A spatial logit model for Bayesian hierarchical semiparametric regression

Spencer Woody*
spencer.woody@utexas.edu

June 12, 2017

Abstract

RNA-seq has emerged as the predominant technology for gene expression analysis. Modeling approaches must take into account the fact that there data are in the form of overdispersed counts. Existing models use the negative binomial distribution. However, there are few existing analysis pipelines which can infer a continuous time course of gene expression. In this paper we present a negative binomial regression model with a logit link on an underlying latent nonparametric time course function taken from a hierarchical Gaussian process to model dynamic gene expression. The advantages of this approach are that it can pool information across replicates and handle the case of samples collected at irregular time intervals. We also present a Gibbs sampler for implementing the model. The model is applied to an RNA-seq time series dataset from an experiment tracking the cellular response of *E. coli* to starvation conditions. We conclude by listing potential extensions of the model and other future work to be done. This paper is written as a final project for the spring 2017 semester course SDS 383D course taught by Professor James Scott at UT-Austin.

1 Introduction

Recently there has been a proliferation of next-generation sequencing technologies used in bioinformatics. RNA-seq has become the new dominant tool for gene expression analysis, which may be used to infer gene regulatory networks and measure cellular response to external stimuli, to name just a few examples. [1] [2]

As opposed to microarray data, where gene expression is measured as a continuous variable, RNA-seq data are in the form of counts and are thus discrete. Furthermore, RNA-seq data are often overdispersed (i.e., the variance is larger than the mean), so the Poisson distribution, which has equal variance and mean, is inappropriate to use. Popular analysis pipelines, such as DESeq2 DESeq2 and edgeR edgeR, use the negative binomial distribution. These techniques often fit a generalized linear model (GLM) with some sort of link function, such as the logarithmic link. This may be well suited for tasks such as differential expression between several conditions (e.g. experiment and control), but they are not adept at modeling the continuous time course of gene expression which is likely nonlinear.

In response I propose a negative binomial regression model with a logit link on a latent hierarchical Gaussian process. The motivation for this model comes from GPmicroarray, where the

*The University of Texas at Austin, Department of Statistics and Data Science. Data and R script used for this analysis available at github.com/spencerwoody/NOLA

authors use a hierarchical Gaussian process to model time series microarray data for samples with multiple replicates. Their approach accounts for hierarchy at the replicate level, but may also be extended to cases like data fusion where an experimenter would also like to pool information across related groups. In our case, we account for the discrete nature of RNA-seq data with the linked negative binomial regression. This allows us to use the same Gaussian process framework in a discrete data context, and the same extensions mentioned in GPmicroarray like clustering and data fusion should be applicable to this case too.

2 Model

We model the probability of reopening store j along street i as $\Pr(y_{ij} = 1) = 1/\{1 + \exp(-\psi_{ij})\}$, using a probit link on a latent value ψ_{ij} which comes from

$$\psi_{ij} = x_{ij}^T \beta + f_i(\mathbf{s}_{ij}),$$

where $\mathbf{s}_{ij}$ is the geographic location ,

$$\begin{aligned} \boldsymbol{\psi}_i &= X_i\beta + \mathbf{f}_i, \\ \mathbf{f}_i &= \left\{ f(\mathbf{s}) \sim \mathrm{GP}(0, k_f(\mathbf{s}, \mathbf{s}')) : \mathbf{s} \in \mathcal{S}_i \right\} \end{aligned}$$

for some covariance function k_f which depends on vector hyperparameters θ. The vector $\mathbf{f}_i$ can be seen as coming from

$$(\mathbf{f}_i|\theta) \sim \mathcal{N}(\mathbf{0}, \mathbf{K}_i),$$

where $\mathbf{K}_i$ is the matrix such that its (j, j') element is $k_f(\mathbf{s}_{ij}, \mathbf{s}_{ij'})$. Then we can see that

$$(\boldsymbol{\psi}_i|\mathbf{f}_i, \theta) \sim \mathcal{N}(X_i\beta, \mathbf{K}_i).$$

The model specification is complete when we assign a noninformative flat prior for β,

$$\pi(\beta) \propto 1$$

2.1 Covariance Function

A good covariance function is the Matérn(5/2),

$$k_f(\mathbf{s}, \mathbf{s}') = \tau^2 \exp\left\{1 + \sqrt{5} \cdot \frac{d}{b} + \frac{5}{3} \cdot \frac{d^2}{b^2}\right\} \exp\left\{-\sqrt{5} \cdot \frac{d}{b}\right\}, \; d = \|\mathbf{s} - \mathbf{s}'\|,$$

with hyperparamters $\theta = (\tau^2, b)$, and d is calculated as some measure of distance between $\mathbf{s}$ and $\mathbf{s}'$. Since the data here are geographic coordinates, we use the distance calculated from the Haversine formula.

3 Inference

Build an MCMC algorithm :)

3.1 Data augmentation using Pólya-Gamma latent variables

As described in [1], we introduce the latent variable

$$\omega_{ij} \sim \mathrm{PG}(1, 0),$$

which allows us to do a trick with the likelihood contribution of y_{ij},

$$\begin{aligned} p(y_{ij}|\psi_{ij}, \omega_{ij}) &\propto \frac{[\exp(\psi_{ij})]^{y_{ij}}}{1 + \exp(\psi_{ij}} \\ &\propto \exp\left[\left(y_{ij} - \frac{1}{2}\right)\psi_{ij}\right] \cdot \mathbb{E}_{\omega_{ij}}\left[\exp\left(-\omega_{ij}\psi_{ij}^2/2\right)\right] \end{aligned}$$

3.2 Gibbs sampler

4 Experiment

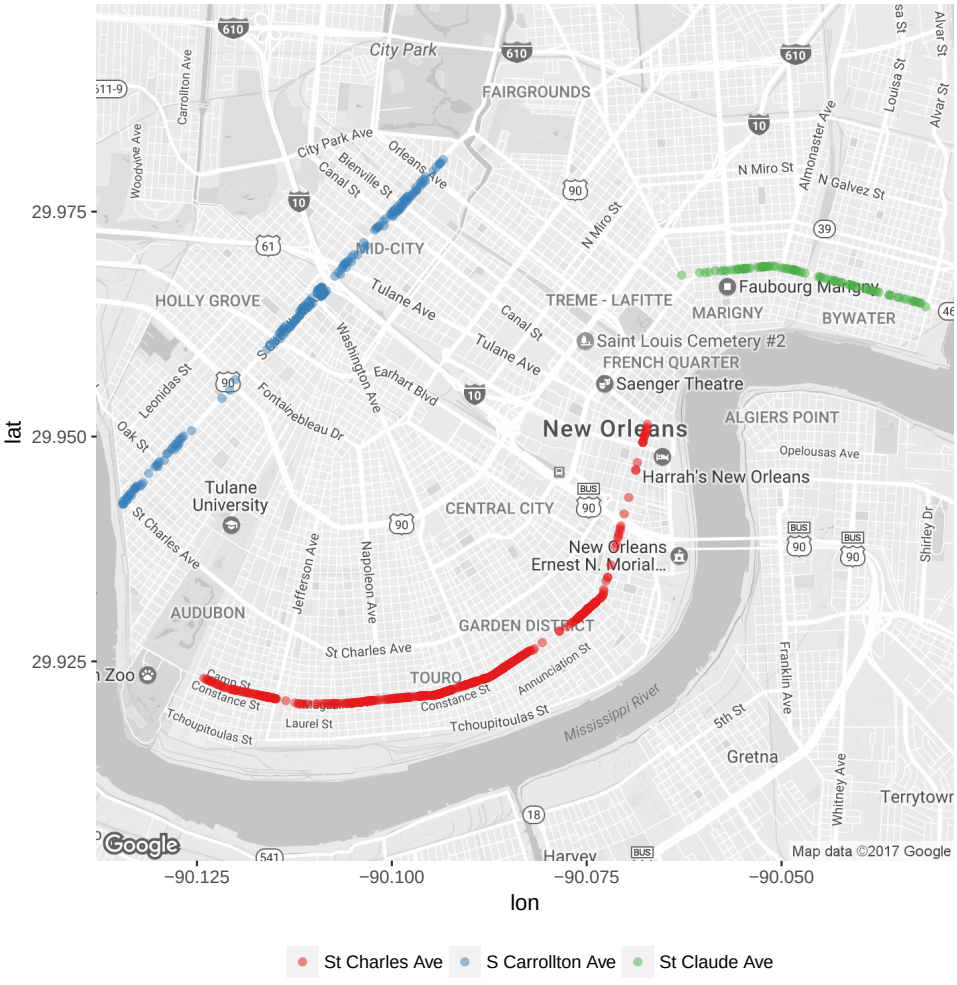

Figure 1: Locations of all sampled stores along their respective streets

5 Conclusion

References

[1] N. G. Polson, J. G. Scott, and J. Windle. Bayesian inference for logistic models using Polya-Gamma latent variables. *ArXiv e-prints*, May 2012.

[2] M. Zhou, L. Li, D. Dunson, and L. Carin. Lognormal and Gamma Mixed Negative Binomial Regression. *ArXiv e-prints*, June 2012.

A Lemma

Lemma A.1. *Define the random vectors x and γ such that the conditional distribution of x given γ and the marginal distribution of γ are, respectively,*

$$\begin{aligned}(x|\gamma) &\sim \mathcal{N}_n(A\gamma, \Sigma)\\ \gamma &\sim \mathcal{N}_p(m, V),\end{aligned}$$

where A is a $n \times p$ matrix. Then the joint distribution of $(x, \gamma)^T$ is

$$\begin{bmatrix} x \\ \gamma \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} Am \\ m \end{bmatrix}, \begin{bmatrix} AVA^T + \Sigma & AV \\ VA^T & \Sigma \end{bmatrix} \right). \tag{1}$$

Proof. Equivalently, x may be written as

$$x = A\gamma + \eta,\ \eta \sim \mathcal{N}_n(0, \Sigma)$$

and then $(x, \gamma)^T$ is multivariate normal because it can be written as an affine transformation of independent multivariate normal variables,

$$\begin{bmatrix} x \\ \gamma \end{bmatrix} = \begin{bmatrix} A \\ \mathcal{I}_p \end{bmatrix} \gamma + \begin{bmatrix} \mathcal{I}_n \\ \mathcal{O}_{p\times n} \end{bmatrix} \eta.$$

From this, the mean and covariance matrix in (1) may be derived from properties of the multivatiate normal distribution. □