Conditioned on n auxiliary gamma random variables, Γ_i, we express the LMA as a scale mixture of normals with gamma variance. For $\alpha = 2$ it was shown in Section (2.3.3) equation (2.14) that $\boldsymbol{K}_2\boldsymbol{w}|\boldsymbol{\Gamma} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Gamma})$, where $\boldsymbol{\Gamma} = diag(\Gamma_1, ..., \Gamma_n)$. The conjugate full-conditional for the weights of the finite element approximation of the LMA model are

$$\left[\boldsymbol{w}|\boldsymbol{y}, \sigma^2, \boldsymbol{\beta}, \boldsymbol{\Gamma}\right] \sim \mathcal{N}\left(\left[\sigma^2\boldsymbol{L}\boldsymbol{\Gamma}^{-1}\boldsymbol{L} + \boldsymbol{A}'\boldsymbol{A}\right]^{-1}\left[\boldsymbol{y} - \boldsymbol{A}\boldsymbol{w}\right], \left[\boldsymbol{L}\boldsymbol{\Gamma}^{-1}\boldsymbol{L} + \left(\frac{1}{\sigma^2}\right)\boldsymbol{A}'\boldsymbol{A}\right]^{-1}\right) \quad (2.16)$$

Let $\boldsymbol{t} = \boldsymbol{K}_\alpha\boldsymbol{w}|\boldsymbol{\Gamma} \sim N(\boldsymbol{0}, \boldsymbol{\Gamma})$ as defined in equation (2.14) of Section 2.3.3. This leads to conjugate generalized inverse Gaussian (GIG) full conditionals for each gamma random variable, $\Gamma_i|t_i, \tau, \lambda \sim GIG(\tau C_{ii} - 1/2, 2/\lambda^2, t_i^2)$, where the *GIG(p,a,b)* density is given by

$$f(x; p, a, b) \propto x^{p-1}\exp\left(-\frac{ax + b/x}{2}\right). \quad (2.17)$$

Thus conjugate updates are available for the FEM weights $\boldsymbol{w}$ and the auxiliary variables (Γ_i) in the LMA model.

In practice we found that Gibbs updates for the conjugate $\Gamma_i's$ resulted in poor-mixing. This is not an uncommon issue, as conjugacy does not always produce an efficient sampler (Schliep and Hoeting, 2015). We found that the overall mixing of the Markov chains were improved by using independent one-at-a-time adaptively tuned Metropolis Hastings updates for each Γ_i. We found that sampling with Gibbs updates from the conditional distribution for Γ_i using the "GIGrvg" package in R for small GIG shape parameters

$(p = \tau C_{ii} - 1/2)$ frequently resulted in numerical underflow, producing samples of Γ_i that were numerically zero.

It is possible that samples from $\Gamma_i|t_i, \tau, \lambda$ are close to zero. This results in a numerically negative-definite precision matrix $\left[\boldsymbol{L}\boldsymbol{\Gamma}^{-1}\boldsymbol{L} + \left(\frac{1}{\sigma^2}\right)\boldsymbol{A}'\boldsymbol{A}\right]^{-1}$ for the full-conditional distribution in (2.16). To overcome this issue, we rejected the $\Gamma_i's$ if $\left[\boldsymbol{L}\boldsymbol{\Gamma}^{-1}\boldsymbol{L} + \left(\frac{1}{\sigma^2}\right)\boldsymbol{A}'\boldsymbol{A}\right]^{-1}$ was numerically rank-deficient. This amounts to a Metropolis Hastings update for $\boldsymbol{w}$ and $\boldsymbol{\Gamma}$ with the constraint that $\left[\boldsymbol{L}\boldsymbol{\Gamma}^{-1}\boldsymbol{L} + \left(\frac{1}{\sigma^2}\right)\boldsymbol{A}'\boldsymbol{A}\right]^{-1}$ is positive-definite.

In binary error response SGLMMs, Gaussian full-conditionals can be constructed using data augmentation (Albert and Chib, 1993). The LMA model can then be fit via MCMC exactly as described above. For Poisson error response distributions, the full-conditionals of $\boldsymbol{w}$ are not Gaussian. In this case, we suggest the use of conditionally independent one-at-a-time Metropolis Hastings updates for each $w_i|\boldsymbol{w}_{-i}$ (see Appendix A.3). We note that one-at-a-time conditionally independent block sampling is applicable for Gaussian and binary responses as well. However, in practice we found that block proposing $\boldsymbol{w}$ and $\boldsymbol{\Gamma}$ produced a faster and more efficient sampler for Gaussian and binary error responses. With the above numerical considerations we were able to fit the LMA models with a sampler that is a simple extension of traditional Gaussian samplers for each of point-referenced data analyses considered in Sections (2.5.3–2.5.4).

In summary, we have shown how LMA models with Matérn covariance can be expressed as an SPDE. We demonstrated that the FEM approximations for the GRF and LMA models result in sparse conditionally Gaussian representations. Following the numerical cautions detailed in this Section, we were able to fit the LMA using MCMC. This allows for Bayesian analyses familiar to traditional SGLMM models with GRFs.

2.4 LMA Models in Discrete Space

In the previous Section, we detailed the construction of the LMA model and provided a method for parameter estimation via MCMC. Wallin and Bolin (2015) acknowledged the potential use for LMAs in discrete space, but no further investigation was pursued. In this Section, we present a novel Bayesian hierarchical formulation for discrete space SGLMMs with LMAs. Our model is shown to be an adaptation of the widely recognized graph trend filtering (GTF) estimates proposed by Wang et al. (2016). Similar to the continuous space model, we introduce auxiliary variables to express the discrete space LMA as a conditionally Gaussian distribution. The resulting model exhibits computing and inference similar to its Gaussian analogue.

2.4.1 Graph Trend Filtering

Heavier tailed alternatives to GRFs have been shown to provide improved predictive power (Tibshirani et al., 2014; Faulkner and Minin, 2018; Wang et al., 2016) for discrete space models. Wang et al. (2016) extended the univariate trend filter of Kim et al. (2009) to irregular graphs. We provide a Bayesian extension of the graph trend filer (GTF) that is analogous to the LMA model for discrete space SGLMMs.

To motivate our discrete space model, we provide a brief overview of GTF estimates. Let G=(V,E) be a graph with vertices $i = \{1, ..., n\}$ and undirected edges $\{e_1, ..., e_m\}$. Suppose that $\boldsymbol{y} = (y_1, ..., y_n)$ are observed at the vertices. The GTF estimates, $\hat{\boldsymbol{\beta}} = (\hat{\beta}_1, ..., \hat{\beta}_n)$, are the solution to

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta} \in \mathbb{R}^n}{\arg\min} \frac{1}{2}||\boldsymbol{y} - \boldsymbol{\beta}||_2^2 + \lambda||\triangle^{(k+1)}\boldsymbol{\beta}||_1, \tag{2.18}$$

where $\triangle^{(k+1)}$ is a recursive graph difference operator of order k, and λ is a shrinkage parameter determined by cross-validation. For the case of $k = 0$, $\triangle^{(1)}$ is defined such that $||\triangle^{(1)}\boldsymbol{\beta}||_1 = \sum_{(i,j)\in E} |\beta_i - \beta_j|$ produces a first order difference penalty over G. The k^{th} order GTF differencing matrix seen in (2.18) is recursively defined as

$$\triangle^{(k+1)} = \begin{cases} \left(\triangle^{(1)}\right)' \triangle^{(k)}, & \text{for odd } k \\ \triangle^{(1)}\triangle^{(k)}, & \text{for even } k \end{cases}. \tag{2.19}$$

Higher orders of k in (2.19) correspond to higher order differencing penalties. For a chain graph, or a graph over a one-dimensional line, the GTF estimates reduce to the trend filter estimates of Kim et al. (2009).

In this Section 2.4.2, we propose a discrete space version of the LMA for SGLMMs. The GTF estimates are shown to be a special case of our model. In turn, our model specification serves as a generalized extension of the GTF to SGLMMs in the Bayesian framework.

2.4.2 LMAs in Discrete Space

Areal datasets are composed of aggregated responses. Examples of common spatial aggregations include cumulative measurements for cities, states, or countries. The areal units at which responses were recorded determine the discretization of space. The spatial discretization can be summarized by a graph with nodes at each areal unit and undirected

edges defined by the spatial neighborhood structure of the areal units.

The graph detailing the spatial neighborhood relationships can be expressed as a graph Laplacian matrix, $\mathbf{A}$, of the form

$$A_{ij} = \begin{cases} -1, & i \text{ is neighboring } j \\ \sum_k |A_{ik}|, & i = j \\ 0, & \text{else} \end{cases} . \tag{2.20}$$

The graph Laplacian ($\mathbf{A}$) serves as the precision matrix of the popular ICAR model for areal spatial random effect models (Besag, 1974). $\mathbf{A}$ is positive semi definite with rank n-1, where n is the number of observations. It is common to add a positive value to the diagonal to ensure that $\mathbf{A}$ is diagonally dominant. Define $\boldsymbol{L} = \kappa^2 \boldsymbol{I} + \boldsymbol{A}$. Let $\mathbf{D}$ be an upper triangular matrix such that $\boldsymbol{L} = \boldsymbol{D}'\boldsymbol{D}$ (i.e., $\mathbf{D}$ could be a Cholesky decomposition). We consider the discrete space precision matrices given by,

$$\boldsymbol{Q}_k = \left(\triangle^{(k)}\right)' \triangle^{(k)}, \tag{2.21}$$

where

$$\triangle^{(k)} = \begin{cases} \boldsymbol{L}^{\frac{(k+1)}{2}}, & \text{for odd } k \\ \boldsymbol{D}\boldsymbol{L}^{\frac{k}{2}}, & \text{for even } k \end{cases} . \tag{2.22}$$

Traditional simultaneously autoregressive (SAR) models (see Appendix A.2) assume a Gaussian prior for the n-dimensional random effect. We assume a Gaussian prior on the weighted sum of differences

$$\triangle^{(k)}\boldsymbol{\eta} | \xi^2, \kappa^2 \sim \mathcal{N}\left(0, \xi^2 \boldsymbol{I}\right), \tag{2.23}$$

which implies that

$$Cov\left(\boldsymbol{\eta} | \xi^2, \kappa^2\right) = \xi^2 \boldsymbol{L}^{-(k+1)} = \xi^2 \boldsymbol{Q}_k^{-1}. \tag{2.24}$$

The discrete space LMA is obtained by placing an *iid* Laplace prior on each sum of weighted differences $\triangle^{(k)}\boldsymbol{\eta} | \lambda, \kappa^2$. That is,

$$\triangle_l'^{(k)}\boldsymbol{\eta} | \lambda, \kappa^2 \overset{iid}{\sim} \mathcal{L}(\lambda), \quad l = 1, ..., n, \tag{2.25}$$

which implies that

$$Cov\left(\boldsymbol{\eta} | \lambda, \kappa^2\right) = \frac{2}{\lambda} \boldsymbol{L}^{-(k+1)} = \frac{2}{\lambda} \boldsymbol{Q}_k^{-1}. \tag{2.26}$$

We observe that the discrete space LMA model is obtained by replacing the Gaussian prior in (2.24) for an *iid* Laplace prior in (2.26).

2.4.3 L2 vs. L1 Penalization

In the proceeding Section we demonstrated that the discrete space LMA is obtained by replacing the Gaussian prior on the weighted sum of differences in (2.23) with *iid* Laplace priors in (2.25). Here we detail the penalization implications that result from altering the prior. The contrast in penalizations is presented to provide intuition as to why the LMA model should be considered in place of the the GRF model.

Let $N(i) := \{j : j \text{ is a neighbor of } i\}$ and $\boldsymbol{\theta}_G = (\xi^2, \kappa^2, \sigma^2)$. The order $k = 1$ discrete space GRF has log full-conditional distribution

$$\log[\boldsymbol{\eta}|\boldsymbol{\beta}, \boldsymbol{y}, \boldsymbol{\theta}_G] \approx \log[\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{\eta}, \boldsymbol{\theta}_G] - \frac{1}{2\xi^2}\sum_{i=1}^{n}(\kappa^2\eta_i + \sum_{j\in N(i)}(\eta_i - \eta_j))^2 + Const. \quad (2.27)$$

Notice that (2.27) resembles penalized regression with a squared penalty term placed on the weighted sum of differenced nodes (η_i). If $\kappa^2 > 0$ in (2.27), $\boldsymbol{\eta}$ is penalized based on the sum of the localized differences relative to the magnitude of each node. We recover the intrinsic conditionally auto-regressive model (ICAR) by taking $\kappa^2 = 1$ and $k = 0$.

Now define $\boldsymbol{\theta}_L = (\lambda, \kappa^2, \sigma^2)$. The order $k = 1$ discrete space LMA log full-conditional is given by

$$\log[\boldsymbol{\eta}|\boldsymbol{\beta}, \boldsymbol{y}, \boldsymbol{\theta}_L] \approx \log[\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{\eta}, \boldsymbol{\theta}_L] - \frac{\lambda}{2}\sum_{i=1}^{n}|\kappa^2\eta_i + \sum_{j\in N(i)}(\eta_i - \eta_j)| + Const. \quad (2.28)$$

Our inclusion of κ^2 allows for a full rank precision matrix, as well as an L-1 penalty based on the magnitude of each node, η_i. The Bayesian equivalent to the k^{th} order GTF estimates proposed by Wang et al. (2016) are obtained by setting $\kappa^2 = 0$ in (2.28) . We also point out that for $\kappa^2 = 1$ and $k = 0$, we recover the Laplace analog to the ICAR model.

We note that the LMA model resembles a LASSO style (L-1) penalty while the GRF model resembles a ridge like (L-2) penalty on the sum of differenced nodes (η_i). There appears to be an analogous extension to be made for penalization in the continuous space. For a discrete space model on a regular grid, lattice refinement results in convergence to the continuous space LMA.

2.4.4 Model Fitting: Discrete Space

We aim to preserve the familiarity of the Gaussian fit for our discrete space model. To do so, we first recognize the Laplace distribution as a scale mixture of normals. This can be seen by taking a Gaussian variable, $Z \sim N(0, S_i)$, where S_i is an independent exponential random variable with rate $\lambda^2/2$. Marginalizing over S_i, it follows that $Z|\lambda \sim \mathcal{L}(\lambda)$.

Recall, the discrete space LMA places an $\mathcal{L}(\lambda)$ prior on the l^{th} weighted difference, $\triangle_l^{(k)}\boldsymbol{\eta}|\lambda, \kappa^2 \overset{iid}{\sim} \mathcal{L}(\lambda)$. By introducing n auxiliary variables $S_i \overset{iid}{\sim} Exp(\lambda^2/2)$, we can express the prior as

$$\triangle^{(k)}\boldsymbol{\eta}|\boldsymbol{S}, \kappa^2 \sim N(0, \boldsymbol{S}), \quad \boldsymbol{S} = diag(S_1, .., S_n). \tag{2.29}$$

Park and Casella (2008) showed that S_i^{-1} are conjugate inverse Gaussian ($InvGauss$). To see this, first define $\boldsymbol{t} = \triangle^{(k)}\boldsymbol{\eta}|\boldsymbol{S}, \kappa^2$. Then $S_i^{-1}|t_i, \lambda \sim InvGauss\left(\lambda^2, \sqrt{\frac{\lambda^2}{t_i^2}}\right)$, where the $InvGauss(a, b)$ has density

$$f(x; a, b) = \left(\frac{b}{2\pi x^3}\right)^{1/2} \exp\left(-\frac{b(x-a)^2}{2b^2 x}\right). \tag{2.30}$$

We note that the *InvGauss* is a special case of the *GIG* with $p = -1/2$. We fit the discrete space LMA and GRF via MCMC as well. From (2.29), we observe that the discrete space LMA model can be expressed as a conditionally sparse Gaussian distribution. Additionally, the full-conditionals for the auxiliary random variables are conjugate. In turn, Gibbs updates can be used for the full-conditionals of the n auxiliary random variables. In Section 2.3.4 we discussed the numerical issues induced by using conjugate updates for the auxiliary mixing variables. This is not an issue in the discrete space sampler, as the *InvGauss* does not require estimation of a shape parameter.

In summary, we have provided a novel discrete space Bayesian hierarchical LMA. The LMA is conditionally Gaussian, as in continuous space, again allowing for inference and computing familiar to GRF models. In Section 2.5 we demonstrate that the discrete space LMA offers improved out-of-sample predictive power in the presence of localized trends.

2.5 Example Analyses

In this section we compare the LMA and GRF models over four datasets; including Gaussian, Poisson, and binary responses observed both on continuous (point referenced) and discrete (areal) support. We perform 10-fold cross validation by randomly splitting the data set into 10 roughly equal sized groups $A_1, ..., A_k$. We withhold a test set $\boldsymbol{y}_{A_k}$ and train the model on the remaining observations $\boldsymbol{y}_{A_k^c}$. We use the Bayesian cross validation scoring criterion (BCVS) of Hooten and Hobbs (2015) given by

$$BCVS = -\sum_{k=1}^{10} \log\left(\frac{\sum_{t=1}^{T}[\boldsymbol{y}_{A_k}|\boldsymbol{y}_{A_k^c}, \boldsymbol{\theta}^{(t)}]}{T}\right), \tag{2.31}$$

where T is the total number of stored MCMC iterations, and $\boldsymbol{\theta}^{(t)}$ is the t^{th} sample of the model's parameters. Note that a smaller value implies a better model fit.

We used effective samples per second (ESS) to compare the computational performance of each model fit. All MCMC iterations had a burn-in phase in which the normal proposals were adaptively tuned according to Roberts and Rosenthal (2009). Every data analysis involved 50,000 post-burn-in MCMC states. The number of burn-in iterations for the GRF and LMA model were held constant for each individual data analysis. The ESS was computed based on the likelihood evaluated at the stored 50,000 states.

2.5.1 Discrete Space Examples: Slovenia Stomach Cancer

We compare models trained on areal count data with the Slovenian stomach cancer outbreak dataset. The dataset consists of 194 responses of aggregated stomach cancer counts in each municipality of Slovenia collected from 1994 to 2001. The model proposed by Hodges and Reich (2010) was fit to investigate the relationship between standardized socioeconomic status, SEc_i, and the occurrence of stomach cancer, y_i.

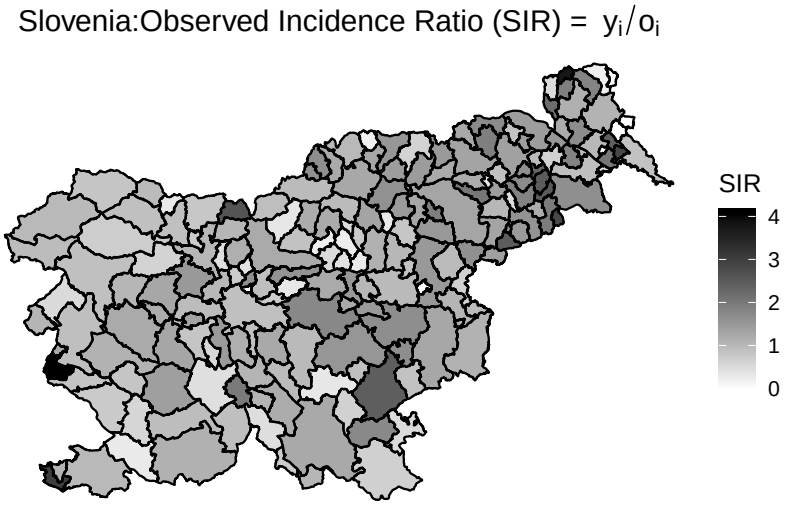

Figure 2.2: Plot of observed incidence ratio of stomach cancer (SIR), reported as the ratio of observed occurrences divided by the expected count in municipality i.

We fit the Poisson spatial model considered by Hodges and Reich (2010) assuming stomach cancer counts are Poisson, $y_i \sim Poisson(\mu_i)$. The mean μ_i is modeled through the log-link function

$$\log(\mu_i) = \log(o_i) + \boldsymbol{x}_i'\boldsymbol{\beta} + \eta_i + \epsilon_i. \tag{2.32}$$

The offset term (o_i) in (2.32), is the expected number of stomach cancer counts for observation i. The covariate matrix ($\boldsymbol{X}$) contains an intercept and the standardized socioeconomic status (SEc).

We use an order $k = 1$ differencing matrix $\triangle^{(1)}$ (see equation (2.22) in Section 2.4.2) to define the priors for the spatially correlated random effects in the GRF and LMA models defined in equations (2.23) and (2.25) of Section 2.4.2. The fixed effects ($\boldsymbol{\beta}$) are assigned *iid* normal priors with variance 10^6. The variance of the spatially homogeneous random effect (σ^2) is assigned an inverse gamma prior with shape and scale of one. The scale parameter for the GRF model (ξ), the scale parameter for the LMA model (λ), and the spatial scale parameter (κ) are all assigned independent half-normal priors with scale one. The full-conditionals for the Poisson response data are detailed in Appendix A.4.

Inference on the fixed effects is similar between the LMA and GRF models (see Table 2.3). Table 2.1 shows that the GRF provides a better fit than the LMA. This is perhaps due to the areal dataset being smooth. We do note that the LMA sampler produces roughly the same ESS.

Model	BCVS	ESS
GRF	**560.61**	**4.12**
LMA	563.04	3.81

Table 2.1: BCVS and ESS for ten-fold cross validation on the Slovenia stomach cancer outbreak dataset.

2.5.2 Discrete Space Examples: Columbus Crime Data

The Columbus crime data are found in the *"spdep"* R package (Bivand et al., 2013). The dataset provides a spatial map with crime rates (y_i) for each county of Ohio. Ver Hoef et al. (2018) suggested modeling the data as an intercept only model with Gaussian response. We include average household income and average household value, as well as an intercept as covariates. Figure 2.3 shows a plot of the crime rates in each county of Ohio. There appears to be a few localized spikes in the data, namely the counties that appear in white. This is sufficient reason to suspect that the LMA model should provide an improved model fit.

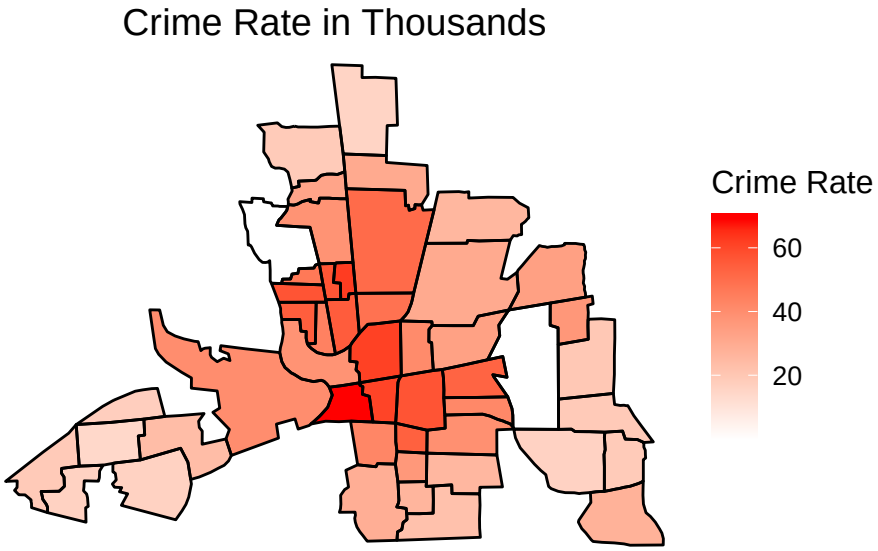

Figure 2.3: Plot of crime rate in thousands in the 49 counties of Columbus, Ohio.

The model is of form

$$y_i = \boldsymbol{x}_i'\boldsymbol{\beta} + \eta_i + \epsilon_i. \tag{2.33}$$

We use the adjacency matrix ($\mathbf{A}$) provided from the "*spdep*" package to construct the

first order differencing matrix $\triangle^{(1)}$ (see equation (2.22) in Section 2.4.2). The precision matrices for the priors of the GRF and LMA models are constructed from $\triangle^{(1)}$ as in Section 2.4.2. We assumed *iid* normal variance 10^3 priors for the fixed effects. For both models κ^2 was assigned a half-normal scale one prior. The variance parameter for the GRF model (ξ^2) and the variance parameter for the LMA model (λ^2) were assigned half normal scale 10 priors. The full-conditionals for all GRF and LMA model parameters can be found in Appendix A.5.

In Table 2.2, we see that the LMA model is the clear favorite in terms of BCVS. Inference on the fixed effects $\boldsymbol{\beta}$ is similar in both cases (see Table 2.3). The ESS for the LMA model seems substantially smaller than the GRF model. We note that the data set is only of size 49, so this difference may be a bit inflated. The ESS for both models is quite large for an MCMC sampler.

Model	BCVS	ESS
GRF	341.58	**18.13**
LRF	**329.85**	14.07

Table 2.2: BCVS from ten-fold cross validation and ESS for Columbus Ohio Crime dataset.

Slovenia Stomach Outbreak					
Predictor	Parameter	GRF Estimate	95% CI	LMA Estimate	95% CI
Intercept	β_0	0.097	$(-0.036, 0.202)$	0.115	$(-0.010, 0.339)$
SEc	β_1	-0.078	$(-0.170, 0.031)$	-0.068	$(-0.162, 0.040)$
	ξ	0.898	$(0.492, 1.561)$	NA	NA
	σ	0.290	$(0.238, 0.349)$	0.282	$(0.230, 0.339)$
	κ	1.427	(0.616, 2.523)	1.608	(0.814, 2.543)
	λ	NA	NA	1.127	(0.718, 1.701)
Columbus Crime Data					
Predictor	Parameter	GRF Estimate	95% CI	LMA Estimate	95% CI
Intercept	β_0	35.107	(-27.661, 76.053)	29.379	(-39.597, 83.936)
Avg. Inc	β_1	-0.321	(-0.391, -0.252)	-0.235	(-0.327, -0.149)
Avg. Value	β_2	-0.981	(-1.247, -0.716)	-1.057	(-1.351, -0.755)
	ξ	4.125	(3.536, 4.676)	NA	NA
	σ	3.370	(3.191, 3.550)	2.803	(2.541, 3.050)
	κ	0.162	(0.051, 0.298)	0.219	(0.081, 0.373)
	λ	NA	NA	4.582	(4.086, 5.040)

Table 2.3: Parameter estimates for discrete space data analysis examples of Sections (2.5.1–2.5.2)

2.5.3 Continuous Space Examples: Malaria in the Gambia, Africa

A model comparison for the binary continuous case is illustrated with the use of presence absence data of malaria in the Gambia, Africa. The data was made publicly available by Diggle and Ribeiro Jr (2007) in the *"geoR"* package of *R* (Bivand and Rundel, 2018). The dataset consists of 2035 children records recorded at 65 village locations denoted $\{\boldsymbol{u}_i\}_{i=1}^{65}$. Each village, i, has n_i, respondents. We consider a model similar to that of Hanks et al. (2015). Let $y_j^{(i)}$ be the indicator for the presence ($y_j^{(i)} = 1$) of malaria in the j^{th} child at village location i. The covariates considered for each child are composed of the child's age, an indicator of whether or not a bed net was used, an indicator for whether or not an insecticide was applied to the bed net, the log normalized difference vegetation index (NDVI) at each village, and an indicator of presence or absence of a health center in the village.

We construct a triangular mesh with $n = 288$ nodes containing all 65 village locations on a node (see Figure 2.4). We specify a binary probit model with $y_j^{(i)} \sim \text{Bernoulli}\left(p_j^{(i)}\right)$, where the probability of malaria presence in the j^{th} child at village location i is linked

through the probit function (standard normal CDF) and auxiliary data

$$z_j^{(i)} = \boldsymbol{x}_j'^{(i)}\boldsymbol{\beta} + \eta(\boldsymbol{u}_i). \tag{2.34}$$

The superscripts for each variable denote the village location $\boldsymbol{u}_i$. We fix $\alpha = 2$ leading to covariance matrices for the GRF and LMA priors for $\eta(\boldsymbol{u})$ in (2.34) given by (2.9) and (2.15) respectively. We assume *iid* normal variance 10 priors for all fixed effects. We follow the data augmentation approach of Albert and Chib (1993) for binary probit GLMs to obtain Gibbs updates for the fixed and random effects. The priors for hyper-parameters $\xi, \kappa, \tau, \lambda$ are all *iid* half normal scale one. Further details pertaining to prior specifications and model fitting are included in Appendix A.6.

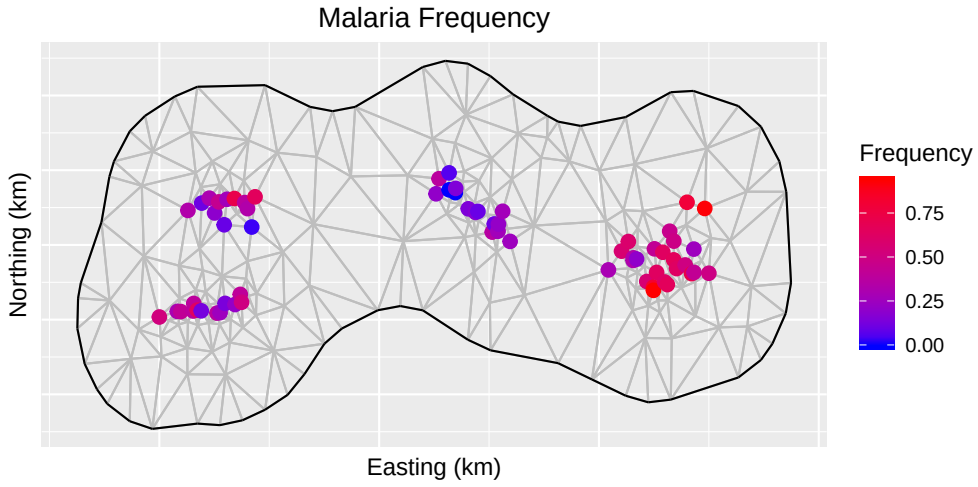

Figure 2.4: Plot of triangular mesh with $n = 288$ nodes and malaria frequency at 65 unique village locations.

For this dataset, we randomly split the dataset into 10 groups of village locations. All observations associated with a given test set were withheld for validation. Table 2.4 shows that the LMA model provides the better fit based on BCVS. In this case the LMA also provided a higher ESS. This may be due to the use of a data augmentation approach suggested by Albert and Chib (1993). Estimates for all model parameters can be found in Table 2.6.

Model	BCVS	ESS
GRF	1292.20	2.66
LMA	**1284.08**	**2.90**

Table 2.4: BCVS from ten-fold cross validation and ESS for malaria in the Gambia, Africa.

2.5.4 Continuous Space Example: LAGOS

Lastly, we compare the performance of the LMA for the continuous response LAGOS lake dataset. The Lake multi-scaled geospatial and temporal database (LAGOS) is a publicly accessible US lake water quality database (Soranno et al., 2017). The dataset used in this paper contains records for 5526 unique lakes observed over Iowa, Missouri, and Illinois at locations $\{\boldsymbol{u}_i\}_{i=1}^{5526}$. We are interested in modeling the log total phosphorus recorded in each lake. First we reduce the many covariates recorded for each lake in the LAGOS database by performing a step-wise regression assuming uncorrelated residuals. The covariates selected by AIC in this stepwise procedure are: an intercept, lake area (in hectares), max depth of the lake (meters), mean runoff (ground-water discharge into streams), the average annual runoff (in/yr), inter lake water shed (IWS) measurements (the area of land that drains directly into a lake) for urban, agricultural, road density, and total wetland regions. We then used these selected covariates to model log total phosphorus ($y(\boldsymbol{u})$) as

$$y(\boldsymbol{u}_i) = \boldsymbol{x}_i'\boldsymbol{\beta} + \eta(\boldsymbol{u}_i) + \epsilon(\boldsymbol{u}_i). \tag{2.35}$$

We fit the LMA model and GRF model by fixing $\alpha = 2$. Again this corresponds to Matérn smoothness parameter $\nu = 1$. We construct a triangular mesh with $n = 671$ nodes according to Section 2.3.2. The fixed effects $\boldsymbol{\beta}$ are assigned a normal prior with variance 10^3. The nugget term (σ^2) and spatial scale parameter (κ) are assumed to follow independent half-normal distribution with scale one. The variance term of the GRF (ξ), as well as the variance (λ) and scale (τ) parameters of the LMA are assigned independent half-normal scale one priors.

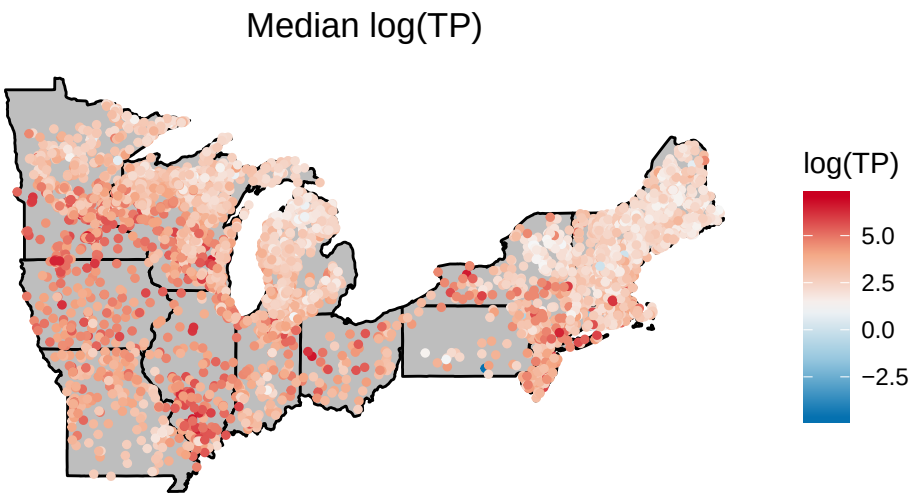

Figure 2.5: Plot of median log total phosphorus (TP) recorded at 5526 unique lake locations.

Table 2.5 shows that the LMA model provides a better fit than the GRF in this case. The ESS is better for the LMA as well, further suggesting the LMA model should be favored. As in all other cases, the fixed effect estimates remain roughly similar. All model parameter estimates can be found in Table 2.6. Details of the LAGOS data analysis and all full-conditionals are included in Appendix A.7.

Model	BCVS	ESS
GRF	5733.309	0.35
LMA	**5727.993**	**0.39**

Table 2.5: BCVS for ten fold cross validation and ESS for the LAGOS dataset.

Malaria of the Gambia					
Predictor	Parameter	GRF Estimate	95% CI	LMA Estimate	95% CI
Intercept	β_0	0.123	(-2.408, 2.677)	-0.975	(-4.563, 2.554)
Age	β_1	0.163	(0.111, 0.215)	0.178	(0.124, 0.232)
Bed Net Use	β_2	-0.262	(-0.413, -0.109)	-0.218	(-0.398, -0.036)
Insecticide Use	β_3	-0.104	(-0.270, 0.062)	-0.211	(-0.417, -0.007)
log(NDVI)	β_4	-0.007	(-0.659, 0.637)	0.283	(-0.633, 1.220)
Health	β_5	-0.253	(-0.397, -0.109)	-0.089	(-0.297, 0.119)
	ξ	2.759	(1.612, 4.203)	NA	NA
	κ	1.413	(0.399, 2.433)	1.192	(0.518, 1.759)
	τ	NA	NA	13.272	(12.074, 14.540)
	λ	NA	NA	4.730	(3.506, 5.824)
LAGOS					
Predictor	Parameter	GRF Estimate	95% CI	LMA Estimate	95% CI
Intercept	β_0	3.93	(3.773,4.087)	3.925	(3.771,4.081)
Area	β_1	6.33e-05	(5.00e-05,7.71e-05)	6.33e-05	(4.99e-05,7.69e-05)
Max depth	β_2	-0.036	(-0.038,-0.034)	-0.036	(-0.038,-0.034)
IWS Urban	β_3	1.296	(1.103,1.489)	1.302	(1.105,1.498)
IWS Ag	β_4	1.594	(1.485,1.701)	1.598	(1.491,1.707)
IWS Wetland	β_5	0.635	(0.448,0.822)	0.639	(0.454,0.826)
Road Density	β_6	-0.002	(-0.003,-0.001)	-0.002	(-0.003,-0.001)
Baseflow	β_7	-0.012	(-0.014,-0.01)	-0.012	(-0.014,-0.01)
Runoff	β_8	-0.026	(-0.03,-0.021)	-0.026	(-0.03,-0.021)
	ξ	0.383	(0.220 , 0.580)	NA	NA
	σ^2	0.464	(0.444 , 0.485)	0.464	(0.444 , 0.485)
	κ	0.772	(0.397 , 1.178)	0.724	(0.477 , 0.988)
	τ	NA	NA	1.751	(1.322 , 2.263)
	λ	NA	NA	0.266	(0.184 , 0.38)

Table 2.6: Parameter estimates for continuous space data analysis examples of Sections (2.5.3–2.5.4)

2.6 Discussion

In this work we proposed a novel discrete space LMA model for irregular lattices and constructed Metropolis Hastings samplers for continuous and discrete space SGLMMs with LMAs. Using the Metropolis Hastings samplers, we provided a broad Bayesian analysis of SGLMMs with LMAs for continuous, binary and Poisson error responses. We found that the LMA offered a better fit than the GRF for datasets which exhibit "spikes" in the response. Provided our MCMC implementation, we saw that the LMA offers similar computational performance to the GRF.

In this work we restricted our attention to the symmetric LMA. Bolin and Wallin (2016) considered more general continuous space LMA which allowed for asymmetric

posterior distributions. We note that the asymmetry parameter can be estimated in the MCMC framework, however we elected to compare symmetric models only. We also note that the extension of the asymmetric LMA to discrete space models on irregular graphs could be considered as well. We leave this for future work, but note that the extension should be straightforward.

The choice of half-normal priors for each of the complexity parameters in the GRF and LMA models, given by ξ, κ, τ, and λ, were made for the sake of comparison. Fuglstad et al. (2015) proposed the use of a joint prior on the variance ξ and scaling parameter κ^2 in the Gaussian case following the work of Simpson et al. (2017). We note that their prior choice was motivated by the desire to provide a weakly informative prior and deal with a partial identifiability issue associated with the two parameters. Though identifiability issues with the complexity parameters persist, one does not usually concern themselves with the value of the parameter. We found that independent half-normal priors resulted in the best mixing.

In summary, we have developed a novel discrete space SGLMM with LMAs. We have proposd the use of Metropolis Hastings samplers to fit the LMA models as a simple extension of the GRF models. Through our extensive data analyses, we have provided evidence of cases in which the LMA model outperforms the traditional GRF SGLMM, while maintaining similar computation efficiency. Due to the comparable computation times and similarity of implementation, we recommend LMA models be considered when modeling correlated error structures.

Chapter 3
Privacy for Spatial Point Process Data

[1]

3.1 Introduction

Individual disease case data offer scientists valuable information about the dynamics of on-going and past disease outbreaks. Due to privacy risks, disease data are often not made publicly available. Synthetic data sets offer reduced disclosure risks while preserving some of the scientific utility of the confidential data set (Wang and Reiter, 2012; Quick et al., 2015). In this work, we consider spatial disease case data: that is we consider data consisting of the spatial locations of each individual with a particular disease. We develop methods for privatizing this fundamental form of disease data. We propose two novel Bayesian methods for generating synthetic individual location data based on log-Gaussian Cox processes (LGCPs). We demonstrate that leave-one-out LGCP densities can easily be approximated by conditional predictive ordinates (CPOs). We propose a novel disclosure risk metric based on a leave-one-out intrusion scenario, which is quickly evaluated by our derived CPO estimates. We propose a model-based utility metric tailored to LGCPs built off the propensity mean square error ($pMSE$) explored by Snoke and Slavković (2018) to assess the scientific quality of a synthetic data set. We motivate our work with Dr. John Snow's renowned cholera outbreak dataset, which is, notably, one of the only openly available datasets consisting of individual disease

[1]This work is in review at *Journal of Computational and Graphical Statistics*. This work is also available on arxiv: Walder, A., Hanks, E. M., & Slavković, A. (2020). Privacy for Spatial Point Process Data. arXiv preprint arXiv:2003.12816.

case locations.

In 1854, the small suburb of Soho, London was plagued by a massive cholera outbreak. When a wave of cholera first hit London in 1831, the scientific community at the time assumed that the disease was spread by “miasma in the atmosphere" (Summers, 1989). Dr. John Snow, an obstetrician, sought to convince the town’s officials that the epidemic was in fact water borne. To do so, Dr. Snow began by constructing a map of Soho (see Figure 3.1). He then went door-to-door collecting, and plotting, each of the 578 reported cholera death incidences. From this map, Dr. Snow concluded that the Broad St. water pump was the source of the outbreak. This analysis was used to convince town officials to close the water pump.

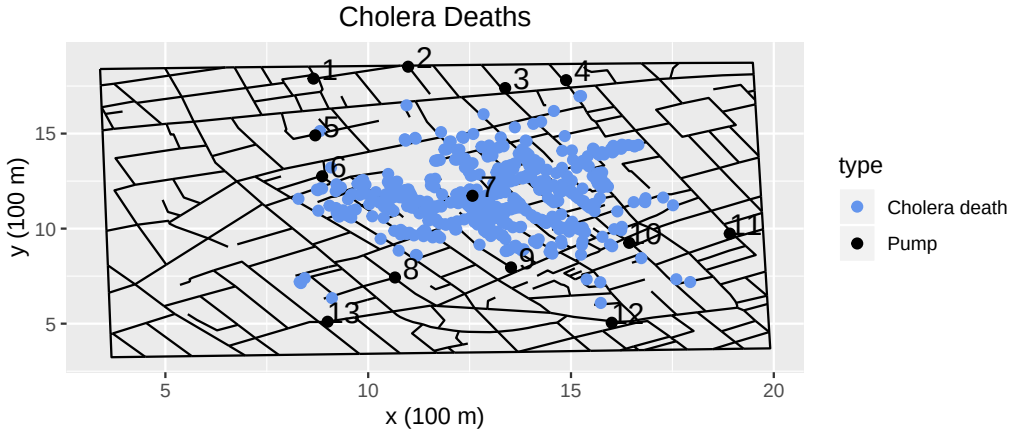

Figure 3.1: A map of the 578 observed cholera deaths, streets, and water pumps in Soho, London. The Broad St. water pump (pump 7) was the source of the outbreak.

Access to individual locations of each disease incidence allows for scientific analyses to accurately capture the geographic trends associated with a disease outbreak. Congress mandated the Health Insurance Portability and Accountability Act of 1996 (HIPAA) establishing standards for the privacy of individually identifiable health information (DHHS, 2016). For disease location data, individual identification is equivalent to knowing one’s individual location. For this reason, agencies are legally obligated to ensure that proper privacy constraints are met before disseminating individual disease occurrence locations.

One of the most common methods used to privatize disease case locations prior to data dissemination is radial perturbation (Armstrong et al., 1999; Wang and Reiter, 2012). In radial perturbation, each location is randomly perturbed within a radius r of their true location. It has been argued that radial perturbation poses two problems; 1) If

perturbations of large radius r are required for strong privacy guarantees, the underlying spatial structure is destroyed; 2) For sparsely populated regions, small perturbations offer weak privacy guarantees (Wang and Reiter, 2012; Quick et al., 2015, 2018). Though the two stated drawbacks of radial perturbation seem to be agreed upon (Raghunathan et al., 2003; Wang and Reiter, 2012; Quick et al., 2015), these points have not been clearly illustrated by any empirical studies. In this work we demonstrate the fallacies of radial perturbation by providing the first Bayesian analysis of risk vs. utility of radially perturbed locations.

Disease case locations which have been perturbed to satisfy legal constraints may be of low scientific quality. An ideal dataset will offer low disclosure risks and provide inference similar to that of the confidential dataset. Wang and Reiter (2012) proposed the use of CART models to generate synthetic locations from approximate conditional distributions of each location given the attributes of individuals in the data set. Since CART models draw locations independently from the response distribution they could produce a synthetic data set that contains two records that are spatially close in the confidential set but distant in the synthetic set (Wang and Reiter, 2012). Although the CART models can preserve some spatial structure in the confidential data, they can miss localized spatial dependencies (Quick et al., 2015). Quick et al. (2015) used LGCPs in place of CART models to preserve the spatial dependence structure of the confidential set. Quick et al. (2015) generated synthetic locations by simulating from the fitted LGCP model with posterior mean estimates for each parameter in the model. We propose a synthesis method that adds random spatial noise to the intensity surface of the LGCP. This method contains the synthesis method of Quick et al. (2015) as a special case when the variance of the random spatial noise is set to zero. We also propose a second synthesis method that includes a resampled spatial random field in the intensity surface. We show that these novel synthesis methods provide improved disclosure risks relative to the methods of Quick et al. (2015) and radial perturbation.

Synthetic data sets should offer quantifiable disclosure risks for every individual involved in the study. Quick et al. (2015) defined a risk metric for fully-synthetic locations by conditioning on the attributes of each individual. This metric is not defined for location-only data sets, which are common for disease outbreaks and are the case we consider in this study. To assess disclosure risks we consider an intrusion scenario in which an intruder attempts to identify a confidential location within a radius r of the truth given 1) knowledge of the synthesis method, 2) unique identification of all but one confidential location, and 3) the released synthetic dataset. In turn, our risk metric

requires the evaluation of a leave-one-out posterior predictive density. Leave-one-out scenarios are commonly used to assess privacy risks and are central to privacy (Wang and Reiter, 2012; Dwork et al., 2014).

Obtaining leave-one-out predictive densities is often a computationally burdensome task for high dimensional data sets and spatial models, as individual model fits are required for every observation in the data set. CPO estimates rely on independent marginal distributions to estimate the leave-one-out posterior predictive density for each observation by Monte Carlo approximation using samples from the posterior density of the full data set (Hooten and Hobbs, 2015). In turn, CPO estimates are easy to obtain when independence is assumed for each observation. In the existing literature, dependence structures commonly assumed in spatial models make CPO estimates computationally expensive (Hooten and Hobbs, 2015). In Section 3.4 we derive CPO estimates for Cox processes with dependence structures, allowing for fast approximations of posterior predictive leave-one-out LGCP densities for spatial point process data. This provides a general approach for model selection for spatial point process data, and in Section 3.7.1 we demonstrate that each individual disclosure risk can be easily approximated with CPO estimates using samples from the posterior conditioned on the full data set. This provides a novel computationally efficient approach to assessing disclosure risks in spatial location data.

Synthetic data sets should satisfy disclosure risk requirements and offer meaningful scientific inference. Woo et al. (2009) and Snoke et al. (2018) developed the propensity mean square error ($pMSE$) to assess the distributional similarity between a synthetic and confidential data set. The $pMSE$ is the mean square error of the predicted probability that a given observation belongs to the confidential set. A synthetic data set that is indistinguishable from the confidential set is said to be of high utility. We propose a model-based utility metric that tailors the $pMSE$ statistic to LGCPs.

In a case study of Dr. John Snow's cholera outbreak dataset, we demonstrate that both of our proposed synthesis methods offer datasets with reduced disclosure risks and higher data utility than radial perturbation. To our knowledge, we offer the first Bayesian analysis of radial perturbation. We also show that our synthesis methods offer improved disclosure risks relative to the approach of Quick et al. (2015). In summary, our contributions in this work include

1. Two novel methods for generating synthetic point process data based on log-Gaussian Cox processes (LGCPs).

2. A derivation of conditional predictive ordinate (CPO) estimates for point process data.

3. A novel metric for evaluating privacy risks based on CPO estimates.

4. An adaptation of the propensity mean square error ($pMSE$) data utility metric tailored to point process data.

The remainder of this manuscript is organized as follows. In Section 3.2 we outline the process of generating and disseminating synthetic locations. In Section 3.3 we provide background information on the LGCPs models considered in this work. In Section 3.4 we demonstrate that the CPO can be readily obtained for any Cox process. In Section 3.5–3.6 we introduce our proposed synthesis methodologies and detail our model based utility metric. In Section 3.7 we introduce our novel risk metric and detail how individual disclosure risks are assessed for each synthesis method using CPO arguments. In Section 3.8 we illustrate our methodology with a case study of Dr. John Snow's cholera dataset. We conclude with a discussion in Section 3.9.

3.2 Data Synthesis and Dissemination

In this section we outline the process of generating, evaluating, and disseminating synthetic location-only data. We begin by fitting the confidential dataset according to a LGCP with intensity surface $\lambda(\cdot)$ as detailed in Section 3.3. The LGCP allows for a model that includes any desired covariates and captures spatial auto-correlation.

Once the model is fitted, synthetic locations are simulated from a LGCP with intensity $\lambda^{\dagger}(\cdot)$, determined from the confidential model fit. The LGCP intensity surface, $\lambda(\cdot)$, includes a zero-mean Gaussian spatial random effect with covariance parameterized by $\boldsymbol{\theta} = (\kappa^2, \tau^2)$. Our first proposed synthesis method re-samples the spatial random effect from a zero-mean Gaussian distribution with covariance determined by plug-in estimates $\hat{\boldsymbol{\theta}} = (\hat{\kappa}^2, \hat{\tau}^2)$. Our second novel synthesis method involves adding random Gaussian noise to the posterior mean estimate of the summary statistic $\hat{\lambda}(\cdot)$ from the confidential set. Synthetic locations are obtained by sampling from a LGCP with synthetic intensity $\lambda^{\dagger}(\cdot)$.

Prior to disseminating data, statistical agencies are required to ensure that subjects' can not be unwillingly identified (DHHS, 2016). We introduce an individual disclosure risk metric to quantify how much an intruder can learn about the confidential locations from a synthetic data set. We assess individual disclosure risks by considering the intrusion scenario in which the intruder has gained knowledge of 1) the synthetic dataset

$\mathcal{S}^{\dagger}$, 2) complete information about the synthesis method, and 3) has uniquely identified all but one confidential location $\mathcal{S}/\boldsymbol{s}_k$. The intruder attempts to identify the final confidential location within a radius r of the truth given the synthetic dataset and all but one confidential location. The risk is taken to be the max individual disclosure risk for $\mathcal{S}/\boldsymbol{s}_k$ given $\mathcal{S}^{\dagger}$ and complete information about the synthesis method. To evaluate this risk, we integrate the marginal density $\pi\Big(\boldsymbol{s}, N|\mathcal{S}/\boldsymbol{s}_k, \mathcal{S}^{\dagger}\Big)$ over a ball of radius r centered at confidential location $\boldsymbol{s}_k$. We require samples from the posterior distributions of $\pi\Big(\lambda|\mathcal{S}^{\dagger}, \mathcal{S}/\boldsymbol{s}_k, N\Big)$ to evaluate the risks for each of the N confidential locations in the Bayesian framework. We use samples from the posterior distribution obtained from a simultaneous fit of all the data, $\pi\Big(\lambda, \lambda^{\dagger}|\mathcal{S}^{\dagger}, \mathcal{S}, N\Big)$, to estimate the CPO $\pi\Big(\boldsymbol{s}, N|\mathcal{S}^{\dagger}, \mathcal{S}/\boldsymbol{s}_k\Big)$. We demonstrate that CPO estimates can be used to approximate the disclosure risks for each synthesis method considered in Section 3.7.

The goal of this work is to provide approaches to disseminate fully synthetic location-only data that offer low disclosure risks and provide scientific inference similar to that of the confidential set. Once a synthetic dataset is generated, we assess the quality (utility) of a synthetic dataset by computing the *pMSE* statistic to quantify how well a given synthetic dataset emulates the confidential dataset. We then evaluate the maximum disclosure risk for our given intrusion scenario. An optimal synthetic dataset will provide a small maximum disclosure risk, while maintaining high data utility. We repeat the processes of data generation and risk/utility assessment until an acceptable balance of data utility vs. maximum disclosure risk has been obtained. A synthetic dataset that satisfies the desired privacy vs. utility trade-off is then released along with all information related to the synthesis process.

3.3 LGCPs

In this section we provide a brief background on LGCPs and the computational details associated with model fitting. A Cox process is a point process governed by a non-negative stochastic process $\Lambda = \{\boldsymbol{s} \in \mathbb{R}^2 : \Lambda(\boldsymbol{s})\}$. Conditioned on the realization $\Lambda(\boldsymbol{s}) = \lambda(\boldsymbol{s})$, the point process is a Poisson process with intensity $\lambda(\boldsymbol{s})$. Cox processes are natural models for point process phenomena that are environmentally driven, such as the spatial locations of infectious disease cases (Diggle et al., 2013). In this work, we model locations according to a Cox process with a spatially varying intensity surface $\lambda(\cdot)$. The number of points inside a region $\Omega \subset \mathbb{R}^2$ is distributed $N|\lambda(\cdot) \sim Pois\left(\int_{\Omega} \lambda(\boldsymbol{s})d\boldsymbol{s}\right)$. The likelihood

for a set of locations $\mathcal{S} = \{\boldsymbol{s}_i\}_i^N$ observed in Ω is given by

$$\pi(\mathcal{S}, N|\lambda(\cdot)) = \exp\left(-\int_\Omega \lambda(\boldsymbol{s})d\boldsymbol{s}\right)\frac{\prod_{i=1}^N \lambda(\boldsymbol{s}_i)}{N!}. \tag{3.1}$$

Møller et al. (1998) introduced the class of LGCPs as a method to describe spatial correlation in point process models. A LGCP with spatially continuous covariates $\boldsymbol{x}(\boldsymbol{s})$ and population density offset $\log(pd(\boldsymbol{s}))$ is a Cox process with intensity $\lambda(\boldsymbol{s})$ given by

$$\log(\lambda(\boldsymbol{s})) = \log(pd(\boldsymbol{s})) + \boldsymbol{x}'(\boldsymbol{s})\boldsymbol{\beta} + \eta(\boldsymbol{s}), \quad \boldsymbol{s} \in \mathbb{R}^2, \tag{3.2}$$

where $\eta(\boldsymbol{s})$ is a Gaussian process with mean zero and some user defined covariance function $C(\cdot, \cdot)$.

3.3.1 Computational Details for Fitting Log-Gaussian Cox Processes

Fitting LGCPs is a computationally burdensome task for high dimensional data. In this work we elect to use the approximation technique of Lindgren et al. (2011) to express the Gaussian process $\eta(\boldsymbol{s})$ as a basis expansion

$$\eta(\boldsymbol{s}) = \sum_{i=1}^n \phi_i(\boldsymbol{s})w_i, \quad \boldsymbol{s} \in \Omega, \tag{3.3}$$

where n is the number of knots placed in Ω and $\{\phi_i(\boldsymbol{s})\}_{i=1}^n$ is a set of piecewise triangular basis functions (see Appendix B.1 for details). Lindgren et al. (2011) showed that the weights of the basis expansion in (3.3) are distributed

$$\boldsymbol{w}|\kappa^2, \xi^2 \sim N(\boldsymbol{0}, \xi^2 Q_{\kappa^2}^{-1}), \tag{3.4}$$

where $Q_{\kappa^2}^{-1}$ is a sparse precision matrix. The distribution in (3.4) is an approximation to a zero-mean Gaussian process with Matérn covariance function

$$C(\boldsymbol{u}, \boldsymbol{v}) = \xi^2(\kappa||\boldsymbol{v} - \boldsymbol{u}||)K_1(\kappa||\boldsymbol{v} - \boldsymbol{u}||), \quad \boldsymbol{v}, \boldsymbol{u} \in \mathbb{R}^2, \tag{3.5}$$

where $||\cdot||$ denotes Euclidean distance, and $K_1(\cdot)$ is an order one Bessel function of the second kind. The marginal variance is given by $\sigma^2 = \xi^2/(4\pi\kappa^2)$, while the approximation of the effective range is given by $\rho = \sqrt{8}/\kappa$ (Lindgren et al., 2011). We refer the reader to Appendix (B.1–B.2) for further details regarding model fitting.

3.4 Conditional Predictive Ordinate Estimates for Cox Processes

Geisser (1980) first introduced the leave-one-out predictive distribution $\pi\Big(y_i|\boldsymbol{y}_{-i}\Big)$, known as the conditional predictive ordinate (CPO), as a diagnostic to detect inconsistent observations from a given model. CPO estimates are commonly used to perform Bayesian model selection for independently distributed error responses, as only one model fit is required to obtain each leave-one-out-predictive density (Pettit, 1990; Hooten and Hobbs, 2015). CPO estimates are obtained by Monte Carlo approximation, using samples from the posterior distribution conditioned on the full data set (Hooten and Hobbs, 2015). In this section we demonstrate that CPO estimates can easily be obtained for Cox processes with dependence structures.

The leave-one-out predictive distribution for each location in a given dataset is

$$CPO_k = \pi(\boldsymbol{s}_k, N|\mathcal{S}_{-k}) = \int \pi(\boldsymbol{s}_k, N|\lambda, \mathcal{S}_{-k})\pi(\lambda|\mathcal{S}_{-k})d\lambda. \tag{3.6}$$

The quantity in (3.6), also known as the conditional predictive ordinate (CPO), represents the density of $\boldsymbol{s}_k$ when a model is fit without $\boldsymbol{s}_k$ (Geisser, 1980). CPO estimates are constructed by utilizing independent marginal distributions to rewrite the likelihood without observation k, $\pi(\mathcal{S}_{-k}|\lambda, N)$, as a scaled factor of the full likelihood $\pi(\mathcal{S}|\lambda, N)$. When $\boldsymbol{s}_k$ is independent of $\mathcal{S}_{-k}$ (i.e. independence), samples from the full posterior $\pi(\lambda|N, \mathcal{S})$ can be used to easily obtain CPO estimates via Monte Carlo approximation of (3.6). The statistic $-2\sum_{i=1}^{N} CPO_i$ is commonly used to perform Bayesian model selection in a fashion similar to leave-one-out cross-validation (Hooten and Hobbs, 2015).

Dependence structures commonly assumed in spatial models are generally computationally expensive, and so CPOs are rarely used for spatial data. Here, we demonstrate that CPO estimates can easily be obtained for Cox processes with spatial dependence structures. The key reason this is possible is that, for a LGCP, the likelihood of each location is independent of all other locations conditioned on the intensity surface $\lambda(\cdot)$, which contains the dependence structure. In Section 3.7, we define a risk metric based on a leave-one-out intrusion scenario. The CPO estimates derived in this section allow us to quickly obtain disclosure risk estimates by avoiding the computationally burdensome task of leave-one-out model fitting.

Assume that the points, $\mathcal{S} = \{\boldsymbol{s}_i\}_{i=1}^{N}$, follow a Cox Process with intensity $\lambda(\cdot)$. CPO_k

is given by

$$CPO_k = \pi(\boldsymbol{s}_k, N|\mathcal{S}_{-k}) = \left(\mathbb{E}_{\pi(\lambda|\mathcal{S},N)}\left[\frac{\int_\Omega \lambda(\boldsymbol{s})}{\lambda(\boldsymbol{s}_k)}\right]\right)^{-1}. \tag{3.7}$$

To see this, we first note that we can express the likelihood $\pi(\mathcal{S}_{-k}, N|\lambda)$ as

$$\begin{aligned}
\pi(\mathcal{S}_{-k}, N|\lambda) &= \int_\Omega \frac{\prod_{i=1}^N \lambda(\boldsymbol{s}_i)}{N!\exp\left(\int_\Omega \lambda(\boldsymbol{s})d\boldsymbol{s}\right)} d\boldsymbol{s}_k \\
&= \frac{\left(\int_\Omega \lambda(\boldsymbol{s})d\boldsymbol{s}\right)\left(\prod_{i\neq k}\lambda(\boldsymbol{s}_i)\right)}{N!\exp\left(\int_\Omega \lambda(\boldsymbol{s})d\boldsymbol{s}\right)} \\
&= \frac{\int_\Omega \lambda(\boldsymbol{s})d\boldsymbol{s}}{\lambda(\boldsymbol{s}_k)}\pi(\mathcal{S}, N|\lambda).
\end{aligned} \tag{3.8}$$

Using (3.8), we obtain the CPO as follows

$$\begin{aligned}
CPO_k = \pi(\boldsymbol{s}_k, N|\mathcal{S}_{-k}) &= \left(\frac{\pi(\mathcal{S}_{-k}, N)}{\pi(\mathcal{S}, N)}\right)^{-1} \\
&= \left(\int \frac{\pi(\mathcal{S}_{-k}, N|\lambda)}{\pi(\mathcal{S}, N)}\pi(\lambda)d\lambda\right)^{-1} \\
&= \left(\int \left(\frac{\int_\Omega \lambda(\boldsymbol{s})d\boldsymbol{s}}{\lambda(\boldsymbol{s}_k)}\right)\frac{\pi(\mathcal{S}, N|\lambda)}{\pi(\mathcal{S}, N)}\pi(\lambda)d\lambda\right)^{-1} \\
&= \left(\int \left(\frac{\int_\Omega \lambda(\boldsymbol{s})d\boldsymbol{s}}{\lambda(\boldsymbol{s}_k)}\right)\pi(\lambda|\mathcal{S}, N)d\lambda\right)^{-1} \\
&= \left(\mathbb{E}_{\pi(\lambda|\mathcal{S},N)}\left[\left(\frac{\int_\Omega \lambda(\boldsymbol{s})d\boldsymbol{s}}{\lambda(\boldsymbol{s}_k)}\right)\right]\right)^{-1}.
\end{aligned}$$

We have shown that the CPO can easily be expressed in terms of expectation with respect to the posterior conditioned on the full data set $\pi(\lambda|\mathcal{S}, N)$ for any Cox process. Thus CPO estimates are easily be obtained by a Monte Carlo approximation for spatial point process data generated by LGCPs! To see this, let $\lambda^{(m)}$ represent the m^{th} sample drawn from $\pi(\lambda|\mathcal{S}, N)$. CPO estimates are obtained by

$$CPO_k = \pi(\boldsymbol{s}_k, N|\mathcal{S}_{-k}) \approx \left(\sum_{m=1}^M \frac{\int_\Omega \lambda^{(m)}(\boldsymbol{s})d\boldsymbol{s}}{\lambda^{(m)}(\boldsymbol{s}_k)}\right)^{-1}, \tag{3.9}$$

where M in (3.9) represents the total number of samples drawn from $\pi(\lambda|\mathcal{S}, N)$. In Section 3.7, we demonstrate that individual disclosure risks can be easily evaluated by

CPO estimates.

3.5 Generating Fully Synthetic Location Data

In this section we detail our proposed Bayesian methods for generating fully synthetic location data based on LGCPs. We also formally introduce radial synthesis as a baseline comparison for our proposed synthesis methods.

3.5.1 Radial Synthesis

Randomly perturbing locations within a radius r of their true location is one of the most common redaction methods (VanWey et al., 2005). It has been claimed that when large perturbations are required to ensure adequate privacy protection, inference related to spatial dependence structures are diminished (Armstrong et al., 1999; Wang and Reiter, 2012; Quick et al., 2015). Intuitively this claim seems reasonable, as large perturbations will destroy localized spatial clusters. Due to its popularity and simplicity, we treat radial perturbation as the baseline synthesis method for comparison. To our knowledge, no Bayesian analyses have been performed to assess the utility and disclosure risks associated with radial perturbation. We perform radial synthesis by drawing N independent synthetic locations from circular uniform distributions with radius r centered at each $\boldsymbol{s}_i$, denoted $\boldsymbol{s}_i^\dagger \sim U(\boldsymbol{s}_i, r)$.

3.5.2 Additive Noise Synthesis

Here, we introduce our first proposed synthesis method, Additive Noise Synthesis (*ANS*). In *ANS*, we alter the global variance parameter of the spatial random field $\eta(\boldsymbol{s})$ contained within the intensity surface $\lambda(\boldsymbol{s})$. We perform *ANS* by adding a noisy spatial random field $\nu(\boldsymbol{s})$ with the same spatial scale to the log-intensity surface. *ANS* proceeds as follows. We first fit an LGCP to the confidential set $\mathcal{S} = \{\boldsymbol{s}_i\}_i^N$ according to Section 3.3. We then obtain posterior mean estimates for the basis expansion weights $\hat{\boldsymbol{w}} = \mathbb{E}_{\pi(\boldsymbol{w},\beta|\mathcal{S})}[\boldsymbol{w}]$, fixed effects $\hat{\boldsymbol{\beta}} = \mathbb{E}_{\pi(\boldsymbol{w},\beta|\mathcal{S})}[\boldsymbol{\beta}]$, and spatial scale parameter $\hat{\kappa}^2 = \mathbb{E}_{\pi(\kappa^2,\xi^2|\mathcal{S})}[\kappa^2]$.

Next, we simulate a Gaussian noise process $\boldsymbol{v} \sim N(\boldsymbol{0}, \sigma^2 \boldsymbol{Q}_{\hat{\kappa}^2}^{-1})$ with precision matrix $\boldsymbol{Q}_{\kappa^2}^{-1}$ (defined in Appendix B.1) for some user-defined noise level σ^2. Following Section 3.3.1, we express the additive noise as a basis expansion $\nu(\boldsymbol{s}) = \sum_{i=1}^n \phi_i(\boldsymbol{s}) v_i$. We obtain

the resulting *ANS* intensity surface

$$\lambda^{\dagger}(\boldsymbol{s}) = \exp\left(\log(pd(\boldsymbol{s})) + \boldsymbol{x}'(\boldsymbol{s})\hat{\boldsymbol{\beta}} + \sum_{i=1}^{n} \phi_i(\boldsymbol{s})\,(\hat{w}_i + v_i)\right). \tag{3.10}$$

Note that $\boldsymbol{w}$ and $\boldsymbol{v}$ are both mean zero normally distributed random vectors with covariance structures given by $\xi^2\boldsymbol{Q}_{\kappa^2}^{-1}$ and $\sigma^2\boldsymbol{Q}_{\kappa^2}^{-1}$ and thus, $\boldsymbol{w}+\boldsymbol{v} \sim N(\boldsymbol{0}, (\xi^2+\sigma^2)\boldsymbol{Q}_{\kappa^2}^{-1})$. Intuitively, the addition of Gaussian random noise in the intensity surface will randomly scale the relative risk of disease occurrence at each location $\boldsymbol{s}$. To see this note that (3.10) can equivalently be expressed as $\lambda^{\dagger}(\boldsymbol{s}) = e^{\nu(\boldsymbol{s})}\hat{\lambda}(\boldsymbol{s})$, where $\hat{\lambda}(\boldsymbol{s})$ denotes the intensity surface with plug-in posterior mean estimates $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{w}}$. We also note that *ANS* can be viewed as an extension of the synthesis method proposed by Quick et al. Quick et al. (2015), in which synthetic locations were generated based on posterior mean plug-in estimates. We obtain the synthesis method of Quick et al. (2015) by simply taking the noise-level to be zero ($\sigma^2 = 0$).

We generate synthetic locations by uniformly sampling N^* points with $N^* >> N$ over the spatial domain Ω. We then assign each of the N^* sampled locations $\{\boldsymbol{s}_i^*\}_{i=1}^{N^*}$ probability weights $\hat{p}_k^* = \frac{\lambda^{\dagger}(\boldsymbol{s}_k^*)}{\sum_{j=1}^{N^*}\lambda^{\dagger}(\boldsymbol{s}_j^*)}$. We obtain an *ANS* synthetic dataset by sampling N of the N^* locations without replacement according to their probabilities $\hat{p}_k^*$.

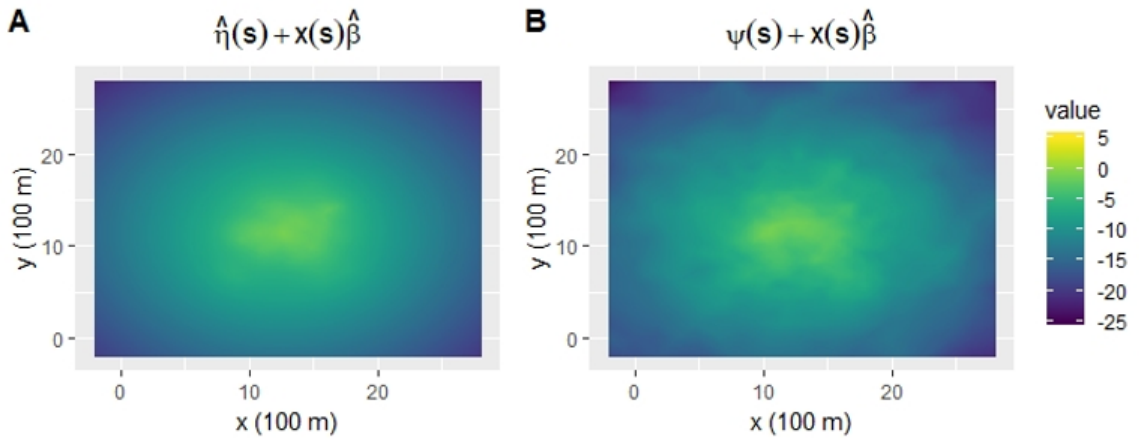

Figure 3.2: (A): An intensity surface plot with the posterior mean estimate for the spatial random field $\hat{\eta}(\boldsymbol{s})$ from the John Snow cholera outbreak dataset. (B): An *ANS* intensity surface plot for an additive noise spatial random field, $\psi(\boldsymbol{s})$, with spatial scale $\hat{\kappa}^2$ and noise-level $\sigma^2 = 10$.

In *ANS*, we alter the marginal variance of the spatial random effect by some user defined noise level σ^2. Therefore, an *ANS* data set produces an intensity surface with the same spatial scale as the confidential data set with larger variability (see Figure 3.2). In turn, we expect *ANS* data sets to offer an inverse relationship between disclosure risks and data utility relative to the user defined noise level. That is, an *ANS* data set with a large user defined noise level will offer lower disclosure risks and lower data utility relative to a data set with a smaller noise level.

3.5.3 Posterior Resampling Synthesis

Here, we propose another novel synthesis method which we call Posterior Resampling Synthesis (*PRS*). Gaussian processes with Matérn covariance given by (3.5) are fully characterized by variance parameter ξ^2 and spatial scale parameter κ^2. For fixed values of ξ^2 and κ^2, realizations of the spatial random field $\eta(\boldsymbol{s})$ will produce random fields with the same marginal variance and spatial range. Intuitively, a resampled spatial random field will relocate spatial clusters while preserving the dependence structure of the confidential intensity surface (see Figure 3.3).

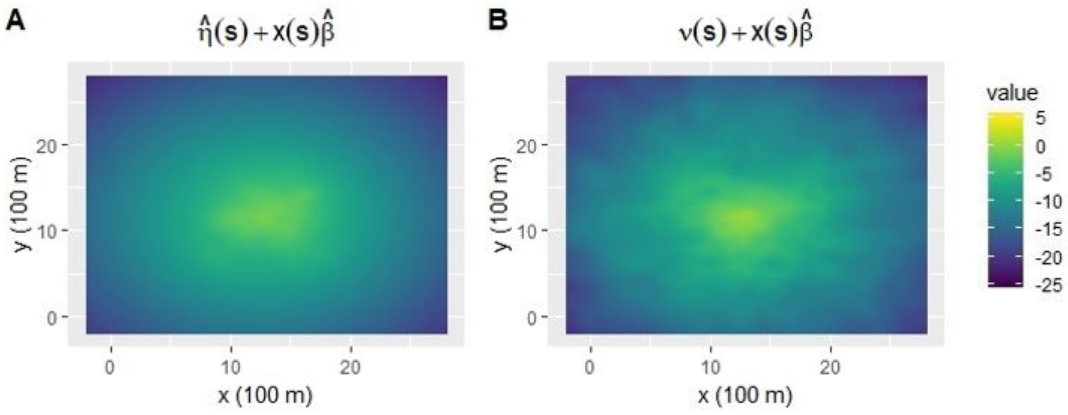

Figure 3.3: (A): An intensity surface plot with the posterior mean estimate for the spatial random field $\hat{\eta}(\boldsymbol{s})$ from the John Snow cholera outbreak dataset. (B): An intensity surface plot for a resampled spatial random field, $\nu(\boldsymbol{s})$, with spatial scale $\hat{\kappa}^2$ and variance parameter $\hat{\xi}^2$.

The posterior resampling synthesis (*PRS*) process is described as follows. Similar to *ANS*, we first fit the confidential dataset according to a LGCP as described in Section 3.3.1 via MCMC. We then compute posterior mean estimates for the fixed effects, $\hat{\boldsymbol{\beta}} = \mathbb{E}_{\pi(\boldsymbol{\beta},\boldsymbol{w}|\mathcal{S})}[\boldsymbol{\beta}]$, spatial scale, $\hat{\kappa}^2 = \mathbb{E}_{\pi(\xi^2,\kappa^2|\mathcal{S})}[\kappa^2]$, and variance parameter,

$\hat{\xi}^2 = \mathbb{E}_{\pi(\xi^2,\kappa^2|\mathcal{S})}[\xi^2]$ from the confidential fit. We generate a new spatial random field based on the posterior mean of the hyperparameters, $\boldsymbol{w}^* \sim N(\mathbf{0}, \hat{\xi}^2 Q^{-1}_{\hat{\kappa}^2})$. We obtain the *PRS* intensity surface by swapping the basis expansion weights $\boldsymbol{w}$ for the resampled weights $\boldsymbol{w}^*$,

$$\lambda^\dagger(\boldsymbol{s}) = \exp\left(\log(pd(\boldsymbol{s})) + \boldsymbol{x}'(\boldsymbol{s})\boldsymbol{\beta} + \sum_{i=1}^{n} \phi_i(\boldsymbol{s}) w_i^*\right). \tag{3.11}$$

We then simulate a *PRS* dataset by sampling N locations from a LGCP with intensity (3.11) as described in Section 3.5.2.

The spatial scale and marginal variance of a *PRS* intensity surface are given by posterior mean estimates from the confidential data set. In turn, *PRS* generated data sets maintain data utility by preserving the correlative structure of the confidential set, and reduce disclosure risks by relocating spatial clusters contained in the intensity surface.

3.6 Evaluating Utility

In Sections (3.5.1–3.5.3) we presented radial synthesis and our two proposed methods for generating fully synthetic location data. Our goal is to disseminate synthetic locations that reduce disclosure risks, while allowing for meaningful analyses to be conducted by secondary analysts. The propensity mean square error (*pMSE*) developed by Woo et al. (2009) and Snoke et al. (2018) is the mean square error of the predicted probability that a given observation belongs to the confidential set. In turn, the *pMSE* is a metric for measuring the distributional similarity between a synthetic and confidential data set (Snoke and Slavković, 2018). In this section, we provide a model-based estimation of the predicted probabilities required to estimate the *pMSE* for LGCPs.

Let $\lambda(\boldsymbol{s})$ and $\lambda^\dagger(\boldsymbol{s})$ denote the intensity surfaces for the true dataset, $\mathcal{S}$, and synthetic dataset, $\mathcal{S}^\dagger$, respectively. Let $\{\boldsymbol{y}_k\}_{k=1}^{2N}$ denote the collection of true and synthetic locations $\mathcal{S} \cup \mathcal{S}^\dagger$. Assuming $P(\boldsymbol{y}_k \in \mathcal{S}) = 0.5$, the probability a point $\boldsymbol{y}_k$ belongs to the true dataset is

$$p_k = P(\boldsymbol{y}_k \in \mathcal{S}|\lambda, \lambda^\dagger) = \frac{\lambda(\boldsymbol{y}_k)}{\lambda(\boldsymbol{y}_k) + \left(\frac{\int_\Omega \lambda(\boldsymbol{s})d\boldsymbol{s}}{\int_\Omega \lambda^{(\dagger)}(\boldsymbol{s})d\boldsymbol{s}}\right)\lambda^\dagger(\boldsymbol{y}_k)}. \tag{3.12}$$

The probability of correct classification for each point in $\mathcal{S} \cup \mathcal{S}^\dagger$ can be approximated

by Monte Carlo simulation. To do so, L samples are drawn from the intensity surface of the true posterior $\lambda^{(l)} \sim \pi(\lambda|\mathcal{S}, N)$ and the synthetic posterior $\lambda^{\dagger(l)} \sim \pi\left(\lambda^{\dagger}|\mathcal{S}^{\dagger}, N\right)$ respectively. Define

$$p_k^{(l)} = P(\boldsymbol{y}_k \in \mathcal{S}|\lambda^{(l)}, \lambda^{\dagger(l)}) = \frac{\lambda^{(l)}(\boldsymbol{y}_k)}{\lambda^{(l)}(\boldsymbol{y}_k) + \left(\frac{\int_\Omega \lambda^{(l)}(\boldsymbol{s})d\boldsymbol{s}}{\int_\Omega \lambda^{\dagger(l)}(\boldsymbol{s})d\boldsymbol{s}}\right)\lambda^{\dagger(l)}(\boldsymbol{y}_k)}. \tag{3.13}$$

The probability of correctly identifying $\boldsymbol{y}_k$ is

$$\hat{p}_k = E[P(\boldsymbol{y}_k \text{ is classified correctly})] \approx \frac{1}{L}\sum_{l=1}^{L}(p_k^{(l)})^{x_k}(1-p_k^{(l)})^{1-x_k}, \tag{3.14}$$

where

$$x_k = \begin{cases} 0, & \boldsymbol{y}_k \in \mathcal{S}^{\dagger} \\ 1, & \boldsymbol{y}_k \in \mathcal{S} \end{cases}.$$

The predicted probabilities given in (3.14) are used to to approximate the *pMSE*,

$$pMSE \approx \widehat{pMSE} = \frac{1}{2N}\sum_{k=1}^{2N}(\hat{p}_k - 0.5)^2. \tag{3.15}$$

We elect to use the *pMSE* to quantify the quality of a given synthetic data set due to its intuitive interpretation. A synthetic dataset, $\mathcal{S}^{\dagger}$, that is indistinguishable from the confidential set, $\mathcal{S}$, produces a *pMSE* score of 0. A synthetic dataset that is systematically distinguishable from the confidential set will offer little scientific inference. Such sets are assigned a worst case *pMSE* score of 0.25.

In this section, we have demonstrated how the *pMSE* can be tailored to LGCPs. In Section 3.8, we use the *pMSE* to assess the quality of synthetic cholera death locations.

3.7 Evaluating Disclosure Risks

In this section we construct a novel risk metric to quantify individual disclosure risks in spatial location data. We provide the computational details required to obtain individual disclosure risk estimates for the synthesis methods discussed in Section 3.5. We also discuss the restrictive a priori assumptions required to obtain differentially private intensity surfaces $\lambda(\cdot)$.

3.7.1 Disclosure Risk Metric

Let $\mathcal{S} = \{\boldsymbol{s}_i\}_{i=1}^N$ be the collection of the N confidential locations and $\mathcal{S}^\dagger = \{\boldsymbol{s}_i^\dagger\}_{i=1}^N$ the collection of N synthetic locations. Assume an adversary has access to the synthetic dataset, knowledge of the synthesis method, population density, and has identified all but the k^{th} confidential location $\boldsymbol{s}_k$. The intruder aims to identify the final k^{th} confidential location within a ball of radius r of the truth, denoted $\mathcal{B}_r(\boldsymbol{s}_k)$. We consider this high risk scenario, as it quantifies how much information a synthetic set offers to an intruder who has identified all but one true location. This is a common scenario in privacy, including empirical differential privacy (Wang and Reiter, 2012; Charest and Hou, 2016). The metric also holds an intuitive interpretation; if a region $\mathcal{B}_r(\boldsymbol{s}_k)$ is of high probability then the synthetic data set provides an intruder with information useful for locating a confidential location.

The probability of identifying person k within a radius r of the true location given the synthetic dataset and all but the k^{th} true location is

$$P(\{\boldsymbol{s} \in \mathcal{B}_r(\boldsymbol{s}_k) \cap \Omega\} | \mathcal{S}_{-k}, \mathcal{S}^\dagger) = \int_{\mathcal{B}_r(\boldsymbol{s}_k)} \pi\Big(\boldsymbol{s}, N | \mathcal{S}_{-k}, \mathcal{S}^\dagger\Big) d\boldsymbol{s}. \tag{3.16}$$

Notice that (3.16) is an integral over a leave-one-out-density $\pi\Big(\boldsymbol{s}, N | \mathcal{S}_{-k}, \mathcal{S}^\dagger\Big)$. We rely on the CPO estimates detailed in Section 3.4 along with a quadrature scheme (detailed in Appendix B.4) to evaluate (3.16). The computational details required for evaluating the CPOs, $\pi\Big(\boldsymbol{s}, N | \mathcal{S}_{-k}, \mathcal{S}^\dagger\Big)$, of each synthesis method are described in the following sections.

3.7.2 Radial Synthesis Disclosure Risks

To fit a LGCP to a radially perturbed synthetic data set and the confidential set simultaneously, we assume a circular uniform prior for the synthetic data, $\boldsymbol{s}_i^\dagger \sim U(\boldsymbol{s}_i, r)$ for $i = 1, 2, ..., N$, and a LGCP with intensity $\lambda(\cdot)$ for the confidential set $\mathcal{S}$. We require an estimate of the CPO $\pi\Big(\boldsymbol{s}, N | \mathcal{S}_{-k}, \mathcal{S}^\dagger\Big)$ to evaluate the risk metric in (3.16). Evaluation of the leave-one-out density $\pi\Big(\boldsymbol{s}, N | \mathcal{S}_{-k}, \mathcal{S}^\dagger\Big)$ requires samples from the posterior distribution $\pi\Big(\lambda | \mathcal{S}_{-k}, \mathcal{S}^\dagger, N\Big)$ for each location $\boldsymbol{s}_k$. We also require a linkage prior to determine which synthetic location $\boldsymbol{s}_i^\dagger$ was uniquely generated from each confidential location $\boldsymbol{s}_i$.

Fortunately we can overcome this computational bottleneck by augmenting the CPO argument presented in Section 3.4. The leave-one-out predictive density for data generated

by radial synthesis is given by

$$\pi\Big(\boldsymbol{s}, N|\mathcal{S}_{-k}, \mathcal{S}^{\dagger}\Big) = \left[\mathbb{E}_{\pi(\lambda|\boldsymbol{S},N)}\left[\frac{\int_{\mathcal{B}_r(\boldsymbol{s}_k^{\dagger})}\lambda(\boldsymbol{s})d\boldsymbol{s}}{\lambda(\boldsymbol{s})}\right]\right]^{-1}. \tag{3.17}$$

Notice that the CPO estimate in (3.17) only requires samples from the marginal distribution of the confidential dataset $\pi(\lambda|\mathcal{S}, N)$. In turn, we do not require samples drawn from the leave-one-out posterior density $\pi\Big(\lambda|\mathcal{S}_{-k}, \mathcal{S}^{\dagger}, N\Big)$. We also avoid the computationally burdensome task of forming a linkage prior between each synthetic location and the confidential location they were generated from. A full derivation of (3.17) is included in Appendix B.3.

We estimate the CPO in (3.17) by Monte Carlo integration. Let L represent the number of samples drawn from the marginal density $\pi(\lambda|\mathcal{S}, N)$. The expectation in (3.17) is estimated by

$$\pi\Big(\boldsymbol{s}, N|\mathcal{S}_{-k}, \mathcal{S}^{\dagger}\Big) \approx \left[\frac{1}{L}\sum_{l=1}^{L}\frac{\int_{\mathcal{B}_r(\boldsymbol{s}_k^{\dagger})}\lambda^{(l)}(\boldsymbol{s})d\boldsymbol{s}}{\lambda^{(l)}(\boldsymbol{s})}\right]^{-1}. \tag{3.18}$$

Note that the integral $\int_{\mathcal{B}_r(\boldsymbol{s}_k^{\dagger})}\lambda^{(l)}(\boldsymbol{s})d\boldsymbol{s}$ in (3.18) must be evaluated for all L samples. We approximate the analytically intractable integral with the quadrature scheme detailed in Appendix B.4.

3.7.3 Additive Noise Synthesis Disclosure Risks

CPO estimates of $\pi\Big(\boldsymbol{s}, N|\mathcal{S}_{-k}, \mathcal{S}^{\dagger}\Big)$ for *ANS* generated sets require samples from the joint distribution $\pi\Big(\lambda, \lambda^{\dagger}|\mathcal{S}, \mathcal{S}^{\dagger}, N\Big)$. We use a Metropolis Hastings sampler to draw samples from the desired joint marginal density. The simultaneous fit assumes that the synthetic data is independent of the confidential data, that is $\pi\Big(\mathcal{S}, \mathcal{S}^{\dagger}, N|\lambda, \lambda^{\dagger}\Big) = \pi(\mathcal{S}, N|\lambda)\pi\Big(\mathcal{S}^{\dagger}, N|\lambda^{\dagger}\Big)$, where $\pi(\mathcal{S}, N|\lambda)$ and $\pi\Big(\mathcal{S}^{\dagger}, N|\lambda^{\dagger}\Big)$ are independent *LGCPs* with intensities $\lambda(\cdot)$ and $\lambda^{\dagger}(\cdot)$ respectively. The intensity surfaces are of the form

$$\log\left(\lambda(\boldsymbol{s})\right) = \boldsymbol{x}'(\boldsymbol{s})\boldsymbol{\beta} + \sum_{i=1}^{n}\phi_i(\boldsymbol{s})w_i + \log\left(pd\left(\boldsymbol{s}\right)\right) \tag{3.19}$$

$$\log\Big(\lambda^{\dagger}(\boldsymbol{s})\Big) = \boldsymbol{x}'(\boldsymbol{s})\boldsymbol{\beta} + \sum_{i=1}^{n}\phi_i(\boldsymbol{s})(w_i + v_i) + \log\left(pd\left(\boldsymbol{s}\right)\right), \tag{3.20}$$

where the basis expansion weights in (3.19) and (3.20) are independently distributed $\pi(\boldsymbol{w}|\xi^2,\kappa^2) \sim N(\boldsymbol{0}, \xi^2 Q_{\kappa^2}^{-1})$ and $\pi(\boldsymbol{v}|\sigma^2,\kappa^2) \sim N(\boldsymbol{0}, \sigma^2 Q_{\kappa^2}^{-1})$. Recall that σ^2 is the noise level, released to the public along with the synthetic dataset. For this reason σ^2 is treated as known and fixed at its true value.

Let $\lambda^{(l)}(\cdot)$ denote the l^{th} draw from $\pi\Big(\lambda, \lambda^\dagger|\mathcal{S}, \mathcal{S}^\dagger, N\Big)$. A Monte Carlo estimate of the CPO, $\pi\Big(\boldsymbol{s}, N|\mathcal{S}_{-k}, \mathcal{S}^\dagger\Big)$, is given by

$$\pi\Big(\boldsymbol{s}, N|\mathcal{S}_{-k}, \mathcal{S}^\dagger\Big) = \left[\mathbb{E}_{\pi\left(\lambda,\lambda^\dagger|\mathcal{S},\mathcal{S}^\dagger,N\right)}\left[\frac{\int_\Omega \lambda(\boldsymbol{s})d\boldsymbol{s}}{\lambda(\boldsymbol{s})}\right]\right]^{-1} \approx \left[\frac{1}{L}\sum_{l=1}^{L}\frac{\int_\Omega \lambda^{(l)}(\boldsymbol{s})d\boldsymbol{s}}{\lambda^{(l)}(\boldsymbol{s})}\right]^{-1}. \quad (3.21)$$

The integral $\int_\Omega \lambda^{(l)}(\boldsymbol{s})d\boldsymbol{s}$ in (3.21) is numerically approximated for all L samples following Simpson et al. (2016) (see (B.5) of Appendix B.2 for details). The individual disclosure risk in (3.16) for location $\boldsymbol{s}_k$ is obtained by using a quadrature scheme (see Appendix B.4) to integrate the CPO over $\mathcal{B}_r(\boldsymbol{s}_k)$.

3.7.4 Posterior Resampling Synthesis Disclosure Risks

Similar to *ANS*, CPO estimates for *PRS* require samples from the joint marginal $\pi\Big(\lambda, \lambda^\dagger|\mathcal{S}, \mathcal{S}^\dagger, N\Big)$. As done with *ANS*, we again assume that the synthetic data is independent of the confidential set, that is $\pi\Big(\lambda, \lambda^\dagger|\mathcal{S}, \mathcal{S}^\dagger, N\Big) = \pi(\lambda|\mathcal{S}, N)\pi\Big(\lambda^\dagger|\mathcal{S}^\dagger, N\Big)$. Recall that the synthetic intensity surface is given by $\log\Big(\lambda^\dagger(\boldsymbol{s})\Big) = \boldsymbol{x}'(\boldsymbol{s})\boldsymbol{\beta} + \sum_{i=1}^{n}\phi_i(\boldsymbol{s})w_i^* + \log\left(pd(\boldsymbol{s})\right)$, where $\boldsymbol{w}^* \sim N(\boldsymbol{0}, \hat{\xi}^2 Q_{\hat{\kappa}^2}^{-1})$.

A Metropolis Hastings sampler is used to obtain samples from the joint distribution by sequentially drawing from the posterior distributions $\pi\Big(\boldsymbol{\beta}|\mathcal{S}, \mathcal{S}^\dagger, \boldsymbol{w}, \boldsymbol{w}^*\Big)$, $\pi(\boldsymbol{w}|\mathcal{S}, \boldsymbol{\beta}, \kappa^2, \xi^2)$, $\pi\Big(\boldsymbol{w}^*|\mathcal{S}^\dagger, \boldsymbol{\beta}, \kappa^2, \xi^2\Big)$ and $\pi(\xi^2, \kappa^2|\boldsymbol{w}, \boldsymbol{w}^*)$. CPO estimates are obtained by evaluating (3.21) using samples from the joint distribution. *PRS* disclosure risks are then obtained by following the quadrature and Monte Carlo scheme as detailed in Section 3.7.3.

3.7.5 Differential Privacy

Dwork et al. (2006) and Dwork (2006) introduced differential privacy (DP) as a measure of confidentiality protection. DP protects the information of every individual in the data set by limiting the influence that any one respondent can have on the released information. DP ensures the confidentiality of each individual in a database, even against an adversary who has gained complete knowledge of the rest of the data set.

Formally, a randomized algorithm $\mathcal{M}(\mathcal{S})$ is said to be ϵ-differentially private if

$$P(\mathcal{M}(\mathcal{S}) \in A) \leq \exp(\epsilon) P(\mathcal{M}(\mathcal{S}^*) \in A), \tag{3.22}$$

for all measurable subsets A of the range of $\mathcal{M}$ and for all datasets $\mathcal{S}, \mathcal{S}^*$ differing by one entry (Dwork et al., 2014). Dimitrakakis et al. (2014) and Foulds et al. (2016) observed that Bayesian posterior sampling provides ϵ-differential privacy under certain prior assumptions. Theorem 2 of Foulds et al. (2016) states that releasing one sample from the posterior distribution, $\pi(\boldsymbol{x}|\theta)$, with any prior, $\pi(\theta)$, is $2C$-differentially private provided $sup_{\boldsymbol{x},\boldsymbol{x}' \in \mathcal{X}, \theta \in \Theta} |\log(\pi(\boldsymbol{x}|\theta))| \leq C$. We describe this result for LGCPs.

Consider the collection of N locations in $\Omega \subset \mathbb{R}^2$ denoted $\mathcal{X}$. Let $\mathcal{S}, \mathcal{S}^* \in \mathcal{X}$ differ by at most one entry, say $\boldsymbol{s}$ and $\boldsymbol{s}^*$. Assume we model the locations according to a $LGCP$ with intensity $\lambda(\cdot)$ given by (3.2) with a spatial random field defined as in (3.4). As shown in Foulds et al. (2016), releasing one sample from the posterior distribution $\pi(\lambda|\mathcal{S})$ is $2C$-differentially private for any prior, provided

$$max_{\mathcal{S},\mathcal{S}' \in \mathcal{X}, \lambda \in \Lambda} |\log \pi(\boldsymbol{s}^*|\lambda) - \log \pi(\boldsymbol{s}|\lambda)| \leq C. \tag{3.23}$$

For LGCPs, (3.23) holds provided $\forall \boldsymbol{s}, \boldsymbol{s}^* \in \Omega$,

$$sup_{\lambda \in \Lambda} |\log \lambda(\boldsymbol{s}^*) - \log \lambda(\boldsymbol{s})| \leq C. \tag{3.24}$$

Equivalently, the left hand side of (3.24) can be expressed as

$$sup_{\lambda \in \Lambda} \left| (\boldsymbol{x}'(\boldsymbol{s}^*) - \boldsymbol{x}'(\boldsymbol{s}))\boldsymbol{\beta} + \sum_{i=1}^{n} (\phi_i(\boldsymbol{s}^*) - \phi_i(\boldsymbol{s}))\, w_i + \log(pd(\boldsymbol{s}^*)) - \log(pd(\boldsymbol{s})) \right|. \tag{3.25}$$

It is clear from (3.25) that the privacy "cost" C in (3.24) is determined by bounding the maximum distance between any continuous covariate measured on Ω and the magnitude of each fixed effect β_i and spatial weight w_i. An a priori constraint on the parameter space of both the fixed effects and spatial random effects is required to produce a desired privacy cost C. Scientists rarely posses enough a priori knowledge to suggest a simultaneous clipping of the fixed effects and spatial random effect. For this reason, we have elected to move away from differential privacy, and instead utilize the disclosure risk metric introduced in Section 3.7.1.

3.8 Case Study: Dr. John Snow's Cholera Outbreak

We now apply our proposed methodology to Dr. John Snow's cholera outbreak. The data set consists of $N = 578$ observed cholera death locations. We fit the locations according to a LGCP with intensity $\lambda(\boldsymbol{s})$ as in (3.2) with the region of Soho, London represented by $\Omega = [200 \text{ m}, 2,200 \text{ m}]^2$. A priori, we assume $\boldsymbol{\beta} \sim N(\mathbf{0}, 2\text{I})$, where $\boldsymbol{\beta}$ consists of an intercept β_0 and the distance to the Broad St. water pump (the source of the outbreak) β_1. The estimated population kernel density estimate (see Figure 3.4) $\log(pd(\boldsymbol{s}))$ serves as the offset. The prior choices for the variance and spatial scale (κ, ξ) of the spatial random field are detailed in Appendix B.5.

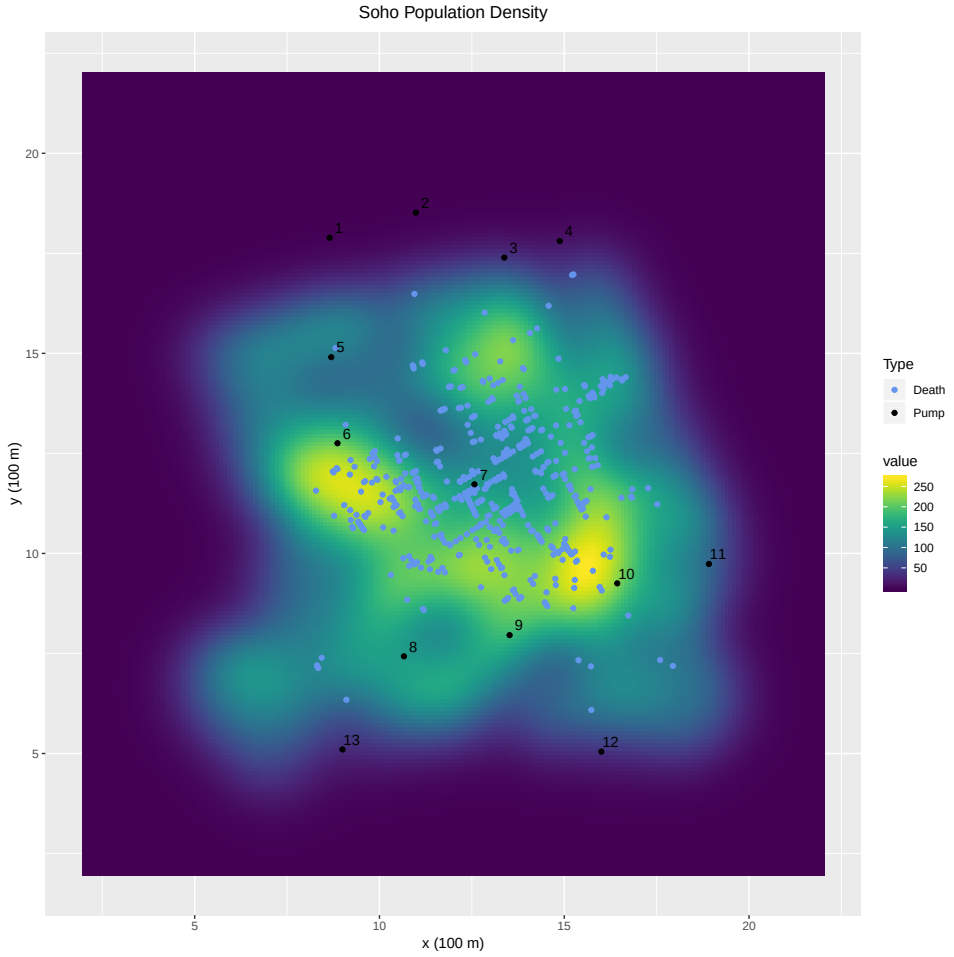

Figure 3.4: A kernel density estimate of the Soho, London population in 1854. Cholera death locations are plotted in blue as well as the water pumps numbered 1-13, with pump 7 being the Broad St. water pump. The estimated total population is 21,345.

We fit the LGCP model via MCMC following the technique described in Appendix B.2.

We tune our Metropolis-Hastings sampler according to the adaptive tuning scheme of Roberts and Rosenthal (2009) for 250,000 burn-in samples. Another 250,000 post burn-in samples are stored for inference. The resulting parameter estimates, including the fixed effects, effective range, and marginal variance of the spatial random field are summarized in Table 3.1. We note that $\hat{\beta}_1 = -0.946$ suggests that the further an individual lives from the Broad St. pump, the less susceptible they are to cholera death.

Using the fitted model, we generate 15 synthetic datasets via *PRS* following Section 3.5.3. Additionally, 20 *ANS* datasets are generated with noise levels $\sigma^2 = 0.5,1,1.5,...,9.5,10$ following Section 3.5.2. 10 synthetic sets via radial synthesis with radii $r = 50$ m, 100 m, ... , 300 m as detailed in Section 3.5.1. Metropolis-Hastings samplers are used to assess utility and disclosure risks following Sections 3.6 and 3.7. Each sampler drew 500,000 samples, with 250,000 discarded as burn-in. The results are summarized in Figure 3.5.

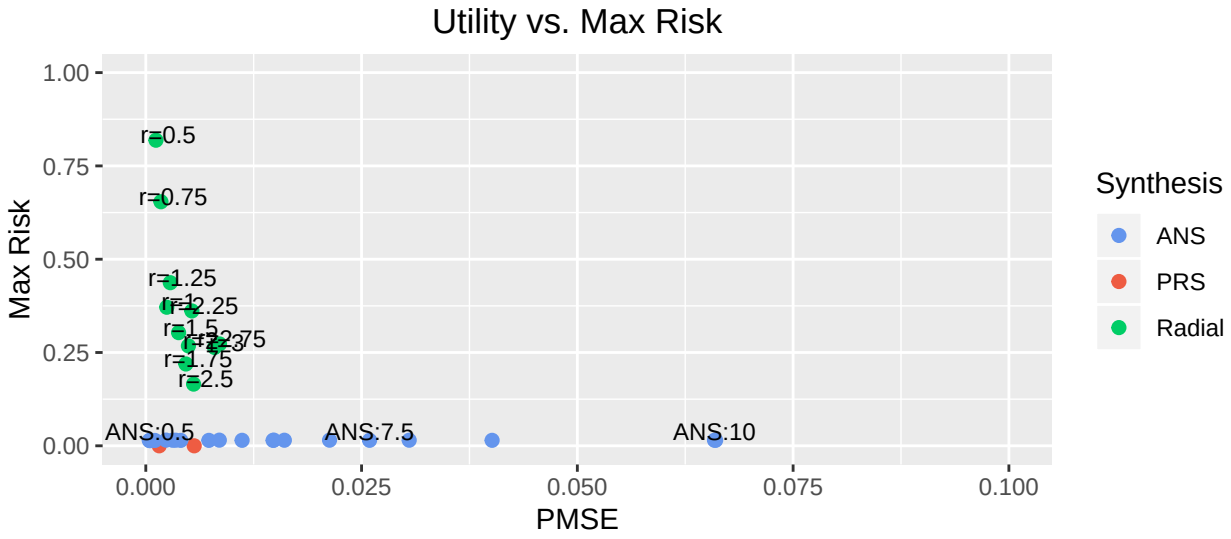

Figure 3.5: Plot of max disclosure risk vs. *pMSE* for *ANS*, radial synthesis, and *PRS* datasets. Note that only the min and max utility are plotted for *PRS*.

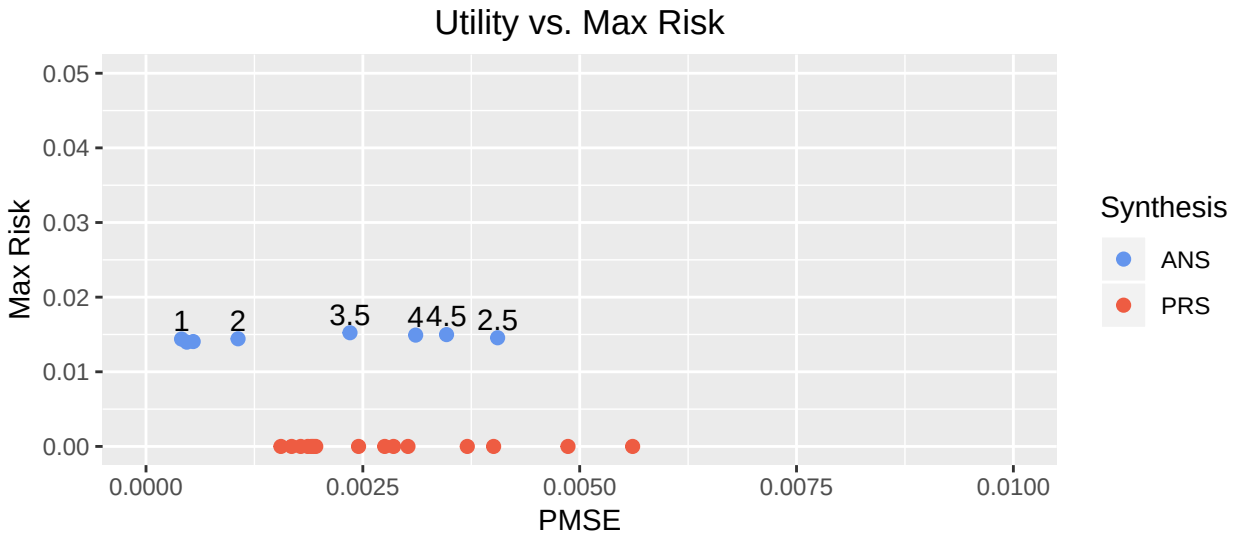

Figure 3.6: Plot of max disclosure risk vs. *pMSE* for *ANS* and *PRS* datasets with max disclosure risks all less than 0.005. The noise level σ^2 is displayed above each *ANS* point. All 15 *PRS* data sets are plotted.

From Figure 3.5 it is clear that synthetic sets generated according to *PRS* offer the lowest max disclosure risks. In Figure 3.5, we only plot the max and min utility scores for *PRS*, as the max disclosure risk for all considered *PRS* datasets was 6.28e-10. Figure 3.5 clearly shows that max disclosure risks for data sets generated from radial synthesis are consistently larger than *ANS* and *PRS* datasets. In Figure 3.6, we see that *ANS* data sets with small noise levels offer an improvement in data utility with an increase in max disclosure risk in comparison to *PRS*. For large noise values, the *ANS* datasets offer small reductions in max disclosure risk at the expense of data utility.

The goal of this work is to generate a synthetic data set that offers low disclosure risks while preserving scientific inference. Since the max disclosure risk for the 15 considered *PRS* data sets were all less than 6.28e-10, we suggest releasing the *PRS* data set that offers the highest data utility with a *pMSE* score of 0.0016. Recall that a small *pMSE* score corresponds to higher data utility. The synthetic intensity surface, synthetic locations, and confidential locations are plotted in Figure 3.7. In Table 3.1, we see that the *PRS* fixed effect estimates, effective range, and marginal variance are similar to the confidential set (see Table 3.1). In turn, Dr. Snow could still have determined the source of the cholera outbreak from the *PRS* data set.

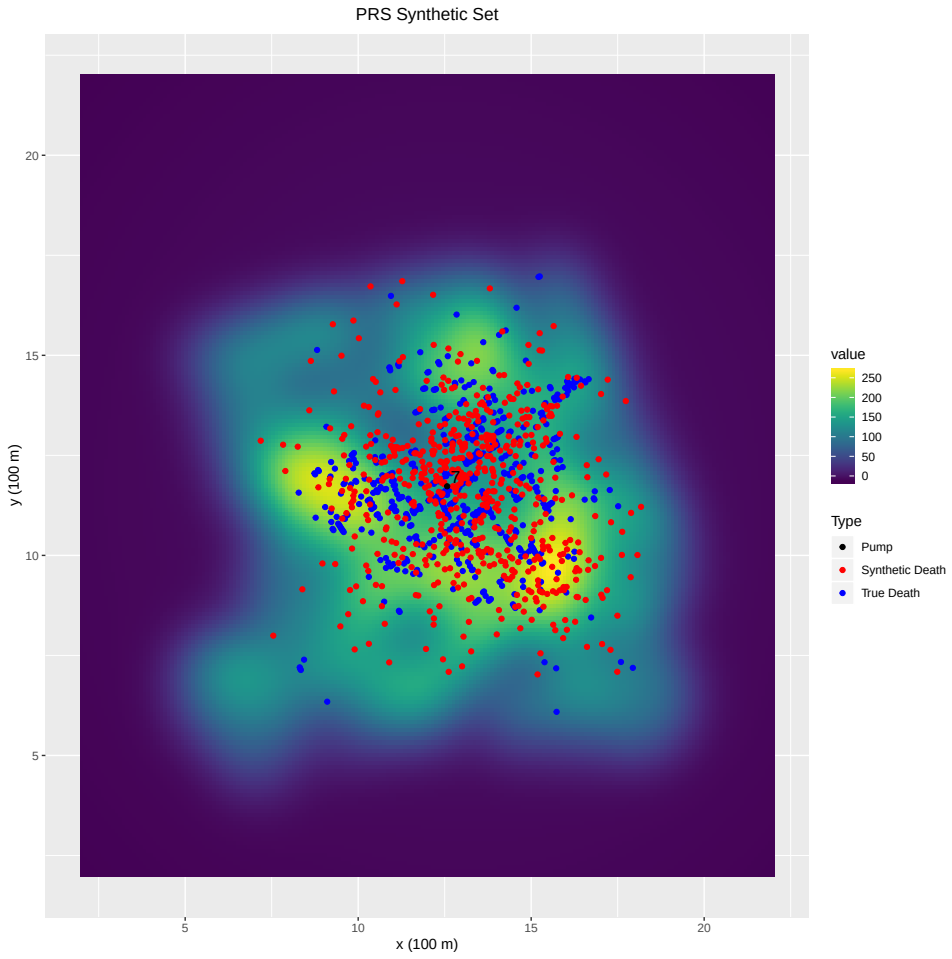

Figure 3.7: A plot of the fitted intensity surface for the optimal *PRS* data set with max disclosure risk of 1.205e-17 and *pMSE* score of 0.0016. The true deaths (blue), *PRS* synthetic locations (red), and the Broad St. pump (black) are plotted as well.

Parameter	Est.	95% CI	*PRS* Est.	*PRS* 95% CI
β_0	-0.771	(-1.928, 0.387)	-0.799	(-1.867, 0.431)
β_1	-0.946	(-1.183,-0.729)	-0.829	(-1.046, -0.642)
ρ	3.487	(2.027, 8.065)	6.506	(3.831, 9.177)
τ	0.725	(0.241, 2.577)	0.554	(0.269,1.143)

Table 3.1: Posterior mean estimates and corresponding 95% credible intervals (CI) for the effective range and marginal variance of the spatial random field for the confidential and *PRS* data sets.

The goal of this work was to produce a synthetic data set that reduces individual disclosure risks while preserving data utility. In this case study, we observed that even

for large radial perturbations (300 m), we could always find a noise level for *ANS* which offers improved data utility and reduced disclosure risks. Similarly, *PRS* datasets always offer reduced disclosure risks and improved data utility in comparison to radial synthesis. In summary, both of our proposed synthesis methods out perform radial perturbation in terms of risk vs. utility. Since the *PRS* data sets considered offered lower max disclosure risks than any of the radial perturbation or *ANS* data sets, we suggested the dissemination of the *PRS* data set that offered the highest data utility. We then showed that the chosen *PRS* data set offered inference similar to the confidential data set.

3.9 Discussion

In this work we proposed two novel Bayesian approaches for generating fully-synthetic location-only datasets. We introduced a novel risk metric for a high intrusion scenario. We demonstrated that CPO estimates can easily be obtained for spatial location data generated from LGCPs, and allow for computationally efficient approximations of individual disclosure risks for each synthesis method. We adapted the *pMSE* statistic for spatial point process data generated from LGCPs to evaluate data utility. We performed a case study of Dr. John Snow's cholera outbreak data set; showing that *PRS* and *ANS* offer an improved risk to utility ratio in comparison to radial perturbation.

In this work, we have conducted a Bayesian analysis of risk vs. utility for radial synthesis. Among other common methods used to privatize disease case locations prior to data dissemination are data aggregation and suppression (Armstrong et al., 1999; Wang and Reiter, 2012). Spatial aggregations attempt to reduce disclosure risks by coarsening the geographic scale; e.g., reporting disease incidences at the county scale. However, any coarsening of the geographic scale diminishes the level at which a spatial analysis can be performed, and localized geographic trends/hazards can easily be missed (Armstrong et al., 1999). Data suppression attempts to reduce disclosure risks by removing high risk individuals, such as spatial outliers, from a dataset. If many individuals are deemed high risk, many locations may need to be suppressed, limiting the quality of any spatial analysis. In this work, we chose to focus on generating spatial disease case location data sets that contained the exact number of observations as the confidential set. Radial perturbation produces data sets consisting of point referenced disease cases with the same amount of observations as the original data set. For this reason, radial perturbation was considered as the baseline case, while aggregation and suppression were not.

We considered an intrusion scenario in which an intruder attempts to identify a

confidential location within a radius r of the truth given 1) the synthetic dataset, 2) knowledge of the synthesis method, and 3) identification of all but the last confidential location. Quick et al. (2015) considered a disclosure risk metric that defines an individual to be at high risk if they are spatially close to other individuals with similar attributes. Clearly this metric is not suited to handle location-only data. Wang and Reiter (2012) consider an intrusion case similar to ours. Their risk metric first computes the expected euclidean distance between an intruder's guess of the confidential geography with respect to $\pi\left(\boldsymbol{s}_k|\mathcal{S}_{-k}, \mathcal{S}^{\dagger}\right)$, denoted R1. The number of actual cases within a radius of R1 of the truth are then counted as a measure of reasonable guesses for the confidential location. Our disclosure risk metric computes the probability that an individual is uniquely identified within a pre-specified radius r of the truth. Our metric also contains population density as an offset, allowing the metric to account for different risk levels in regions of dense and sparse populations.

We proposed a utility metric based on the *pMSE* statistic to assess the quality of a synthetic data. Previous work by Quick et al. (2015) assessed data utility with the K-function. The K function is a measure of spatial dependence that computes the expected number of events within a radius h for each observation in the data set (Bartlett, 1964; Ripley, 1976). An *ANS* generated data set will have a dependence structure with similar spatial scale and marginal variance that differs from the confidential set by some user defined noise level. *PRS* data sets have the same spatial scale and marginal variance as the confidential set. Both of our proposed synthesis methods produce data sets with correlative structures similar to the confidential data set. In turn, they will likely produce similar K-function estimates. The *pMSE* is a data-based utility metric that determines the utility of a synthetic data set based on how well a synthetic set emulates the true data set. We tailored the predicted probabilities used to classify conditional data in the *pMSE* to LGCPs. In turn our metric accounts for the spatial dependence structure of the synthetic data set as well. We elected to use the *pMSE* due to its intuitive interpretation; as the *pMSE* can be viewed as the mean square misclassification score between synthetic and confidential locations.

In summary we have developed two novel methods for generating fully-synthetic location-only data. We demonstrated that CPOs are easily obtainable for spatial point process data generated by LGCPs. We proposed a disclosure risk metric and model based data utility metric suited for synthetic location data with no attributes that could easily be evaluated with CPO estimates. We showed that our proposed methodology outperforms the common approach of radial synthesis in a case study of Dr. John Snow's

cholera outbreak dataset. We illustrated that *PRS* offers small max disclosure risk while preserving data utility. Our second proposed synthesis method *ANS* was shown to offer the best utility for small noise levels. For this reason we believe both of our proposed synthesis methods should be used with the goal of balancing the risk vs data utility trade off.

Chapter 4
Laplace Approximations for Fast Bayesian Inference on Markov Population Models

[1]

4.1 Introduction

Statistical inference methods for population process models are required to understand the dynamics of disease outbreaks such as the current COVID-19 global pandemic (Chen et al., 2020; He et al., 2020; Mwalili et al., 2020; Lai et al., 2021). Markov population models are commonly used to model disease dynamics in small to moderately sized populations (Sun et al., 2015; Allen, 2017; Fricks and Hanks, 2018). Simulation and inference for Markov population models can be computationally burdensome often requiring approximate methods (Fricks and Hanks, 2018). Kurtz (1978) derived the linear noise approximation (LNA) as a limiting Gaussian distribution for Markov population models which allows for simulation and inference based on the solutions to systems of ordinary differential equations (ODEs). The LNA of Kurtz (1978) has been used to help facilitate Bayesian inference on Markov population models (Komorowski et al., 2009; Fearnhead et al., 2014) but remains computationally expensive for some systems. In this work, we derive a Laplace approximation for LNA models that offers fast Bayesian inference on Markov population models.

A common framework for inference on Markov population models is to consider a

[1]This work is in preparation for submission. This is a joint work with Ephraim M. Hanks.

Gaussian approximation of the Markov population model using the functional central limit theorem (FCLT). This approximates the Markov population model as a system of linear stochastic differential equations (SDEs). Except in a few specific cases, systems of linear SDEs with time-varying coefficients do not have analytical solutions. In the absence of an analytic solution, numerical methods are required to solve the system of SDEs. LNAs offer a Gaussian density for the approximate solution to SDEs, whose mean and covariance can be obtained by solving a system of ODEs. LNAs have been shown to be more accurate than ODE models for Markov population process (Giagos, 2010; Fearnhead et al., 2014), and offer a substantially easier method for inference in comparison to general SDE models in which no analytic solution exists. Since LNA likelihood evaluations rely solely on the solutions to ODEs, Inference with LNAs allows for inference on parameters in the SDE using simpler ODE solving techniques.

The main novel contribution of this work is the construction of a Laplace approximation for performing Bayesian inference with LNAs. We demonstrate that the Gaussian distribution for the LNA has a sparse precision matrix constructed from the solution to a system of ODEs. We then tailor the work of Rue et al. (2009) to LNAs to obtain an approximate marginal distribution for the parameters governing the dynamics of the Markov population process. We show how to obtain marginal distributions for each individual parameter from the approximate joint marginal density by performing a numerical integration following Rue et al. (2009). We show that each marginal distributions can be obtained through independent evaluations of the LNA likelihood. In turn, Laplace approximations can be parallelized providing a vast computational speed up in comparison to standard LNA inference methods such as sequential Monte Carlo (Fearnhead et al., 2014) and MCMC Komorowski et al. (2009).

This is not the first presentation of methods for Bayesian inference on Markov population models with LNAs. Komorowski et al. (2009) utilized the joint distribution of the LNA to perform inference on Markov population models via MCMC. Fearnhead et al. (2014) partitioned the observation times into subintervals and fit a sequence of LNAs with a particle filter. Both methods require stochastic infill to estimate latent population states and sequential constructions of the LNA distribution. We demonstrate that inference with Laplace approximations does not require stochastic infill to perform inference on LNAs and can be fit in parallel. We demonstrate that LNAs fit with Laplace approximations offer similar parameter inference to LNAs fit via MCMC with stochastic infill on a simulated Susceptible-Infected-Susceptible (SIS) data set. In turn, Laplace approximations are shown to preserve parameter inference while reducing the

computational complexity of model fitting in comparison to MCMC.

The remainder of the manuscript is organized as follows. In Section 4.2 we introduce Markov population models and the diffusion approximations required to construct the LNA distribution. In Section 4.3 we summarize the results of Kurtz (1978) and derive a sparse precision matrix for the LNA. We derive the Laplace approximation for LNAs and demonstrate how to perform inference on Markov population models in Section 4.4. In Section 4.5, we compare the Laplace approximation to MCMC on a simulated SIS data set. In Section 4.6, we use Laplace approximations to fit a stochastic SEIR model to the Princess Diamond Cruise COVID-19 data set. We conclude with a discussion in Section 4.7.

4.2 Markov Population Models

Deterministic population models are widely used in fields such as chemistry, ecology, and epidemiology to model large scale population dynamics such as disease outbreaks (Keeling and Rohani, 2011; Fricks and Hanks, 2018). While deterministic models work well for very large populations, stochastic methods are needed to capture fine scale dynamics for populations with few individuals (Fricks and Hanks, 2018). In this section we formulate the Markov population model in terms of stochastic reactions and rates. Our treatment follows the formulation of Kurtz (1978) and Fricks and Hanks (2018).

Let $\boldsymbol{X}(t) = (X_1(t), X_2(t), ..., X_d(t))'$ be a d-dimensional random vector on the non-negative integers, with each entry $X_i(t)$ representing the number of individuals in a population of size N belonging to group i at time t (e.g. the number of infected individuals in a population). A population reaction occurs when one individual moves from one group to another. We allow for n possible reactions among the d groups in any given model. We let $\boldsymbol{R}_i$ be a d-dimensional vector denoting the i^{th} reaction. For example, if an individual can move from group 1 to 3, the reaction vector will contain -1 as the first element, 1 as the third element, and 0 for all other $d-2$ elements.

Each individual reaction is assumed to occur stochastically at a rate which depends on the current state $\boldsymbol{X}(t)$ and unknown rate parameters $\boldsymbol{\theta}$. We denote the rate corresponding to reaction $\boldsymbol{R}_i$ by $\lambda^i_{\boldsymbol{\theta}}(\boldsymbol{X}(t))$. Let $Y_i(\lambda(\cdot))$ be an independent Poisson process with rate $\lambda(\cdot)$. We define the stochastic population model as the sum of Poisson processes in terms of reactions vectors and reaction rates given by

$$\boldsymbol{X}(t) = \boldsymbol{X}(0) + \sum_{i=1}^{n} \boldsymbol{R}_i Y_i \left(\int_0^t \lambda^i_{\boldsymbol{\theta}} (\boldsymbol{X}(s))\, ds \right). \tag{4.1}$$

4.2.1 Example: The Stochastic SIS Model

A classic example of a Markov population model is the stochastic SIS model with a closed population of size N. The SIS model tracks the proportions of susceptible and infected individuals. We let $\boldsymbol{X}_N(t) = (S_N(t), I_N(t))'$, denote the population scaled proportions of susceptible and infected individuals at time t. There are two possible reactions; a susceptible individual becomes infected, $\boldsymbol{R}_1 = (-1, 1)'$, or an infected individual recovers, $\boldsymbol{R}_2 = (1, -1)'$. Susceptible individuals become infected at rate of $\lambda^1_{\boldsymbol{\theta}}(\boldsymbol{X}_N(t)) = \beta S_N(t) I_N(t)$, and infected individuals are assumed to recover at a rate of $\lambda^2_{\boldsymbol{\theta}}(\boldsymbol{X}_N(t)) = \gamma I_N(t)$. We denote $\boldsymbol{\theta} = (\beta, \gamma)$, where β is the contact rate, and γ is the recovery rate, both having support on the positive reals.

We note that this system can equivalently be reformulated as a one-dimensional system that tracks the proportions of infected $P(t) = I_N(t)$. We simply write reaction vector $\boldsymbol{R} = [1, -1]$, and take the re-parameterized rates to be $\lambda^1_{\boldsymbol{\theta}}(P(t)) = \beta(1 - P(t))P(t)$ and $\lambda^2_{\boldsymbol{\theta}}(P(t)) = \gamma P(t)$. Simulation methods such as the Gillespie algorithm (Gillespie, 1977) and tau-leaping (Cao et al., 2006) can be used to generate stochastic realizations from (4.1) for fixed values of $\boldsymbol{\theta}$. A trajectory of the proportion of infected individuals simulated from a stochastic SIS model with a population of size $N = 1000$, contact rate $\beta = 0.50$, recovery rate $\gamma = 0.33$, and initial state $P(0) = 0.05$ is shown in Figure 4.1. We return to this data set for a simulation study in Section 4.5.

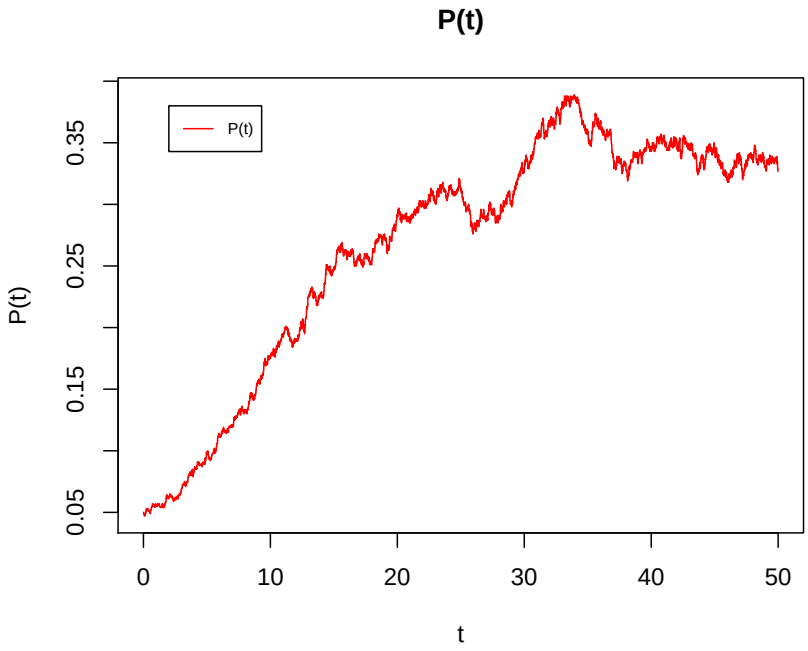

Figure 4.1: A plot of the stochastic trajectory of $P(t)$ (red) generated from an SIS model with $N = 1000$, $P(0) = 0.05$, $\beta = 0.50$, and $\gamma = 0.33$.

4.2.2 Diffusion Approximations

Markov population model dynamics are controlled by the rate parameters $\boldsymbol{\theta}$ which are often unknown and need to be estimated from data. The most common approaches for inference on $\boldsymbol{\theta}$ are based on diffusion approximations to (4.1). We utilize the FCLT as done in the works of Baxendale and Greenwood (2011) and Fricks and Hanks (2018) to construct a system of linear SDEs to help facilitate inference on (4.1).

Let $\boldsymbol{X}_N(t) = \frac{1}{N}\boldsymbol{X}(t)$ be the normalized population process. We scale (4.1) by N to obtain

$$\boldsymbol{X}_N(t) = \boldsymbol{X}_N(0) + \sum_{i=1}^{n} \boldsymbol{R}_i \frac{1}{N} Y_i \left(N \int_0^t \lambda_{\boldsymbol{\theta}}^i \left(\boldsymbol{X}_N(s) \right) ds \right). \tag{4.2}$$

We apply the FCLT for Poisson processes (see Appendix C.1) to each scaled Poisson process in (4.2) to obtain the Gaussian approximation

$$\boldsymbol{X}_N(t) \approx \boldsymbol{X}_N(t) + \sum_{i=1}^{n} \boldsymbol{R}_i \left(\int_0^t \lambda_{\boldsymbol{\theta}}^i \left(\boldsymbol{X}_N(s) \right) ds + \frac{1}{\sqrt{N}} B_i \left(\int_0^t \lambda_{\boldsymbol{\theta}}^i \left(\boldsymbol{X}_N(s) \right) ds \right) \right), \tag{4.3}$$

where $B_i(t)$ are independent Brownian motions with variance t. We denote $\boldsymbol{F}_{\boldsymbol{\theta}}\left(\boldsymbol{X}_N(t)\right) = \sum_{i=1}^{n} \boldsymbol{R}_i \lambda^i_{\boldsymbol{\theta}}\left(\boldsymbol{X}_N(t)\right)$ and let $\mathrm{G}_{\boldsymbol{\theta}}\left(\boldsymbol{X}_N(t)\right)$ be a matrix with columns $\boldsymbol{g}_i\left(\boldsymbol{X}_N(s)\right) = \boldsymbol{R}_i \sqrt{\lambda^i_{\boldsymbol{\theta}}\left(\boldsymbol{X}_N(s)\right)}$. We take a stochastic derivative of (4.3) to obtain

$$d\boldsymbol{X}_N(t) = \boldsymbol{F}_{\boldsymbol{\theta}}\left(\boldsymbol{X}_N(t)\right) dt + G_{\boldsymbol{\theta}}\left(\boldsymbol{X}_N(t)\right) d\boldsymbol{B}(t), \tag{4.4}$$

where $d\boldsymbol{B}(t) = (dB_1(t), dB_2(t), ..., dB_d(t))'$ is a d-dimensional vector of differentiated independent Brownian motions $B_i(t)$. In some cases, the system of SDEs in (4.4) yields an analytic solution (Øksendal, 2003). However, in most cases, numerical methods are needed to provide approximate solutions.

The remainder of this work focuses on approximate inference methods for Markov population models that rely on numerical solutions to the system of SDEs in (4.4). In the next section we derive the LNA distribution, which is an approximate joint Gaussian likelihood for (4.4) constructed from the solution to a system of ODEs. We demonstrate how to perform fast inference on LNAs in Section 4.4.

4.3 Linear Noise Approximations to Stochastic Population Models

In this section we construct the joint Gaussian likelihood of the LNA. We follow the results of Kurtz (1978) to show that the SDE approximation of the Markov population model converges to a system of ODEs. We propose a novel solution to the LNA which results in a sparse representation of the LNA covariance which is constructed from the solution to a system of ODEs.

4.3.1 Limiting Deterministic Systems

We summarize the results of Theorem 8.1 in Kurtz (1978), which details the sufficient conditions required for the system of SDEs in (4.4) to converge to a deterministic system. Assume that $\boldsymbol{x}$ is an element in some compact subset $K \subset \mathbb{R}^d$ and the following conditions hold: (1) $\sum_{i=1}^{n} |\boldsymbol{R}_i| \sup \lambda^i_{\boldsymbol{\theta}}(\boldsymbol{x}) < \infty$, (2) $\boldsymbol{F}(\cdot) = \sum_{i=1}^{n} \boldsymbol{R}_i \lambda^i_{\boldsymbol{\theta}}(\cdot)$ is Lipshitz on K,and (3) $\lim_{n\to\infty} \boldsymbol{x}_n = \boldsymbol{x}^{\dagger}$. Theorem 8.1 of Kurtz (1981) states that for each $t > 0$, $\lim_{n\to\infty} \sup_{s\leq t} |\boldsymbol{X}_n(t) - \boldsymbol{X}^{\dagger}_{\boldsymbol{\theta}}(t)| \to 0$, where $\boldsymbol{X}^{\dagger}_{\boldsymbol{\theta}}(t)$ is the solution to the system of ordinary

differential equations

$$d\boldsymbol{X}_{\boldsymbol{\theta}}^{\dagger}(t) = \boldsymbol{F}_{\boldsymbol{\theta}}\left(\boldsymbol{X}_{\boldsymbol{\theta}}^{\dagger}(t)\right)dt = \sum_{i=1}^{n} \boldsymbol{R}_i \lambda_{\boldsymbol{\theta}}^{i}\left(\boldsymbol{X}_{\boldsymbol{\theta}}^{\dagger}(t)\right)dt, \tag{4.5}$$

with initial state $\boldsymbol{X}_{\boldsymbol{\theta}}^{\dagger}(0)$. That is, the diffusion approximation of the Markov population process in (4.4) converges to a system of ODEs as the population size N tends towards infinity.

4.3.2 Approximate Gaussian Processes

Kurtz (1978) constructed an approximate distribution for the system of SDEs in (4.4) from the limiting deterministic system in (4.5). We present the results of Kurtz (1978) required to obtain an approximate distribution for (4.1). We begin by considering the difference between $\boldsymbol{X}_N(t)$ and its corresponding infinite population limit $\boldsymbol{X}_{\boldsymbol{\theta}}^{\dagger}(t)$. From equations (4.4) and (4.5)

$$d\left(\boldsymbol{X}_N(t) - \boldsymbol{X}_{\boldsymbol{\theta}}^{\dagger}(t)\right) \approx \left(\boldsymbol{F}_{\boldsymbol{\theta}}\left(\boldsymbol{X}_N(t)\right) - \boldsymbol{F}_{\boldsymbol{\theta}}\left(\boldsymbol{X}_{\boldsymbol{\theta}}^{\dagger}(t)\right)\right)dt + \frac{1}{\sqrt{N}} G_{\boldsymbol{\theta}}\left(\boldsymbol{X}_N(t)\right)d\boldsymbol{B}(t). \tag{4.6}$$

Using a first-order Taylor expansion of $\boldsymbol{F}_{\boldsymbol{\theta}}(\boldsymbol{X}_N(s))$ about $\boldsymbol{X}^{\dagger}(s)$, (4.6) becomes

$$d\left(\boldsymbol{X}_N(t) - \boldsymbol{X}_{\boldsymbol{\theta}}^{\dagger}(t)\right) \approx \partial\boldsymbol{F}_{\boldsymbol{\theta}}\left(\boldsymbol{X}_N(t)\right)\left(\boldsymbol{X}_N(t) - \boldsymbol{X}_{\boldsymbol{\theta}}^{\dagger}(t)\right)dt + \frac{1}{\sqrt{N}} G_{\boldsymbol{\theta}}\left(\boldsymbol{X}_N(t)\right)d\boldsymbol{B}(t). \tag{4.7}$$

The FCLT ensures that as the population size $N \to \infty$, $\sqrt{N}\left(\boldsymbol{X}_N(t) - \boldsymbol{X}_{\boldsymbol{\theta}}^{\dagger}(t)\right)$ converges in distribution to some zero-mean Gaussian process $\boldsymbol{V}(t)$ Kurtz (1978). Applying the limit to (4.7) we have

$$d\boldsymbol{V}(t) = \boldsymbol{F}_{\boldsymbol{\theta}}\left(\boldsymbol{X}_{\boldsymbol{\theta}}^{\dagger}(t)\right)\boldsymbol{V}(t)ds + G_{\boldsymbol{\theta}}\left(\boldsymbol{X}_{\boldsymbol{\theta}}^{\dagger}(t)\right)d\boldsymbol{B}(t). \tag{4.8}$$

We approximate the finite sample process by the LNA

$$\boldsymbol{X}_N(t) \approx \boldsymbol{X}_{\boldsymbol{\theta}}^{\dagger}(t) + \frac{1}{\sqrt{N}}\boldsymbol{V}(t). \tag{4.9}$$

We note that the approximate distribution in (4.8), which was first constructed by Kurtz (1978), is centered about the solution to the system of ODEs $\boldsymbol{X}_{\boldsymbol{\theta}}^{\dagger}(t)$ in (4.5) (i.e. $\mathbb{E}\left(\boldsymbol{X}_N(t)\right) = \boldsymbol{X}_{\boldsymbol{\theta}}^{\dagger}(t)$). Equation (4.9) states that deviations from the mean ODE solution $\boldsymbol{X}_{\boldsymbol{\theta}}^{\dagger}(t)$ can be approximated with a Gaussian process with covariance $Cov\left(\boldsymbol{V}(t)\right)$. In the

following section we propose a novel solution to (4.8) that results in a sparse precision matrix for the joint likelihood of (4.9).

4.3.3 Sparse Precision Matrices

We construct the $Cov\left(\boldsymbol{V}(t)\right)$ by seeking separable solutions to (4.8) of the form $\boldsymbol{V}(t) = U(t)\boldsymbol{W}(t)$, where $\boldsymbol{W}(t)$ is an n-dimensional zero mean Gaussian process and $U(t)$ is a d by d matrix. We note (4.8) is a linear SDE with deterministic time varying coefficients $\partial \boldsymbol{F}_{\boldsymbol{\theta}}\left(\boldsymbol{X}^{\dagger}_{\boldsymbol{\theta}}(t)\right)$ and $G_{\boldsymbol{\theta}}\left(\boldsymbol{X}^{\dagger}_{\boldsymbol{\theta}}(t)\right)$ for which solutions are well known (Klebaner, 2012). The solution to (4.8) (see Appendix C.2) is given by

$$\begin{aligned} dU(t) &= \partial \boldsymbol{F}_{\boldsymbol{\theta}}\left(\boldsymbol{X}^{\dagger}_{\boldsymbol{\theta}}(t)\right) U(t)dt, U(0) = \mathbb{I}_{d\times d} && (4.10)\\ \boldsymbol{W}(t) &= \int_0^t U^{-1}(s) G_{\boldsymbol{\theta}}\left(\boldsymbol{X}^{\dagger}_{\boldsymbol{\theta}}(s)\right) ds, \boldsymbol{W}(0) = \boldsymbol{0}. && (4.11) \end{aligned}$$

We define $A(s) = U^{-1}(s) G_{\boldsymbol{\theta}}\left(\boldsymbol{X}^{\dagger}_{\boldsymbol{\theta}}(s)\right)$, giving $Cov\left(\boldsymbol{W}(t)\right) = \int_0^t A(s)A'(s)ds$. For time points $t_0 < t_1 < t_2 < ... < t_T$, we define the auto-regressive differencing matrix S such that $S\boldsymbol{W} = (\boldsymbol{W}(t_1), \boldsymbol{W}(t_2) - \boldsymbol{W}(t_1), ..., \boldsymbol{W}(t_T) - \boldsymbol{W}(t_{T-1}))$. We define the covariance of each temporal increment

$$M_i = Cov\left(\boldsymbol{W}(t_i) - \boldsymbol{W}(t_{i-1})\right) = \int_{t_{i-1}}^{t_i} A(s)A'(s)ds. \tag{4.12}$$

Let U and M be a block diagonal matrix composed of d by d matrices $U(t_i)$ and M_i respectively. Denote the time referenced observations of $\boldsymbol{X}_N = (\boldsymbol{X}_N(t_1), \boldsymbol{X}_N(t_2), ..., \boldsymbol{X}_N(t_T))'$ and $\boldsymbol{X}^{\dagger}_{\boldsymbol{\theta}} = \left(\boldsymbol{X}^{\dagger}_{\boldsymbol{\theta}}(t_1), \boldsymbol{X}^{\dagger}_{\boldsymbol{\theta}}(t_2), ..., \boldsymbol{X}^{\dagger}_{\boldsymbol{\theta}}(t_T)\right)'$. The LNA distribution is given by

$$\pi(\boldsymbol{X}_N|\boldsymbol{\theta}) \sim N\left(\boldsymbol{X}^{\dagger}_{\boldsymbol{\theta}}, \frac{1}{N} Q^{-1}_{\boldsymbol{\theta}}\right), \tag{4.13}$$

where $Q_{\boldsymbol{\theta}} = \left(SU^{-1}\right)' M^{-1} \left(SU^{-1}\right)$ is a sparse precision matrix that depends solely on the deterministic solution $\boldsymbol{X}^{\dagger}_{\boldsymbol{\theta}}(t)$. We note that other formulations of $\boldsymbol{Q}^{-1}_{\boldsymbol{\theta}}$ exist (Komorowski et al., 2009; Giagos, 2010; Fearnhead et al., 2014). Our novel formulation offers a sparse precision matrix that relies on sequentially solving the deterministic systems in (4.10) and (4.12). We will show that this formulation provides the basis for computationally efficient statistical inference.

Observation models for infectious disease dynamics are often generalized linear models (GLMs) with link functions that depend only on one component of the unobserved Markov

population process $\boldsymbol{X}_N(t)$. For example, consider a Markov population model that tracks the number of susceptible, exposed, and infected individuals in a population $\boldsymbol{X}_N(t) = (S_N(t), E_N(t), I_N(t))$. Further, assume the observation process $y(t) \sim f\left(I_N(t), \boldsymbol{\theta}\right)$, where $f(\cdot)$ is some GLM, relies only on the number of infected individuals $\boldsymbol{I}_N$. In such cases, we would like to obtain a marginal distribution for $\pi(\boldsymbol{I}_N|\boldsymbol{\theta})$ from the LNA in (4.13). The marginal distribution is obtained by forming a permutation matrix P, such that $P\boldsymbol{X}_N = (\boldsymbol{I}_N, \boldsymbol{S}_N, \boldsymbol{E}_N)'$. We then have $\pi(P\boldsymbol{X}_N|\boldsymbol{\theta}) \sim N\left(P\boldsymbol{X}^{\dagger}_{\boldsymbol{\theta}}, \frac{1}{N}PQ^{-1}_{\boldsymbol{\theta}}P'\right)$ which gives $\pi(\boldsymbol{I}_N|\boldsymbol{\theta}) \sim N\left(\boldsymbol{I}^{\dagger}_{\boldsymbol{\theta}}, \frac{1}{N}\tilde{Q}^{-1}_{\boldsymbol{\theta}}\right)$, where $\tilde{Q}_{\boldsymbol{\theta}}$ can be obtained from Schur compliment of precision matrix $PQ_{\boldsymbol{\theta}}P'$. We demonstrate the usefulness of this fact in our data analysis of Section 4.6.

4.4 Laplace Approximations for Inference with LNAs

LNA approximations of Section 4.3 have been used to perform inference on Markov population model with GLM link functions by relying on sampling methods such as MCMC and sequential Monte Carlo (Komorowski et al., 2009; Fearnhead et al., 2014). Computing in these methods is complicated by, among other things, the need for stochastic infill of time points and latent states without observations. In this section, we demonstrate how to perform Bayesian inference on with LNAs with Laplace approximations that offer parallel computing and require no stochastic infill.

4.4.1 Approximate Marginal Distributions for $\pi(\boldsymbol{\theta}|\boldsymbol{y})$

We consider a hidden Markov population model with time referenced observations where $\boldsymbol{y} = (y_{t_1}, y_{t_2}, ..., y_{t_T})'$ are independently distributed according to some GLM f, $y_t \sim f(\boldsymbol{X}_N(t), \boldsymbol{\theta}_2)$. We model the unobserved Markov population process with an LNA distribution $\pi\left(\boldsymbol{X}_N|\boldsymbol{\theta}_1\right) \sim N\left(\boldsymbol{X}^{\dagger}_{\boldsymbol{\theta}_1}, Q^{-1}_{\boldsymbol{\theta}_1}\right)$ as defined in (4.9). We complete the hierarchical model by assuming a prior distribution on $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) \sim \pi(\boldsymbol{\theta})$.

Our goal is to obtain marginal posterior distributions $\pi\left(\theta_i|\boldsymbol{y}\right)$ and thus perform inference on parameters in the Markov population model. We obtain $\pi\left(\theta_i|\boldsymbol{y}\right)$ by integrating the joint marginal $\pi\left(\boldsymbol{\theta}|\boldsymbol{y}\right)$ over $\boldsymbol{\theta}_{-i}$. To perform this integration, we must approximate the joint marginal density

$$\pi\left(\boldsymbol{\theta}|\boldsymbol{y}\right) = \frac{\pi\left(\boldsymbol{X}_N, \boldsymbol{\theta}|\boldsymbol{y}\right)}{\pi\left(\boldsymbol{X}_N|\boldsymbol{\theta}, \boldsymbol{y}\right)}. \tag{4.14}$$

Following Rue et al. (2009), we obtain a Laplace approximation to (4.14) by first performing a Gaussian approximation to $\tilde{\pi}\left(\boldsymbol{X}_N|\boldsymbol{\theta},\boldsymbol{y}\right)$ denoted $\tilde{\pi}\left(\boldsymbol{X}_N|\boldsymbol{\theta},\boldsymbol{y}\right)$. In doing so, we approximate (4.14) by

$$\tilde{\pi}\left(\boldsymbol{\theta}|\boldsymbol{y}\right) \propto \left.\frac{\pi\left(\boldsymbol{y}|\boldsymbol{X}_N,\boldsymbol{\theta}_2\right)\pi\left(\boldsymbol{X}_N|\boldsymbol{\theta}_1\right)\pi(\boldsymbol{\theta})}{\tilde{\pi}\left(\boldsymbol{X}_N|\boldsymbol{\theta},\boldsymbol{y}\right)}\right|_{\boldsymbol{X}_N=\boldsymbol{\mu}_{\boldsymbol{\theta}}}, \tag{4.15}$$

where $\boldsymbol{\mu}_{\boldsymbol{\theta}}$ is the mode of $\tilde{\pi}\left(\boldsymbol{X}_N|\boldsymbol{\theta},\boldsymbol{y}\right)$. Details on locating the mode $\boldsymbol{\mu}_{\boldsymbol{\theta}}$ are included in the following section. We obtain approximate marginal densities $\tilde{\pi}\left(\theta_i|\boldsymbol{y}\right)$ by numerically integrating the approximate joint density $\tilde{\pi}\left(\boldsymbol{\theta}|\boldsymbol{y}\right)$ over $\boldsymbol{\theta}_{-i}$.

4.4.2 Gaussian Approximations of $\pi\left(\boldsymbol{X}_N|\boldsymbol{\theta},\boldsymbol{y}\right)$

We use a second-order Taylor expansion to obtain a Gaussian approximation of the posterior density $\pi\left(\boldsymbol{X}_N|\boldsymbol{\theta},\boldsymbol{y}\right)$ as performed in (Rue et al., 2009). We let $g\left(\boldsymbol{X}_N\right) = \sum_i \log\left(\pi\left(y_{t_i}|\boldsymbol{X}_N(t_i),\boldsymbol{\theta}_2\right)\right)$. We use a second order Taylor expansion of $g\left(\boldsymbol{X}_N\right)$ about $\boldsymbol{\mu}_{\boldsymbol{\theta}}$, to obtain

$$g\left(\boldsymbol{X}_N\right) \approx g(\boldsymbol{\mu}_{\boldsymbol{\theta}}) + Dg\left(\boldsymbol{\mu}_{\boldsymbol{\theta}}\right)\left(\boldsymbol{X}_N - \boldsymbol{\mu}_{\boldsymbol{\theta}}\right) - \frac{1}{2}\left(\boldsymbol{X}_N - \boldsymbol{\mu}_{\boldsymbol{\theta}}\right)'\left[-H(\boldsymbol{\mu}_{\boldsymbol{\theta}})\right]\left(\boldsymbol{X}_N - \boldsymbol{\mu}_{\boldsymbol{\theta}}\right), \tag{4.16}$$

where $H\left(\boldsymbol{\mu}_{\boldsymbol{\theta}}\right)$ is the Hessian and $Dg\left(\boldsymbol{\mu}_{\boldsymbol{\theta}}\right)$ denotes the gradient of $g\left(\boldsymbol{X}_N\right)$ evaluated at $\boldsymbol{\mu}_{\boldsymbol{\theta}}$. From (4.16), we have $\exp\left(g(\boldsymbol{X}_N)\right) \sim N\left(-H^{-1}(\boldsymbol{\mu}_{\boldsymbol{\theta}})\left(H(\boldsymbol{\mu}_{\boldsymbol{\theta}})\boldsymbol{\mu}_{\boldsymbol{\theta}} + Dg\left(\boldsymbol{\mu}_{\boldsymbol{\theta}}\right)\right), -H^{-1}(\boldsymbol{\mu}_{\boldsymbol{\theta}})\right)$.

In most cases, $\boldsymbol{\mu}_{\theta}$ must be found numerically. We use the Fisher scoring algorithm (Fahrmeir and Tutz, 2013) to match the modal configuration of $\tilde{\pi}\left(\boldsymbol{X}_N|\boldsymbol{\theta},\boldsymbol{y}\right)$ by iteratively solving

$$\boldsymbol{\mu}_{\boldsymbol{\theta}}^{(n+1)} = \left[Q_{\boldsymbol{\theta}_1} - H\left(\boldsymbol{\mu}_{\boldsymbol{\theta}}^{(n)}\right)\right]^{-1}\left[Q_{\boldsymbol{\theta}_1}\boldsymbol{X}_{\boldsymbol{\theta}_1}^{\dagger} + Dg\left(\boldsymbol{\mu}_{\boldsymbol{\theta}}^{(n)}\right) - H\left(\boldsymbol{\mu}_{\boldsymbol{\theta}}^{(n)}\right)\boldsymbol{\mu}_{\boldsymbol{\theta}}^{(n)}\right], \tag{4.17}$$

until $\boldsymbol{\mu}_{\boldsymbol{\theta}}^{(n)}$ converges to an estimate of the mode $\boldsymbol{\mu}_{\theta}^{*}$. We then have approximate Gaussian density $\tilde{\pi}(\boldsymbol{X_N}|\boldsymbol{\theta},\boldsymbol{y}) \sim N\left(\boldsymbol{\mu}_{\theta}^{*}, \left[Q_{\boldsymbol{\theta}_1} - H\left(\boldsymbol{\mu}_{\boldsymbol{\theta}}^{*}\right)\right]^{-1}\right)$, which is used to evaluate (4.15) as

$$\tilde{\pi}\left(\boldsymbol{\theta}|\boldsymbol{y}\right) \propto \frac{\pi\left(\boldsymbol{y}|\boldsymbol{\mu}_{\theta}^{*},\boldsymbol{\theta}_2\right)\pi\left(\boldsymbol{\mu}_{\theta}^{*}|\boldsymbol{\theta}_1\right)\pi(\boldsymbol{\theta})}{\left|Q_{\boldsymbol{\theta}_1} - H\left(\boldsymbol{\mu}_{\boldsymbol{\theta}}^{*}\right)\right|^{-1/2}}. \tag{4.18}$$

4.4.3 Estimating $\pi\left(\theta_i|\boldsymbol{y}\right)$

We use the approximate joint marginal density in (4.18) to approximate the individual marginal densities by solving

$$\tilde{\pi}\left(\theta_i|\boldsymbol{y}\right) = \int \tilde{\pi}\left(\boldsymbol{\theta}|\boldsymbol{y}\right) d\boldsymbol{\theta}_{-i}, \tag{4.19}$$

for each parameter θ_i. We note that any numerical scheme used to approximate (4.19) will require a set of density evaluations $\left\{\boldsymbol{\theta}^{(j)}, \tilde{\pi}\left(\boldsymbol{\theta}^{(j)}|\boldsymbol{y}\right)\right\}_{j=1}^{J}$. We further note that the computational cost of evaluating $\tilde{\pi}\left(\boldsymbol{\theta}|\boldsymbol{y}\right)$ is dominated by the ODE solves required to construct of the LNA distribution.

We reduce the number of density evaluations required for numerical integration by obtaining values of $\boldsymbol{\theta}^{(j)}$ that provide good integration weights $w^{(j)} = \tilde{\pi}\left(\boldsymbol{\theta}^{(j)}|\boldsymbol{y}\right)$. We follow Section 3.1 of Rue et al. (2009) to explore the parameter space of $\boldsymbol{\theta}$. We begin by locating the mode of $\log\left(\tilde{\pi}\left(\boldsymbol{\theta}|\boldsymbol{y}\right)\right)$ denoted $\boldsymbol{\theta}^*$. We then perform a singular value decomposition, $V\Lambda V'$, of the inverse negative Hessian matrix of $\log\left(\tilde{\pi}\left(\boldsymbol{\theta}|\boldsymbol{y}\right)\right)$ at $\boldsymbol{\theta}^*$. We explore the parameter space by scaling and rotating $\boldsymbol{\theta}$ to obtain, $\boldsymbol{\theta}(\boldsymbol{z}) = \boldsymbol{\theta}^* + V\Lambda^{1/2}\boldsymbol{z}$.

Integration weights $w^{(j)}$ are obtained by the following algorithm: Start at the mode ($\boldsymbol{\theta}^* = \boldsymbol{\theta}(\boldsymbol{0})$). For each component, move in the positive direction of z_i with step-length δ_z, storing $(\theta(\boldsymbol{z}), \tilde{\pi}\left(\boldsymbol{\theta}(\boldsymbol{z})|\boldsymbol{y}\right))$ until

$$\left|\log\left(\tilde{\pi}\left(\boldsymbol{\theta}^*|\boldsymbol{y}\right)\right) - \log\left(\tilde{\pi}\left(\boldsymbol{\theta}(\boldsymbol{z})|\boldsymbol{y}\right)\right)\right| > \delta_\pi. \tag{4.20}$$

Repeat this process in the negative direction of z_i. We then fill in our grid by and considering all the combinations of the $\boldsymbol{z}$ accepted in the exploration algorithm. We perform a nonlinear interpolation of all the collected grid locations and weights $\left\{\boldsymbol{\theta}^{(j)}, w^{(j)}\right\}_{j=1}^{J}$ to assist in numerically integrating (4.19).

We note that the task of exploring the parameter space outlined in this section can become computationally taxing for small step-sizes of δ_z and/or large values of δ_π. We are assisted by the fact that the exploration in each component direction can be performed independent of the other components. Further, evaluation of the infill locations can be performed in parallel as well. We also note that the latent states $\boldsymbol{X}_N$ were estimated at the modal configuration $\boldsymbol{\mu}_{\boldsymbol{\theta}}$, in stark contrast to the MCMC and sequential Monte Carlo methods of (Fearnhead et al., 2014; Komorowski et al., 2009) which require $\boldsymbol{X}_N$ to be stochastically estimated. In summary, inference with Laplace approximations offer a fixed number of LNA evaluations which which can be performed in parallel and does not

require stochastic infill.

4.5 Simulation Study

In this section we demonstrate that LNAs fit with Laplace approximations offer similar inference to MCMC. We simulate a stochastic infected proportion $P(t)$ trajectory from an SIS model (see Section 4.2.1) with population size $N = 1000$ and $\boldsymbol{\theta} = (0.50, 0.33)$ using the Gillespie algorithm over the time interval $[0, 50]$ with $P(0) = 0.05$. We assume the process $y_t \sim Binomial\,(50, P(t))$ is observed at time points $t = 1, 2, 3, ..., 50$. We denote the time-referenced observations and latent proportions of infected by $\boldsymbol{y} = (y_1, y_2, ..., y_{50})'$ and $\boldsymbol{P} = (P(1), P(2), ..., P(50))'$ respectively.

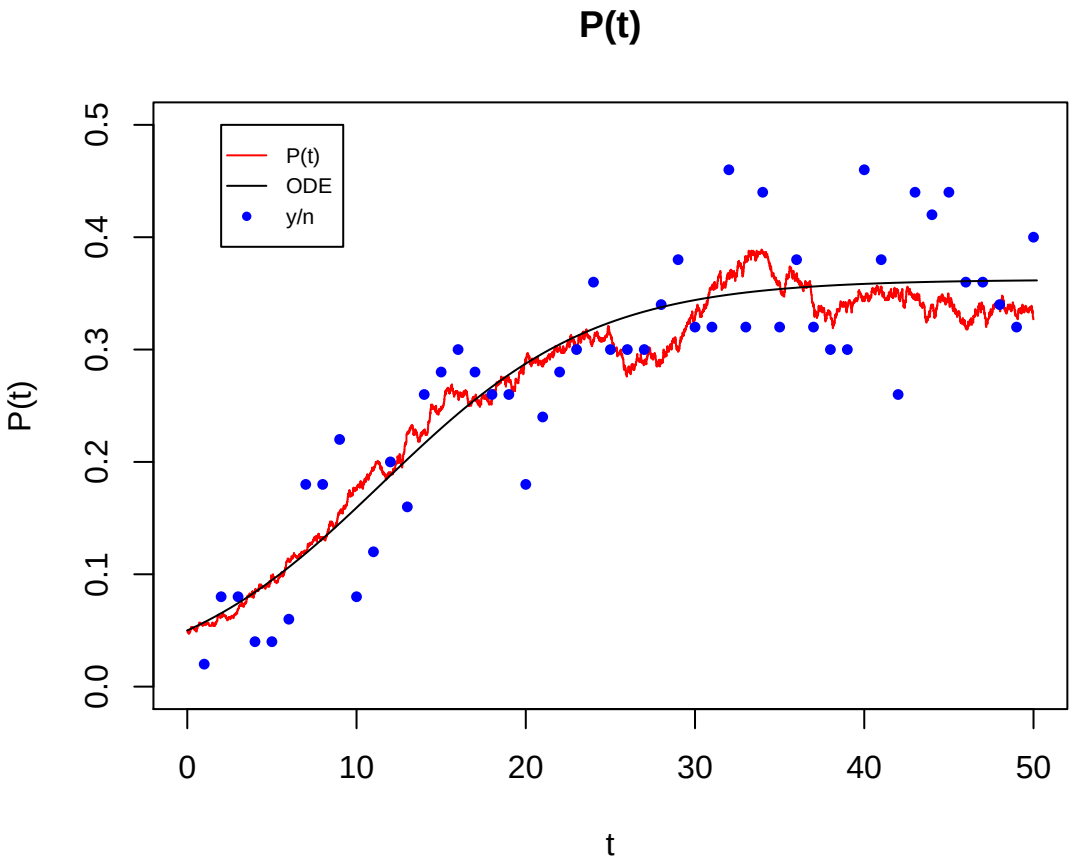

Figure 4.2: A plot of the stochastic trajectory of $P(t)$ (red), the ODE fit $P^{\dagger}_{\hat{\boldsymbol{\theta}}}(t)$ (black) for the posterior mean estimates of the Laplace approximation $\hat{\boldsymbol{\theta}} = (0.4383, 0.2798)$, and the observed sample proportions $y_t/50$ (blue).

We model the latent proportions of infected individuals with an LNA distribution $\pi(\boldsymbol{P}|\boldsymbol{\theta}) \sim N\left(\boldsymbol{P}^{\dagger}_{\boldsymbol{\theta}}, \frac{1}{N}Q^{-1}_{\boldsymbol{\theta}}\right)$, where $\boldsymbol{P}^{\dagger}_{\boldsymbol{\theta}} = \left(P^{\dagger}_{\boldsymbol{\theta}}(1), P^{\dagger}_{\boldsymbol{\theta}}(2), ..., P^{\dagger}_{\boldsymbol{\theta}}(50)\right)'$ and $Q_{\boldsymbol{\theta}}$ are obtained by solving $dP^{\dagger}_{\boldsymbol{\theta}}(t) = \boldsymbol{F}_{\boldsymbol{\theta}}\left(P^{\dagger}_{\boldsymbol{\theta}}(t)\right)dt$ and following Section 4.3.3. We also assume half normal priors with centrality parameter 0 and scale parameters of 1 and 0.25 respectively.

We first fit the model by estimating each latent proportion of infected individuals as a baseline to compare with the Laplace approximations of Section 4.4. We sequentially draw Metropolis-Hasting samples form $\pi(\log(\boldsymbol{\theta})|\boldsymbol{y}, \boldsymbol{P}) \propto \pi(\boldsymbol{P}|\boldsymbol{\theta})\pi(\log(\boldsymbol{\theta}))$, and then draw one-at-a-time Metropolis Hastings samples from the posterior conditional distribution of each latent infected proportion $\pi(P(t)|P(t-1), P(t+1), \boldsymbol{\theta}, \boldsymbol{y}) \propto \pi(y_t|P(t))\pi(P(t)|P(t-1), P(t+1), \boldsymbol{\theta})$. We note that every iteration of MCMC requires the LNA distribution to be constructed to evaluate the Metropolis-Hastings accept/reject ratio for $\pi(\log(\boldsymbol{\theta})|\boldsymbol{y}, \boldsymbol{P})$. Further, we must iterate until convergence of the Markov chain. Long running chains become costly, as we must evaluate the LNA likelihood at every iteration of MCMC until we have reached a stationary distribution. For this study we drew 200,000 Metropolis-Hastings samples, storing 100,000 for the analysis.

Next, we fit the SIS model using the Laplace approximation detailed in Section 4.4. We used *R*'s built in optimizer "*optim*" (R Core Team, 2021) with Nelder-Meade setting to locate the mode of $\log\left(\tilde{\pi}\left(\log(\boldsymbol{\theta})|\boldsymbol{y}\right)\right)$ denoted $\log(\boldsymbol{\theta}^*)$. The optimization algorithm converged after 59 LNA likelihood evaluations. We performed the grid search detailed in Section 4.4.3 with $\delta_z = 1/4$ and $\delta_\pi = 5$ for $\log(\boldsymbol{\theta})$ which required 1140 LNA likelihood evaluations, which were evaluated in parallel using 6 cores. We constructed an interpolate of $\tilde{\pi}(\boldsymbol{\theta}|\boldsymbol{y})$ by fitting a cubic-spline to the grid locations. We used the interpolate to assist in numerically evaluating (4.19) to obtain the desired marginal densities $\tilde{\pi}(\beta|\boldsymbol{y})$ and $\tilde{\pi}(\gamma|\boldsymbol{y})$.

We observe similar posterior mean estimates for β and γ for the Laplace approximation and MCMC model fits. We note that the 95% highest posterior density intervals (HPDs) for the marginal distributions produced by both methods cover the true parameter values. We observe smaller HPD intervals for the Laplace approximation fits, suggesting the Laplace approximation produces less variance in its parameter estimates. We restate that the Laplace approximation required 1199 LNA likelihood evaluations, with 1140 of the grid-search evaluations performed in parallel. In comparison, MCMC required 200,000 LNA likelihood evaluations. Further, no stochastic infill was required for the Laplace approximation fit. In summary Laplace approximations offer similar inference accuracy to MCMC without the need for stochastic infill at a highly reduced computational cost.

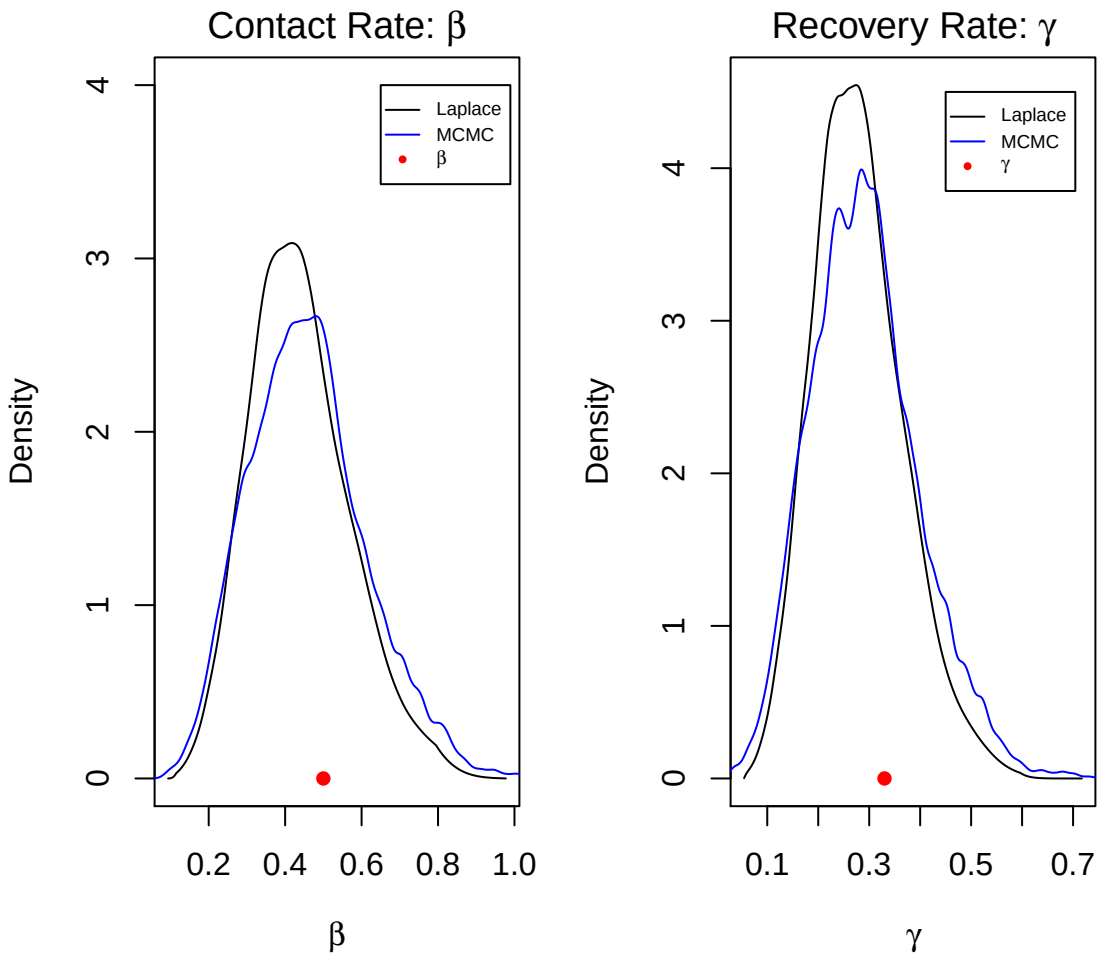

Figure 4.3: Marginal density plots of $\pi(\beta|\boldsymbol{y})$ and $\pi(\gamma|\boldsymbol{y})$ fit by Laplace approximations (black) and MCMC (blue). The true values of $\beta = 0.50$ and $\gamma = 0.33$ are plotted in red.

Parameter	Method	Posterior Mean	95% HPD	HPD Width
β	Laplace	0.4383	(0.1919, 0.7035)	0.5116
β	MCMC	0.4587	(0.1811 0.7671)	0.5860
γ	Laplace	0.2798	(0.1125, 0.4571)	0.3446
γ	MCMC	0.2917	(0.0930 0.5019)	0.4089

Table 4.1: Posterior mean estimates and 95% highest posterior density (HPD) intervals for the Laplace approximation (Laplace) and MCMC model fits on the simulated SIS data set. True parameter values are $\beta = 0.50$ and $\gamma = 0.33$.

4.6 Data Analysis: COVID Cruise Ship

On January 25^{th} 2020, a COVID-19 positive passenger boarded the Princess Diamond cruise ship (Lai et al., 2021). On 5 February 2020, the cruise ship hosting 3711 people docked for a 2-week quarantine in Yokohama, Japan after 10 passengers tested positive for the coronavirus disease (COVID-19) (Mizumoto et al., 2020). Random testing of passengers and crew members began on February 5^{th} and continued through February 20^{th}. We note that on February 11^{th} and 14^{th} no testing occurred and denote the fourteen observed times $T^{obs} = \{16, 17, , .., 21, 23, 24, 26, 27, ..., 31\}$. We let January 25^{th} denote the index case of $t = 0$, in which the first infected individual boarded the ship. The number of tests administered per day n_t, positive tests per day y_t, and total number of individuals remaining on the ship N_t are shown in Table 4.2. A plot of the observed infected proportions $\frac{y_t}{n_t}$ can be seen in Figure 4.4.

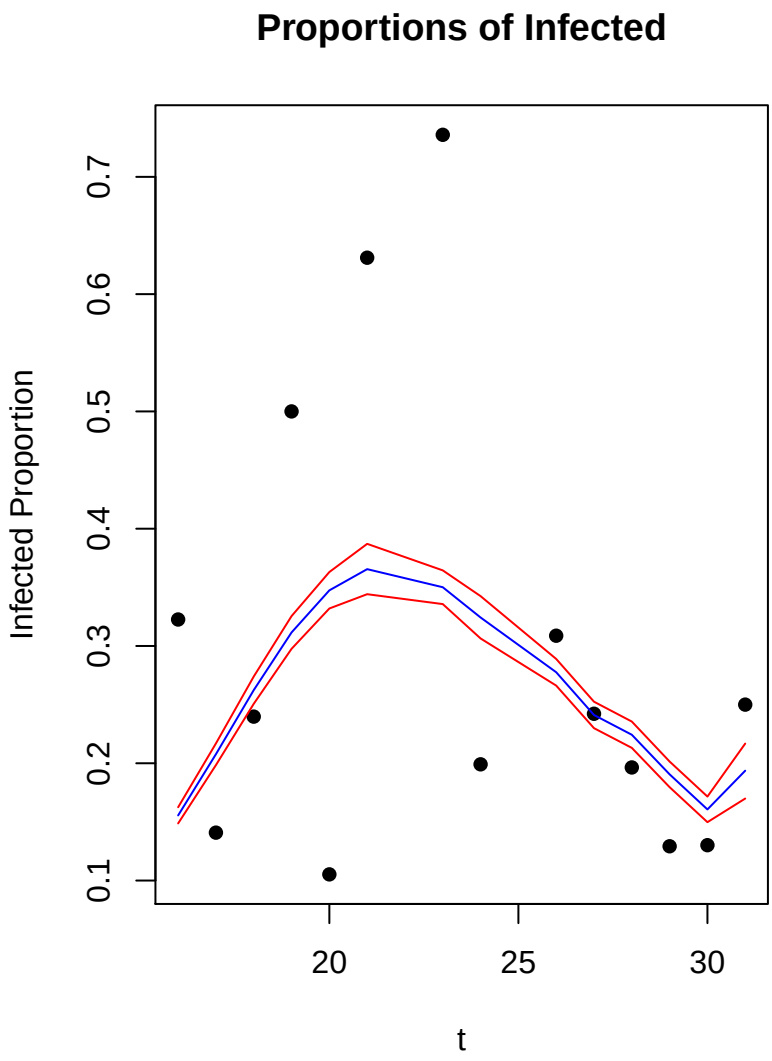

Figure 4.4: A plot of the observed proportions of seropositive individuals aboard the Princess Diamond cruise ship (black), the ODE solution $P^{\dagger}_{\hat{\boldsymbol{\theta}}}(t)$ (blue) for posterior mean estimates of $\boldsymbol{\theta}$ from fitting a Laplace approximation, and 95% credible intervals of $P^{\dagger}_{\hat{\boldsymbol{\theta}}}(t)$ generated by simulating from the LNA (red).

Once exposed to COVID, susceptible individuals experience an incubation period prior to becoming infectious (Chen et al., 2020). Lai et al. (2021) fit a deterministic susceptible-exposed-infected-removed (SEIR) model to the Princess Diamond cruise ship COVID-19 outbreak data set to account for the incubation period. We fit a stochastic

SEIR model by first specifying the system of ODEs governing our SEIR model

$$\frac{d}{dt}S^{\dagger}(t) = -\beta I^{\dagger}(t)S^{\dagger}(t) - \mu_S(t)S^{\dagger}(t), \tag{4.21}$$

$$\frac{d}{dt}E^{\dagger}(t) = \beta I^{\dagger}(t)S^{\dagger}(t) - \alpha E^{\dagger}(t), \tag{4.22}$$

$$\frac{d}{dt}I^{\dagger}(t) = \alpha E^{\dagger}(t) - \gamma I^{\dagger}(t), \tag{4.23}$$

$$\frac{d}{dt}R^{\dagger}(t) = \gamma I^{\dagger}(t) + \mu_S(t)S^{\dagger}(t), \tag{4.24}$$

where $I^{\dagger}(t)+S^{\dagger}(t)+E^{\dagger}(t)+R^{\dagger}(t) = 1$. We assume one individual is infected at time $t = 0$ and all other 3710 passengers are considered susceptible. We note that the system contains three compartments $\boldsymbol{X}^{\dagger}_{\boldsymbol{\theta}}(t) = \left(S^{\dagger}(t), E^{\dagger}(t), I^{\dagger}(t)\right)'$, since $R^{\dagger}(t) = 1 - S^{\dagger}(t) - E^{\dagger}(t) - I^{\dagger}(t)$. We account for the disembarkment of susceptible passengers with the inclusion of a fixed time-varying rate $\alpha_S(t)$ estimated from passenger records. The unknown parameters $\boldsymbol{\theta} = (\beta, \alpha, \gamma)$, consist of the contact rate β, incubation rate α, and the recovery rate γ, all with support on the positive reals.

We model the number of observed seropositive individuals on day t as

$$\pi(y_t|I_N(t)) \sim Binom\left(n_t, P(t) = I_N(t)\frac{N}{N_t}\right). \tag{4.25}$$

We define the four reaction vectors; a susceptible individual becomes exposed $\boldsymbol{R}_1 = (-1, 1, 0)'$, a susceptible individual disembarks from the ship $\boldsymbol{R}_2 = (-1, 0, 0)'$, an exposed individual becomes infected $\boldsymbol{R}_3 = (0, -1, 1)'$, or an infected individual recovers $\boldsymbol{R}_4 = (0, 0, -1)'$. The corresponding reaction rates are denoted $\lambda^1_{\boldsymbol{\theta}}\left(\boldsymbol{X}^{\dagger}(t)\right) = \beta I^{\dagger}(t)S^{\dagger}(t)$, $\lambda^2_{\boldsymbol{\theta}}\left(\boldsymbol{X}^{\dagger}(t)\right) = \mu_S(t)S^{\dagger}(t)$, $\lambda^3_{\boldsymbol{\theta}}\left(\boldsymbol{X}^{\dagger}(t)\right) = \alpha E^{\dagger}(t)$, and $\lambda^4_{\boldsymbol{\theta}}\left(\boldsymbol{X}^{\dagger}(t)\right) = \gamma I^{\dagger}(t)$. Let $\boldsymbol{X}_N$ and $\boldsymbol{X}^{\dagger}_{\boldsymbol{\theta}}$ denote the latent population states and the ODE solution at times T^{obs} respectively. We construct an LNA distribution for the latent proportions $\pi(\boldsymbol{X}_N|\boldsymbol{\theta}) \sim N\left(\boldsymbol{X}^{\dagger}_{\boldsymbol{\theta}}, \frac{1}{N}Q^{-1}_{\boldsymbol{\theta}}\right)$ from the system of ODEs in (4.21–4.23) as described in Section 4.3. We assign independent mean-zero half-normal priors to β, α, and γ with scale parameters 0.50, 0.10 and 0.10 respectively.

We note that the likelihood in (4.25) only requires $I_N(t)$ from $\boldsymbol{X}_N(t) = (S_N(t), E_N(t), I_N(t))'$. We define P such that $P\boldsymbol{X}_N = (\boldsymbol{I}_N, \boldsymbol{S}_N, \boldsymbol{E}_N)'$. Then the LNA distribution for the proportion of infected individuals is given by $\pi(\boldsymbol{I}_N|\boldsymbol{\theta}) \sim N\left(\boldsymbol{I}^{\dagger}_{\boldsymbol{\theta}}, \frac{1}{N}\tilde{Q}^{-1}_{\boldsymbol{\theta}}\right)$, where $\tilde{Q}_{\boldsymbol{\theta}}$ is the Schur compliment of $PQ^{-1}_{\boldsymbol{\theta}}P'$. We locate the mode of the Gaussian approximation to $\pi(\boldsymbol{I}_N|\boldsymbol{\theta}, \boldsymbol{y})$ using the Nelder-Meade setting in R's built-in optimizer "*optim*". We explore the marginal distribution $\log\left(\tilde{\pi}(\log\boldsymbol{\theta}|\boldsymbol{y})\right)$ for integration nodes by following Section

4.4.3 with $\delta_z = 1/2$ and $\delta_\pi = 5$. We fit a cubic spline model to interpolate the in-fill grid locations and numerically integrate $\tilde{\pi}(\boldsymbol{\theta}|\boldsymbol{y})$ using a quadrature scheme as done in Section 4.4.3 to obtain$\tilde{\pi}(\theta_i|\boldsymbol{y})$. A full derivation of the LNA distribution and Laplace approximation are included in the Appendix.

We observe that the ODE solution evaluated from the posterior mean estimates for $\boldsymbol{\theta}$ captures the mean proportion of the observed infected well (see Figure 4.4). We report posterior mean estimates and corresponding 95% HPD intervals for all SEIR parameters in Table 4.3. We note that our posterior mean estimate of the incubation period $\hat{\alpha}^{-1} \approx 2$ days is lower than incubation estimates of roughly five to six days estimated in other studies (Chen et al., 2020). Our estimate is most likely lower due to the fact that several tests were administered to a small population (i.e. many were tested positive before symptoms began). We also note that the posterior mean estimate for the recovery/removal time $\hat{\gamma}^{-1}$ is roughly 5 days. This estimate is lower than the known time to recover from SARS-COV2, as individuals on the ship were not tested once found positive, and likely confined to their rooms (i.e. quarantined), and are thus functionally removed from the population, despite remaining infectious.

The outbreak on the Princess Diamond cruise ship provided an opportunity to estimate transmissibility and the basic reproductive rate of COVID-19 at an early stage of the global pandemic (Lai et al., 2021). The basic reproductive rate $R_0 = \frac{\beta}{\gamma}$ is the expected number of infections directly generated by one infectious individual in a population where all individuals are assumed susceptible to infection (Keeling and Rohani, 2011). We obtain an approximate density for R_0 by using rejection sampling to draw independent samples from $\tilde{\pi}(\beta|\boldsymbol{y})$ and $\tilde{\pi}(\gamma|\boldsymbol{y})$ (see Figure 4.5). Our estimate for R_0 is roughly 6.3458, with 95% HPD of $(5.0454, 7.7608)$. This is alarmingly high, given the basic reproductive number of SARS has been estimated to be between $2 - 3$ (Riley et al., 2003; Lai et al., 2021), suggesting COVID-19 is highly contagious and easily transmitted in closed populations.

t	Date	Positive tests (y_t)	Number of Tests (n_t)	On Ship (N_t)
0	Jan, 25	1	1	3711
16	Feb, 5	10	31	3711
17	Feb, 6	10	71	3711
18	Feb, 7	41	171	3711
19	Feb, 8	3	6	3711
20	Feb, 9	6	57	3711
21	Feb, 10	65	103	3711
22	Feb, 11	NA	NA	3711
23	Feb, 12	39	53	3711
24	Feb, 13	44	221	3711
25	Feb, 14	NA	NA	3451
26	Feb, 15	67	217	3451
27	Feb, 16	70	289	3451
28	Feb, 17	99	504	3183
29	Feb, 18	88	681	3183
30	Feb, 19	79	607	3183
31	Feb, 20	13	52	2213

Table 4.2: Covid data

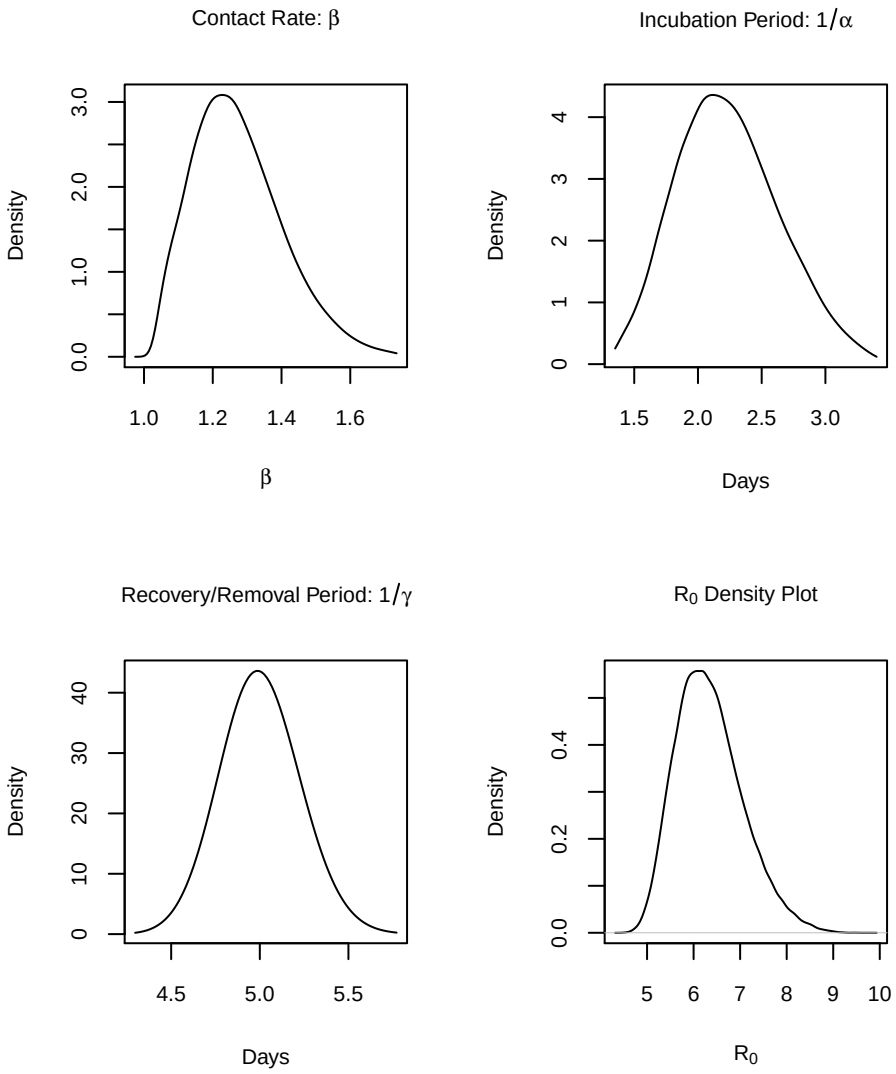

Figure 4.5: Marginal density plots ($\pi(\theta_i|\boldsymbol{y})$).

Parameter	Posterior Mean	95% HPD
β	1.2739	(1.0401 , 1.5296)
$\frac{1}{\alpha}$	2.098	(1.386 , 2.8244)
$\frac{1}{\gamma}$	4.9813	(4.5421 , 5.4262)
R_0	6.3458	(5.0454 , 7.7608)

Table 4.3: Posterior mean estimates and 95% HPD intervals for the SEIR model fit to the Princess Diamond cruise ship data set.

4.7 Discussion

In this work we derived a Laplace approximation for approximate Bayesian inference on Markov population models. We constructed a sparse covariance matrix for the LNA likelihood based on the solution a system of ODEs. We demonstrated that Laplace approximations do not require stochastic infill and can easily be fit using parallel computations. We showed that Laplace approximations offer similar inference accuracy to MCMC with stochastic infill at a highly reduced computational cost on a simulated SIS data set. In summary, Laplace approximations for LNAs provide a fast method for approximate Bayesian inference on Markov population models.

Komorowski et al. (2009) used the full joint distribution of the LNA to perform inference on Markov population models by estimating the latent population states via MCMC similar to Section 4.5. MCMC samplers must be sequentially run until convergence to stationary distribution. Every iteration of MCMC requires a numerical solve of the system of differential equations must be solved to construct the LNA distribution. We considered a comparison with MCMC using the joint LNA likelihood in Section 4.5 because the marginal distributions estimated Laplace approximations are our target distributions. We clearly showed that the Laplace approximations only require a fixed number of likelihood evaluations, most of which can be performed in parallel. To fit the LNA via MCMC, we were not only required to construct the LNA likelihood, but we were also required to draw samples of the latent proportions of infected. For these reasons, LNAs with Laplace approximations is a less computationally intensive task than fitting LNAs via MCMC as considered in Komorowski et al. (2009).

The accuracy of LNA models that rely on the full joint distribution of the latent states was investigated by Giagos (2010) and Fearnhead et al. (2014). LNAs have been shown to perform poorly on systems in which the ODE model becomes a poor estimate of the mean behavior of the stochastic model over long periods of time (Giagos, 2010; Fearnhead et al., 2014). Fearnhead et al. (2014) considered fitting the LNA over each partitioned time interval $[t_{i-1}, t_i]$, in contrast to fitting an LNA over the entire time interval $[0, t_T]$. Fearnhead et al. (2014) then used a particle filter to perform inference. The length of the time series was not a concern in the data analysis performed in Section 4.6 as data was collected over a short time interval. For this reason, we did not compare our method to Fearnhead et al. (2014).

Lai et al. (2021) fit a deterministic SEIR model to the Princess Diamond cruise ship data set. Lai et al. (2021) modeled the total cumulative counts of confirmed infectious

status and unconfirmed status. We used a binomial likelihood in our analysis, as we had knowledge of both the number of tests administered and positive tests. The model of Lai et al. (2021) attempted to estimate transmission rates on each deck level of the cruise ship. We did not include this in our model as the data surrounding this information was quite sparse. Lai et al. (2021) did not account for the change in populations due to passengers disembarking from the ship. We accounted for disembarkment in both our likelihood and SEIR dynamics.

We note that Rue et al. (2009) have explored nested Laplace approximations for Gauss-Markov processes. Their method for approximate Bayesian inference begins by performing a Laplace approximation for $\pi(\boldsymbol{\theta}|\boldsymbol{y})$ as done in Section 4.4. The integrated nested Laplace (INLA) algorithm of Rue et al. (2009) focuses on obtaining marginal distributions for the latent random effects. The computational speed and accuracy of the INLA algorithm is centered around the assumption that $|\boldsymbol{\theta}|$ is small and the random effects (or latent states) are a Gauss-Markov random field. In Section 4.3, we constructed a sparse precision matrix for the LNA distribution. Further Markov population models are commonly driven by a small number of parameters. For these reasons, we should be able to employ methods similar to INLA to obtain approximate marginal distributions for the latent population states. We also note that more accurate Gaussian approximations have been performed than suggested in Section 4.4 which correct for skewness (Rue et al., 2009; Ferkingstad et al., 2015). We did incorporate this here and leave this and other improvements to future work.

In summary, we demonstrated that Laplace approximations for LNAs provide a fast method for approximate Bayesian inference on Markov population models. We derived a sparse precision matrix for the LNA distribution, which was used to perform a Gaussian approximation of the posterior distribution of latent states. We showed that Laplace approximations offer similar inference to MCMC methods that rely on stochastic infill for latent population states. We demonstrated that Laplace approximations can be fit in parallel and do not require stochastic infill to perform inference on Markov population models. In turn, our method offers a fast alternative to the inference methods for LNAs that rely on stochastic infill used in Komorowski et al. (2009) and Fearnhead et al. (2014).

Chapter 5
Conclusion

In this work we presented three projects on Bayesian methods for SGLMMs, data privacy, and Markov population models. Each work proposed novel statistical and computational advances in penalization methods for SGLMMs, Bayesian privacy for spatial point process data, and Bayesian inference for Markov population models. In the following sections we address three major topics surrounding each project: (1) What scientific problems can be solved by this work, (2) What are the novel statistical contributions of this work, and (3) What are the limitations and/or future contributions of this work.

5.1 Concluding Remarks for Chapter 2

In Chapter 2, we developed a Bayesian hierarchical model for fitting SGLMMs with LMAs in both discrete and continuous space. Our implementation offers computing and inference that is analogous to SGLMMs fit with GRFs. In our extensive data analysis, we demonstrated that the computational complexity of the LMA model fit is similar to the GRF model fit. Since parameter inference is directly analogous for both LMAs and GRFs, we suggest the use of LMAs in any case where a GRF SGLMM is considered.

The argument for the use of LMAs in place of GRF models was due to an assumption of a localized trend of “spikes" in the observed data. The LMA model can capture a wider range of behavior than the GRF, and also contains the GRF as a limiting case. Thus it provides a general purpose model appropriate for a wide range of spatial data, as it can capture behavior that is not easily captured by the GRF. We note that an SGLMM fit with an LMA that produces a better fit than a GRF is not sufficient evidence to suggest that there is an underlying mechanism responsible for the localized “spikes". Instead, we argue that the LMA better captures the spatially correlated missing co-variate in the SGLMM. Since, the computational trade off and overall model complexity is between

LMAs and GRFs is negligible, we should favor the model that produces the best out-of-sample predictive power. Further, swapping a GRF for an LMA has no impact on the interpretation of fixed effect estimates, which is often the primary goal of inference.

5.2 Concluding Remarks for Chapter 3

In Chapter 3, we developed an approach for generating and disseminating fully-synthetic location-only data. In this work, we constructed a risk metric suited for location-only data sets and tailored the widely used *pMSE* utility metric to LGCPs. Prior to this work, there were no existing model based metrics for evaluating individual disclosure risks and assessing data utility on location-only data sets. Our work offers an improvement in data stewards ability to ensure proper privacy constraints are satisfied before disseminating private location-only data such as crime occurrences and disease case locations for public use.

This work had several novel methodological and computational contributions. We derived CPO estimates for LGCPs that allowed for a computationally efficient evaluation of our individual disclosure risk metric for location-only data. We also provided a Monte Carlo approach for estimating the *pMSE* for synthetic data sets generated from LGCPs. To our knowledge, we were also the first to provide empirical evidence illustrating the fallacies of radial perturbation (Armstrong et al., 1999; Wang and Reiter, 2012; Quick et al., 2015) for anonymizing location-only data sets.

We elected to use John Snow's Cholera outbreak data set in our data analysis section. We would have liked to have use a more recent data set, but were unable to obtain a workable set which could be publicly released due to privacy constraints. We also note that in our analysis we knew the source of the Cholera outbreak was the Broad St. water pump. This was included as a co-variate in the data-generating model. A more non-parametric model may be preferred in cases where data stewards are unaware of the source of a disease outbreak. In future work, we would like to extend our methodology to allow for fully non-parametric data synthesis approaches. We believe this would allow data stewards more flexibility when generating synthetic data sets.

We also note that LGCP intensity surfaces with spatially smooth covariates are generally subject to spatial confounding (Hodges and Reich, 2010; Hanks et al., 2015). We would like to provide a more rigorous study on the bias introduced by spatial confounding when synthesizing data according to the *PRS* and *ANS* methods. Finally, we were unable to make formal guarantees related to differential privacy in this work. In

future works, we would like to construct a synthesis method that satisfies differential privacy.

5.3 Concluding Remarks for Chapter 4

In Chapter 4, we derived Laplace approximations for fast Bayesian inference on Markov population models. We derived the LNA as an approximate joint Gaussian distribution for the time-referenced observations of the Markov population model compartments. Our proposed solution to the LNA resulted in a joint Gaussian density with a sparse precision matrix and mean structure constructed from the solution to a system of ODEs.

This work allows for approximate inference on Markov population models which offers parallel computing and does not require stochastic estimation of the unobserved population states. Models for chronic wasting disease (CWD) and brucella in white tailed deer and elk can be modeled as a SIS process. Spatial movement models are required to accurately describe the disease dynamics of CWD and brucella data. In spatial SIS models, seropositive counts for CWD or brucella are aggregated into regional locations. Markov population models are needed to accurately describe the fine scale dynamics of the small population sizes contained in the spatial regions.

Markov population models that contain spatial movement are composed of several unobserved compartments. For spatial SIS models, susceptible and infected individuals must be stochastically estimated in each spatial region. In practice, we have observed that the stochastic estimation of the latent susceptible and infected compartments produces poor MCMC mixing. In future work, we would use the Laplace approximations discussed in Chapter 4 to fit the spatial SIS model to avoid stochastic infill for the unobserved compartments of the Markov population model.

The extension to spatial movement models is not the only avenue for future work with the Laplace approximations discussed in Chapter 4. We note that marginal distributions for each latent population state can be obtained by following the work of Rue et al. (2009). Marginal distribution for Gaussian random effects in SGLMMs have been obtained by Laplace approximations in the *INLA* package developed by Rue et al. (2009). We plan to investigate this further in future works.

Appendix A
Chapter 2 Appendices

A.1 CAR Models

Definition: (Conditionally Autoregressive Model) A Conditionally Autoregressive Model (CAR) takes on the form

$$\eta_i|\boldsymbol{\eta}_{-i} \sim \mathcal{N}\left(\sum_{\forall C_{ij}\neq 0} C_{ij}\eta_j, M_{ii}\right), \tag{A.1}$$

where $\mathbf{C}$ is the spatial dependence matrix with $C_{ii} = 0$, and $\mathbf{M}$ is a diagonal matrix with entries M_{ii}. The conditional mean of each η_i is determined by a weighted sum of neighboring η_j's. Note that each marginal variance, M_{ii}, varies, so $\mathbf{M}$ is often non stationary.

The CAR model in (A.1) was shown to lead to the full distribution of $\boldsymbol{\eta}$ by Besag (1974). For positive definite $\boldsymbol{Q}^{-1} = (\boldsymbol{I} - \boldsymbol{C})^{-1}\boldsymbol{M}$, (A.1) leads to the full distribution $\boldsymbol{\eta} \sim N(0, \boldsymbol{Q}^{-1})$. Matrices $\mathbf{M}$ and $\mathbf{C}$ are defined from $\mathbf{Q}$ as follows: Write, $\boldsymbol{Q} = \boldsymbol{D} - \boldsymbol{R}$ with

$$R_{ij} = \begin{cases} 0, & \text{if } i = j \\ -Q_{ij}, & \text{if } i \neq j \end{cases}, \qquad \boldsymbol{D} = \begin{cases} Q_{ii}, & \text{if } i = j \\ 0, & \text{if } i \neq j \end{cases}.$$

This gives $\boldsymbol{M} = \boldsymbol{D}^{-1}$ and $\boldsymbol{C} = \boldsymbol{D}\boldsymbol{R}$ in (A.1).

A.2 SAR Models

Ver Hoef et al. (2018) summarized the relationships between SAR and CAR models. CAR and SAR models are widely used in both temporal and spatial statistics due to

their intuitive dependence structures. We provide a brief summary of the SAR model for the unfamiliar audience.

Consider a collection of random variables at n spatial locations or graph nodes, $\boldsymbol{Y} = (Y_1..., Y_n)$. Let $\boldsymbol{\Lambda}$ be a positive diagonal matrix. A SAR model imposes an explicit spatial dependence structure,

$$\boldsymbol{Y} = \boldsymbol{B}\boldsymbol{Y} + \boldsymbol{\nu}, \quad \boldsymbol{\nu} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Lambda}), \tag{A.2}$$

where $\mathbf{B}$ is a spatial dependency matrix that is not necessarily symmetric. Note that $\mathbf{B}$ relates $\boldsymbol{Y}$ to itself, and no site can depend on itself, so B_{ii} must be zero for all i. Solving for $\boldsymbol{Y}$ in (A.2) we have, $(\boldsymbol{I} - \boldsymbol{B})\boldsymbol{Y} \sim \mathcal{N}(0, \boldsymbol{\Lambda})$. The covariance of the SAR model can then be written as $(\boldsymbol{I} - \boldsymbol{B})^{-1}\boldsymbol{\Lambda}[(\boldsymbol{I} - \boldsymbol{B})']^{-1}$, provided $(\boldsymbol{I} - \boldsymbol{B})$ is invertible. For a thorough comparison of SAR and CAR models see Ver Hoef et al. (2018).

A.3 Conditionally Independent Block Proposals

Consider $\boldsymbol{\eta} \sim N(\mathbf{0}, \boldsymbol{Q}^{-1})$, where $\boldsymbol{Q}$ is a GMRF. Define $N(i) := \{j : Q_{ij} \neq 0\}$. $N(i)$ is the collection of indices j, such that i and j are neighboring points in the spatial lattice. η can be expressed as a CAR model with sparse $\mathbf{Q}$

$$\eta_i | \boldsymbol{\eta}_{-i} = \eta_i | \boldsymbol{\eta}_{N(i)} \sim \mathcal{N}\left(\sum_{j \neq i} C_{ij}\eta_j, M_{ii}\right), \tag{A.3}$$

where the procedure to obtain matrices $\mathbf{M}$ and $\mathbf{C}$ are described in Appendix A.1.

To produce one at a time Metropolis Hastings samples, we consider grouping subsets of conditionally independent η_i's into blocks. Let A_k be the collection of indices such that, for all $i, j \in A_k$ we have $\eta_i | \boldsymbol{\eta}_{N(i)} \perp\!\!\!\perp \eta_j | \boldsymbol{\eta}_{N(j)}$ and $N(i) \cup N(j) \subset A_k^c$. We can now perform one at a time Metropolis Hastings updates for each individual η_i within each block A_k.

A.4 Details of Slovenia Data Analysis

We consider the Poisson SGLMM of form

$$\begin{aligned} y_i &\sim Poisson(\mu_i), \\ \log(\mu_i) &= \log(o_i) + \boldsymbol{x}_i'\boldsymbol{\beta} + \eta_i + \epsilon_i, \end{aligned}$$

where o_i is an offset for individual i.

We use an order $k = 1$ differencing matrix for the GRF and LMA model (see equation (2.22) of Section 2.4.2). We use the conditional distribution form of the CAR model to perform one at a time block Metropolis Hastings updates following the results of Appendix A.1 and A.3. Matrices $\mathbf{M}$ and $\mathbf{C}$ of Appendix A.1 are defined for the GRF and LMA models from sparse $\boldsymbol{Q}_1$ of equation (2.21). Denote $\eta_{\mu_i} = \sum_{C_{ij} \neq 0} C_{ij}\eta_j$. The full-conditionals for $\boldsymbol{\eta}$ are given by

$$\begin{aligned} \log\Big([\eta_i|\boldsymbol{\eta}_{N(i)}, y_i, \boldsymbol{\beta}, \epsilon_i, \boldsymbol{\theta}]\Big) &\approx \log[y_i|\mu_i] + \log[\eta_i|\boldsymbol{\eta}_{N(i)}, \boldsymbol{\theta}] + Const & \text{(A.4)} \\ &\approx y_i \log(\mu_i) - \mu_i - \frac{(\eta_i - \eta_{\mu_i})^2}{2m_{ii}} + Const. & \text{(A.5)} \end{aligned}$$

In (A.4), $\boldsymbol{\theta} = (\kappa^2, \xi)$ for the GRF model and $\boldsymbol{\theta} = (\kappa^2, \lambda, \boldsymbol{S})$ for the LMA.

A normal prior with variance $\sigma^2_\beta = 10^6$ is assumed for the fixed effects giving log full-conditionals

$$\log[\boldsymbol{\beta}|\boldsymbol{y}, \boldsymbol{\eta}, \boldsymbol{\epsilon}] \approx \sum_{i=1}^{194} y_i \log(\mu_i) - \mu_i - \frac{\boldsymbol{\beta}'\boldsymbol{\beta}}{2\sigma^2_\beta} + Const.$$

The prior for the homogeneous spatial random effect is assumed to be $iid\, \mathcal{N}(0, \sigma^2)$, giving log full-conditionals

$$\log[\epsilon_i|\eta_i, \boldsymbol{\beta}, y_i, \sigma^2] \approx y_i \log(\mu_i) - \mu_i - \frac{\epsilon_i^2}{2\sigma^2} + Const.$$

An inverse gamma prior with scale and shape one is assumed for σ^2 giving conjugate full-conditional

$$\sigma^2 \sim InvGamma(98, \frac{||\boldsymbol{\epsilon}||^2}{2} + 1).$$

The priors for the variance parameter (ξ) and κ for the GRF model are assumed to be independent scale one half-normals. The log full-conditionals are

$$\log([\xi|\boldsymbol{w}, \kappa^2]) \approx -97\log(\xi^2) - \frac{1}{2\xi^2}\boldsymbol{w}'\boldsymbol{L}\boldsymbol{L}\boldsymbol{w} - \frac{\xi^2}{2} + Const,$$

and

$$\log\left([\kappa|\xi, \boldsymbol{\eta}]\right) \approx 2\sum_{i=1}^{194} \log(U_{ii}) - \frac{1}{2\xi^2}\boldsymbol{w}'\boldsymbol{L}\boldsymbol{L}\boldsymbol{w} - \frac{\kappa^2}{2} + Const. \quad \text{(A.6)}$$

where U_{ii} in (A.6) is the i^{th} diagonal entry of the Cholesky decomposition of $\boldsymbol{L}$. The scale parameter (λ) and κ of the LMA have log full-conditionals

$$\log([\lambda|\boldsymbol{S}]) \approx -194\log(\lambda^2) - \frac{1}{2\lambda^2}\sum_{i=1}^{n} s_{ii} - \frac{\lambda^2}{2} + Const,$$

and

$$\log\left([\kappa|\boldsymbol{S},\boldsymbol{\eta}]\right) \approx 2\sum_{i=1}^{194}\log(U_{ii}) - \frac{1}{2}\boldsymbol{w}'\boldsymbol{L}\boldsymbol{S}^{-1}\boldsymbol{L}\boldsymbol{w} - \frac{\kappa^2}{2} + Const.$$

A.5 Details of the Columbus Crime Dataset Analysis

The model is of form

$$\log(y_i) = \boldsymbol{x}_i'\boldsymbol{\beta} + \eta_i + \epsilon_i, \tag{A.7}$$

where the covariates and response are detailed in Section 2.5.2. The fixed effects were assigned a normal prior with variance 10^6 giving conjugate full-conditionals

$$[\boldsymbol{\beta}|\boldsymbol{y},\boldsymbol{\eta},\sigma^2] \sim \mathcal{N}\left(\left(\left(\frac{\sigma^2}{10^6}\right)\mathbf{I} + \boldsymbol{X}'\boldsymbol{X}\right)^{-1}(\boldsymbol{y}-\boldsymbol{\eta}), \left(\left(\frac{1}{10^6}\right)\mathbf{I} + \left(\frac{1}{\sigma^2}\right)\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\right).$$

The variance of the spatially homogeneous random effect (σ^2) is given a half-normal scale one prior. The log full-conditional is

$$\log\left(\left[\sigma^2|\boldsymbol{y},\boldsymbol{\eta},\boldsymbol{\beta}\right]\right) \approx -\frac{49}{2}\log(\sigma^2) - \frac{||\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}-\boldsymbol{\eta}||^2}{2\sigma^2} - \frac{\sigma^2}{2} + Const.$$

We use an order $k=1$ differencing matrix to define the covariance structure of the GRF and LMA model (see equation (2.22) of Section 2.4.2). The conjugate full-conditionals for the GRF random effects are

$$[\boldsymbol{\eta}|\boldsymbol{\beta},\boldsymbol{y},\kappa^2,\xi] \sim \mathcal{N}\left(\left(\left(\frac{\sigma^2}{\xi^2}\right)\boldsymbol{L}\boldsymbol{L} + \mathbf{I}\right)^{-1}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}), \left(\left(\frac{1}{\xi^2}\right)\boldsymbol{L}\boldsymbol{L} + \left(\frac{1}{\sigma^2}\right)\mathbf{I}\right)^{-1}\right).$$

The scale parameter for the GRF (ξ) is given a half-normal scale 10 prior while κ^2 is given an independent half-normal scale one prior leading to log full-conditionals

$$\log\left(\left[\xi^2|\boldsymbol{\eta},\kappa^2\right]\right) \approx -\frac{49}{2}\log(\xi^2) - \frac{1}{2\xi^2}\boldsymbol{\eta}'\boldsymbol{L}\boldsymbol{L}\boldsymbol{\eta} - \frac{\xi^4}{20} + Const,$$

and

$$\log\left(\left[\kappa^2|\boldsymbol{\eta},\xi\right]\right) \approx 2\sum_{i=1}^{49}\log(U_{ii}) - \frac{1}{2\xi^2}\boldsymbol{\eta}'\boldsymbol{L}\boldsymbol{L}\boldsymbol{\eta} - \frac{\kappa^4}{2} + Const.$$

where U_{ii} is the Cholesky decomposition of $\boldsymbol{L} = \triangle^{(1)}$. The conjugate full-conditionals for the LMA random effects are

$$\left[\boldsymbol{\eta}|\boldsymbol{\beta},\boldsymbol{y},\kappa^2,\boldsymbol{S}\right] \sim \mathcal{N}\left(\left(\sigma^2\boldsymbol{L}\boldsymbol{S}^{-1}\boldsymbol{L}+\mathbf{I}\right)^{-1}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}), \left(\boldsymbol{L}\boldsymbol{S}^{-1}\boldsymbol{L}+\left(\frac{1}{\sigma^2}\right)\mathbf{I}\right)^{-1}\right).$$

The scale parameter (λ) for the LMA is given a half-normal scale 10 prior and κ^2 is given an independent scale one half-normal prior. The log full-conditionals for LMA model parameters are

$$\log([\lambda^2|\boldsymbol{S}]) \approx -49\log\left(\lambda^2\right) - \frac{1}{2\lambda^2}\sum_{i=1}^{49}S_{ii} - \frac{\lambda^4}{20} + Const,$$

and,

$$\log([\kappa^2|\boldsymbol{\eta},\boldsymbol{S}]) \approx 2\sum_{i=1}^{49}\log(U_{ii}) - \frac{1}{2}\boldsymbol{\eta}'\boldsymbol{L}\boldsymbol{S}^{-1}\boldsymbol{L}\boldsymbol{\eta} - \frac{\kappa^4}{2} + Const.$$

We observed spatial confounding among the random effects and the intercept. This is not uncommon, however to assess convergence we analyzed the trace plots of $\beta_0\mathbf{1}+\boldsymbol{\eta}$.

A.6 Details of Malaria Data Analyses

We follow the auxiliary data approach of Albert and Chib (1993). Let $\Phi(\cdot)$ denote the standard normal CDF. Consider the continuous space binary response model $y_j^{(i)} \sim Bernoulli(p_j^{(i)})$ where $p_j^{(i)}$ is the probability that the j^{th} child in the i^{th} village has malaria. We model $p_j^{(i)}$ through the probit link function by introducing auxiliary data $z_j^{(i)}$ as

follows

$$p_j^{(i)} = \Phi(z_j^{(i)}), \qquad z_j^{(i)} = \boldsymbol{x}'^{(i)}_j \boldsymbol{\beta} + \eta(\boldsymbol{u}_i).$$

The covariates are as described in Section 2.5.3. Define $A_{ij} = \phi_j(\boldsymbol{u}_i)$, where $\{\phi_l(\boldsymbol{u})\}_{l=1}^n$ are the basis functions corresponding to the triangular mesh with $n = 288$ mesh nodes formed in Section 2.5.3. Define the 2035 by 65 matrix $\mathbf{B}$ such that n_i entries of column $\boldsymbol{b}_i$ corresponding to responses $y_j^{(i)}$ are 1, and the remaining entries are 0. The auxiliary variables can be equivalently expressed in matrix form as

$$\boldsymbol{z} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{B}\boldsymbol{A}\boldsymbol{w}. \tag{A.8}$$

The priors for $\boldsymbol{w}$ are constructed following Sections (2.3.2-2.3.3) with $\alpha = 2$. We place a normal prior on the fixed effects with variance 100. In both models the spatial scale parameter κ was assigned a half-normal prior with scale one. Additionally, in the Gaussian model, the scale parameter (ξ) was assigned a half-normal scale one prior as well. We used joint MH proposals for κ and ξ in the Gaussian model. For the LMA, shape parameter (τ) and scale parameter (λ) were jointly proposed with independent scale one half-normal priors.

Full-Conditionals

Let $TN_{(a,b)(x)}$ be the density of a standard normal truncated such that its support is (a, b). For both models, we have conjugate truncated normal updates for the auxiliary variables,

$$\left[z_j^{(i)} | y_j^{(i)}, \boldsymbol{x}'^{(i)}_j, w_i, \boldsymbol{\beta}\right] \sim \begin{cases} TN_{(0,\infty)}(z_j^{(i)} - \boldsymbol{x}'^{(i)}_j \boldsymbol{\beta} - w_i), & y_j^{(i)} = 1 \\ TN_{(-\infty,0)}(z_j^{(i)} - \boldsymbol{x}'^{(i)}_j \boldsymbol{\beta} - w_i), & y_j^{(i)} = 0 \end{cases},$$

and conjugate normal updates for the fixed effects,

$$[\boldsymbol{\beta}|\boldsymbol{w}, \boldsymbol{y}, \boldsymbol{z}] \sim \mathcal{N}\left(\left[\boldsymbol{X}'\boldsymbol{X} + \left(\frac{1}{100}\right)\mathbf{I}\right]^{-1}(\boldsymbol{y} - \boldsymbol{B}\boldsymbol{A}\boldsymbol{w}), \left[\boldsymbol{X}'\boldsymbol{X} + \left(\frac{1}{100}\right)\mathbf{I}\right]^{-1}\right).$$

The weights of the basis expansion for the GRF are given by

$$[\boldsymbol{w}|\boldsymbol{\beta}, \boldsymbol{y}, \boldsymbol{z}, \xi, \kappa] \sim \mathcal{N}\left(\left[\xi^{-2}\boldsymbol{L}\boldsymbol{L} + \boldsymbol{A}'\boldsymbol{B}'\boldsymbol{B}\boldsymbol{A}\right]^{-1}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}), \left[\xi^{-2}\boldsymbol{L}\boldsymbol{L} + \boldsymbol{A}'\boldsymbol{B}'\boldsymbol{B}\boldsymbol{A}\right]^{-1}\right).$$

The log full-conditional for the spatial scale (κ) and variance parameter (ξ) are given by

$$\begin{aligned}\log([\kappa,\xi|\boldsymbol{w}]) &\approx -\left(\frac{1}{2\xi^2}\right)\boldsymbol{w}'\boldsymbol{L}\boldsymbol{L}\boldsymbol{w} - \left(\frac{n}{2}\right)\log(\xi^2) + 2\sum_{i=1}^{n}\log(U_{ii}) \\ &\quad -\frac{\kappa^2}{2} - \frac{\xi^2}{2} + Const, \end{aligned} \tag{A.9}$$

where U_{ii} denotes the i^{th} diagonal entry of the Cholesky of $\boldsymbol{L} = \kappa^2\boldsymbol{C} + \boldsymbol{G}$. The weights of the basis expansion for the LMA are given by

$$[\boldsymbol{w}|\boldsymbol{\beta},\boldsymbol{y},\boldsymbol{z},\boldsymbol{\Gamma},\kappa] \sim \mathcal{N}\left(\left[\boldsymbol{L}\boldsymbol{\Gamma}^{-1}\boldsymbol{L} + \boldsymbol{A}'\boldsymbol{B}'\boldsymbol{B}\boldsymbol{A}\right]^{-1}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}), \left[\boldsymbol{L}\boldsymbol{\Gamma}^{-1}\boldsymbol{L} + \boldsymbol{A}'\boldsymbol{B}'\boldsymbol{B}\boldsymbol{A}\right]^{-1}\right).$$

The spatial scale (κ), shape (τ), and variance parameter (λ), have log full-conditionals

$$\begin{aligned}\log\left([\kappa,\tau,\lambda|\boldsymbol{\Gamma},\boldsymbol{w}]\right) &\approx \sum_{i=1}^{n}\left((\tau C_{ii})\left(\log(\Gamma_i) - \log\left(\lambda^2\right)\right) - \log(\Gamma(\tau C_{ii})) - \frac{\Gamma_{ii}}{\lambda^2} + 2\log(U_{ii})\right) \\ &\quad -\frac{1}{2}\boldsymbol{w}\boldsymbol{L}\boldsymbol{\Gamma}^{-1}\boldsymbol{L}\boldsymbol{w} - \frac{\lambda^2}{2} - \frac{\kappa^2}{2} - \frac{\tau}{2} + Const.\end{aligned}$$

A.7 Details of LAGOS Analysis

We fit a continuous response point referenced model with 5526 unique lake locations denoted $\{\boldsymbol{u}_i\}_{i=1}^{5526}$. We form a mesh with $n = 671$ nodes. The model considered is of the form

$$y(\boldsymbol{u}_i) = \boldsymbol{x}_i'\boldsymbol{\beta} + \eta(\boldsymbol{u}_i) + \epsilon(\boldsymbol{u}_i). \tag{A.10}$$

Define the 5526 by 671 projection matrix ($\boldsymbol{A}$) with entries $A_{ij} = \phi_j(\boldsymbol{u}_i)$, where $\{\phi_l(\boldsymbol{u})\}_{l=1}^{671}$ are the basis functions corresponding to the mesh formed in Section 2.5.3. Using the resulting basis expansion of $\eta(\boldsymbol{u})$ (see (2.7) Section 2.3.2) and assuming $\epsilon(\boldsymbol{u}_i) \overset{iid}{\sim} \mathcal{N}(0,\sigma^2)$, we can re-write the discretized likelihood as follows

$$[\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{w},\sigma^2] \sim \mathcal{N}\left(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{A}\boldsymbol{w}, \sigma^2\boldsymbol{I}\right). \tag{A.11}$$

The priors for $\boldsymbol{w}$ are constructed following Sections (2.3.2-2.3.3) with $\alpha = 2$. We have assumed half-normal scale one priors for $\sigma^2, \kappa, \xi, \tau$, and λ.

Full-Conditionals

The fixed effects for both models have conjugate full-conditionals

$$[\boldsymbol{\beta}|\boldsymbol{y},\sigma^2,\boldsymbol{w}] \sim \mathcal{N}\left(\left[\left(\frac{\sigma^2}{1000}\right)\boldsymbol{I}+\boldsymbol{X}'\boldsymbol{X}\right]^{-1}[\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}],\left[\left(\frac{1}{1000}\right)\boldsymbol{I}+\left(\frac{1}{\sigma^2}\right)\boldsymbol{X}'\boldsymbol{X}\right]^{-1}\right).$$

The conjugate full-conditionals for the weights of the GRF are

$$\left[\boldsymbol{w}|\boldsymbol{y},\sigma^2,\boldsymbol{\beta},\xi\right] \sim \mathcal{N}\left(\left[\left(\frac{\sigma^2}{\xi^2}\right)\boldsymbol{L}\boldsymbol{C}^{-1}\boldsymbol{L}+\boldsymbol{A}'\boldsymbol{A}\right]^{-1}[\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}],\left[\left(\frac{1}{\xi^2}\right)\boldsymbol{L}\boldsymbol{C}^{-1}\boldsymbol{L}+\left(\frac{1}{\sigma^2}\right)\boldsymbol{A}'\boldsymbol{A}\right]^{-1}\right).$$

The log full-conditionals for κ and ξ of the GRF model are as seen in equation (A.9) of Appendix A.6. The conjugate full-conditionals for the weights of the LMA are

$$\left[\boldsymbol{w}|\boldsymbol{y},\sigma^2,\boldsymbol{\beta},\boldsymbol{\Gamma}\right] \sim \mathcal{N}\left(\left[\sigma^2\boldsymbol{L}\boldsymbol{\Gamma}^{-1}\boldsymbol{L}+\boldsymbol{A}'\boldsymbol{A}\right]^{-1}[\boldsymbol{y}-\boldsymbol{A}\boldsymbol{w}],\left[\boldsymbol{L}\boldsymbol{\Gamma}^{-1}\boldsymbol{L}+\left(\frac{1}{\sigma^2}\right)\boldsymbol{A}'\boldsymbol{A}\right]^{-1}\right).$$

The log full-conditionals for the parameters κ, τ and λ are as seen in (A.9) of Appendix A.6. For model fitting we used normal proposals for all parameters. τ and λ were jointly proposed for the LMA model, while κ and ξ were jointly proposed for the GRF.

Appendix B
Chapter 3 Appendices

B.1 Finite Element Approximations for Matérn GRFs

Stationary Matérn random fields arise as stationary solutions to the stochastic partial differential equation

$$\left(\kappa^2 - \triangle\right) \eta(\boldsymbol{s}) = \xi \mathcal{W}(\boldsymbol{s}), \quad \boldsymbol{s} \in \mathbb{R}^2, \tag{B.1}$$

where $\triangle = \sum_{i=1}^{2} \frac{\partial^2}{\partial s_i^2}$ is the Laplacian operator in 2-dimensions, $\kappa > 0$ is a spatial scale parameter, $\tau > 0$ is a variance parameter, and $\mathcal{W}(\boldsymbol{s})$ is Gaussian white noise. Whittle (1954, 1963) showed that solutions to (B.1) are GRF with Matérn covariance given by

$$C(\boldsymbol{u}, \boldsymbol{v}) = \xi^2 (\kappa ||\boldsymbol{v} - \boldsymbol{u}||) K_1(\kappa ||\boldsymbol{v} - \boldsymbol{u}||), \quad \boldsymbol{v}, \boldsymbol{u} \in \mathbb{R}^2, \tag{B.2}$$

where $|| \cdot ||$ denotes Euclidean distance, and $K_1(\cdot)$ is an order one Bessel function of the second kind. The marginal variance is given by $\sigma^2 = \xi^2 / (4\pi\kappa^2)$. The effective range is approximated by $\rho = \sqrt{8}/\kappa$ (Lindgren et al., 2011).

Though the analytic solution provides useful insights, model fitting and parameter estimation are often facilitated by considering a numerical approximation. Lindgren et al. (2011) proposed the use of a finite element approximation to the stochastic weak formulation of the SPDE $(\kappa^2 - \triangle)^{\alpha/2} \eta(\psi) = \xi \mathcal{W}(\psi)$, where $\{\psi\}$ is a set of test functions. The finite element method (FEM) solution begins by expressing the solution, $\eta(\boldsymbol{u})$, as a basis expansion

$$\eta(\boldsymbol{s}) = \sum_{i=1}^{n} \phi_i(\boldsymbol{s}) w_i, \quad \boldsymbol{s} \in \Omega, \tag{B.3}$$

where $\{\phi_i(\boldsymbol{s})\}_{i=1}^n$ is a set of basis functions on Ω. The solution is only required to hold for a finite collection of ψ_i. The Galerkin method approximate solution is obtained by setting $\{\psi_i\}_{i=1}^n = \{\phi_i\}_{i=1}^n$.

Lindgren et al. (2011) formulated an FEM approximation by considering $\{\phi_i(\boldsymbol{s})\}_{i=1}^n$ to be piecewise triangular basis functions. The basis functions are constructed by partitioning the spatial region of interest, $\Omega \subset \mathbb{R}^d$, into non-overlapping triangular regions. The corners of the triangles, referred to as vertices, are assigned n Gaussian weights, denoted w_i. Each ϕ_i is defined to be 1 at vertex i and 0 at all other vertices. Lindgren et al. (2011) derived the distribution of the weights

$$\boldsymbol{w}|\xi,\kappa \sim \mathcal{N}\left(\mathbf{0}, \xi^2 \boldsymbol{Q}_{\kappa^2}^{-1}\right), \tag{B.4}$$

where $\boldsymbol{Q}_{\kappa^2}^{-1} = \boldsymbol{L}^{-1}\boldsymbol{C}\boldsymbol{L}^{-1}$, with $\boldsymbol{L} = \kappa^2\boldsymbol{C} + \boldsymbol{G}$. The matrices used to define $\mathbf{L}$ are given by $C_{ij} = \int_\Omega \phi_i(\boldsymbol{s})d\boldsymbol{s}$, and $G_{ij} = \int_\Omega \nabla\phi_i(\boldsymbol{s})\nabla\phi_j(\boldsymbol{s})d\boldsymbol{s}$. Under lattice refinement, the FEM solution converges to the true solution (see Appendix C.5 of Lindgren et al. (2011)).

B.2 Approximate Inference for LGCPs

The stochastic integral, $\int_\Omega \lambda(\boldsymbol{s})d\boldsymbol{s}$, in the likelihood of (3.1) cannot be computed analytically. Numerical integration schemes are used to approximate Cox process likelihoods. A common approach is to grid up the spatial domain into rectangular regions and evaluate the integral as a weighted sum (Diggle et al., 2013). Approximations are improved by refining the lattice into rectangles of smaller area. This becomes computationally burdensome, as rectangular lattice refinements often produce a mesh with fine partitions in regions containing few observations.

A desirable mesh possess a finer partition in regions with many observations, and a sparse partition in regions of few observations; a feature difficult to obtain for regularly spaced lattices. Simpson et al. (2016) proposed the use of a second order approximate dual-cell mesh, known as the Voronoi mesh, by joining the triangular elements required for fitting the Gaussian process described in Appendix B.1 at their centroids. The integral can now be approximated by

$$\int_\Omega \lambda(\boldsymbol{s})d\boldsymbol{s} \approx \sum_{i=1}^n \tilde{\alpha}_i \exp(\log(o(\tilde{\boldsymbol{s}}_i)) + \boldsymbol{x}'(\tilde{\boldsymbol{s}}_i)\boldsymbol{\beta} + \sum_{j=1}^n \phi_j(\tilde{\boldsymbol{s}}_i)), \tag{B.5}$$

where $\{\tilde{\boldsymbol{s}}_i\}_i^n$ are mesh nodes corresponding to the FEM mesh, and $\tilde{\alpha}_i$ is the volume of

the i^{th} dual cell produced by the Voronoi mesh.

To construct the approximate likelihood first define the projection matrix $P_{ij} = \phi_j(\boldsymbol{s}_i)$, where $\{\boldsymbol{s}_i\}^N$ are the N observed locations. Let the spatially continuous covariates at the observed locations and be denoted by $\boldsymbol{X} = [\mathbf{1}_{Nx1}, \boldsymbol{x}(\boldsymbol{s}_1)...\boldsymbol{x}(\boldsymbol{s_N})]$, and $\tilde{\boldsymbol{X}} = [\mathbf{1}_{nx1}, \tilde{\boldsymbol{x}}(\boldsymbol{s}_1)...\tilde{\boldsymbol{x}}(\boldsymbol{s}_n)]$ denote the mesh node covariates. We then define $\log(\boldsymbol{\eta}) = \left(\boldsymbol{w}' + \boldsymbol{\beta}'\boldsymbol{X}', \boldsymbol{w}'\boldsymbol{P}' + \boldsymbol{\beta}'\tilde{\boldsymbol{X}}'\right)$, $\boldsymbol{\alpha} = (\tilde{\alpha}, \mathbf{0}'_{Nx1})$, and $\boldsymbol{y} = (\mathbf{0}'_{Nx1}, \mathbf{1}'_{nx1})$. The approximate log-likelihood is given by

$$\log(\mathcal{S}, N|\lambda) = \sum_{i=1}^{N+n} y_i \log(\eta_i) - \alpha_i \eta_i. \tag{B.6}$$

We note that (B.6) now resembles the sum of $N + n$ independent Poisson random variables with rate $y_i\alpha_i$. Model fitting can now be performed similarly to Poisson spatial generalized linear mixed models.

B.3 Circular Synthesis Disclosure Risk Details

Here we derive the CPO estimate used to evaluate disclosure risks for radial synthesis given in (3.17) of Section 3.7.2. First note that the joint likelihood can be written as follows,

$$\pi\Big(\mathcal{S}, \mathcal{S}^\dagger, N|\lambda\Big) = \pi\Big(\mathcal{S}^\dagger, N|\mathcal{S}\Big)\pi(\mathcal{S}, N|\lambda) = \left(\frac{1}{\pi r^2}\right)^N \left(\prod_{k=1}^{N} \mathbb{I}_{\{||\boldsymbol{s}_k - \boldsymbol{s}|| < r\}}(\boldsymbol{s}_k^\dagger)\right) \pi(\mathcal{S}, N|\lambda). \tag{B.7}$$

Next, observe that $\pi(\mathcal{S}_{-k}, \mathcal{S}, N|\lambda)$ is proportional to $\pi\Big(\mathcal{S}^\dagger, \mathcal{S}, N|\lambda\Big)$.

$$\begin{aligned}
\pi\Big(\mathcal{S}_{-k}, \mathcal{S}^\dagger, N|\lambda\Big) &= \int_\Omega \pi\Big(\mathcal{S}^\dagger, N|\mathcal{S}\Big)\pi(\mathcal{S}, N|\lambda) d\boldsymbol{s}_k \\
&= \left(\frac{1}{\pi r^2}\right)^N \frac{\prod_{i\neq k} \lambda(\boldsymbol{s}_i)}{N! \exp\left(\int_\Omega \lambda(\boldsymbol{s}) d\boldsymbol{s}\right)} \int_\Omega \mathbb{I}_{\{||\boldsymbol{s}_k - \boldsymbol{s}|| < r\}}(\boldsymbol{s}_k^\dagger)\lambda(\boldsymbol{s}_k) d\boldsymbol{s}_k \\
&= \left(\frac{1}{\pi r^2}\right)^N \frac{\prod_{i\neq k} \lambda(\boldsymbol{s}_i)}{N! \exp\left(\int_\Omega \lambda(\boldsymbol{s}) d\boldsymbol{s}\right)} \int_{\mathcal{B}_r(\boldsymbol{s}_k^\dagger)} \lambda(\boldsymbol{s}) d\boldsymbol{s} \\
&= \left(\frac{\lambda(\boldsymbol{s}_k)}{\int_{\mathcal{B}_r(\boldsymbol{s}_k^\dagger)} \lambda(\boldsymbol{s}) d\boldsymbol{s}}\right) \pi\Big(\mathcal{S}, \mathcal{S}^\dagger, N|\lambda\Big).
\end{aligned} \tag{B.8}$$

We use (B.7) and (B.8) to obtain the CPO estimate

$$\begin{aligned}
\pi\Big(\boldsymbol{s}_k, N|\mathcal{S}_{-k}, \mathcal{S}^\dagger\Big) &= \left[\frac{\pi\Big(\mathcal{S}_{-k}, \mathcal{S}^\dagger, N\Big)}{\pi(\mathcal{S}, \mathcal{S}^\dagger, N)}\right]^{-1} \\
&= \left[\frac{\int \pi\Big(\mathcal{S}_{-k}, \mathcal{S}^\dagger, N|\lambda\Big)\pi(\lambda)d\lambda}{\pi(\mathcal{S}, \mathcal{S}^\dagger, N)}\right]^{-1} \\
&= \left[\frac{\int \left(\frac{\int_{\mathcal{B}_r\left(\boldsymbol{s}_k^\dagger\right)} \lambda(\boldsymbol{s})d\boldsymbol{s}}{\lambda(\boldsymbol{s}_k)}\right) \pi\Big(\mathcal{S}, \mathcal{S}^\dagger, N|\lambda\Big)\pi(\lambda)d\lambda}{\pi(\mathcal{S}, \mathcal{S}^\dagger, N)}\right]^{-1} \\
&= \left[\frac{\int \left(\frac{\int_{\mathcal{B}_r\left(\boldsymbol{s}_k^\dagger\right)} \lambda(\boldsymbol{s})d\boldsymbol{s}}{\lambda(\boldsymbol{s}_k)}\right) \pi\Big(\mathcal{S}^\dagger, N|\mathcal{S}\Big)\pi(\mathcal{S}, N|\lambda)\pi(\lambda)d\lambda}{\pi(\mathcal{S}^\dagger, N|\mathcal{S})\pi(\mathcal{S}, N)}\right]^{-1} \\
&= \left[\int \left(\frac{\int_{\mathcal{B}_r\left(\boldsymbol{s}_k^\dagger\right)} \lambda(\boldsymbol{s})d\boldsymbol{s}}{\lambda(\boldsymbol{s}_k)}\right) \pi(\lambda|\mathcal{S}, N)d\lambda\right]^{-1} \\
&= \left[\mathbb{E}_{\pi(\lambda|\mathcal{S},N)}\left[\left(\frac{\int_{\mathcal{B}_r\left(\boldsymbol{s}_k^\dagger\right)} \lambda(\boldsymbol{s})d\boldsymbol{s}}{\lambda(\boldsymbol{s}_k)}\right)\right]\right]^{-1}. \qquad \text{(B.9)}
\end{aligned}$$

B.4 Quadrature Scheme For Circular Domains

Here we detail the quadrature scheme for numerical integration a over circular domain used to obtain disclosure risk estimates in Section 3.7. Let $\boldsymbol{s}_k = (x_k, y_k)$ represent a location such that $x_k, y_k > 0$. We wish to integrate

$$\int_{\mathcal{B}_r(\boldsymbol{s}_k)} \lambda(x, y)dxdy = \int_{-r}^{r} \int_{x_k - g(y)}^{x_k + g(y)} \lambda(x, y)dxdy \qquad \text{(B.10)}$$

where $g(y) = \sqrt{(r^2 - (y - y_k)^2)}$. Making the change of variables $\tilde{y} = \frac{(y-y_k)}{r}$ and $\tilde{x} = \frac{(x-x_k)}{g(r\tilde{y}+y_k)}$, the integral in (B.10) becomes

$$\int_{-1}^{1} \int_{-1}^{1} \lambda(g(r\tilde{y} + y_k)\tilde{x} + x_k, r\tilde{y} + y_k) * r * g(r\tilde{y} + y_k)d\tilde{x}d\tilde{y}. \qquad \text{(B.11)}$$

Notice that the integral in (B.11) is now computed over the square $[-1,1]^2$. We now partition $[-1,1]^2$ into M equally sized squares of area A_{xy}. Let $(\tilde{x}_m, \tilde{y}_m)$ represent the center of each of the M squares. The integral in (B.11) is now approximated by the sum

$$\sum_{m=1}^{M} \lambda(g(r\tilde{y}_m + y_k)\tilde{x}_m + x_k, r\tilde{y}_m + y_k) * r * g(r\tilde{y}_m + y_k) * A_{xy}. \tag{B.12}$$

B.5 Prior Choice for (κ, ξ)

Here we summarize the prior choice for (κ, ξ) suggested by Lindgren et al. (2015). Following the Appendix B.1 we obtain an approximation to a Gaussian random field with Matérn covariance. The basis expansion weights have variance and spatial scale hyperparameters ξ^2 and κ^2 (see Appendix B.1 equation (B.4)). The covariance is summarized by these parameters relationship to the spatial range ρ and marginal variance σ^2 given by

$$\rho = \frac{\sqrt{8}}{\kappa}, \quad \sigma^2 = \frac{\xi^2}{4\pi\kappa^2}. \tag{B.13}$$

From (B.13) it is clear that the marginal variance is influenced by both κ and ξ. We would like a prior that captures this relationship. To do so, assume $(\theta_1, \theta_2) \sim N(\mathbf{0}, \Sigma_\theta)$. Suppose we want the parameterization

$$\log(\rho) = \log(\rho_0) + \theta_1 \tag{B.14}$$

$$\log(\sigma) = \log(\sigma_0) + \theta_2, \tag{B.15}$$

where $\log(\rho_0)$ and $\log(\sigma_0)$ are the baseline range an marginal standard deviation. Using equations (B.13) and (B.14) we can write

$$\log(\kappa) = \frac{1}{2}\log(8) - \log(\rho_0) - \theta_1 = \log(\kappa_0) - \theta_1, \tag{B.16}$$

where $\log(\kappa_0) = \frac{1}{2}\log(8) - \log(\rho_0)$. It follows from equations (B.13) and (B.15) that

$$\log(\xi) = \log(\sigma_0) + \frac{1}{2}\log(4\pi) + \log(\kappa_0) + \theta_2 - \theta_1 \tag{B.17}$$

$$= \log(\xi_0) + \theta_2 - \theta_1, \tag{B.18}$$

where $\log(\phi_0) = \log(\sigma_0) + \frac{1}{2}\log(4\pi) + \log(\kappa_0)$. Equations (B.16) and (B.17) now give a joint prior on (κ, ξ) that captures their dependent influence on the marginal variance. In Section 3.8 we took $\log(\rho_0) = 0$ and $\log(\sigma_0) = 0$.

Appendix C
Chapter 4 Appendices

C.1 FCLT for Poisson Processes

We state the FCLT for Poisson processes used in Section 4.2.2 to perform a Gaussian approximation of the Markov population model in (4.1). The FCLT states that for a Poisson process with rate λt, denoted $Y(\lambda t)$, as $n \to \infty$ we have

$$\sqrt{n}\left(\frac{1}{n}Y(n\lambda t) - \lambda t\right) \Rightarrow B(\lambda t), \tag{C.1}$$

where $B(t)$ is a standard Brownian motion, and " $\Rightarrow$ " denotes convergence in distribution (Kurtz, 1978; Van Kampen, 1992). We use (C.1) to perform a Gaussian approximation of a Poisson process

$$\frac{1}{n}Y(n\lambda t) \approx \lambda t + \frac{1}{\sqrt{n}}B(\lambda t), \tag{C.2}$$

for sufficiently large n.

C.2 Solving for V(t)

In Section 4.3.3 formed a sparse precision matrix by seeking solutions to the Itødiffusion in (4.8) of the form $\boldsymbol{V}(t) = U(t)\boldsymbol{W}(t)$, with $\boldsymbol{V}(0) = \mathbf{0}$. From equation (4.8) we then have

$$d\boldsymbol{V}(t) = dU(t)\boldsymbol{W}(t)dt + U(t)d\boldsymbol{W}(t). \tag{C.3}$$

From (C.3) and (4.8), we require

$$\begin{aligned} dU(t)\boldsymbol{W}(t) &= \partial \boldsymbol{F}_{\boldsymbol{\theta}}\left(\boldsymbol{X}^{\dagger}_{\boldsymbol{\theta}}(t)\right) U(t)\boldsymbol{W}(t)dt & \text{(C.4)} \\ U(t)d\boldsymbol{W}(t) &= G_{\boldsymbol{\theta}}\left(\boldsymbol{X}^{\dagger}_{\boldsymbol{\theta}}(t)\right) d\boldsymbol{B}(t). & \text{(C.5)} \end{aligned}$$

From (C.5), we obtain the solution

$$\begin{aligned} dU(t) &= \partial \boldsymbol{F}_{\boldsymbol{\theta}}\left(\boldsymbol{X}^{\dagger}_{\boldsymbol{\theta}}(t)\right) U(t)dt,\ U(0) = \mathbb{I}_{dxd} & \text{(C.6)} \\ d\boldsymbol{W}(t) &= U^{-1}(t)G_{\boldsymbol{\theta}}\left(\boldsymbol{X}^{\dagger}_{\boldsymbol{\theta}}(t)\right) d\boldsymbol{B}(t),\ \boldsymbol{W}(0) = \boldsymbol{0}. & \text{(C.7)} \end{aligned}$$

We recognize (C.7) as a mean-zero Brownian motion

$$\boldsymbol{W}(t) = \int_0^t A(s)d\boldsymbol{B}(s),\ A(s) = U^{-1}(s)G_{\boldsymbol{\theta}}\left(\boldsymbol{X}^{\dagger}_{\boldsymbol{\theta}}(s)\right), \tag{C.8}$$

with covariance $Cov\left(\boldsymbol{W}(s)\right) = \int_0^t A(s)A'(s)ds$.

C.3 SIS LNA

Here we write out the algorithm for deriving the LNA distribution for the LNA in Section 4.5. For details on the stochastic SIS model see Section 4.2. We first demonstrate how to construct the LNA for a given value of $\boldsymbol{\theta} = (\beta, \gamma)$. We denote $\boldsymbol{\lambda}_{\boldsymbol{\theta}}\left(P(t)\right) = \left(\lambda^1_{\boldsymbol{\theta}}\left(P(t)\right), \lambda^2_{\boldsymbol{\theta}}\left(P(t)\right)\right)'$. We begin by solving the deterministic set of equations

$$\dot{P}^{\dagger}_{\boldsymbol{\theta}}(t) = \boldsymbol{R}\boldsymbol{\lambda}_{\boldsymbol{\theta}}\left(P(t)\right), P^{\dagger}_{\boldsymbol{\theta}}(0) = 0.05, \tag{C.9}$$

for $\boldsymbol{P}^{\dagger}_{\boldsymbol{\theta}} = \left(P^{\dagger}_{\boldsymbol{\theta}}(1), P^{\dagger}_{\boldsymbol{\theta}}(2), ..., P^{\dagger}_{\boldsymbol{\theta}}(20)\right)'$. We note that $\partial \boldsymbol{F}_{\boldsymbol{\theta}}\left(P^{\dagger}_{\boldsymbol{\theta}}(t)\right) = \left(\beta\left(1 - 2P^{\dagger}_{\boldsymbol{\theta}}(t)\right) - \gamma\right)$ and solve

$$dU(t) = \partial \boldsymbol{F}_{\boldsymbol{\theta}}\left(P^{\dagger}_{\boldsymbol{\theta}}(t)\right) U(t), U(0) = 1, \tag{C.10}$$

over $(0, 20]$. Next we define $G_{\boldsymbol{\theta}}\left(P^{\dagger}_{\boldsymbol{\theta}}(t)\right) = \left(\sqrt{\lambda^1_{\boldsymbol{\theta}}\left(P^{\dagger}(t)\right)}, -\sqrt{\lambda^2_{\boldsymbol{\theta}}\left(P^{\dagger}(t)\right)}\right)$, and let $A_{\boldsymbol{\theta}}(s) = U^{-1}(s)G_{\boldsymbol{\theta}}\left(P^{\dagger}_{\boldsymbol{\theta}}(t)\right)$. We solve for the covariance of $S\boldsymbol{Y}$ as defined in Section 4.3.3 by solving

$$M_t = \int_{t-1}^{t} A_{\boldsymbol{\theta}}(s)A'_{\boldsymbol{\theta}}(s)ds = \int_{t-1}^{t} U^{-2}(s)\left(\lambda^1_{\boldsymbol{\theta}}\left(P^{\dagger}(t)\right) + \lambda^2_{\boldsymbol{\theta}}\left(P^{\dagger}(t)\right)\right) ds \tag{C.11}$$

for $t = 1, 2, ..., 20$. We then form diagonal matrices $U = diag\,(U(1), U(2), ..., U(20))$, $M = diag\,(M_1, M_2, ..., M_{20})$, and let $C = SU^{-1}$. We then define $Q_{\boldsymbol{\theta}} = C'M^{-1}C$ and evaluate the LNA likelihood $\pi(\boldsymbol{P}|\boldsymbol{\theta}) \sim N\left(\boldsymbol{P}^{\dagger}_{\boldsymbol{\theta}}, \frac{1}{N}Q^{-1}_{\boldsymbol{\theta}}\right)$.

C.4 SEIR LNA

We derive the LNA distribution for the SEIR model in Section 4.6. First we define $\boldsymbol{\lambda}_{\boldsymbol{\theta}}\left(\boldsymbol{X}^{\dagger}_{\boldsymbol{\theta}}(t)\right) = \left(\lambda^1_{\boldsymbol{\theta}}\left(\boldsymbol{X}^{\dagger}_{\boldsymbol{\theta}}(t)\right), \lambda^2_{\boldsymbol{\theta}}\left(\boldsymbol{X}^{\dagger}_{\boldsymbol{\theta}}(t)\right), \lambda^3_{\boldsymbol{\theta}}\left(\boldsymbol{X}^{\dagger}_{\boldsymbol{\theta}}(t)\right)\right)'$. The LNA distribution is constructed by solving

$$\frac{d}{dt}\boldsymbol{X}^{\dagger}_{\boldsymbol{\theta}}(t) = \boldsymbol{\lambda}_{\boldsymbol{\theta}}\left(\boldsymbol{X}^{\dagger}_{\boldsymbol{\theta}}(t)\right)\boldsymbol{X}^{\dagger}_{\boldsymbol{\theta}}(t), \tag{C.12}$$

with initial conditions $\boldsymbol{X}^{\dagger}_{\boldsymbol{\theta}}(0) = \left(\frac{N-1}{N}, 0, \frac{1}{N}\right)'$. We define

$$\partial \boldsymbol{F}_{\boldsymbol{\theta}}\left(\boldsymbol{X}^{\dagger}_{\boldsymbol{\theta}}(t)\right) = \begin{bmatrix} -\left(\beta I^{\dagger}_{\boldsymbol{\theta}}(t) + \mu_S(t)\right) & 0 & -\beta S^{\dagger}_{\boldsymbol{\theta}}(t) \\ \beta I^{\dagger}_{\boldsymbol{\theta}}(t) & -\alpha & \beta S^{\dagger}_{\boldsymbol{\theta}}(t) \\ 0 & \alpha & -\gamma \end{bmatrix}$$

and solve equation (4.10) for $U(t)$ in Section 4.3.3 with $U(16) = \mathbb{I}_{3\times 3}$. We solve for the covariance of the LNA by considering $\boldsymbol{W}(16) = \boldsymbol{0}$ in (4.11) in Section 4.3. We form block diagonal matrices U and M from the 3×3 of $U(t_i)$ and M_i for T^{obs} to construct the LNA distribution for $\boldsymbol{X}_N|\boldsymbol{\theta}$ in (4.13).

We use the LNA distribution for $\pi(\boldsymbol{I}_N|\boldsymbol{\theta}) \sim N\left(\boldsymbol{I}^{\dagger}_{\boldsymbol{\theta}}(t), \frac{1}{N}\tilde{Q}^{-1}_{\boldsymbol{\theta}}\right)$ in the analysis of Section 4.6. We obtain $\tilde{Q}^{-1}_{\boldsymbol{\theta}}$ by taking the Schur compliment of $PQ_{\boldsymbol{\theta}}P'$. To do this consider

$$PQ_{\boldsymbol{\theta}}P' = \begin{bmatrix} Q_1 & Q_2 \\ Q_2' & Q_3 \end{bmatrix}.$$

The Schur compliment of $PQ_{\boldsymbol{\theta}}P'$ is given by $\tilde{Q}^{-1}_{\boldsymbol{\theta}} = Q_1 - Q_2Q_3^{-1}Q_2'$. We then perform a Laplace approximation for $\pi(\boldsymbol{\theta}|\boldsymbol{y})$ using $\pi(\boldsymbol{I}_N|\boldsymbol{\theta})$ as detailed in Section 4.4.

Bibliography

Åberg, S. and K. Podgórski (2011). A class of non-Gaussian second order random fields. *Extremes 14*(2), 187–222.

Albert, J. H. and S. Chib (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American statistical Association 88*(422), 669–679.

Allen, L. J. (2017). A primer on stochastic epidemic models: Formulation, numerical simulation, and analysis. *Infectious Disease Modelling 2*(2), 128–142.

Armstrong, M. P., G. Rushton, and D. L. Zimmerman (1999). Geographically masking health data to preserve confidentiality. *Statistics in medicine 18*(5), 497–525.

Bartlett, M. (1964). The spectral analysis of two-dimensional point processes. *Biometrika 51*(3/4), 299–311.

Baxendale, P. H. and P. E. Greenwood (2011). Sustained oscillations for density dependent markov processes. *Journal of mathematical biology 63*(3), 433–457.

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society. Series B (Methodological)*, 192–236.

Bivand, R., J. Hauke, and T. Kossowski (2013). Computing the Jacobian in Gaussian spatial autoregressive models: An illustrated comparison of available methods. *Geographical Analysis 45*(2), 150–179.

Bivand, R. and C. Rundel (2018). *rgeos: Interface to Geometry Engine - Open Source ('GEOS')*. R package version 0.3-28.

Bolin, D. (2014). Spatial Matérn fields driven by non-Gaussian noise. *Scandinavian journal of statistics 41*(3), 557–579.

Bolin, D. and J. Wallin (2016). Multivariate Type-G Matérn fields. *arXiv preprint arXiv:1606.08298*.

Cao, Y., D. T. Gillespie, and L. R. Petzold (2006). Efficient step size selection for the tau-leaping simulation method. *The Journal of chemical physics 124*(4), 044109.

Charest, A.-S. and Y. Hou (2016). On the meaning and limits of empirical differential privacy. *Journal of Privacy and Confidentiality 7*(3), 53–66.

Chen, T.-M., J. Rui, Q.-P. Wang, Z.-Y. Zhao, J.-A. Cui, and L. Yin (2020). A mathematical model for simulating the phase-based transmissibility of a novel coronavirus. *Infectious diseases of poverty 9*(1), 1–8.

DHHS (2016, Oct). Standards for privacy of individually identifiable health information. *U.S. Department of Health and Human Services*.

Diggle, P. and P. Ribeiro Jr (2007). Model-based geostatistics.,(springer: New york).

Diggle, P. J., P. Moraga, B. Rowlingson, B. M. Taylor, et al. (2013). Spatial and spatio-temporal log-Gaussian Cox processes: extending the geostatistical paradigm. *Statistical Science 28*(4), 542–563.

Dimitrakakis, C., B. Nelson, A. Mitrokotsa, and B. I. Rubinstein (2014). Robust and private Bayesian inference. In *International Conference on Algorithmic Learning Theory*, pp. 291–305. Springer.

Dwork, C. (2006). Differential privacy, in automata, languages and programming. *ser. Lecture Notes in Computer Scienc 4052*, 112.

Dwork, C., F. McSherry, K. Nissim, and A. Smith (2006). Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography conference*, pp. 265–284. Springer.

Dwork, C., A. Roth, et al. (2014). The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science 9*(3–4), 211–407.

Fahrmeir, L. and G. Tutz (2013). *Multivariate statistical modelling based on generalized linear models*. Springer Science & Business Media.

Faulkner, J. R. and V. N. Minin (2018). Locally adaptive smoothing with Markov random fields and shrinkage priors. *Bayesian analysis 13*(1), 225.

Fearnhead, P., V. Giagos, and C. Sherlock (2014). Inference for reaction networks using the linear noise approximation. *Biometrics 70*(2), 457–466.

Ferkingstad, E., H. Rue, et al. (2015). Improving the inla approach for approximate bayesian inference for latent gaussian models. *Electronic Journal of Statistics 9*(2), 2706–2731.

Foulds, J., J. Geumlek, M. Welling, and K. Chaudhuri (2016). On the theory and practice of privacy-preserving bayesian data analysis. *arXiv preprint arXiv:1603.07294*.

Fricks, J. and E. Hanks (2018). Stochastic population models. In *Handbook of statistics*, Volume 39, pp. 443–480. Elsevier.

Fuglstad, G.-A., D. Simpson, F. Lindgren, and H. Rue (2015). Interpretable priors for hyperparameters for Gaussian random fields. *arXiv preprint arXiv:1503.00256*.

Geisser, S. (1980). Discussion on sampling and Bayes' inference in scientific modeling and robustness. *Journal of the Royal Statistical Society A 143*, 416–417.

Giagos, V. (2010). *Inference for auto-regulatory genetic networks using diffusion process approximations*. Ph. D. thesis, Lancaster University.

Gillespie, D. T. (1977). Exact stochastic simulation of coupled chemical reactions. *The journal of physical chemistry 81*(25), 2340–2361.

Hanks, E. M., E. M. Schliep, M. B. Hooten, and J. A. Hoeting (2015). Restricted spatial regression in practice: geostatistical models, confounding, and robustness under model misspecification. *Environmetrics 26*(4), 243–254.

He, S., Y. Peng, and K. Sun (2020). Seir modeling of the covid-19 and its dynamics. *Nonlinear Dynamics*, 1–14.

Hodges, J. S. and B. J. Reich (2010). Adding spatially-correlated errors can mess up the fixed effect you love. *The American Statistician 64*(4), 325–334.

Hooten, M. B. and N. T. Hobbs (2015). A guide to Bayesian model selection for ecologists. *Ecological Monographs 85*(1), 3–28.

Keeling, M. J. and P. Rohani (2011). *Modeling infectious diseases in humans and animals*. Princeton University Press.

Kim, S.-J., K. Koh, S. Boyd, and D. Gorinevsky (2009). ℓ_1 trend filtering. *SIAM review 51*(2), 339–360.

Klebaner, F. C. (2012). pp. 123–145. Imperial College Press.

Komorowski, M., B. Finkenstädt, C. V. Harper, and D. A. Rand (2009). Bayesian inference of biochemical kinetic parameters using the linear noise approximation. *BMC bioinformatics 10*(1), 1–10.

Kurtz, T. G. (1978). Strong approximation theorems for density dependent markov chains. *Stochastic Processes and their Applications 6*(3), 223–240.

Kurtz, T. G. (1981). *Approximation of population processes*. SIAM.

Lai, C.-C., C.-Y. Hsu, H.-H. Jen, A. M.-F. Yen, C.-C. Chan, and H.-H. Chen (2021). The bayesian susceptible-exposed-infected-recovered model for the outbreak of covid-19 on the diamond princess cruise ship. *Stochastic Environmental Research and Risk Assessment*, 1–15.

Lindgren, F., H. Rue, et al. (2015). Bayesian spatial modelling with R-INLA. *Journal of Statistical Software 63*(19), 1–25.

Lindgren, F., H. Rue, and J. Lindström (2011). An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 73*(4), 423–498.

Mizumoto, K., K. Kagaya, A. Zarebski, and G. Chowell (2020). Estimating the asymptomatic proportion of coronavirus disease 2019 (covid-19) cases on board the diamond princess cruise ship, yokohama, japan, 2020. *Eurosurveillance 25*(10), 2000180.

Møller, J., A. R. Syversveen, and R. P. Waagepetersen (1998). Log-Gaussian Cox processes. *Scandinavian journal of statistics 25*(3), 451–482.

Mwalili, S., M. Kimanthi, V. Ojiambo, D. Gathungu, and R. W. Mbogo (2020). Seir model for covid-19 dynamics incorporating the environment and social distancing.

Øksendal, B. (2003). Stochastic differential equations. In *Stochastic differential equations*, pp. 65–84. Springer.

Opitz, T. (2016). Modeling asymptotically independent spatial extremes based on Laplace random fields. *Spatial Statistics 16*, 1–18.

Paciorek, C. J. and M. J. Schervish (2004). Nonstationary covariance functions for Gaussian process regression. In *Advances in neural information processing systems*, pp. 273–280.

Park, T. and G. Casella (2008). The Bayesian LASSO. *Journal of the American Statistical Association 103*(482), 681–686.

Pettit, L. (1990). The conditional predictive ordinate for the normal distribution. *Journal of the Royal Statistical Society: Series B (Methodological) 52*(1), 175–184.

Podgórski, K. and J. Wegener (2011). Estimation for stochastic models driven by Laplace motion. *Communications in Statistics-Theory and Methods 40*(18), 3281–3302.

Quick, H., S. H. Holan, and C. K. Wikle (2018). Generating partially synthetic geocoded public use data with decreased disclosure risk by using differential smoothing. *Journal of the Royal Statistical Society: Series A (Statistics in Society) 181*(3), 649–661.

Quick, H., S. H. Holan, C. K. Wikle, and J. P. Reiter (2015). Bayesian marked point process modeling for generating fully synthetic public use data with point-referenced geography. *Spatial Statistics 14*, 439–451.

R Core Team (2021). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Raghunathan, T. E., J. P. Reiter, and D. B. Rubin (2003). Multiple imputation for statistical disclosure limitation. *Journal of official statistics 19*(1), 1.

Riley, S., F. Christophe, A. Christl, et al. (2003). Transmission dynamics of the etiological agent of sars in hong kong: impact of public health interventions. science.

Ripley, B. D. (1976). The second-order analysis of stationary point processes. *Journal of applied probability 13*(2), 255–266.

Roberts, G. O. and J. S. Rosenthal (2009). Examples of adaptive mcmc. *Journal of Computational and Graphical Statistics 18*(2), 349–367.

Rue, H. (2001). Fast sampling of Gaussian Markov random fields. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 63*(2), 325–338.

Rue, H., S. Martino, and N. Chopin (2009). Approximate bayesian inference for latent gaussian models by using integrated nested laplace approximations. *Journal of the royal statistical society: Series b (statistical methodology) 71*(2), 319–392.

Schliep, E. M. and J. A. Hoeting (2015). Data augmentation and parameter expansion for independent or spatially correlated ordinal data. *Computational Statistics & Data Analysis 90*, 1–14.

Simpson, D., J. B. Illian, F. Lindgren, S. H. Sørbye, and H. Rue (2016). Going off grid: Computationally efficient inference for log-Gaussian Cox processes. *Biometrika 103*(1), 49–70.

Simpson, D., H. Rue, A. Riebler, T. G. Martins, S. H. Sørbye, et al. (2017). Penalising model component complexity: A principled, practical approach to constructing priors. *Statistical Science 32*(1), 1–28.

Snoke, J., G. M. Raab, B. Nowok, C. Dibben, and A. Slavkovic (2018). General and specific utility measures for synthetic data. *Journal of the Royal Statistical Society: Series A (Statistics in Society) 181*(3), 663–688.

Snoke, J. and A. Slavković (2018). pMSE mechanism: Differentially private synthetic data with maximal distributional similarity. In *International Conference on Privacy in Statistical Databases*, pp. 138–159. Springer.

Soranno, P. A., L. C. Bacon, M. Beauchene, K. E. Bednar, E. G. Bissell, C. K. Boudreau, M. G. Boyer, M. T. Bremigan, S. R. Carpenter, J. W. Carr, et al. (2017). Lagos-ne: a multi-scaled geospatial and temporal database of lake ecological context and water quality for thousands of US lakes. *GigaScience 6*(12), gix101.

Summers, J. (1989). *A History of London's Most Colourful Neighborhood*. Bloomsbury, London.

Sun, L., C. Lee, and J. A. Hoeting (2015). Parameter inference and model selection in deterministic and stochastic dynamical models via approximate bayesian computation: modeling a wildlife epidemic. *Environmetrics 26*(7), 451–462.

Tibshirani, R. J. et al. (2014). Adaptive piecewise polynomial estimation via trend filtering. *The Annals of Statistics 42*(1), 285–323.

Van Kampen, N. G. (1992). *Stochastic processes in physics and chemistry*, Volume 1. Elsevier.

VanWey, L. K., R. R. Rindfuss, M. P. Gutmann, B. Entwisle, and D. L. Balk (2005). Confidentiality and spatially explicit data: Concerns and challenges. *Proceedings of the National Academy of Sciences 102*(43), 15337–15342.

Ver Hoef, J. M., E. M. Hanks, and M. B. Hooten (2018). On the relationship between conditional (CAR) and simultaneous (SAR) autoregressive models. *Spatial Statistics 25*, 68–85.

Wallin, J. and D. Bolin (2015). Geostatistical modelling using non-Gaussian Matérn fields. *Scandinavian Journal of Statistics 42*(3), 872–890.

Wang, H. and J. P. Reiter (2012). Multiple imputation for sharing precise geographies in public use data. *The annals of applied statistics 6*(1), 229.

Wang, Y.-X., J. Sharpnack, A. J. Smola, and R. J. Tibshirani (2016). Trend filtering on graphs. *The Journal of Machine Learning Research 17*(1), 3651–3691.

Whittle, P. (1954). On stationary processes in the plane. *Biometrika*, 434–449.

Whittle, P. (1963). Prediction and regulation by linear least-square methods.

Woo, M.-J., J. P. Reiter, A. Oganian, and A. F. Karr (2009). Global measures of data utility for microdata masked for disclosure limitation. *Journal of Privacy and Confidentiality 1*(1).

Vita

Adam Walder

Email: adamwalder24@gmail.com **Phone:** (714) 235-7785
LinkedIn: https://www.linkedin.com/in/walderadam

Education

Pennsylvania State University — **State College, PA**
PhD in Statistics, GPA: 3.8/4.0 — August 2016 – August 2021 (Expected)

California State University of Fullerton — **Fullerton, California**
Bachelor of the Arts in Applied Mathematics, GPA: 3.9/4.0 — August 2011 – May 2016

Professional Experience

Pennsylvania State University — **State College, PA**
Instructor — July 2017 – August 2017
Teaching/Research Assistant — August 2016 – May 2021

Honors and Awards

- J. Keith Ord Scholarship in Statistics. — May 2021
- ASA student paper competition winner, Government Statistics Section. — August 2020
- Stiel prize for excellence in mathematics. — May 2016

Publications

- Walder A. & Hanks E.M. (2021). Laplace Approximations for Fast Bayesian Inference on Markov Population Models (in preparation).
- Walder A., Hanks E.M., & Slavković, A. (2020). Privacy for spatial point process data. (Submitted to Computational and Graphical Statistics) *arxiv preprint https://arxiv.ord/abs/2003.12816*
- Walder, A & Hanks, E. M. (2020). Bayesian analysis of spatial generalized linear mixed models with Laplace moving average random fields. *Computational Statistics & Data Analysis, 144, 106861*

Research Interests

- Bayesian hierarchical modeling, Markov Population Models, Stochastic Differential Equations, Spatial Statistics, Spatial Privacy.