

Bayesian Network Regularized Regression for Modeling Urban Crime Occurrences

Elizabeth Upton and Luis Carvalho

Department of Mathematics and Statistics, Boston University

August 18, 2017

Abstract

This paper considers the problem of statistical inference and prediction for processes defined on networks. We assume that the network is known and measures similarity, and our goal is to learn about an attribute associated with its vertices. Classical regression methods are not immediately applicable to this setting, as we would like our model to incorporate information from both network structure and pertinent covariates. Our proposed model consists of a generalized linear model with vertex indexed predictors and a basis expansion of their coefficients, allowing the coefficients to vary over the network. We employ a regularization procedure, cast as a prior distribution on the regression coefficients under a Bayesian setup, so that the predicted responses vary smoothly according to the topology of the network. We motivate the need for this model by examining occurrences of residential burglary in Boston, Massachusetts. Noting that crime rates are not spatially homogeneous, and that the rates appear to vary sharply across regions in the city, we construct a hierarchical model that addresses these issues and gives insight into spatial patterns of crime occurrences. Furthermore, we examine efficient expectation-maximization fitting algorithms and provide computationally-friendly methods for eliciting hyper-prior parameters.

1 Introduction

Given a network where vertices are connected by weighted edges, we observe vertex indexed data in the shape of vertex attributes. As common in many applications, we distinguish

between a response attribute of interest and a set of predictor attributes, and let the edge weights capture a measure of similarity between vertex attributes. Some examples include the infection status of individuals in a network of injection drug users given their drug use habits and other covariates such as age and gender; the political party affiliation of web blog authors in a network of hyperlinked connected blogs given the distribution of post topics and readership ideological inclinations; and the functional classes of proteins in a network of protein-protein interactions given gene pathway and other biological information (Leskovec and Krevl, 2014). Our main interest is then to model the response attribute using regression, but in a way that explores the topology and vertex similarity information in the network. Before discussing our proposed solution we describe a motivating example.

1.1 Modeling Residential Burglary: Main Considerations

We are interested in a specific type of urban crime, residential burglary. Burglary can be legally defined as “the act of breaking and entering a building with the intent to commit a felony” (Garner, 2001). Our network of interest, the street network of Boston, Massachusetts, contains 18,889 streets segments (edges) and 13,308 intersections (vertices) forming an undirected simple graph. We pooled 7,012 occurrences of residential burglary in the city from July of 2012 through October of 2015 (Open Data, 2016) over time, and mapped each occurrence to its closest intersection. Figure 1 pictures each vertex color-coded to indicate the value of our attribute of interest, counts of residential burglary occurrences. We model these counts with the objective of gaining an understanding as to what makes certain areas of the city more susceptible to burglary. With this information in hand, local law enforcement could, for instance, direct their crime prevention efforts to narrowly defined regions of the city and identify specific interventions to decrease the occurrence rate of residential burglary.

A naive approach to modeling crime counts on each intersection would entail identifying a set of predictor attributes describing each intersection and performing count regression, for instance, Poisson or negative binomial regression. That is, if Y_v and \mathbf{x}_v are the crime occurrence counts and covariate attributes at vertex v , we could assume $Y_v \stackrel{\text{ind}}{\sim} \text{Po}[\exp(\mathbf{x}_v^T \beta)]$. However, this specification assumes that crime effects β are constant over the network and thus spatially homogeneous. Empirical evidence suggests otherwise; for example, Figure 1 displays gross tax information for the city of Boston in 2015. Comparing this attribute to the counts of residential burglary in the same year, we see that in some areas of the city higher taxes, indicative of wealth, correspond to larger crime rates, while in other locations lower taxes, identifying poorer neighborhoods, correlate with higher crime rates. Thus,

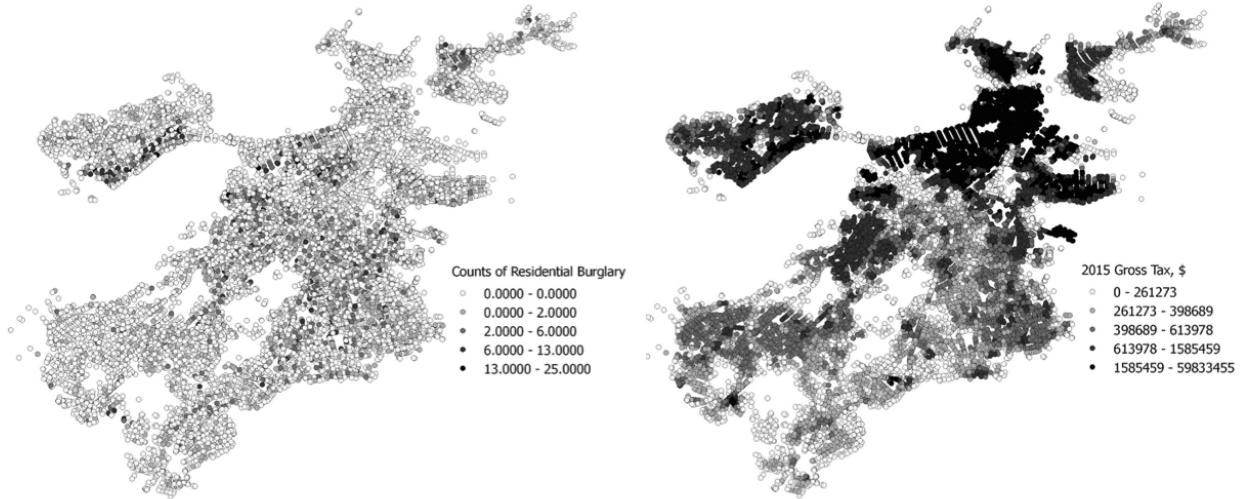


Figure 1: Residential burglary occurrence counts (left) and average wealth estimate (right) for intersections in the streets network of Boston, MA.

even when including informative covariates, a simple regression cannot explain all of the variability in burglary occurrences across a city or neighborhood. Moreover, the existence of crime “hot spots”, areas of concentrated crime counts, is widely acknowledged in crime theory literature (Eck et al., 2005). These zones can be identified in Figure 1. Given that crime rates can vary sharply across the boundaries of hot zones, it is not enough to account for gradual or smooth variations in crime occurrence rates in the network; an appropriate model must aim at capturing two types of change in process rates, gradual and abrupt.

1.2 Main Contributions

We address the issue of effect homogeneity by vertex indexing our coefficients, allowing them to vary across the network. To avoid overfitting the model we impose smoothness on the linear predictor using a penalty based on a discrete differential operator induced by the network Laplacian matrix (Ramsay and Silverman, 2005). We examine the eigendecomposition of the network Laplacian and adopt basis expansions of varying sizes composed of subsets of the eigenvectors of the Laplacian matrix for each coefficient. The rank of each basis expansion is determined via a spike-and-slab prior on the basis expansion coefficients, a Bayesian variable selection technique (George and McCulloch, 1993). In this manner, we incorporate information from both the network topology and meaningful predictors into

our regression model. To address abrupt changes in crime rates, we use a latent network-indexed indicator to identify residential burglary hot spots. The indicator attribute assigns to each intersection its hot spot status, and is designed to vary smoothly over the network in the same manner as the other network predictors. In effect, this hierarchical specification results in a zero-inflated count regression since the model now defines a “background” crime rate and a separate hot zone crime rate for each intersection. The procedure estimates the number of crimes occurring at a specific location by weighting these rates by the probability of that location being in a hot spot or not. The model is relatively easy to formalize in a Bayesian setup with Gaussian priors on the parameter sets. Due to the potentially large parameter space, we propose a computationally efficient expectation-maximization (EM) algorithm (Dempster et al., 1977) to fit this model, that is, to find maximum *a posteriori* estimates. Finally, we also provide suggestions for eliciting hyper-prior parameter values.

The outline of this paper is as follows: next, we examine a regression model based only on the topology of the network, and we compare it to current methods of network regression. We then extend these ideas to include predictor attributes and cater our model to the application of crime prediction. In Section 3, we provide guidelines for eliciting hyper-prior parameters and outline how to fit the model using two EM algorithms. In section 4, we illustrate the performance of the model on both the Boston residential burglary data and simulated crime data. Finally, we conclude with a brief summary and discussion on future extensions.

2 Proposed Model and Related Work

We state our problem of interest as follows: consider a weighted graph $G = (V, E, w)$ with vertex set V , edge set E , and positive weights w , that is, $w_{ij} > 0$ whenever $(i, j) \in E$ for all $i, j \in V$ and $w_{ij} = 0$ otherwise. We wish to regress a response attribute Y on a set of predictor attributes X using a generalized linear model: for each vertex $v \in V$, $Y_v \stackrel{\text{ind}}{\sim} F[g^{-1}(\mathbf{x}_v^\top \beta(v))]$, where F belongs to the exponential family, g is a link, and, following our discussion, the network effects β are also vertex indexed and thus non-homogeneous.

2.1 Single-intercept model

Let us assume, for now, that we have a single-intercept model, $Y_v \stackrel{\text{ind}}{\sim} F[g^{-1}(\beta(v))]$. To avoid overfitting, we impose smoothness on β through an informative prior that measures the roughness of β using a differential operator M_w and a roughness penalty $\lambda > 0$: $\beta \sim$

$N(0, \lambda^{-1}(M_w^\top M_w)^{-})$. The maximum *a posteriori* (MAP) estimate $\hat{\beta}$ for β is then

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \left\{ \mathcal{D}(Y, g^{-1}(\beta)) + \lambda \beta^\top M_w^\top M_w \beta \right\}, \quad (1)$$

where \mathcal{D} is the model deviance.

We take M_w to be the oriented weighted incidence matrix of the network: for $e = (i, j) \in E$, $M_{w,ei} = \sqrt{w_{ij}}$, $M_{w,ej} = -\sqrt{w_{ij}}$ and $M_{w,ev} = 0$ for all $v \in V$, $v \neq i$ and $v \neq j$. Thus $L_w = M_w^\top M_w = D_w - W$ is the weighted graph Laplacian with $W = [w_{ij}]$, the weighted adjacency matrix, and $D_w = \operatorname{Diag}_{i \in V} \{\sum_{j \in V} w_{ij}\}$, a diagonal matrix with the weighted degrees. Our choice of M_w is deliberate and designed to exploit basic identities in spectral graph theory to leverage the network topology and weights in the prior for regularization. This approach has been followed before in the context of regularized least squares and kernel regression in graphs (Smola and Kondor, 2003; Belkin et al., 2004), where the log prior is interpreted as a penalty for a (log) Gaussian likelihood. It is implicit that the edge weights are assumed to indicate vertex affinity and that adjacent or close vertices in the network are similar on some characteristic level relevant to the vertex attribute of interest. In fact, expanding the log prior we have

$$\beta^\top M_w^\top M_w \beta = \beta^\top L_w \beta = \sum_{(i,j) \in E} w_{ij} (\beta(i) - \beta(j))^2,$$

that is, the term penalizes the weighted sum of squares of the difference of the coefficients between adjacent vertices in the network (Kolaczyk, 2009). Thus, as usual in Bayesian inference, we seek estimates of $\beta(v)$ that balance representativeness with respect to our observed data Y and X in the likelihood with smoothness with respect to the network topology in the prior.

The network effects β are more conveniently represented using a basis expansion with respect to the eigenvectors of the operator L_w , a common approach in functional data analysis (Ramsay and Silverman, 2005). More specifically, since L_w is symmetric it realizes an eigen-decomposition $L_w = \Phi \Xi \Phi^\top$, and we can take τ eigenvectors to represent β as $\beta = \Phi_{1:\tau} \theta$, where $\Phi_{1:\tau}$ contains the first τ eigenvectors of L_w , ordered by the eigenvalues $\xi_1 < \dots < \xi_\tau$. Using this formulation, the log prior becomes

$$\lambda \beta^\top L_w \beta = \lambda \theta^\top \Phi_{1:\tau}^\top \Phi \Xi \Phi^\top \Phi_{1:\tau} \theta = \lambda \theta^\top \operatorname{Diag}_{i=1,\dots,\tau} \{\xi_i\} \theta.$$

2.2 Prior and related work

As previously stated, this formulation is akin to kernel based regression methods commonly used to operate on discrete input spaces, such as graphs (e.g. Kolaczyk, 2009, Section 8.4).

Kernels can be thought of as functions that define similarity relationships between pairs of entities. The graph Laplacian is often used in the formation of these kernels when the goal is to approximate data on a graph; see Smola and Kondor (2003) for a number of examples of kernels defined via the Laplacian. Kernel methods employ a penalized regression strategy, as in Eq. (1), where the predictor variables are derived from the kernel (Kolaczyk and Csárdi, 2014). Similarly, Belkin et al. (2004) consider the problem of labeling a partially labeled graph through regularization algorithms using a smoothing matrix, such as the Laplacian, and discuss theoretical guarantees for the generalization error of the presented regularization framework.

While these model formulations capture information in the network topology, they do not allow us to easily incorporate pertinent covariates. Kernel methods have been extended to include information from multiple kernel functions, each arising from a different data source (Lanckriet et al., 2004). In this case, the problem is often redefined as determining an optimal set of weights used to merge the various kernel matrices (Kolaczyk, 2009). However, these methods often lack interpretability and suffer from computational issues. Research has also been performed on variable selection for graph-structure covariates (Li and Li, 2008). The introduced procedure involves a smoothness penalty on the coefficients derived from the Laplacian, however, in this particular application it is the predictor variables that represent the vertices in a graph, and the presence or absence of an edge identifies correlated features. The question of interest revolves around identifying grouping effects for predictors that are linked in the network. While the machinery employed is similar to that previously discussed, the question of interest is essentially different.

With the increased popularity and availability of network data, developing a framework for regression models specific to network indexed data has become a focus of recent research. Li et al. (2016) discuss network prediction models that incorporate network cohesion, the idea that linked nodes act similarly, and node covariates. They develop the theoretical properties of their estimator and demonstrate its advantage over regressions that ignore network information. Our method differs in that the coefficients are designed to vary over the network, addressing the nonhomogeneity of covariate effects. Furthermore, our hierarchical structure allows for both smooth and abrupt changes in the process rate. Similar to Li et al., we focus on interpretability and generalization; learning about the network and the vertex attribute of interest by examining the covariate values and introducing a flexible framework adaptable to a variety of GLM settings.

2.3 Extending the intercept model

We want our model to include vertex indexed covariate information leading to better predictive power and further understanding of the vertex attribute process. That is, given p predictors, we wish to regress on

$$\eta_v = g(\mathbb{E}[Y_v]) = \beta_0(v) + x_{1v}\beta_1(v) + \cdots + x_{pv}\beta_p(v).$$

We perform the same basis expansion described in the intercept model on each coefficient, using the first τ_j eigenvectors of L , yielding

$$\eta_v = \sum_{j=0}^p x_{jv}\beta_j(v) = \sum_{j=0}^p x_{jv}\phi_{\tau_j}^\top \theta_j$$

that is, with $\beta_j(v) = \phi_{\tau_j}^\top \theta_j$ where ϕ_{τ_j} is the v -th row in $\Phi_{1:\tau_j}$ and we identify $x_{0v} = 1$ for the intercept. Notice, we can write $\eta = D_X \theta$ where

$$D_X = [\Phi_{1:\tau_0} \text{Diag}_{v \in V}\{x_1\}\Phi_{1:\tau_1} \cdots \text{Diag}_{v \in V}\{x_p\}\Phi_{1:\tau_p}]$$

and $\theta = [\theta_0 \ \theta_1 \ \dots \ \theta_p]$. We choose to smooth this entire term over the network, resulting in predictions for the vertex attribute that vary smoothly over the topology of G . This new, more general specification extends the posterior estimate in (1) to accommodate the roughness penalty $\lambda \eta^\top L_w \eta = \lambda \theta^\top D_X^\top L_w D_X \theta$:

$$\hat{\theta} = \underset{\theta}{\operatorname{argmin}} \left\{ \mathcal{D}(Y, g^{-1}(D_X \theta)) + \lambda \theta^\top D_X^\top L_w D_X \theta \right\}. \quad (2)$$

Returning to a Bayesian perspective, we have the prior on the basis expansion coefficients $\theta \sim N(0, \lambda^{-1}(D_X^\top L_w D_X)^{-})$.

2.4 Determining the basis rank, τ_j

To determine τ_j , the number of weighted Laplacian eigenvectors to include in the basis expansion of coefficient β_j , we perform Bayesian variable selection via a spike-and-slab prior. With τ_j as the latent variable, we define the prior conditionally, that is:

$$\theta_j | \tau_j \stackrel{\text{ind}}{\sim} N\left(0, \lambda^{-1} M_{\tau_j}^{1/2} (D_{X_j}^\top L_w D_{X_j})^{-} M_{\tau_j}^{1/2}\right) \quad (3)$$

where $M_{\tau_j} = \text{Diag}_{i=1 \dots K} \{I(i > \tau_j)V_0 + I(i \leq \tau_j)\}$, K is the maximum basis rank, and $V_0 < 1$ is small distinguishing the separation between the variance of the spike and slab components. Further,

$$\mathbb{P}(\tau_j) = \begin{cases} 1 - \alpha_0 & \text{for } \tau_j = 0 \\ \alpha_0(1 - \alpha_1) & \text{for } \tau_j = 1 \\ \alpha_0\alpha_1 \frac{\rho^{i-1}}{\sum_{k=2}^K \rho^{k-1}} & \text{for } \tau_j = i. \end{cases} \quad (4)$$

We use an EM algorithm coupled with a thresholding rule to perform variable selection on each predictor, continuously rotating through the model’s predictor set before converging to a final model where the rank of each β ’s basis expansion has been determined. We note here that the Laplacian eigenvectors are ordered by increasing eigenvalue magnitude as small eigenvalues have associated eigenvectors that vary little between connected vertices (Smola and Kondor, 2003). That is if $\tau_j = t$, the first t eigenvectors corresponding to the smallest t eigenvalues are used in the basis expansion of β_j . In this manner, the model captures the degree to which different crime effects vary over the topology of the graph.

2.5 Incorporating abrupt changes

We next add flexibility to the model, allowing it to detect abrupt changes in the vertex attribute over the topology of the graph. To this end, we introduce a network-indexed latent binary variable Z ,

$$Z_v | \gamma \stackrel{\text{ind}}{\sim} \text{Bern} \left[\text{logit}^{-1}(U_v^\top \gamma(v)) \right], \quad (5)$$

where a set of predictor attributes U and network-smoothed effects γ determine the odds of v belonging to a normal or changed state. In the context of modeling residential burglary, this variable allows us to discriminate between two crime rates: if $Z_v = 0$ the vertex is considered to be located in a crime hot zone where the crime rate is described by the predictors included in our model; if $Z_v = 1$ the vertex lies in an area of the city with a flat crime rate, a “background” rate ζ . The latent effects γ assume the same basis expansion as previously described for the main effects β , that is, $\gamma_j(v) = \phi_{\tau v}^\top \omega_j$ and so the linear effects

are $U_v^\top \gamma(v) = D_U(v)^\top \omega$, with D_U defined similarly to D_X . Together we have

$$\begin{aligned}
Y_v \mid \zeta, \theta, Z_v &\stackrel{\text{ind}}{\sim} \text{F} \left[g^{-1} (Z_v \zeta + (1 - Z_v) (D_X(v)^\top \theta)) \right] \\
Z_v \mid \omega &\stackrel{\text{ind}}{\sim} \text{Bern} \left[\text{logit}^{-1} (D_U(v)^\top \omega) \right] \\
\theta &\stackrel{\text{ind}}{\sim} N \left(0, \lambda_\theta^{-1} (D_X^\top L_w D_X)^- \right) \\
\omega &\stackrel{\text{ind}}{\sim} N \left(0, \lambda_\omega^{-1} (D_U^\top L_w D_U)^- \right).
\end{aligned} \tag{6}$$

3 Prior Elicitation and Inference

Fitting the aforementioned model to our Boston crime data requires three main steps: hyper-prior parameter selection, estimation of the basis expansion ranks, and estimation of linear coefficients. The three sets of hyper-prior parameters to elicit include:

Network range ψ : the weights defining the Laplacian L_w measure similarity. However, in some contexts only distances between adjacent nodes are available, as in our application, and so they need to be translated into similarities. A network range parameter is introduced to calibrate this transformation.

$V_0, \alpha_0, \alpha_1, \rho$: components of the spike-and-slab prior used in (3) and (4) to determine the number of weighted Laplacian eigenvectors to include in the basis expansion of latent and main coefficients γ and β .

Roughness penalties λ_ω and λ_θ : tuning parameters controlling the amount of regularization or smoothing of coefficients γ and β , respectively.

We discuss methods to select these hyper-priors next.

3.1 Selecting ψ

Our model is constructed on the premise that adjacent or similar vertices in the network are somehow related in terms of crime counts. Given that the edges in our network are streets, a natural similarity measure between adjacent intersections is some inverse of street distance. The pure reciprocal function relating similarity to distance will tend to infinity as distance tends to zero, causing computational problems. Borrowing ideas from spatial statistics (Banerjee et al., 2014), we employ an exponential decay function to define weights,

$w_{ij} \propto \exp\{-d(i, j)/\psi\}$, where ψ is defined as the range parameter and $\max\{w_{ij}\} = 1$. As a guideline, we suggest defining ψ such that the similarity weight distribution is not too peaked. For instance, in our case study in Section 4.2, we defined ϕ such that the median distance maps to 80% similarity. This pragmatic approach allows the model to effectively differentiate between intersections that are close together and far apart, ensuring that the range of distances in the network corresponds to an appropriate range of weights.

3.2 Selecting α_0, α_1, ρ

We would like each τ_j to be large enough so that the combination of the τ_j vectors reflects characteristics of our attribute process, however we wish to maintain the interpretability of the coefficients and keep the computational expense of our model in check (Ramsay and Silverman, 2005). To this end, we recommend choosing α_0, α_1 , and ρ so that the expectation of τ_j , see (4), is reasonable. A specific suggestion involves examining the smallest eigenvalues of the graph Laplacian (omitting the zero eigenvalue), and choosing the first inverse eigenvalue that is some small percentage in magnitude of the largest inverse eigenvalue. That is, for example, solving for t where $\mathbb{E}[\tau_j] = t$ in the following, $.05 = \frac{\lambda_2}{\lambda_t}$. Given this guideline, sensibly calibrating α_0, α_1 and ρ is not difficult. Note (4) is equivalent for all j .

3.3 Selecting $V_0, \lambda_\theta, \lambda_\omega$

Here we use a leave-one-out cross validation PRESS statistic defined on the working responses in the final step of the iterative reweighted least squares (IRLS) algorithm, the usual computational routine used to fit generalized linear models (McCullagh and Nelder, 1989). We call this the LOOP (leave-one-out-proxy) statistic,

$$\text{LOOP} = \sum_{v \in V} \frac{(Y_v - \hat{\mu}_{(v),v})^2}{V(\hat{\mu}_{(v),v})} \approx \sum_{v \in V} \frac{r_v^2}{1 - h_v},$$

where $r_v^2 = (Y_v - \hat{\mu}_v)^2 / V(\hat{\mu}_v)$ is the Pearson residual and h_v is the leverage at v . A better approximation is provided by Williams (1987), but our formulation is more computationally convenient. For now, we set $\lambda_\omega = \lambda_\theta = \lambda$, and define V_0 jointly with λ . That is, for a particular V_0 value, we find the corresponding basis expansions, τ , via the spike-and-slab prior (see section 3.4). We use τ to define D_X , and then minimize the LOOP to determine λ . Using this updated λ , we find new values for τ . The process repeats until there is no

change in τ and λ between iterations. The procedure is completed for a range of V_0 values, and the combination of V_0 and λ that minimizes the LOOP (i.e., the prediction errors with respect to Y) determines these parameters in the final model. This process can be viewed as an empirical Bayes procedure where we are optimizing the prediction accuracy rather than the likelihood.

3.4 Estimating τ_j

Following the work of Ročková and George (2014), we employ an EM algorithm, with $\nu_i = \mathbb{E}[\mathbb{I}(i > \tau_j)]$ as the latent variable, to perform modified spike-and-slab variable selection. Namely, rather than explore the possible 2^p subsets of eigenvectors of the graph Laplacian, we only need to determine the basis rank τ_j . For the E-step we compute, at the t -th iteration,

$$\begin{aligned}\nu_i^{(t)} &= \mathbb{E}[\mathbb{I}(i > \tau_j)] = \sum_{l=1}^{i-1} \mathbb{P}(\tau_j = l | Y, \theta_j^{(t)}) \quad \text{where} \\ \mathbb{P}(\tau_j = l | Y, \theta_j^{(t)}) &= \frac{\mathbb{P}(\theta_j^{(t)} | \tau_j = l) \mathbb{P}(\tau_j = l)}{\sum_{l=1}^K \mathbb{P}(\theta_j^{(t)} | \tau_j = l) \mathbb{P}(\tau_j = l)}.\end{aligned}$$

For the M-step we assume that F is Poisson and $g = \log$, the canonical link in model (6). We set $\theta_j^{(t+1)}$ by Poisson regression $Y \sim D_{X_j}$ with prior precision $\lambda^{-1} M_{\tau_j}^{1/2} (D_{X_j}^\top L_w D_{x_j})^{-1} M_{\tau_j}^{1/2}$ where $D_{X_j} = [\text{Diag}\{X_j\} \Phi_{1:K}]$ and an offset composed of the information from the remaining p predictors in the model. We iterate between the E and M steps until the difference in the deviance of successive regressions in the M step is smaller than a previously set tolerance level. We then adopt a *sequential centroid estimator* for τ_j selecting τ_j to be the minimum number of eigenvectors such that the cumulative posterior is less than some threshold. We repeatedly cycle through all $(p + 1)$ predictors, continuously updating the offset, until τ does not change between consecutive cycles. See the appendix for thorough presentations of the EM algorithm and the *sequential centroid estimator*.

3.5 Estimating β and γ

Due to the potentially large scale of data in common applications, we forgo the usual MCMC methods used to fit Bayesian GLMs (Dey et al., 2000) and employ an EM algorithm with

Z as a latent variable. For the E-step we compute, at the t -th iteration,

$$\begin{aligned}\pi_v^{(t)} &= \mathbb{E}[Z_v | Y, \theta^{(t)}, \omega^{(t)}, \zeta^{(t)}] = \mathbb{P}(Z_v = 1 | Y, \theta^{(t)}, \omega^{(t)}, \zeta^{(t)}) \\ &= \frac{\mathbb{P}(Y_v | Z_v = 1, \theta^{(t)})\mathbb{P}(Z_v = 1 | \omega^{(t)})}{\sum_{\tilde{Z}_v \in \{0,1\}} \mathbb{P}(Y_v | \tilde{Z}_v, \theta^{(t)}, \zeta^{(t)})\mathbb{P}(\tilde{Z}_v | \omega^{(t)})}.\end{aligned}$$

Following our previous discussion, we assume that F is Poisson and $g = \log$. The three M-steps are then as follows:

M-step for ζ : set $\zeta^{(t+1)}$ by quasi-Poisson regressing $\pi^{(t)}Y \sim 1$ with offset $\log \pi^{(t)}$;

M-step for θ : set $\theta^{(t+1)}$ by quasi-Poisson regressing $(1 - \pi^{(t)})Y \sim D_X$ with offset $\log(1 - \pi^{(t)})$ and $\lambda_\theta D_X^\top L_{w(\psi)} D_X$ as prior precision;

M-step for ω : set $\omega^{(t+1)}$ by quasi-binomial regressing $\pi^{(t)} \sim D_U$ with $\lambda_\omega D_U^\top L_{w(\psi)} D_U$ as prior precision.

Convergence for this process is defined as when the change in the combined deviance of the three GLM regressions in the M steps between successive EM iterations is smaller than our set tolerance. While we are unable to approximate the entire posterior distributions of the parameters, this process provides us with the posterior modes under the model in (6). Again, further details on the derivation of these steps are outlined in the appendix.

4 Data Analysis and Results

We now conduct two studies around residential burglary occurrences in Boston, MA: a simulation study and a more detailed case study. The data, provided by the city of Boston and available to the public (City of Boston, 2016; Open Data, 2016) contain information on the 7,012 instances of residential burglary occurring between July 2012 and October 2015. The attribute covariates gathered and included in our final model are:

Distance from each intersection to the nearest police station.

Sub-district designation of intersection location (business, residential, or other).

Gross tax amount for each parcel in the city of Boston in 2015. To convert gross tax into a vertex indexed covariate, we construct a buffer around each intersection and then aggregate the fractions of gross tax from each parcel in proportion to the parcel area covered by the buffer.

Our choice of analyzing the aforementioned covariates was driven by established crime theory (Bernasco and Block, 2009) and available data.

4.1 Simulation study

In order to assess the performance of our model (**Mod4**, in (6)) we compare its output against three competing methods:

Mod1: Intercept-only Poisson regression, akin to kernel regression, $Y_v | \theta \stackrel{\text{ind}}{\sim} \text{Po}[\exp(\phi_{\tau_v}^\top \theta)]$ and $\theta \sim N(0, \lambda^{-1} \text{Diag}_{i=1, \dots, \tau} \{\xi_i\}^-)$.

Mod2: Poisson regression using the covariates, but ignoring the network topology, $Y_v | \theta \stackrel{\text{ind}}{\sim} \text{Po}[\exp(\mathbf{x}_v^\top \beta)]$.

Mod3: Poisson regression defining D_X and smoothing the linear predictor, but ignoring the abrupt changes in the network, $Y_v | \theta \stackrel{\text{ind}}{\sim} \text{Po}[\exp(D_X(v)^\top \theta)]$ and $\theta \sim N(0, \lambda^{-1} (D_X^\top L_w D_X)^-)$.

We first generate a connected subgraph of Boston consisting of 818 vertices, the average neighborhood size of Boston's 16 neighborhoods, by randomly choosing a source intersection and employing a breadth-first search algorithm (Cormen et al., 2001). Given this smaller network, we elicit ϕ , the network range, and L_w following the guidelines outlined in Section 3. We set a universal τ_j , construct D_X , sample θ from the prior distribution given in (6), then set ζ , the background crime rate, equal to -2.5. Using the breadth-first search algorithm three more times, we create hot zones, each of size 40 and originating at a randomly chosen intersection, within our subgraph. For the intersections reached by the search we set $Z_v = 0$. Lastly, we generate random crime counts using a negative binomial distribution with mean $\mu = \exp(Z_v \zeta + (1 - Z_v)(D_X(v)^\top \theta))$ and variance $\mu + \mu^2$.

Now that we have a simulated network loaded with crime counts, we perform the four regressions of interest and compare model performance. For **Mod1** we systematically choose τ to be the maximum number of eigenvectors included in the basis expansion that allows $N_k = D_X^\top Y D_X$ to be invertible. For **Mod3** and **Mod4** we define τ_j and λ using the procedure described in 3.4. We compare the performance of each model using the sum of the relative errors, that is $\sum_v |Y_v - \hat{\mu}_v| / \sum_v Y_v$. As shown in Figure 2, **Mod4** significantly outperforms the three competing methods for intersections in and out of the three hot zones. Furthermore, **Mod3** shows slight improvement over **Mod1** and **Mod2**. By construction, the coefficient effects on the simulated networks are more homogenous, and the minimal improvement from allowing the coefficients to vary over these small networks is expected.

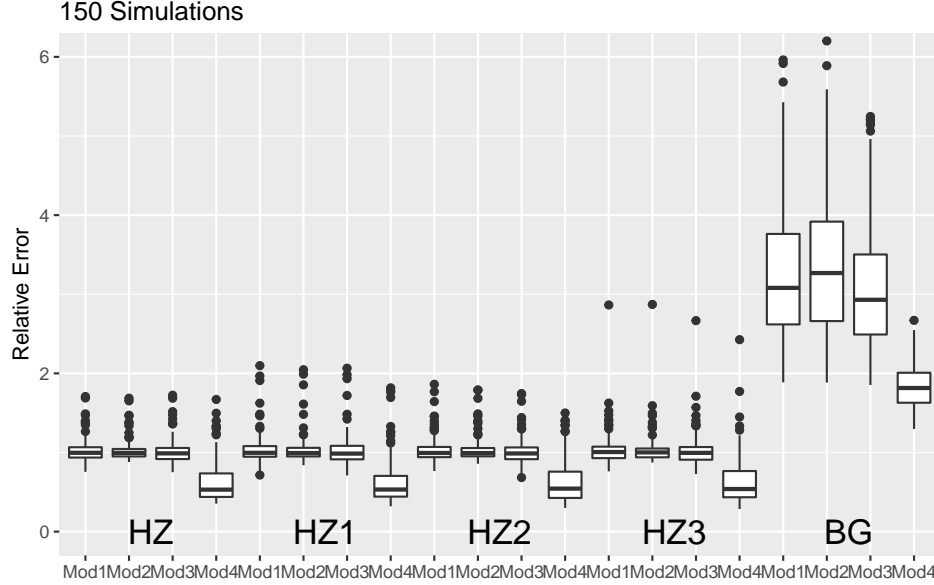


Figure 2: We compare the relative error from four models for intersections in each hot zone (**HZ**) and those intersections not in a hot zone (**BG**).

4.2 Case study: Boston, Massachusetts

We now analyze the entire metropolitan region of Boston. Following the guidelines in Section 3, we set the value of ϕ to 0.1617 and use the EM algorithm in 3.4 to find V_o (0.4) and τ . Figure 3 displays the cumulative posterior results, $\mathbb{P}(\tau_j | Y, \theta_j)$ for each predictor. As posited, the *Gross tax* effect varies over the network and is best captured via a basis expansion of large rank; conversely, the *Distance* effect is close to uniform. We construct D_X and, using the PRESS statistic, determine λ (0.36) (see Figure 4).

The predicted crime counts from model (2) define the initial values of Z_v , θ , ω , and ζ in the second EM algorithm. Specifically, for the vertices in the upper quartile of predicted crime occurrences, we set Z_v equal to 0; this subset of intersections is used via model (2) to find initial estimates of θ and ω . The remaining points define the initial value for ζ . In our final model, the EM algorithm requires 52 iterations to converge.

The results of our final model can be seen in Figure 5; our predicted crime counts vary smoothly over the network and the model captures the overall pattern of residential burglary in the city reasonably well. The deviance plot is close to a null plot given our dependent variable is discrete. In Figure 5 we see that some intersections qualify as outliers

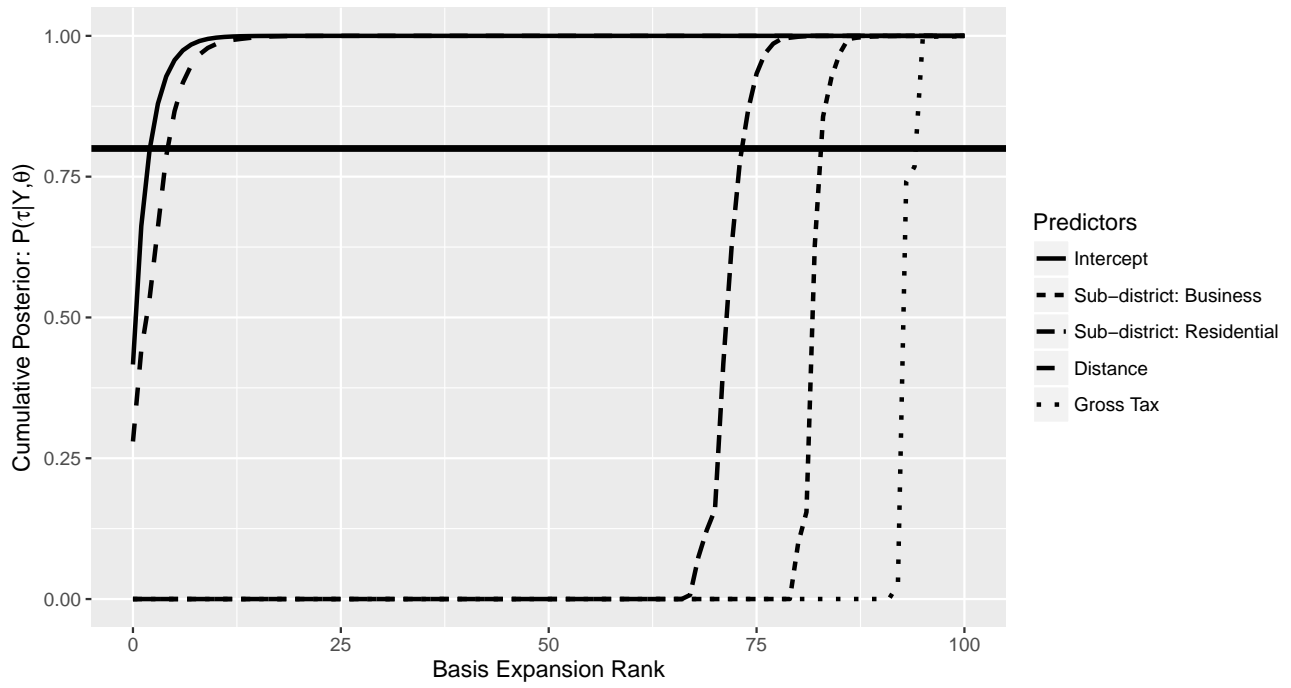


Figure 3: Choosing τ based on a threshold of 0.8.

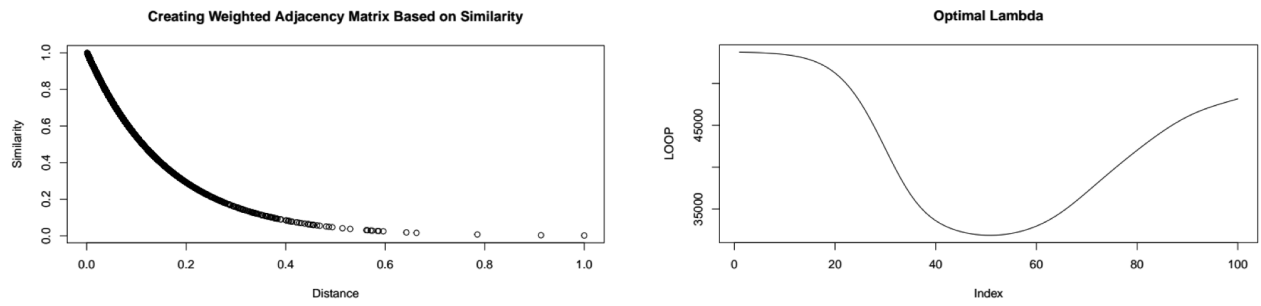


Figure 4: Inverse relationship of similarity and distance used to define the weights (left); the value of λ is found by minimizing the LOOP statistic (right).

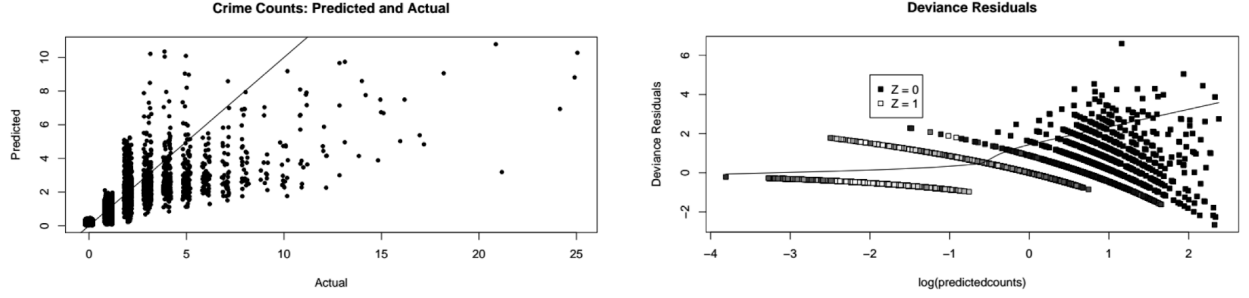


Figure 5: Left: predicted versus actual crime counts for the 13,307 intersections in Boston; Right: deviance residuals color coded by the value of π .

with crime counts over 20. While our model predicts relatively high crime for these intersections, it is not able to capture these extremes. Further inspection reveals that the majority of these points come from one neighborhood in Boston, Allston. The neighborhood is differentiated from the rest of Boston in two particular ways. Firstly, in the network sense, Allston is separated from Boston proper (see the Northwest region in Figure 1) because its southern border town is not part of Boston, effectively making the neighborhood an island. Secondly, 78.3% of Allston’s population is composed of young adults (age 18-34), compared to 39.4% for the city of Boston as a whole (Lima et al., 2015), due to its large student population. The young demographics coupled with a constant population turnover suggests a target rich environment for criminals. The outlined methodology, including the EM algorithms, is easily adapted to a negative binomial regression. However, we found that the additional precision parameter in the negative binomial distribution introduces too much flexibility. That is, using the Poisson distribution, the distribution of hot zone probabilities π is bi-modal, with peaks near 0 and 1; see Figure 6. If we use the negative binomial distribution throughout the analysis, the distribution of π has greater mass towards 0.5, creating “luke warm” zones. These probabilities decrease the predicted crime rates for intersections located in hot zones and exacerbate the aforementioned problem of underestimating extremely high crime counts.

Figure 6 summarizes the effects of wealth and income on predicted crime counts. Of note, we see that in the Southern and Northeast regions of Boston the wealth coefficients are highly correlated with counts of residential burglary. Given that these areas are not considered affluent this relationship appears curious; however, these neighborhoods have been identified as undergoing gentrification and displacement (Governing Data, 2013). Burglars may be attracted to these areas of new wealth and wish to take advantage of changing

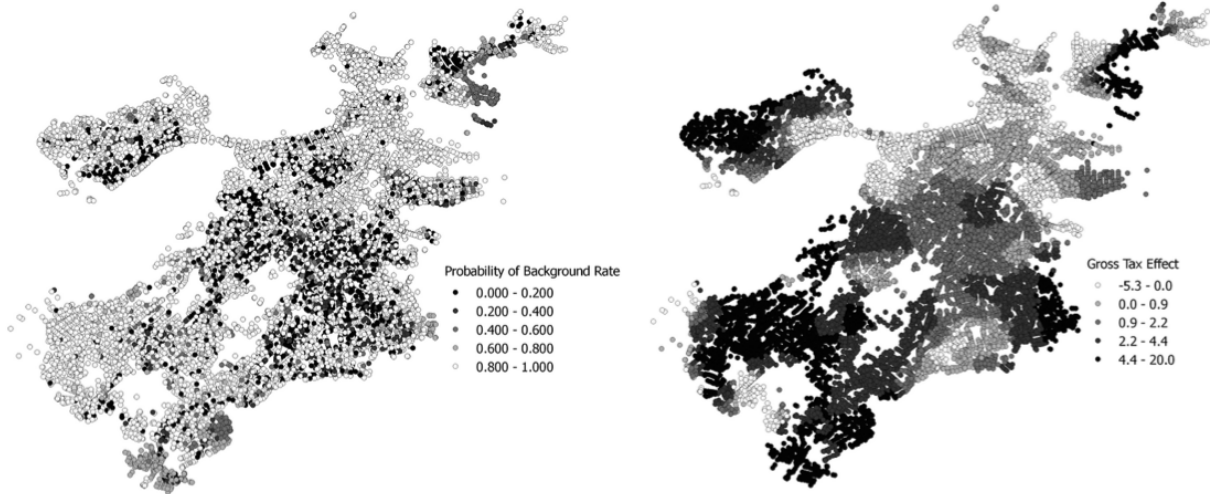


Figure 6: Left: the value of π_v , indicating probability of background status. Right: wealth effect for each intersection.

neighborhood dynamics. This information is beneficial to local law enforcement agencies seeking to identify and address patterns of residential burglary in the city of Boston. Furthermore, this knowledge may prove useful in the continued pursuit to predict accurately occurrences of crime.

5 Conclusions

We have presented a method of Bayesian network regularized regression for modeling vertex attributes that incorporates both relevant covariates and topological information in its construction. The resulting model is composed of node indexed coefficients, regularized using the Laplacian matrix, producing fitted values that vary smoothly over the network. This machinery is widely applicable and easily adaptable to a variety of GLM settings. Furthermore, as seen in the motivating example of modeling counts of residential burglary occurrences, the described method is easily modified to include specific characteristics of the network being analyzed, such as hot spots.

As shown in the simulation study, our model outperforms current methods of network regression. The model is intuitive in its construction and provides interpretable results. However, there are further gains to be had. For example, modeling vertex attributes

over time using a Bayesian dynamic model is an area that will be further explored. The motivating example, modeling residential burglary, will be improved upon given that the burglary counts were pooled over a three year period. Also, extending the methodology to topology inference, that is regression to identify the structure of a partially unknown network, is a proposed and intriguing topic of future research.

References

- Banerjee, S., B. P. Carlin, and A. E. Gelfand (2014). *Hierarchical Modeling and Analysis for Spatial Data*. CRC Press.
- Belkin, M., I. Matveeva, and P. Niyogi (2004). Regularization and semi-supervised learning on large graphs. In *International Conference on Computational Learning Theory*, pp. 624–638. Springer.
- Bernasco, W. and R. Block (2009). Where offenders choose to attack: A discrete choice model of robberies in Chicago. *Criminology* 47(1), 93–130.
- City of Boston (2016). Data Boston. <https://data.cityofboston.gov/>. Retrieved February 16, 2016.
- Cormen, T. H., C. E. Leiserson, R. L. Rivest, and C. Stein (2001). *Introduction to Algorithms*, Volume 6. MIT press Cambridge.
- Dempster, A. P., N. M. Laird, and D. B. Rubin (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society Series B* 39(1), 1–38.
- Dey, D. K., S. K. Ghosh, and B. K. Mallick (2000). *Generalized Linear Models: a Bayesian Perspective*. CRC Press.
- Eck, J., S. Chainey, J. Cameron, and R. Wilson (2005). Mapping crime: Understanding hotspots. Technical report, National Institute of Justice.
- Garner, B. A. (2001). *A Dictionary of Modern Legal Usage*. Oxford University Press, USA.
- George, E. I. and R. E. McCulloch (1993). Variable selection via gibbs sampling. *Journal of the American Statistical Association* 88(423), 881–889.

- Governing Data (2013). Boston Gentrification Maps and Data. <http://www.governing.com/gov-data/boston-gentrification-maps-demographic-data.html>. Retrieved March 28, 2017.
- Kolaczyk, E. D. (2009). *Statistical Analysis of Network Data*. Springer.
- Kolaczyk, E. D. and G. Csárdi (2014). *Statistical Analysis of Network Data with R*. Springer.
- Lanckriet, G. R., T. De Bie, N. Cristianini, M. I. Jordan, and W. S. Noble (2004). A statistical framework for genomic data fusion. *Bioinformatics* 20(16), 2626–2635.
- Leskovec, J. and A. Krevl (2014). SNAP Datasets: Stanford Large Network Dataset Collection. <http://snap.stanford.edu/data>. Retrieved March 23, 2017.
- Li, C. and H. Li (2008). Network-constrained regularization and variable selection for analysis of genomic data. *Bioinformatics* 24(9), 1175–1182.
- Li, T., E. Levina, and J. Zhu (2016). Prediction models for network-linked data. *arXiv preprint arXiv:1602.01192*.
- Lima, A., J. Lee, and C. Kim (2015). Boston by the numbers 2015. Technical report, Boston Redevelopment Authority Research Division.
- McCullagh, P. and J. A. Nelder (1989). *Generalized Linear Models*. Chapman & Hall.
- Open Data (2016). Boston Maps. <http://bostonopendata-boston.opendata.arcgis.com/>. Retrieved February 16, 2016.
- Ramsay, J. and B. W. Silverman (2005). *Functional Data Analysis*. Springer.
- Ročková, V. and E. I. George (2014). Emvs: The em approach to bayesian variable selection. *Journal of the American Statistical Association* 109(506), 828–846.
- Smola, A. J. and R. Kondor (2003). Kernels and regularization on graphs. In *Learning Theory and Kernel Machines*, pp. 144–158. Springer.
- Williams, D. (1987). Generalized linear model diagnostics using the deviance and single case deletions. *Journal of the Royal Statistical Society Series C* 36(2), 181–191.

A

A.1 EM Specifics- Spike and Slab Variable Selection

For a particular θ_j and τ_j we have: $\mathbb{P}(Y | \theta, \tau) \stackrel{\text{ind}}{\sim} \text{Po}(\exp(D_{X_j}(v)^\top \theta))$ and $\theta | \tau \sim N(0, \Sigma)$ where $\Sigma = \lambda^{-1} M_\tau^{1/2} (D_{X_j}^\top L_w D_{X_j})^- M_\tau^{1/2}$ and $M_\tau = \text{Diag}_{i=1 \dots K} \{I(i > \tau) V_0 + I(i \leq \tau)\}$. τ is as defined in (4). We wish to optimize the expected log joint (Dempster et al., 1977):

$$Q(\theta; \theta^{(t)}) = \mathbb{E}_{\tau | Y; \theta^{(t)}} [\log \mathbb{P}(\theta, \tau | y)] \quad (7)$$

Thus, for the E-step we need,

$$\begin{aligned} \nu_i^{(t)} &= \mathbb{E}[I(i > \tau)] = \sum_{l=1}^{i-1} \mathbb{P}(\tau = l | Y, \theta^{(t)}) \\ \text{where } \mathbb{P}(\tau = l | Y, \theta^{(t)}) &= \frac{\mathbb{P}(\theta^{(t)} | \tau = l) \mathbb{P}(\tau = l)}{\sum_{l=1}^K \mathbb{P}(\theta^{(t)} | \tau = l) \mathbb{P}(\tau = l)}. \end{aligned}$$

Of note, $\mathbb{P}(\theta | \tau = l + 1)$ is quickly found given $\mathbb{P}(\theta | \tau = l)$ making ν_i relatively easy to calculate. Next we update θ by maximizing the expected log likelihood given in (A.1).

$$Q = c - v \exp(D_{X_j}^\top \theta^{(t)}) + \sum_v Y_v \log(D_{X_j}^\top \theta^{(t)}) - \frac{\theta^{(t)\top} \mathbb{E}[\Sigma^{-1}] \theta^{(t)}}{2}$$

We see that updating θ is equivalent to fitting a Poisson regression with prior precision on the θ s of $\mathbb{E}[\Sigma] = \lambda^{-1} N_\tau^{1/2} (D_{X_j}^\top L_w D_{X_j})^- N_\tau^{1/2}$ and $N_\tau = \text{Diag}_{i=1 \dots K} \{\nu_i V_0 + (1 - \nu_i)\}$.

A.2 EM Specifics- Hot Zone Identification

Let δ be a vector of the current parameters: ζ , ω , and θ . We have: $\mathbb{P}(Y | \delta) = \sum_z \mathbb{P}(Y, Z | \delta)$, with $Y_v | Z_v, \delta \stackrel{\text{ind}}{\sim} \text{Po}(\exp(Z_v \zeta + (1 - Z_v) D_X(v)^\top \theta))$ and $Z_v | \delta \stackrel{\text{ind}}{\sim} \text{Bern}(\text{logit}^{-1}(D_U(v)^\top \omega))$. Again, we wish to maximize the expected log joint:

$$\begin{aligned} Q(\delta; \delta^{(t)}) &= \mathbb{E}_{Z | Y; \delta^{(t)}} [\log \mathbb{P}(Y, Z | \delta)] \\ &= \underbrace{\mathbb{E}_{Z | Y; \delta^{(t)}} [\log \mathbb{P}(Y | Z, \delta)]}_{Q_1} + \underbrace{\mathbb{E}_{Z | Y; \delta^{(t)}} [\log \mathbb{P}(Z | \delta)]}_{Q_2}. \end{aligned} \quad (8)$$

For the E-step we need $\pi_v^{(t)} \doteq \mathbb{E}_{Z|Y;\delta^{(t)}}[Z_v]$, that is,

$$\pi_v^{(t)} = \frac{\mathbb{P}(Y_v | Z_v = 1, \theta^{(t)})\mathbb{P}(Z_v = 1 | \omega^{(t)})}{\sum_{\tilde{Z}_v \in \{0,1\}} \mathbb{P}(Y_v | \tilde{Z}_v, \theta^{(t)}, \zeta^{(t)})\mathbb{P}(\tilde{Z}_v | \omega^{(t)})}.$$

It follows that

$$\begin{aligned} -\log \pi_v^{(t)} = \log \Big(& 1 + \exp \left(D_X(v)^\top \theta^{(t)} Y_v - \zeta^{(t)} Y_v \right. \\ & \left. - \exp \left(D_X(v)^\top \theta^{(t)} \right) + \exp(\zeta^{(t)}) - D_U(v)^\top \omega^{(t)} \right) \Big). \end{aligned}$$

Next, we update ζ , θ , and ω by maximizing the expected log likelihood given in (8). From the first part:

$$\begin{aligned} Q_1 = & \mathbb{E}_{Z|Y;\delta^{(t)}} \left[\sum_v -\exp \left(-Z_v \zeta^{(t)} + (1 - Z_v) D_X(v)^\top \theta^{(t)} \right) \right. \\ & \left. + Y_v \left(Z_v \zeta^{(t)} + (1 - Z_v) D_X(v)^\top \theta^{(t)} \right) \right] \\ = & \mathbb{E}_{Z|Y;\delta^{(t)}} \left[\sum_v Z_v (Y_v \zeta^{(t)} - \exp(\zeta^{(t)})) \right. \\ & \left. + (1 - Z_v) (Y_v D_X(v)^\top \theta^{(t)} - \exp(D_X(v)^\top \theta^{(t)})) \right] \\ = & \sum_v \pi_v^{(t)} (Y_v \zeta^{(t)} - \exp(\zeta^{(t)})) \\ & + (1 - \pi_v^{(t)}) (Y_v D_X(v)^\top \theta^{(t)} - \exp(D_X(v)^\top \theta^{(t)})) \\ = & \sum_v \pi_v^{(t)} Y_v (\zeta^{(t)} + \log \pi_v^{(t)}) - \pi_v^{(t)} Y_v \log \pi_v^{(t)} \\ & - \exp(\zeta^{(t)} + \log \pi_v^{(t)}) \\ & + (1 - \pi_v^{(t)}) Y_v (D_X(v)^\top \theta^{(t)} + \log(1 - \pi_v^{(t)})) \\ & - (1 - \pi_v^{(t)}) Y_v \log(1 - \pi_v^{(t)}) \\ & - \exp(D_X(v)^\top \theta^{(t)} + \log(1 - \pi_v^{(t)})). \end{aligned}$$

Analyzing the terms that contain ζ , we see that updating ζ is equivalent to fitting a quasi-Poisson regression with non-integer response, $\pi^{(t)} Y$. We have $\pi^{(t)} Y \sim \text{Quasi-Po} \left[\exp(\zeta + \right.$

$\log \pi^{(t)})]$. Similarly, we update θ where $(1 - \pi^{(t)})Y \sim \text{Quasi-Po}[\exp(D_X\theta + \log(1 - \pi^{(t)}))]$ and we use prior precision $\lambda_\theta D_X^\top L_{w(\psi)} D_X$. Now, from the second part:

$$\begin{aligned} Q_2 &= \mathbb{E}_{Z|Y, \delta^{(t)}} \left[Z_v \log(\text{logit}^{-1}(D_U(v)^\top \omega^{(t)})) \right. \\ &\quad \left. + (1 - Z_v) \left(1 - \log(\text{logit}^{-1}(D_U(v)^\top \omega^{(t)})) \right) \right] \\ &= \sum_v \pi_v^{(t)} \log(\text{logit}^{-1}(D_U(v)^\top \omega^{(t)})) \\ &\quad + (1 - \pi_v^{(t)}) \left(1 - \log(\text{logit}^{-1}(D_U(v)^\top \omega^{(t)})) \right). \end{aligned}$$

Using similar reasoning as in the previous step, we update ω using a quasi-Bernoulli regression. That is, $\pi^{(t)} \sim \text{Quasi-Bern}[\text{logit}^{-1}(D_U\omega)]$. We use prior precision $\lambda_\omega D_U^\top L_{w(\psi)} D_U$.

A.3 Selecting τ_j

To select τ_j (we drop the subscript j for the remainder of this discussion), the rank of the basis expansion for each covariate in the model, we adopt a *sequential centroid estimator*. Let us first define an auxiliary variable $\omega(\tau)$ that represents τ as an indicator vector: $\omega(\tau)_i = I(i \leq \tau)$, for $i = 1, \dots, K$. For instance, $\omega(0) = (0, 0, \dots, 0)$, $\omega(1) = (1, 0, \dots, 0)$, and so on, with $\omega(K) = \mathbf{1}_K$. Note the one-to-one correspondence between τ and ω , and thus, while $\tau \in \mathcal{T} \doteq \{0, \dots, K\}$, ω only takes values in $\Omega \doteq \cup_{j \in \mathcal{T}} \omega(j)$.

Now, given the marginal posteriors $\mathbb{P}(\tau | Y)$ or the EM-conditional posteriors $\mathbb{P}(\tau | Y, \theta^{(t)})$, which we denote in general by π_τ , we define a Bayes estimator $\hat{\tau}$ according to a generalized Hamming gain G on the ω -map:

$$\hat{\tau} \doteq \arg \max_{\tilde{\tau} \in \mathcal{T}} \sum_{\tau \in \mathcal{T}} G(\omega(\tilde{\tau}), \omega(\tau)) \pi_\tau.$$

When comparing two indicator ranks, the gain function G assigns zero gain to each discrepancy between them, a unit gain to matched zeroes (true negatives) and a gain of $\kappa > 0$ to matched ones (true positives). For example, if $K = 7$, then $G(\omega(3), \omega(5)) = 2 + 3\kappa$ since there are three matched ones from positions 1 through 3, two mismatches from positions 4 and 5, and two matched zeros from the last two positions, 6 and 7. Thus,

$$G(\omega(\tau_1), \omega(\tau_2)) = K - \max\{\tau_1, \tau_2\} + \kappa \min\{\tau_1, \tau_2\}.$$

Then,

$$\begin{aligned}
\hat{\tau} &= \arg \max_{\tilde{\tau} \in \mathcal{T}} \sum_{\tau \in \mathcal{T}} \left(\kappa \min\{\tau, \tilde{\tau}\} - \max\{\tau, \tilde{\tau}\} \right) \pi_{\tau} \\
&= \arg \max_{\tilde{\tau} \in \mathcal{T}} \left\{ \sum_{\tau \leq \tilde{\tau}} (\kappa \tau - \tilde{\tau}) \pi_{\tau} + \sum_{\tau > \tilde{\tau}} (\kappa \tilde{\tau} - \tau) \pi_{\tau} \right\} \\
&= \arg \max_{\tilde{\tau} \in \mathcal{T}} \left\{ \sum_{\tau \leq \tilde{\tau}} ((\kappa + 1) \tau - \tilde{\tau}) \pi_{\tau} + \sum_{\tau > \tilde{\tau}} \kappa \tilde{\tau} \pi_{\tau} \right\} \\
&= \arg \max_{\tilde{\tau} \in \mathcal{T}} \left\{ (\kappa + 1) \sum_{\tau \leq \tilde{\tau}} \tau \pi_{\tau} - \tilde{\tau} \mathbb{P}(\tau \leq \tilde{\tau} | Y) + \kappa \tilde{\tau} (1 - \mathbb{P}(\tau \leq \tilde{\tau} | Y)) \right\} \\
&= \arg \max_{\tilde{\tau} \in \mathcal{T}} \left\{ (\kappa + 1) \mathbb{E}[\tau | \tau \leq \tilde{\tau}, Y] + \tilde{\tau} [\kappa - (\kappa + 1) \mathbb{P}(\tau \leq \tilde{\tau} | Y)] \right\},
\end{aligned}$$

that is, $\hat{\tau} = \arg \max_{\tilde{\tau} \in \mathcal{T}} g(\tilde{\tau})$, where g is last expression within brackets above. Clearly, $g(0) = 0$; in general,

$$g(j) = (\kappa + 1) \sum_{i=0}^j i \pi_i + j [\kappa - (\kappa + 1) \sum_{i=0}^j \pi_i] = \kappa j + (\kappa + 1) \underbrace{\sum_{i=0}^j (j - i) \pi_i}_{\doteq s_j}.$$

But since $s_j = j \sum_{i=0}^j \pi_i - \sum_{i=0}^j i \pi_i$, it follows that

$$s_{j+1} = (j + 1) \left(\sum_{i=0}^j \pi_i + \pi_{j+1} \right) - \sum_{i=0}^j i \pi_i - (j + 1) \pi_{j+1} = s_j + \sum_{i=0}^j \pi_i.$$

Thus,

$$g(j + 1) = \kappa(j + 1) - (\kappa + 1) \left(s_j + \sum_{i=0}^j \pi_i \right) = g(j) + \kappa - (\kappa + 1) \sum_{i=0}^j \pi_i,$$

and so $g(j + 1) > g(j)$ if and only if $\kappa - (\kappa + 1) \sum_{i=0}^j \pi_i$, that is, $\kappa > \sum_{i=0}^j \pi_i / (1 - \sum_{i=0}^j \pi_i)$, when κ exceeds the cumulative odds. Thus, since $\sum_{i=0}^j \pi_i$ is non-decreasing, we conclude that

$$\hat{\tau} = \max \left\{ \tilde{\tau} \in \mathcal{T} : \sum_{\tau=0}^{\tilde{\tau}} \pi_{\tau} < \frac{\kappa}{1 + \kappa} \right\},$$

so we propose to expand the basis expansion up to when the cumulative posterior exceeds the $\kappa/(1 + \kappa)$ threshold.