Bayesian Spatial Clustered Regression for Count Value Data

Peng Zhao
Department of Statistics
Texas A&M University
pzhao@tamu.edu

Hou-Cheng Yang
FDA
hy15e@my.fsu.edu

Dipak Dey
Department of Statistics
University of Connecticut
dipak.dey@uconn.edu

Guanyu Hu
Department of Statistics
University of Missouri - Columbia
guanyu.hu@missouri.edu

Abstract

Investigating relationships between response variables and covariates in environmental science, geoscience, and public health is an important endeavor. Based on a Bayesian mixture of finite mixtures model, we present a novel spatially clustered coefficients regression model for count value data. The proposed method detects the spatial homogeneity of the Poisson regression coefficients. A Markov random field constrained mixture of finite mixtures prior provides a regularized estimator of the number of clusters of regression coefficients with geographical neighborhood information. An efficient Markov chain Monte Carlo algorithm is developed using multivariate log gamma distribution as a base distribution. Simulation studies are carried out to examine the empirical performance of the proposed method. Finally, we analyze Georgia's premature death data as an illustration of the effectiveness of our approach.

1 Introduction

Spatial regression models have been universally used in many different fields such as environmental science (Hu & Bradley, 2018; Yang et al., 2019; Yang & Bradley, 2021), biological science (Zhang & Lawson, 2011), and econometrics (Brunsdon et al., 1996; Yang et al., 2022) to explore the relationship between a response variable and a set of predictors over a region. One of the most important tasks for a spatial regression model is to capture the spatial dependence structure for the response variable. Spatial random effects are accounted for by the intercepts, and the regression coefficients are assumed to be constant over space under both linear models (Cressie, 1992) and generalized linear models (Diggle et al., 1998). Brunsdon et al. (1996) proposed a geographically weighted regression (GWR) to capture spatially varying patterns in regression coefficients. The idea of GWR has been subsequently extended to various works by Hu & Huffer (2019); Ma et al. (2020a); Xue et al. (2019). Furthermore, Gelfand et al. (2003) incorporated spatial Gaussian process to linear regressions to build a spatially varying coefficients regression model. The aforementioned works all assume that each location has its own set of regression parameters, which sometimes leads to overfitting. The detection of clustered covariate effects has significant benefits in various fields, including environmental science, spatial econometrics, and disease mapping. For instance, different parts of a country may have different economic conditions and development patterns. From a modeling perspective, grouping more advanced and less developed regions into separate clusters produces a more parsimonious model.

2022 NeurIPS Workshop on Gaussian Processes, Spatiotemporal Modeling, and Decision-making Systems.

Spatial cluster detection methods, such as the scan statistic-based method (Kulldorff & Nagarwalla, 1995; Jung et al., 2007), provide a remedy for spatial heterogeneity detection. Another important approach for spatial heterogeneity detection is to use the Bayesian framework to pursue spatial clusters (Carlin et al., 2014; Li et al., 2010). These two important approaches mainly focus on estimating cluster configurations of spatial responses. Recently, methods for cluster detection of spatial regression coefficients have been proposed to detect the homogeneity of the covariate effects among sub-areas (Lee et al., 2017, 2019) under spatial scan statistics. From a graph theory perspective, Li & Sang (2019) incorporated spatial neighborhood information based on minimum spanning trees in a penalized approach to detect spatially clustered coefficients. The existing literature focuses on Gaussian data under the linear model framework. For many social and environmental applications, Poisson regression for count response plays an important role (Bradley et al., 2018).

Several major challenges exist in developing clustering algorithms for regression coefficients under the Poisson model. First, specific spatial contiguity constraints must be imposed on the clustering configuration to facilitate interpretations in the spatially clustered coefficients regression. Furthermore, spatially contiguous constraints in many regional science applications should not dominate the global clustering configuration. In other words, the clustering results should contain the spatially contiguous and spatially disconnected patterns. The aforementioned methods (Lee et al., 2017, 2019; Li & Sang, 2019) guarantee spatial contiguity, but fail to obtain globally discontiguous clusters that allow two clusters with long geographical distances to belong to the same cluster. In addition, Anderson et al. (2017) discusses Poisson regression with a spatially clustered intercept and a spatially clustered slope but does not impose a spatial contiguity constraint.

Second, an important consideration in the clustering algorithm is to estimate the number of clusters. Bayesian inference provides a probabilistic framework for simultaneous inference of the number of clusters and the clustering configurations. Nonparametric Bayesian approaches, such as the Dirichlet process mixture (DPM) model (Ferguson, 1973), offer choices to estimate the number of clusters and the clustering configurations simultaneously. Ma et al. (2020b) proposed a Bayesian clustered regression for spatially dependent data based on Dirichlet process mixture model. However, their methods do not contain a consistent estimator of the number of the clusters due to inconsistency of the Dirichlet process mixture model (Miller & Harrison, 2013). In order to solve this over clustering problem of DPM, rich literature Miller & Harrison (2018); Xie & Xu (2019); Lu et al. (2018) propose several different ideas to obtain consistent estimators of the number of clusters. While existing works try to mitigate the over-clustering problem, no spatial information, such as neighborhood relationships, is utilized, while these have great potential for improving the clustering performance.

To address these challenges, in this paper, we develop a Markov random field (MRF) constrained MFM (MRF-MFM) model to capture the spatial homogeneity in regression coefficients for the Poisson model. Specifically, we develop a new Bayesian method for spatially clustered coefficients Poisson regression which leverages geographical information based on Markov random field constrained MFM model. The proposed methods leverage geographical information in Bayesian model-based clustering algorithm for Poisson regression. MRF-MFM can capture both locally spatially contiguous clusters and globally discontiguous clusters simultaneously. We develop a Gibbs sampler that enables efficient full Bayesian inference on the number of clusters, mixture probabilities, and other modeling parameters for Poisson regression with the help of multivariate log gamma (MLG) process (Bradley et al., 2018). We demonstrate the excellent numerical performance of proposed mixture models through simulations and an analysis of the premature death data in the state of Georgia.

2 Methodology

Consider a Poisson regression model with spatially varying coefficients as follows

$$y(\boldsymbol{s}_i) \sim \text{Poisson}(\exp(\boldsymbol{X}(\boldsymbol{s}_i)\boldsymbol{\beta}_{z_i})), \quad i = 1, \cdots, n, \tag{1}$$

where $z_i \in \{1, \cdots, k\}$ are labels of clusters, $\boldsymbol{\beta}_{z_i} = \boldsymbol{\beta}(s_i)$ is a p dimensional regression coefficients at location $\boldsymbol{s}_i$. From Gelfand et al. (2003), a Gaussian process prior can be assigned on regression coefficients to obtain spatially varying pattern. Compared with spatially varying pattern, heterogeneity pattern of covariate effects over subareas is also universally discussed in many different fields, such as real estate applications, spatial econometrics, and environmental science.

We consider the following Bayesian hierarchical model

$$\begin{aligned}
&K \sim p_K,\ \text{ where } p_K \text{ is a p.m.f. on}\{1,2,\ldots\}\\
&(\pi_1,\ldots,\pi_k) \sim \text{Dirichlet}_k(\gamma,\ldots,\gamma),\ \text{ given } K=k; \quad z_1,\ldots,z_n \overset{\text{iid}}{\sim} \pi,\ \text{ given } \pi\\
&(\boldsymbol{\beta}_1,\ldots,\boldsymbol{\beta}_k) \sim H,\ \text{ given } K=k\\
&y_j \sim f_{\beta_{z_j}} \text{ independently for } j=1,\ldots,n,\ \text{ given } \boldsymbol{\beta}_{1:K}, z_{1:n},
\end{aligned} \tag{2}$$

where H is the joint distribution for $\boldsymbol{\beta}_1,\ldots,\boldsymbol{\beta}_k$. The main differences between our approach and the mixture of finite mixtures (MFM) is that MFM assumes $\boldsymbol{\beta}_1,\ldots,\boldsymbol{\beta}_k$ are i.i.d sampled from a based distribution, which fails to incorporate any dependency structure. Inspired by Orbanz & Buhmann (2008), we apply the pairwise MRF in the level of coefficients to bring in interactions. With the assistance of Markov random field modeling, our MRF-MFM can incorporate more broad types of base measures with introducing dependence. Consider an undirected random graph $G=(V,E,W)$, where $V=\{v_1,\ldots,v_n\}$ is the vertex set while E is the set of graph edges, with weights W on the corresponding edges. Each vertex v_i is associated with a random variable $\boldsymbol{\beta}_i$ for $i=1,2,\ldots,k$. The pairwise MRF model is defined as

$$\begin{aligned}
\Pi\left(\boldsymbol{\beta}_1,\ldots,\boldsymbol{\beta}_k\right) &= \exp\left\{\sum_{i\in E} H_i\left(\boldsymbol{\beta}_i\right) + \sum_{(i,j)\in E, j\neq i} H_{ij}(\boldsymbol{\beta}_i\boldsymbol{\beta}_j) - A\left(W\right)\right\}\\
&= \frac{1}{Z_H}\exp\left(H\left(\boldsymbol{\beta}_1,\ldots,\boldsymbol{\beta}_k\right)\right),
\end{aligned}$$

where Z_H is the normalizing constant. The Theorem A.1 in the supplement provides the generalized urn-model induced by MRF-MFM, thus a collapsed Gibbs sampler can be applied. Consider the pairwise interactions, we model the conditional cost functions as

$$H_{i|-i}(\boldsymbol{\beta}_i \mid \boldsymbol{\beta}_{-i}) = \lambda \sum_{\{j\in\partial(i)\}} I(\boldsymbol{\beta}_i = \boldsymbol{\beta}_j), \tag{3}$$

where λ is the smoothness parameter, $\partial(i)$ denotes the set of the neighbors of observation i. The spatial smoothness can be controlled by the magnitude of λ. When $\lambda = 0$, the MRF-MFM reduces to MFM (Miller & Harrison, 2018). The conditional cost function in (3) is used in the data analysis of the paper.

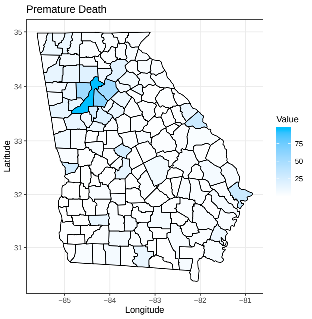
In the MRF-MFM, a natural choice for the base distribution of $\boldsymbol{\beta}_1,\cdots,\boldsymbol{\beta}_k$ is the multivariate normal distribution. However, since the multivariate normal distribution is not a conjugate prior for Poisson regression, if it is to be used as the base distribution, it must be updated with Metropolis-Hastings or auxiliary parameters such as (Neal, 2000) in Gibbs sampling algorithms. Bradley et al. (2018) constructed a multivariate log-gamma distribution (MLG) which is conjugate with a Poisson distribution. We propose an MRF-MFM for spatial clustered coefficients in Poisson regression based on the MLG prior. The multivariate log-gamma random variable with four parameters $\boldsymbol{\mu}, \boldsymbol{V}, \boldsymbol{\alpha}, \boldsymbol{\kappa}$ has the following probability density function:

$$f(\boldsymbol{q} \mid \boldsymbol{\mu}, \boldsymbol{V}, \boldsymbol{\alpha}, \boldsymbol{\kappa}) = \frac{1}{\det(\boldsymbol{V})}\left(\prod_{i=1}^{n} \frac{\kappa_i^{\alpha_i}}{\Gamma(\alpha_i)}\right)\exp[\boldsymbol{\alpha}'\boldsymbol{V}^{-1}(\boldsymbol{q}-\boldsymbol{\mu}) - \boldsymbol{\kappa}'\exp\{\boldsymbol{V}^{-1}(\boldsymbol{q}-\boldsymbol{\mu})\}],$$

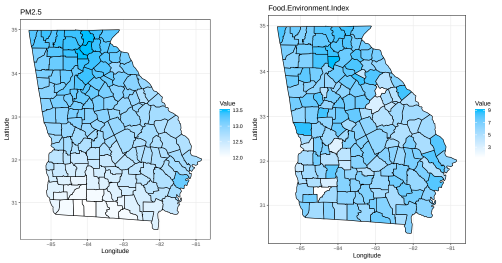
where "det" represents the determinant function. As a shorthand, we use the notation, $\text{MLG}\left(\boldsymbol{\mu}, \boldsymbol{V}, \boldsymbol{\alpha}, \boldsymbol{\kappa}\right)$ for a MLG random variable. We adapt the MRF-MFM in conjunction with MLG to a spatial Poisson regression setting, focusing on the clustering of spatially-varying coefficients $\boldsymbol{\beta}(\boldsymbol{s}_1),\cdots,\boldsymbol{\beta}(\boldsymbol{s}_n)$, where $\boldsymbol{\beta}(\boldsymbol{s}_i)$ is the p-dimensional coefficient vector for location $\boldsymbol{s}_i$.

3 Illustration: Premature Deaths in Georgia

In this study, the proposed methods are used to analyze the factors that influence the number of premature deaths in Georgia. The objective of this study is to investigate the relationship between premature deaths and environmental factors such as PM 2.5 and food environment index. The dataset is available at `www.countyhealthrankings.org` with 159 observations corresponding to the 159 counties in state of Georgia in 2015. For each county, the dependent variable is the number of the



(a) Count of the Premature Death (in hundreds)



(b) Visualizations of PM 2.5 and Food Environment Index

premature death in each county. The premature death is the death that occurs before the average age of death in a certain population. In the United States, the average age of death is about 75 years. The dependent variable is the number of lives lost per 100,000 population before age 75 in each county. The two covariates we consider in this paper are PM 2.5 (X_1) and food environment index (X_2). PM 2.5 is the average daily density of fine particulate matter in micrograms per cubic meter. The food environment index is the index of factors that contribute to a healthy food environment, 0 (worst) to 10 (best). Figures 1a and 1b present a visualization of the response and two covariates on the Georgia map.

We apply the proposed methodology to present a detailed analysis of premature death data in the state of Georgia. First, we rescale the data to a decent range as the variance in the Poisson distribution is equal to the mean. The count of the premature death is scaled to hundreds. We run 25,000 MCMC iterations and burn-in the first 15,000 iterations. The smoothing parameter is tuned over the grid $\{0.1, 0.2, \ldots, 1\}$. All other parameters are set to be consistent with the simulation study. The final clustering result corresponds to the largest Logarithm of the Pseudo-Marginal Likelihood (LPML) (Ibrahim et al., 2013), hence we choose the smoothing parameter equal 0.3. The 159 counties turned out to be put into four clusters as illustrated in Figure 2. The number of the counties in each cluster are 150, 3, 5 and 1, respectively. We also compare our model with the best LPML to vanilla MFM, Latent Gaussian Process (LGP) (Hadfield et al., 2010), conditional autoregressive (CAR) (Lee, 2013) models and Bayesian spatially varying coefficient models (SVC) (Gelfand et al., 2003; Wheeler & Calder, 2007; Finley et al., 2013). The LPML values for candidate models are: -2221.45 (MRF-MFM), -3614.38 (MFM), -2461.31 (LGP), -5015.93 (CAR), -3123.47 (SVC). Based on the LPML results, our proposed model outperforms other models. In contrast, there are 15 different clusters identified by vanilla MFM. From the estimation results shown in Table 1, we see that all the counties with higher PM 2.5 will have higher premature deaths. For Cobb County, PM 2.5 has the largest effect on premature death.

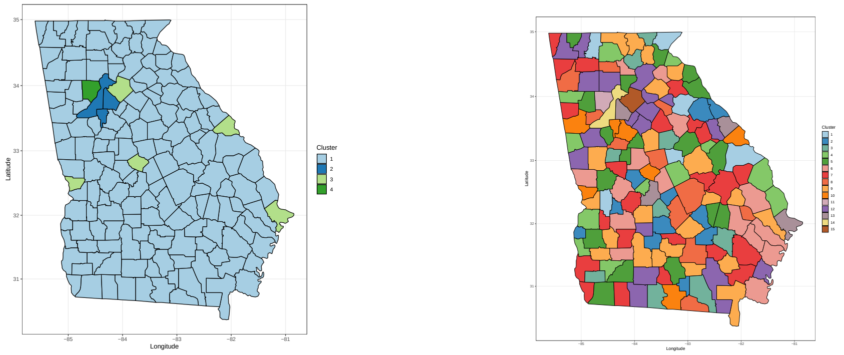


Figure 2: Left: Illustration of 4 clusters identified by the proposed method for counties. Right: Illustration of 15 clusters identified by vanilla MFM for counties.

Table 1: Dahl's method estimates for the four clusters of Georgia Data

Cluster	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$
1	-1.134	0.077	0.209
2	-3.644	0.060	1.222
3	-1.325	0.476	-0.249
4	-0.188	1.446	-2.093

4 Supplementary Material

The supplementary material contains a detailed derivation of the proposed sampling algorithm, a theoretical analysis of the proposed model under a specific Markov random field structure, and additional simulations and real data analyses.

5 Acknowledgement

Dr. Hu's work was supported by National Science Foundation grant BCS-2152822 and National Science Foundation grant DMS-2210371. Dr. Zhao's work was partially supported by National Science Foundation grant CCF-1934904. We would like to thank the meta reviewer and two anonymous reviewers for their careful comments and helpful suggestions that significantly improved the quality of the manuscript.

References

Anderson, C., Lee, D., and Dean, N. Spatial clustering of average risks and risk trends in Bayesian disease mapping. *Biometrical Journal*, 59(1):41–56, 2017.

Bradley, J. R., Holan, S. H., Wikle, C. K., et al. Computationally efficient multivariate spatio-temporal models for high-dimensional count-valued data (with discussion). *Bayesian Analysis*, 13 (1):253–310, 2018.

Brunsdon, C., Fotheringham, A. S., and Charlton, M. E. Geographically weighted regression: A method for exploring spatial nonstationarity. *Geographical Analysis*, 28(4):281–298, 1996.

Carlin, B. P., Gelfand, A. E., and Banerjee, S. *Hierarchical Modeling and Analysis for Spatial Data*. Chapman and Hall/CRC, 2014.

Cressie, N. Statistics for spatial data. *Terra Nova*, 4(5):613–617, 1992.

Diggle, P. J., Tawn, J. A., and Moyeed, R. Model-based geostatistics. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 47(3):299–350, 1998.

Ferguson, T. S. A Bayesian analysis of some nonparametric problems. *The Annals of Statistics*, 1(2): 209–230, 1973.

Finley, A. O., Banerjee, S., and Gelfand, A. E. spbayes for large univariate and multivariate point-referenced spatio-temporal data models. *arXiv preprint arXiv:1310.8192*, 2013.

Gelfand, A. E., Kim, H.-J., Sirmans, C., and Banerjee, S. Spatial modeling with spatially varying coefficient processes. *Journal of the American Statistical Association*, 98(462):387–396, 2003.

Hadfield, J. D. et al. Mcmc methods for multi-response generalized linear mixed models: the MCMCglmm R package. *Journal of Statistical Software*, 33(2):1–22, 2010.

Hu, G. and Bradley, J. A Bayesian spatial-temporal model with latent multivariate log-gamma random effects with application to earthquake magnitudes. *Stat*, 7(1):e179, 2018. e179 sta4.179.

Hu, G. and Huffer, F. Modified Kaplan–Meier estimator and Nelson–Aalen estimator with geographical weighting for survival data. *Geographical Analysis*, 52(1):28–48, 2019.

Ibrahim, J. G., Chen, M.-H., and Sinha, D. *Bayesian Survival Analysis*. Springer Science & Business Media, 2013.

Jung, I., Kulldorff, M., and Klassen, A. C. A spatial scan statistic for ordinal data. *Statistics in Medicine*, 26(7):1594–1607, 2007.

Kulldorff, M. and Nagarwalla, N. Spatial disease clusters: Detection and inference. *Statistics in Medicine*, 14(8):799–810, 1995.

Lee, D. CARBayes: an R package for Bayesian spatial modeling with conditional autoregressive priors. *Journal of Statistical Software*, 55(13):1–24, 2013.

Lee, J., Gangnon, R. E., and Zhu, J. Cluster detection of spatial regression coefficients. *Statistics in Medicine*, 36(7):1118–1133, 2017.

Lee, J., Sun, Y., and Chang, H. H. Spatial cluster detection of regression coefficients in a mixed-effects model. *Environmetrics*, pp. e2578, 2019.

Li, F. and Sang, H. Spatial homogeneity pursuit of regression coefficients for large datasets. *Journal of the American Statistical Association*, pp. 1–21, 2019.

Li, P., Banerjee, S., Hanson, T. E., and McBean, A. M. Nonparametric hierarchical modeling for detecting boundaries in areally referenced spatial datasets. Technical report, Technical Report rr2010-014, Divison of Biostatistics, School of Public , 2010.

Lu, J., Li, M., and Dunson, D. Reducing over-clustering via the powered Chinese restaurant process. *arXiv preprint*, pp. arXiv:1802.05392, 2018.

Ma, Z., Xue, Y., and Hu, G. Geographically weighted regression analysis for spatial economics data: A Bayesian recourse. *arXiv preprint arXiv:2007.02222*, 2020a.

Ma, Z., Xue, Y., and Hu, G. Heterogeneous regression models for clusters of spatial dependent data. *Spatial Economic Analysis*, pp. 1–17, 2020b.

Miller, J. W. and Harrison, M. T. A simple example of Dirichlet process mixture inconsistency for the number of components. In *Advances in Neural Information Processing Systems*, pp. 199–206, 2013.

Miller, J. W. and Harrison, M. T. Mixture models with a prior on the number of components. *Journal of the American Statistical Association*, 113(521):340–356, 2018.

Neal, R. M. Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, 9(2):249–265, 2000.

Orbanz, P. and Buhmann, J. M. Nonparametric Bayesian image segmentation. *International Journal of Computer Vision*, 77(1-3):25–45, 2008.

Wheeler, D. C. and Calder, C. A. An assessment of coefficient accuracy in linear regression models with spatially varying coefficients. *Journal of Geographical Systems*, 9(2):145–166, 2007.

Xie, F. and Xu, Y. Bayesian repulsive Gaussian mixture model. *Journal of the American Statistical Association*, pp. 1–29, 2019.

Xue, Y., Schifano, E. D., and Hu, G. Geographically weighted Cox regression for prostate cancer survival data in Louisiana. *Geographical Analysis*, 2019. Forthcoming.

Yang, H.-C. and Bradley, J. R. Bayesian inference for big spatial data using non-stationary spectral simulation. *Spatial Statistics*, 43:100507, 2021.

Yang, H.-C., Hu, G., and Chen, M.-H. Bayesian variable selection for pareto regression models with latent multivariate log gamma process with applications to earthquake magnitudes. *Geosciences*, 9 (4):169, 2019.

Yang, H.-C., Xue, Y., Geng, L., and Hu, G. Spatial weibull regression with multivariate log gamma process and its applications to china earthquake economic loss. *Statistics and Its Interface*, 15(1): 29–38, 2022.

Zhang, J. and Lawson, A. B. Bayesian parametric accelerated failure time spatial model and its application to prostate cancer. *Journal of Applied Statistics*, 38(3):591–603, 2011.

A Methodology

A.1 Clustered Poisson Regression and Mixture of Finite Mixtures

In the popular Chinese restaurant process, z_i, $i = 2, \dots, n$ are defined through the following conditional distribution Ferguson (1973):

$$P(z_i = c \mid z_1, \dots, z_{i-1}) \propto \begin{cases} |c|, & \text{at an existing table labeled } c \\ \alpha, & \text{if c is a new table} \end{cases}, \tag{4}$$

where $|c|$ is the size of cluster c.

While CRP has a very attractive feature of simultaneous estimation on the number of clusters and the cluster configuration, a striking consequence of this has been recently discovered Miller & Harrison (2013) where it is shown that the CRP produces extraneous clusters in the posterior, leading to inconsistent estimation of the *number of clusters* even when the sample size grows to infinity. A modification of the CRP called Mixture of finite mixtures (MFM) model is proposed to circumvent this issue Miller & Harrison (2018):

$$k \sim p(\cdot), \quad (\pi_1, \dots, \pi_k) \mid k \sim \text{Dir}(\gamma, \dots, \gamma), \quad z_i \mid k, \pi \sim \sum_{h=1}^{k} \pi_h \delta_h, \quad i = 1, \dots, n, \tag{5}$$

where $p(\cdot)$ is a proper probability mass function on $\{1, 2, \dots, \}$ and δ_h is a point-mass at h. Compared to the CRP, the introduction of new tables is slowed down by the factor $V_n(t+1)/V_n(t)$, which allows a model-based pruning of the tiny extraneous clusters.

The coefficient $V_n(t)$ is precomputed as:

$$V_n(t) = \sum_{k=1}^{+\infty} \frac{k_{(t)}}{(\gamma k)^{(n)}} p(k), \tag{6}$$

where $k_{(t)} = k(k-1)\dots(k-t+1)$, and $(\gamma k)^{(n)} = \gamma k(\gamma k+1)\dots(\gamma k+n-1)$. $z_i, i = 2, \dots, n$ under (5) can be defined in a Pólya urn scheme similar to CRP:

$$P(z_i = c \mid z_1, \dots, z_{i-1}) \propto \begin{cases} |c| + \gamma, & \text{at an existing table labeled } c. \\ \gamma V_n(t+1)/V_n(t), & \text{if } c \text{ is a new table.} \end{cases}, \tag{7}$$

where t is the number of existing clusters.

A.2 Introducing Dependency on the Base Measure

Recall that the full model for MFM is

$$\begin{aligned} & K \sim p_K, \ \text{where } p_K \text{ is a p.m.f. on} \{1, 2, \dots\} \\ & (\pi_1, \dots, \pi_k) \sim \text{Dirichlet}_k(\gamma, \dots, \gamma), \ \text{given } K = k; \quad z_1, \dots, z_n \overset{\text{iid}}{\sim} \pi, \ \text{given } \pi \\ & \boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_k \overset{\text{iid}}{\sim} H, \ \text{given } K = k \\ & y_j \sim f_{\beta_{z_j}} \ \text{independently for } j = 1, \dots, n, \ \text{given } \boldsymbol{\beta}_{1:K}, z_{1:n}, \end{aligned} \tag{8}$$

where H is the base distribution for β. The main insight of MFM is introducing a prior on the length of the Dirichlet distribution, and thus renders some regularization on the number of clusters created. However, the fourth step in the model, where i.i.d. samples are obtained from a base measure, fails to incorporate any dependency structure.

Inspired by Orbanz & Buhmann (2008), we apply the pairwise MRF in the level of coefficients to bring in interactions. With the assistance of Markov random field modeling, our MRF-MFM can incorporate more broad types of base measures. Consider an undirected random graph $G = (V, E, W)$, where $V = \{v_1, \dots, v_n\}$ is the vertex set while E is the set of graph edges, with weights W on the corresponding edges. Each vertex v_i is associated with a random variable $\boldsymbol{\beta}_i$ for $i = 1, 2, \dots, k$. The pairwise MRF model is defined as

$$\begin{aligned} \Pi(\boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_k) &= \exp\left\{ \sum_{i \in E} H_i(\boldsymbol{\beta}_i) + \sum_{(i,j) \in E, j \neq i} H_{ij}(\boldsymbol{\beta}_i \boldsymbol{\beta}_j) - A(W) \right\} \\ &= \frac{1}{Z_H} \exp\left(H(\boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_k) \right), \end{aligned} \tag{9}$$

where Z_H is the normalizing constant. For example, for a Gaussian MRF, $H_i(\boldsymbol{\beta}_i) = -W_{ii}\boldsymbol{\beta}_i^2/2$ and $H_{ij}(\boldsymbol{\beta}_i\boldsymbol{\beta}_j) = -W_{ij}\boldsymbol{\beta}_i\boldsymbol{\beta}_j/2$; while for a binary MRF, i.e., the celebrated Ising model, $H_i(\boldsymbol{\beta}_i) = W_{ii}\boldsymbol{\beta}_i$ and $H_{ij}(\boldsymbol{\beta}_i\boldsymbol{\beta}_j) = W_{ij}\boldsymbol{\beta}_i\boldsymbol{\beta}_j$. We can then decompose the pairwise MRF into a vertex-wise term P and an interaction term M, then

$$\begin{aligned}&\Pi\left(\boldsymbol{\beta}_1,\ldots,\boldsymbol{\beta}_k\right) \propto P\left(\boldsymbol{\beta}_1,\ldots,\boldsymbol{\beta}_k\right) M\left(\boldsymbol{\beta}_1,\ldots,\boldsymbol{\beta}_k\right), \text{ with}\\ &P\left(\boldsymbol{\beta}_1,\ldots,\boldsymbol{\beta}_k\right) := \frac{1}{Z_P}\exp\left\{\sum_i H_i\left(\boldsymbol{\beta}_i\right)\right\}; M\left(\boldsymbol{\beta}_1,\ldots,\boldsymbol{\beta}_k\right) := \frac{1}{Z_M}\exp\left\{\sum_{C\in\mathcal{C}_2} H_C\left(\boldsymbol{\beta}_C\right)\right\},\end{aligned} \tag{10}$$

where $\mathcal{C}_2 := \{C \in \mathcal{C} \mid s.t. : |C| = 2\}$ and $\mathcal{C}$ is the set of all cliques for the random graph (V, E, W). For the spatial clustered coefficient regression, we study the component P defined in equation (10) with a MFM prior. Our next theorem provides the generalized urn-model induced by MRF-MFM, thus a collapsed Gibbs sampler can be applied.

Theorem A.1. *Suppose the data generating process follows equation* (8) *with H replaced by the Markov random field $\Pi(\boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_k)$ in equation* (10). *If P is a continuous distribution and $n_0 > 1$, the distributions of $\boldsymbol{\beta}_{n_0}$ given $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_{n_0-1}$ is proportional to*

$$\frac{V_{n_0}(t+1)\gamma}{V_{n_0}(t)} P(\boldsymbol{\beta}) + \sum_{i=1}^{t} \exp\left(H_{i|-i}\left(\boldsymbol{\beta}_i \mid \boldsymbol{\beta}_{-i}\right)\right)\left(n_i + \gamma\right)\delta_{\boldsymbol{\beta}_i^*},$$

with

$$V_{n_0}(t) = \sum_{k=1}^{\infty} \frac{k_{(t)}}{(\gamma k)^{(n_0)}} p_K(k); \quad H_{i|-i}\left(\boldsymbol{\beta}_i \mid \boldsymbol{\beta}_{-i}\right) = \sum_{\{j:(i,j)\in E, j\neq i\}} H_{ij}(\boldsymbol{\beta}_i\boldsymbol{\beta}_j), \tag{11}$$

where $\boldsymbol{\beta}_1^*, \ldots, \boldsymbol{\beta}_t^*$, $t \le n_0 - 1$ are the distinct values taken by $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_{n_0-1}$ and $n_i = \#\{j \in \{1, 2, \ldots, n_0 - 1\} : \boldsymbol{\beta}_j = \boldsymbol{\beta}_i^*\}$, $x^{(m)} = x(x+1)\cdots(x+m-1)$ and $x_{(m)} = x(x-1)\cdots(x-m+1)$.

This theorem shows how the MRF constraints directly affect the urn sampling scheme compared with MFM.

A.3 Theoretical Properties under the exchangeable structure

In this section, we assume the covariates $\boldsymbol{X}(s_i)$ are generated from random homogenous distribution so it is marginalized. The incorporation of proper dependency structures into the estimation process and assessing uncertainty is always an interesting subject. However, complex dependency structures may destroy the consistency of MFM. Therefore, to maintain theoretical consistency, this paper considers the case in which samples from the base measure are *a subset of an infinite sequence of exchangeable variables*.

In Bayesian Statistics, the infinite sequence of exchangeable random variables is an important concept. When $\boldsymbol{\beta}_1, \ldots$ are infinite exchangeable, for any finite k,

$$\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_k \overset{\mathcal{D}}{=} \boldsymbol{\beta}_{\pi(1)}, \ldots, \boldsymbol{\beta}_{\pi(k)} \text{ for all } \pi \in S(k), \tag{12}$$

where $S(k)$ is the set of all permutations for the index set $\{1, \ldots, k\}$. If $\boldsymbol{\beta}_1, \ldots$ are i.i.d. sampled from a distribution $P(\boldsymbol{\beta})$, then they are exchangeable, but the reverse is not always true. Some widely used models are based on exchangeable random variables that are not independent, like the Pólya's Urn (Blackwell et al., 1973) and Gaussian random variables that have the same marginal distribution and the same correlation between any two of them.

The famous de Finetti's Theorem (De Finetti, 1929) reveals the intrinsic characterization of exchangeable random variables: there is a latent random variable $\boldsymbol{\theta}$, such that $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_n$ are a subset of a infinite sequence of exchangeable variables sampled from $\Pi(\boldsymbol{\beta}_1, \ldots)$. It is summarized into the following sampling procedure:

$$\boldsymbol{\theta} \sim \boldsymbol{\Theta}, \quad \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_k \overset{i.i.d.}{\sim} \Pi(\boldsymbol{\beta}|\boldsymbol{\theta}), \tag{13}$$

where $\boldsymbol{\Theta}$ only depends on $\Pi(\boldsymbol{\beta}_1, \ldots)$. In other words, a subset of an infinite sequence of exchangeable variables are conditionally i.i.d. given their latent labels. We refer to Bernardo & Smith (2009) for more details on exchangeable sequences.

Theorem A.2. *Suppose the data generating process follows equation* (8) *with H replaced by the hierarchical distribution in equation* (13)*, and the distribution is correctly specified. If $p_K(1), \ldots, p_K(k) > 0$, denote T as the random variable for the number of clusters. Then we have*

$$|p\,(T = t \mid \boldsymbol{y}) - p\,(K = k \mid \boldsymbol{y})| \longrightarrow 0 \tag{14}$$

as $n \to \infty$.

Theorem A.2 provide some insight into our proposed MRF-MFM, compared to Dirichlet process mixture model with the above Markov random fields (DP-MRF; Orbanz & Buhmann, 2008). For DP-MRF, there could be a lot of small spurious clusters due to inconsistency of the Dirichlet process mixture even in the i.i.d. case (Miller & Harrison, 2013). Due to the fact that we specify a prior distribution for the number of components, the number of components in the posterior is appropriately regularized. Even though the consistency result only holds for the exchangeable structure, we believe that the regularization effect holds for all types of structures. Theorem A.2 is an extension of Theorem 5.2 in Miller & Harrison (2018) to the case of an exchangeable base measure. The limitation of the above theorem is that it does not explore the frequentist property of the posterior, where the number of clusters is assumed to be a fixed truth.

B Simulation

B.1 Settings

Our goal is to sample from the posterior distribution of the unknown parameters k, $z = (z_1, ..., z_n) \in \{1, ..., k\}$ and $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_k)$. We choose $k - 1 \sim \text{Poisson}(1)$ and $\gamma = 1$ in (36), $\boldsymbol{\mu} = \mathbf{0}_n$, $\boldsymbol{V} = 100\boldsymbol{I}_n$ and $\boldsymbol{\alpha} = \boldsymbol{\kappa} = 10000\mathbf{1}_n$ for all the simulations and real data analysis, where $\mathbf{0}_n$ is an n-dimensional vector with 0, $\mathbf{1}_n$ is an n-dimensional vector of 1's, and $\boldsymbol{I}_n$ is an n-dimensional identity matrix. The computing algorithm and full conditional distributions are presented in Appendix G, which efficiently cycles through the full conditional distributions of $z_i|z_{-i}$ for $i = 1, 2, \ldots, n$ and $\boldsymbol{\beta}$, where $z_{-i} = z \setminus z_i$. The marginalization over k can avoid complicated reversible jump MCMC algorithms or even allocation samplers. The posterior sampling algorithm is given in Algorithm 1 in Appendix G. The details of the deviations of full conditionals are also given in Appendix G.

It is not appropriate to use the posterior mean or median of clustering configurations $[z]$. Dahl's method Dahl (2006) provides a remedy for posterior inference of clustering configurations based on the squared error loss. There are also alternative loss functions in Wade et al. (2018) that do not involve squared errors such as those in Dahl (2006). The Rand Index (RI) Rand (1971) is used to measure the accuracy of clustering. The tuning parameter in Markov random fields needs to be selected in our proposed model. The Logarithm of the Pseudo-Marginal Likelihood (LPML) Ibrahim et al. (2013) is applied for tuning parameter selection, where a model with a larger LPML value is more preferred.

B.2 Simulation Setting and Evaluation Metrics

Our analysis is based on the spatial structure of the state of Georgia, which contains 159 counties. Using the county-level data, we build the graph using an adjacency matrix among different counties. 159 counties represent 159 vertices in this graph, and if a county shares a boundary with another county, then v_i and v_j are connected. This graph is used for both simulation studies and real data analysis. We consider two different spatial cluster designs shown in Figure 3. The first design consists of two disjoint parts located in the top and bottom parts of Georgia. A second cluster comprises the counties in the middle. The second design comprises three major spatial clusters. It is designed to mimic a common premature death pattern in which geographically distant areas can share a similar distribution pattern, and geographical proximity is not considered the only factor responsible for homogeneity in premature death rates.

Two different scenarios are considered for each design. The first scenario does not take into account spatial random effects, while in the second scenario, spatial random effects are included for each design. The spatial random effects are assumed to follow a multivariate normal distribution with a mean zero and exponential covariogram. Our simulation study consists of four scenarios in total. The details of the data generation process are given in Appendix H. The four scenarios are for two cluster

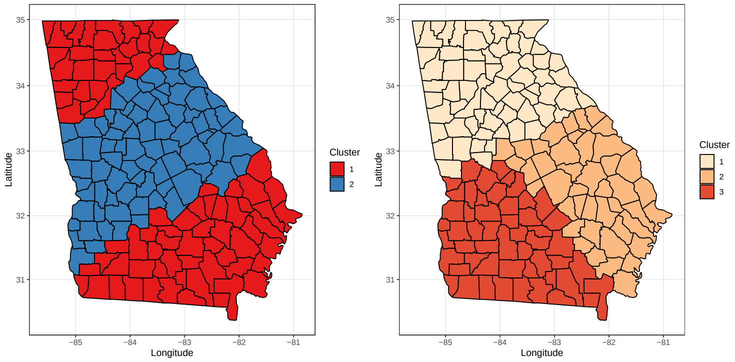


Figure 3: Simulation design with two and three cluster assignments

design without spatial random effect, two cluster design with spatial random effect, three cluster design without spatial random effect, and three cluster design with spatial random effect, respectively. In the three clusters design, the original regression coefficients are set to be 0.5, 1 and 1.5 for each cluster correspondingly. On the other hand, in two clusters design, the original regression coefficient set to be 1 and 1.5 for each cluster, respectively. For each case, we add the spatial random effect with the intensity. We use the centroid coordinate in each county to represent that county then construct the spatial random effect. Also, the range parameter and spatial variance parameter are both fixed in each simulation. In each case, we avoid the zero count value to prevent numerical instability. Based on the estimated number of clusters and Rand Index (RI), the clustering performance is evaluated. Each replicate is also used to calculate the final number of clusters estimated. A total of 100 sets of data are generated under different scenarios. We run 5000 iterations of the MCMC chain and burn-in the first 1000 for each replicate.

B.3 Simulation Results

For each replicated data set, we fit MFM and MRF-MFM with different values of the smoothness parameter and select the best smoothness parameter for each replicate based on LPML. We see that our model outperforms the MFM model in terms of LPML in all four different scenarios. We also evaluate the performance in terms of estimation results of the number of clusters. We report the proportion of times the true cluster recovered among the 100 replicates. For the two-cluster without spatial random effect design, we find out our model can recover the true number of clusters 100% of the replicates. And the MFM model can recover 85% of the replicates. In this case, both models perform well in the number of clusters estimation. But our model outperforms the MFM model in terms of LPML value. For the two-cluster design with spatial random effects, we see that our model can recover the true number of clusters 97% of the replicates, but the MFM model did not recover the true cluster for any replicates. For the three-cluster without spatial random effect design, we find out our model can recover the true number of clusters 88% of the replicates. On the other hand, MFM recovers 62% of the replicates. Finally, for the three-cluster design with spatial random effects, we find out our model can recover the true number of clusters 73% of the replicates. However, MFM did not recover the true cluster for all replicates.

The results of the comparison of LPML, Rand index, and estimation of the number of clusters for each design can be found in Table 2. Our method can effectively estimate the true number of clusters based on the results shown in Table 2. However, if spatial random effects exist, MFM will overestimate the number of clusters. Our proposed method also outperforms vanilla MFM with respect to model fitness and clustering, as demonstrated by the LPML values and Rand index.

Furthermore, we show the average mean square error (AMSE) of our proposed method and MFM in Table 3. We see that in all four different scenarios, our proposed method outperforms MFM in

Table 2: Simulation Results for Four Scenarios including LPML, Rand Index (RI), and number of true cluster cover rate (CR) by MRF-MFM (optimal) model and MFM model. We provide mean and standard deviation for both LPML and RI.

Method	Scenario	LPML	RI	CR	Scenario	LPML	RI	CR
Optimal	1	-544.29	0.9970	100%	3	-752.76	0.9470	88%
		(12.06)	(0.0062)			(235.91)	0.0389	
MFM		-1146.32	0.9901	85%		-2201.69	0.9570	62%
		(593.33)	(0.0233)			(830.66)	(0.0231)	
Optimal	2	-690.91	0.9875	97%	4	-1297.58	0.8469	73%
		(34.36)	(0.0129)			(278.33)	0.0434	
MFM		-7632.18	0.8348	0%		-8890.92	0.8350	0%
		(1947.31)	(0.0597)			(2028.92)	(0.0431)	

Table 3: AMSE for β Estimation under All Scenarios

Method		No Spatial Random effect		With Spatial Random effect	
		Two Clusters	Three Clusters	Two Clusters	Three Clusters
MRF-MFM-MLG	$\hat{\beta}_1$	0.0848	0.2508	0.0966	0.3918
	$\hat{\beta}_2$	0.0839	0.2435	0.0967	0.3814
MFM-MLG	$\hat{\beta}_1$	0.1170	0.2841	0.3675	0.6996
	$\hat{\beta}_2$	0.1164	0.2781	0.3668	0.6898

terms of coefficients estimations. The improvement of our proposed methods is evident for the data generated from the model with spatial random effect.

C Discussion

Some topics beyond the scope of this paper are worth further investigations. First, in our MCMC algorithm, one numerical integration is required for Gibbs sampling. Proposing an efficient calculation algorithm of the numerical integration will broaden the applications of our proposed methods. Furthermore, the proposed algorithm is numerically unstable when zero counts are observed, which should be addressed in the future. Furthermore, different clusters may have different sparsity patterns of the covariates. Incorporating spatial clustered sparsity structure of regression coefficients into the model will enable the selection and identification of the most important covariates. One parameter of the Markov random field is required to be selected. Proposing a hierarchical model for the tuning parameter is also an interesting future work. The frequentist property of the posterior distribution is also expected to be explored in the future. The inconsistency issue for the number of components due to model misspecification (Miller & Harrison, 2018; Cai et al., 2021) is also worth addressing.

D Proof of the Theorem A.1

By Bayes' theorem, we have:

$$\Pi\left(\boldsymbol{\beta}_i|\boldsymbol{\beta}_{-i}\right) \propto \Pi\left(\boldsymbol{\beta}_1,\ldots,\boldsymbol{\beta}_{n_0}\right) = P\left(\boldsymbol{\beta}_1,\ldots,\boldsymbol{\beta}_{n_0}\right) M\left(\boldsymbol{\beta}_1,\ldots,\boldsymbol{\beta}_{n_0}\right) \propto P\left(\boldsymbol{\beta}_i|\boldsymbol{\beta}_{-i}\right) M\left(\boldsymbol{\beta}_i|\boldsymbol{\beta}_{-i}\right). \tag{15}$$

As shown in Miller & Harrison (2018), by conditioning on the different possible situations of the cluster for the new observations, we have

$$P\left(\boldsymbol{\beta}_i|\boldsymbol{\beta}_{-i}\right) \propto \frac{V_{n_0}(t+1)\gamma}{V_{n_0}(t)} P(\boldsymbol{\beta}_i) + \sum_{i=1}^{t}\left(n_i+\gamma\right)\delta_{\boldsymbol{\beta}_i^*}. \tag{16}$$

Let $\partial(i) := \{j : (i,j) \in E\}$. When considering the full conditional distribution

$$M\left(\boldsymbol{\beta}_i|\boldsymbol{\beta}_{-i}\right) \propto \exp\left(H_{i|-i}(\boldsymbol{\beta}_i|\boldsymbol{\beta}_{-i})\right), \tag{17}$$

where $H_{i|-i}(\boldsymbol{\beta}_i|\boldsymbol{\beta}_{-i})$ only depends on $H_{ij}(\boldsymbol{\beta}_i\boldsymbol{\beta}_j)$ for $(i,j)\in E$. Note that

$$H_{i|-i}\left(\boldsymbol{\beta}_i|\boldsymbol{\beta}_{-i}\right)=0 \quad \text{if } S_i \notin S_{\partial(i)} \ \ , \tag{18}$$

where $\boldsymbol{s}_i$ specifies the cluster that $\boldsymbol{\beta}_i$ belongs to. With the property in equation (18) and the assumption that P is continuous, $\exp\left(H_{i|-i}\left(\boldsymbol{\beta}_i|\boldsymbol{\beta}_{-i}\right)\right)=1$ almost surely for $\boldsymbol{\beta}_i \sim P$. Then given any measurable function f for $P(\boldsymbol{\beta})$ and any subset A for the domain of $\boldsymbol{\beta}_i$,

$$\begin{aligned}\int_A f\left(\boldsymbol{\beta}_i\right)M\left(\boldsymbol{\beta}_i|\boldsymbol{\beta}_{-i}\right)P\left(\boldsymbol{\beta}_i\right)d\boldsymbol{\beta}_i = &\int_A f\left(\boldsymbol{\beta}_i\right)\frac{1}{Z_{H'}}\exp\left(H_{i|-i}\left(\boldsymbol{\beta}_i|\boldsymbol{\beta}_{-i}\right)\right)P\left(\boldsymbol{\beta}_i\right)d\boldsymbol{\beta}_i\\ &= \int_A f\left(\boldsymbol{\beta}_i\right)\frac{1}{Z_{H'}}P\left(\boldsymbol{\beta}_i\right)d\boldsymbol{\beta}_i,\end{aligned} \tag{19}$$

where the constant $Z_{H'}$ only depends on the constant part of $M\left(\boldsymbol{\beta}_i|\boldsymbol{\beta}_{-i}\right)$. Hence, the full conditional of Π can be derived

$$\Pi\left(\boldsymbol{\beta}_i|\boldsymbol{\beta}_{-i}\right)\propto\frac{V_{n_0}(t+1)\gamma}{V_{n_0}(t)}P(\boldsymbol{\beta}_i)+\sum_{i=1}^{t}\exp\left(H_{i|-i}\left(\boldsymbol{\beta}_i|\boldsymbol{\beta}_{-i}\right)\right)\left(n_i+\gamma\right)\delta_{\boldsymbol{\beta}_i^*}. \tag{20}$$

E Proof of the Theorem A.2

Proposition E.1. *If the data generating process follows equation (8) with H replaced by the hierarchical distribution in equation (13), then we have*

$$\begin{aligned}p(\mathcal{C})=V_n(t)\prod_{c\in\mathcal{C}}\gamma^{(|c|)},\quad p(\mathcal{C}\mid k)=\frac{k_{(t)}}{(\gamma k)^{(n)}}\prod_{c\in\mathcal{C}}\gamma^{(|c|)},\\ p(K=t\mid T=t)=\frac{t_{(t)}}{V_n(t)(rt)^{(n)}}p_K(t)\to 1,\quad \mathcal{C}\perp K\mid T,\end{aligned} \tag{21}$$

where $t=|\mathcal{C}|$ is the number of clusters while T is the corresponding random variable of t and $V_n(t)$ is defined in equation (6).

The proof of this proposition directly follows from Miller & Harrison (2018), since all conclusions only involves on $\mathcal{C}$, K and T, while the i.i.d assumption on $\boldsymbol{\beta}$ is not used.

Lemma E.2. *Suppose the data generating process in Proposition E.1, such that the distribution is correctly specified. Given the cluster configuration $\mathcal{C}$, the data $\boldsymbol{y}$ and the number of components K are independent.*

Remark E.3. As with MFM, we generalize the same result to exchangeable cases. Since the dependence between $\boldsymbol{y}$ is totally decided by $\boldsymbol{\beta}$, when $\boldsymbol{\beta}$ are exchangeable, all the $\boldsymbol{\beta}$ play the same role in generating $\boldsymbol{y}$. When $\boldsymbol{\beta}$ are marginalized, the cluster configuration $\mathcal{C}$ covers the same information with the number of components K and the latent labels $\boldsymbol{z}$.

E.1 Proof of the Lemma E.2

Proof. We show that the conditional independence among data $\boldsymbol{y}$ and the number of components $K=k$ given the cluster configuration $\mathcal{C}$ still holds when all $\boldsymbol{\beta}$ are exchangeable.
Let $E_i=\{j:z_j=i\}$, based on the definition of E_i and $\boldsymbol{z}$, we have

$$p(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{z},k)=\prod_{i=1}^{k}\prod_{j\in E_i}p(y_j|\boldsymbol{\beta}_i)=\prod_{i=1}^{t}\prod_{j\in E_i}p(y_j|\boldsymbol{\beta}_i^*), \tag{22}$$

where $\boldsymbol{\beta}_i^*, i=1,2,\dots,t$ are the distinct values of $\boldsymbol{\beta}_{1:k}$ decided by $\boldsymbol{z}$ and $\boldsymbol{y}$, and $\boldsymbol{\beta}_{1:k}=(\boldsymbol{\beta}_1,...,\boldsymbol{\beta}_k)^\top$. Given $\boldsymbol{z}$, the transformation from variable $\boldsymbol{\beta}_{1:k}$ to $\boldsymbol{\beta}_{1:t}^*$, is totally decided, so when marginalizing the unused $\boldsymbol{\beta}_{(t+1):k}^*$, given any function $g(\boldsymbol{\beta}_{1:t}^*)$, we have the identity

$$\int_{\Theta^k}g(\boldsymbol{\beta}_{1:t}^*)p(\boldsymbol{\beta}|\boldsymbol{z},k)\left(d\boldsymbol{\beta}\right)=\int_{\Theta^t}g(\boldsymbol{\beta}_{1:t}^*)p(\boldsymbol{\beta}_{1:t}^*)d\boldsymbol{\beta}^*. \tag{23}$$

Note that $\boldsymbol{\beta}^*_{1:t}$ are exchangeable based on assumption, then the density after marginalizing $\boldsymbol{\beta}$ can be seen

$$\begin{aligned}
p(\boldsymbol{y}|\boldsymbol{z},k) &= \int_{\Theta^k} p(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{z},k)p(\boldsymbol{\beta}|\boldsymbol{z},k)d\boldsymbol{\beta} = \int_{\Theta^k} \prod_{i=1}^{k}\prod_{j\in E_i} p(y_j|\boldsymbol{\beta}_i)p(\boldsymbol{\beta}|\boldsymbol{z},k)\,(d\boldsymbol{\beta}) \\
&= \int_{\Theta^t} \prod_{i=1}^{t}\prod_{j\in E_i} p(y_j|\boldsymbol{\beta}^*_i)p(\boldsymbol{\beta}^*_{1:t})d\boldsymbol{\beta}^* \\
&\overset{(i)}{=} \int_{\Theta^t} \prod_{i=1}^{t} p(\boldsymbol{y}_{E_i}|\boldsymbol{\beta}^*_i) \int \prod_{i=1}^{t} p(\boldsymbol{\beta}^*_i|\boldsymbol{\theta})dF(\boldsymbol{\theta})d\boldsymbol{\beta}^* \qquad (24) \\
&\overset{(ii)}{=} \int\int_{\Theta^t} \prod_{i=1}^{t} \left[p(\boldsymbol{y}_{E_i}|\boldsymbol{\beta}^*_i)p(\boldsymbol{\beta}^*_i|\boldsymbol{\theta})\right] d\boldsymbol{\beta}^* dF(\boldsymbol{\theta}) \\
&\overset{(iii)}{=} \int \prod_{i=1}^{t} m_i(\boldsymbol{y}_{E_i},\boldsymbol{\theta})dF(\boldsymbol{\theta}),
\end{aligned}$$

where $m_i(\boldsymbol{y}_{E_i},\boldsymbol{\theta})$ is a function only depends on $\boldsymbol{y}_{E_i}$ and $\boldsymbol{\theta}$. In addition, (i) directly follows from de Finetti's Theorem; for (ii), we apply the Fubini's theorem; (iii) is because the expression depends only on $\boldsymbol{z}, k$ through $\mathcal{C} = \mathcal{C}(\boldsymbol{z})$ since there is no correspondence between E_i and $\boldsymbol{\beta}^*_i$ after integrating out $\boldsymbol{\beta}^*$. From the last expression, we can see $p(\boldsymbol{y}|\boldsymbol{z},k)$ can be represented as a function of $\mathcal{C}, \boldsymbol{y}$, which implies that $\boldsymbol{y}$ and K are conditional independent given the cluster configuration $\mathcal{C}$. □

Based on the fourth line in Proposition 1 and Lemma E.2, we have $\mathcal{C} \perp K \,|\, T$ and $\boldsymbol{y} \perp K \,|\, \mathcal{C}$. Then we have

$$p(\boldsymbol{y}|t,k) = \sum_{\mathcal{C}:|\mathcal{C}|=t} p(\boldsymbol{y}|\mathcal{C},t,k)p(\mathcal{C}|t,k) = \sum_{\mathcal{C}:|\mathcal{C}|=t} p(\boldsymbol{y}|\mathcal{C},t)p(\mathcal{C}|t) = p(\boldsymbol{y}|t), \tag{25}$$

which implies $\boldsymbol{y} \perp K \,|\, T$. Then for any $n \geq k$,

$$p\left(K=k\,|\,\boldsymbol{y}\right) = \sum_{t=1}^{k} p\left(K=k\,|\,T=t,\boldsymbol{y}\right)p\left(T=t\,|\,\boldsymbol{y}\right) = \sum_{t=1}^{k} p\left(K=k\,|\,T=t\right)p\left(T=t\,|\,\boldsymbol{y}\right). \tag{26}$$

In addition, $p(K=t|T=t) = 1/V_n(t) \longrightarrow 1$ as $n \to \infty$ based on the third equation in Proposition 1. Thus

$$p\left(K=k\,|\,\boldsymbol{y}\right) \to \sum_{t=1}^{k} I(k=t)p(T=t\,|\,\boldsymbol{y}) = p(T=t\,|\,\boldsymbol{y}). \tag{27}$$

F Review Multivariate Log-Gamma Distribution

F.1 Probability Density Function for Multivariate Log-Gamma Distribution

We first review the multivariate log-gamma distribution from Bradley et al. (2018). We define the n-dimensional random vector $\boldsymbol{\phi} = (\phi_1, ..., \phi_n)'$, which consists of n mutually independent log-gamma random variables with shape and scale parameters organized into the n-dimensional vectors $\boldsymbol{\alpha} \equiv (\alpha_1, ..., \alpha_n)'$, and $\boldsymbol{\kappa} \equiv (\kappa_1, ..., \kappa_n)'$, respectively. Then define the n-dimensional random vector $\boldsymbol{q}$ as follows

$$\boldsymbol{q} = \boldsymbol{\mu} + \boldsymbol{V}\boldsymbol{\phi}, \tag{28}$$

where the matrix $\boldsymbol{V} \in \mathcal{R}^n \times \mathcal{R}^n$ and $\boldsymbol{\mu} \in \mathcal{R}^n$. Bradley et al. (2018) called $\boldsymbol{q}$ the multivariate log-gamma random vector. The random vector $\boldsymbol{q}$ has the following probability density function:

$$f(\boldsymbol{q} \mid \boldsymbol{c}, \boldsymbol{V}, \boldsymbol{\alpha}, \boldsymbol{\kappa}) = \frac{1}{\text{det}(\boldsymbol{V})}\left(\prod_{i=1}^{n}\frac{\kappa_i^{\alpha_i}}{\Gamma(\alpha_i)}\right)\exp[\boldsymbol{\alpha}'\boldsymbol{V}^{-1}(\boldsymbol{q}-\boldsymbol{\mu}) - \boldsymbol{\kappa}'\exp\{\boldsymbol{V}^{-1}(\boldsymbol{q}-\boldsymbol{\mu})\}];\;\; \boldsymbol{q} \in \mathcal{R}^n, \tag{29}$$

where "det" represents the determinant function. As a shorthand we use the notation, MLG $(\boldsymbol{\mu}, \boldsymbol{V}, \boldsymbol{\alpha}, \boldsymbol{\kappa})$, for the probability density function in (29).

F.2 Conditional Distributions for Multivariate Log-Gamma Random Vectors

Gibbs sampling from full-conditional distributions will require simulating from conditional distributions of multivariate log-gamma random vectors. Here, we provide a review for the technical results needed to simulate from these conditional distributions.

We first look at the Proposition 1 from Bradley et al. (2018). Let $\boldsymbol{q} \sim \text{MLG}\,(\boldsymbol{c}, \boldsymbol{V}, \boldsymbol{\alpha}, \boldsymbol{\kappa})$, and let $\boldsymbol{q} = (\boldsymbol{q}_1', \boldsymbol{q}_2')'$, where $\boldsymbol{q}_1$ is g-dimensional and $\boldsymbol{q}_2$ is $(n-g)$-dimensional. In a similar manner, partition $\boldsymbol{V}^{-1} = [H\ B]$ into an $m \times g$ matrix H and an $m \times (m-g)$ matrix B. Then, the conditional pdf of is given by

$$f(\boldsymbol{q}_1 \mid \boldsymbol{q}_2 = \boldsymbol{d}, \boldsymbol{c}, \boldsymbol{\alpha}, \boldsymbol{\kappa}) = M \exp(\boldsymbol{\alpha}' \boldsymbol{H} \boldsymbol{q}_1 - \boldsymbol{\kappa}_{1.2}' \exp(\boldsymbol{H} \boldsymbol{q}_1)). \tag{30}$$

where $\boldsymbol{\kappa}_{1.2} \equiv \exp(\boldsymbol{B}\boldsymbol{d} - \boldsymbol{V}^{-1}\boldsymbol{c} - log(\boldsymbol{\kappa}))$ and the normalizeing constant M is

$$M = \frac{1}{det(\boldsymbol{V}\boldsymbol{V}')^{\frac{1}{2}}} \left(\prod_{i=1}^{n} \frac{\kappa_i^{\alpha_i}}{\Gamma(\alpha_i)} \right) \frac{\exp \boldsymbol{\alpha}' \boldsymbol{B} \boldsymbol{d} - \boldsymbol{\alpha}' \boldsymbol{V}^{-1} \boldsymbol{c}}{\left[\int f(\boldsymbol{q} \mid \boldsymbol{c}, \boldsymbol{V}, \boldsymbol{\alpha}, \boldsymbol{\kappa}) d\boldsymbol{q}_1 \right]_{\boldsymbol{q}_2 = \boldsymbol{d}}}, \tag{31}$$

so the cMLG$(\boldsymbol{H}, \boldsymbol{\alpha}, \boldsymbol{\kappa}_{1.2})$ is equal to the pdf in equation (34), where "cMLG" stands for "conditional multivariate log-gamma." In Bradley et al. (2018), it indicates that cMLG does not fall within the same class of pdfs given in (30). This is primarily due to the fact that the real-valued matrix H, within the expression of cMLG, is not square. Thus, we require an additional result that allows us to simulate from cMLG.

Next, we look at the Theorem 2 from Bradley et al. (2018). Let $\boldsymbol{q} \sim \text{MLG}\,(\boldsymbol{0}_n, \boldsymbol{V}, \boldsymbol{\alpha}, \boldsymbol{\kappa})$, and partition this n-dimensional random vector so that $\boldsymbol{q} = (\boldsymbol{q}_1', \boldsymbol{q}_2')'$, where $\boldsymbol{q}_1$ is g-dimensional and $\boldsymbol{q}_2$ is $(n-g)$-dimensional. Additionally, consider the class of MLG random vectors that satisfy the following:

$$\boldsymbol{V}^{-1} = [Q_1\ Q_2] \begin{bmatrix} R_1 & \boldsymbol{0}_{g,n-g} \\ \boldsymbol{0}_{n-g,g} & \frac{1}{\sigma_2} \boldsymbol{I}_{n-g} \end{bmatrix} \tag{32}$$

where in general $\boldsymbol{0}_{r,t}$ is a $r \times t$ matrix of zeros; $\boldsymbol{I}_{n-g}$ is an $(n-g) \times (n-g)$ identity matrix;

$$\boldsymbol{H} = [Q_1\ Q_2] \begin{bmatrix} R_1 \\ \boldsymbol{0}_{n-g,g} \end{bmatrix} \tag{33}$$

is the QR decomposition of the $n \times g$ matrix H; the $n \times g$ matrix Q_1 satisfies $Q_1'Q_1 = \boldsymbol{I}_g$, the $n \times (n-g)$ matrix Q_2 satisfies $Q_2'Q_2 = \boldsymbol{I}_{n-g}$, and $Q_2'Q_1 = \boldsymbol{0}_{n-g,g}$; R_1 is a $g \times g$ upper triangular matrix; and $\sigma_2 > 0$. Hence, the marginal distribution of the g-dimensional random vector q_1 is given by

$$f(\boldsymbol{q}_1 \mid \boldsymbol{H}, \boldsymbol{\alpha}, \boldsymbol{\kappa}) = M_1 \exp(\boldsymbol{\alpha}' \boldsymbol{H} \boldsymbol{q}_1 - \boldsymbol{\kappa}' \exp(\boldsymbol{H} \boldsymbol{q}_1)). \tag{34}$$

where the normalizing constant M_1 is

$$M_1 = det([\boldsymbol{H} \quad Q_2]) \left(\prod_{i=1}^{n} \frac{\kappa_i^{\alpha_i}}{\Gamma(\alpha_i)} \right) \frac{1}{\int f(\boldsymbol{q} \mid \boldsymbol{0}_n, \boldsymbol{V} = [\boldsymbol{H}\ Q_2]^{-1}, \boldsymbol{\alpha}, \boldsymbol{\kappa}) d\boldsymbol{q}_1}. \tag{35}$$

And, the g-dimensional random vector $\boldsymbol{q}_1$ is equal in distribution to $(\boldsymbol{H}'\boldsymbol{H})^{-1}\boldsymbol{H}'\boldsymbol{\omega}$, where the n-dimensional random vector $\boldsymbol{\omega} \sim \text{MLG}(\boldsymbol{0}_n, \boldsymbol{I}_n, \boldsymbol{\alpha}, \boldsymbol{\kappa})$.

In Bradley et al. (2018), it is evident that this particular class of marginal distributions (defined in Theorem 2 in Bradley et al. (2018)) falls into the same class of distributions as the conditional distribution of $\boldsymbol{q}_1$ given $\boldsymbol{q}_2$. And Theorem 2 in Bradley et al. (2018) provides a way to simulate from cMLG. Furthermore, it shows that it is (computationally) easy to simulate from cMLG provided that $g \ll n$. Recall that $\boldsymbol{H}$ is $n \times g$, which implies that computing the $g \times g$ matrix $(\boldsymbol{H}'\boldsymbol{H})^{-1}$ is computationally feasible when g is "small." We refer the readers to see Bradley et al. (2018) for a comprehensive discussion.

Table 4: Parameters of the full conditional distribution

Parameter	Form
$\boldsymbol{H}_\beta$	$\begin{bmatrix} \boldsymbol{V^{-1}} \\ \boldsymbol{X}(\boldsymbol{s}_i) \end{bmatrix}$
$\boldsymbol{\alpha}_\beta$	$\begin{bmatrix} \alpha \\ \sum_{z_i=r} y(\boldsymbol{s}_i) \end{bmatrix}$
$\boldsymbol{\kappa}_\beta$	$\begin{bmatrix} \kappa \\ \sum_{z_i=r} I_{(z_i=r)} \end{bmatrix}$

G Full Conditional Distributions and Algorithm

In general, the hierarchical model can be expressed as follows

$$
\begin{aligned}
\textbf{Data Model: } & y(\boldsymbol{s}_i) \mid \boldsymbol{\beta}(\boldsymbol{s}_i) \sim \text{Poisson}(\exp\left(\boldsymbol{X}(\boldsymbol{s}_i)\boldsymbol{\beta}(\boldsymbol{s}_i)\right)) \\
\textbf{MRF: } & (\boldsymbol{\beta}(\boldsymbol{s}_1), \cdots, \boldsymbol{\beta}(\boldsymbol{s}_n)) \sim M(\boldsymbol{\beta}(\boldsymbol{s}_1), \cdots, \boldsymbol{\beta}(\boldsymbol{s}_n)) \prod_{i=1}^{n} G(\boldsymbol{\beta}(\boldsymbol{s}_i)) \\
\textbf{MLG: } & \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_k \sim \text{MLG}(\boldsymbol{\mu}, \boldsymbol{V}, \boldsymbol{\alpha}, \boldsymbol{\kappa}) \\
\textbf{MFM: } & G(\boldsymbol{\beta}(\boldsymbol{s}_i)) = \sum_{j=1}^{k} \pi_j \boldsymbol{\beta}_j, \pi_1, \ldots, \pi_k \mid k \sim \text{Dirichlet}(\gamma, \ldots, \gamma), \\
& k \sim p(\cdot), \text{where } p(\cdot) \text{ is a p.m.f on } \{1, 2, \ldots\}.
\end{aligned}
\tag{36}
$$

The full conditional distributions in Markov chain Monte Carlo (MCMC) sampling of MRF-MFM are given as follow.

For each term $\boldsymbol{\beta}_r$ in $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_k)$, the full conditional distribution is:

$$
\begin{aligned}
f(\boldsymbol{\beta}_r \mid -) \propto\ & \text{MLG}(\boldsymbol{0}_p, \boldsymbol{V}, \boldsymbol{\alpha}, \boldsymbol{\kappa}) \prod_{z_i=r} \text{Poisson}(\exp\left(\boldsymbol{X}(\boldsymbol{s}_i)\boldsymbol{\beta}(s_{z_i})\right) \\
& \propto \exp(\boldsymbol{\alpha}'\boldsymbol{V}^{-1}\boldsymbol{\beta}_r - \boldsymbol{\kappa}' \exp(\boldsymbol{V}^{-1}\boldsymbol{\beta}_r)) \prod_{z_i=r} \exp\left(\boldsymbol{X}(\boldsymbol{s}_i)\boldsymbol{\beta}(s_{z_i})\right)^{y(\boldsymbol{s}_i)} \exp(-\exp\left(\boldsymbol{X}(\boldsymbol{s}_i)\boldsymbol{\beta}(s_{z_i})\right)) \\
& \propto \exp\left(\boldsymbol{\alpha}'\boldsymbol{V}^{-1}\boldsymbol{\beta}_r + \sum_{z_i=r} y(\boldsymbol{s}_i)\boldsymbol{X}(\boldsymbol{s}_i)\boldsymbol{\beta}(s_{z_i})\right) \\
& \exp\left(-\boldsymbol{\kappa}' \exp(\boldsymbol{V}^{-1}\boldsymbol{\beta}_r) - \sum_{z_i=r} I_{(z_i=r)} \exp(\boldsymbol{X}(\boldsymbol{s}_i)\boldsymbol{\beta}(s_{z_i}))\right) \\
& \propto \exp\left[(\alpha, \sum_{z_i=r} y(\boldsymbol{s}_i))' \begin{bmatrix} \boldsymbol{V^{-1}} \\ \boldsymbol{X}(\boldsymbol{s}_i) \end{bmatrix} \beta_r\right] \exp\left[-(\kappa, \sum_{z_i=r} I_{(z_i=r)})' \exp(\begin{bmatrix} \boldsymbol{V^{-1}} \\ \boldsymbol{X}(\boldsymbol{s}_i) \end{bmatrix} \beta_r)\right]
\end{aligned}
\tag{37}
$$

This implies that $f(\boldsymbol{\beta}_r \mid -) \sim \text{cMLG}(\boldsymbol{H}_\beta, \boldsymbol{\alpha}_\beta, \boldsymbol{\kappa}_\beta)$.

For each term z_i in $z = (z_i, \ldots, z_n)$, the full conditional distribution is:

$$
P(z_i = c \mid z_1, \ldots, z_{i-1}) \propto \begin{cases} P(z_i = c \mid z_{-i}) d\text{Poisson}(y(\boldsymbol{s}_i), \exp(\boldsymbol{X}(\boldsymbol{s}_i)\boldsymbol{\beta}_r)), & \text{at table labeled } c \\ \frac{V_n(|C_{-i}|+1)}{V_n(|C_{-i}|)} \gamma m(y(\boldsymbol{s}_i)), & \text{if } c \text{ is a new table} \end{cases}.
$$

where

$$
\begin{aligned}
m(y(\boldsymbol{s}_i)) &= \int \text{MLG}(\boldsymbol{0}_p, \boldsymbol{V}, \boldsymbol{\alpha}, \boldsymbol{\kappa})\text{Poisson}(y(\boldsymbol{s}_i) \mid \boldsymbol{\beta}_r)d\boldsymbol{\beta}_r \\
&\propto \int \frac{1}{det(\boldsymbol{V}\boldsymbol{V}')^{\frac{1}{2}}} \left(\prod_{i=1}^{p} \frac{\kappa_i^{\alpha_i}}{\Gamma(\alpha_i)}\right) \exp(\boldsymbol{\alpha}'\boldsymbol{V}^{-1}\boldsymbol{\beta}_r - \ +\kappa' \exp(\boldsymbol{V}^{-1}\boldsymbol{\beta}_r)) \\
&\exp\left[\boldsymbol{X}(\boldsymbol{s}_i)\boldsymbol{\beta}_r\right]^{y(\boldsymbol{s}_i)} \exp\left[-\exp(\boldsymbol{X}(\boldsymbol{s}_i)\boldsymbol{\beta}_r)\right] \\
&= \frac{1}{det(\boldsymbol{V}\boldsymbol{V}')^{\frac{1}{2}}} \left(\prod_{i=1}^{p} \frac{\kappa_i^{\alpha_i}}{\Gamma(\alpha_i)}\right) \\
&\int \exp\left[(\alpha, \sum_{z_i=r} y(\boldsymbol{s}_i))' \begin{bmatrix} \boldsymbol{V^{-1}} \\ \boldsymbol{X}(\boldsymbol{s}_i) \end{bmatrix} \boldsymbol{\beta}_r\right] \exp\left[-(\kappa, \sum_{z_i=r} I_{(z_i=r)})' \exp(\begin{bmatrix} \boldsymbol{V^{-1}} \\ \boldsymbol{X}(\boldsymbol{s}_i) \end{bmatrix} \boldsymbol{\beta}_r)\right] \\
&= \frac{1}{det(\boldsymbol{V}\boldsymbol{V}')^{\frac{1}{2}}} \left(\prod_{i=1}^{p} \frac{\kappa_i^{\alpha_i}}{\Gamma(\alpha_i)}\right) \frac{1}{M_1}
\end{aligned}
$$

and

$$
M_1 = det([\boldsymbol{H}_\beta \quad Q_2]) \left(\prod_{i=1}^{n+p} \frac{\kappa_i^{\alpha_i}}{\Gamma(\alpha_i)}\right) \frac{1}{\int f(y(\boldsymbol{s}_i) \mid \boldsymbol{0}_{n+p}, \boldsymbol{V} = \left[\boldsymbol{H}_\beta \ Q_2\right]^{-1}, \boldsymbol{\alpha}, \boldsymbol{\kappa})}
$$

and "det" is a short hand as determinant of a matrix.

Algorithm 1 Collapsed sampler for MRF-MFM

```
Initialize: $z = (z_1, \ldots, z_n)$ and $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_k)$
for each iteration $= 1$ to $B$ do
  Update $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_k)$ conditional on $z$ in a closed form as
    $f(\boldsymbol{\beta}_r \mid -) \sim \text{cMLG}(\boldsymbol{H}_\beta, \boldsymbol{\alpha}_\beta, \boldsymbol{\kappa}_\beta)$
  where,
  $\boldsymbol{H_\beta} = \begin{bmatrix} \boldsymbol{V^{-1}} \\ \boldsymbol{X}(\boldsymbol{s}_i) \end{bmatrix}$ $\boldsymbol{\alpha}_\beta = \begin{bmatrix} \alpha \\ \sum_{z_i=r} y(\boldsymbol{s}_i) \end{bmatrix}$ $\boldsymbol{\kappa}_\beta = \begin{bmatrix} \kappa \\ \sum_{z_i=r} I_{(z_i=r)} \end{bmatrix}$
  Update $z = (z_1, \ldots, z_n)$ conditional on $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_k)$ for each $i$ in (1,...,n), we can get
  closed form expression for $P(z_i = c|z_{-i}, \boldsymbol{\beta})$:
    $\propto \begin{cases} P(z_i = c \mid z_{-i})d\text{Poisson}(y(\boldsymbol{s}_i), \exp(\boldsymbol{X}(\boldsymbol{s}_i)\boldsymbol{\beta}_c)), & \text{at an existing table labeled } c \\ \frac{V_n(|C_{-i}+1)|}{V_n(|C_{-i}|)}\gamma m(y(\boldsymbol{s}_i)), & \text{if } c \text{ is a new table} \end{cases}.$
  where $C_{-i}$ denotes the partition obtained by removing $z_i$ and
    $m(y(\boldsymbol{s}_i)) = \frac{1}{\text{det}(\boldsymbol{V}\boldsymbol{V}')^{\frac{1}{2}}} \left(\prod_{i=1}^{p} \frac{\kappa_i^{\alpha_i}}{\Gamma(\alpha_i)}\right) \frac{1}{M_1}$
  where,
    $M_1 = \text{det}([\boldsymbol{H}_\beta \quad Q_2]) \left(\prod_{i=1}^{n+p} \frac{\kappa_i^{\alpha_i}}{\Gamma(\alpha_i)}\right) \frac{1}{\int f(y(\boldsymbol{s}_i) \mid \boldsymbol{0}, \boldsymbol{V} = \left[\boldsymbol{H}_\beta, Q_2\right]^{-1}, \boldsymbol{\alpha}, \boldsymbol{\kappa})}$
end for
```

H Data Generation Processes

The four data generation processes are given as

1. $y(\boldsymbol{s}_i) \sim \text{Poisson}(X_1(\boldsymbol{s}_i)\beta_{1z_i} + X_2(\boldsymbol{s}_i)\beta_{2z_i})$, where $X_1(\boldsymbol{s}_i), X_2(\boldsymbol{s}_i) \overset{\text{ind}}{\sim} \text{Unif}(1, 2)$, $i = 1, \ldots, n$, $(\beta_{11}, \beta_{21}) = (1, 1)$, $(\beta_{12}, \beta_{22}) = (1.5, 1.5)$.

2. $y(\boldsymbol{s}_i) \sim \text{Poisson}(X_1(\boldsymbol{s}_i)\beta_{1z_i} + X_2(\boldsymbol{s}_i)\beta_{2z_i} + w(\boldsymbol{s}_i))$, where $X_1(\boldsymbol{s}_i), X_2(\boldsymbol{s}_i) \overset{\text{ind}}{\sim} \text{Unif}(1, 2)$, $i = 1, \ldots, n$, $(\beta_{11}, \beta_{21}) = (1, 1)$, $(\beta_{12}, \beta_{22}) = (1.5, 1.5)$. $\omega \sim \text{N}(0, \sigma^2_\omega H(\phi))$, where $H(\phi) = \exp(-\phi\|s_i - s_j\|)$, we set $\sigma^2_\omega = 0.3$ and $\phi = 0.05$.

3. $y(\boldsymbol{s}_i) \sim \text{Poisson}(X_1(\boldsymbol{s}_i)\beta_{1z_i} + X_2(\boldsymbol{s}_i)\beta_{2z_i})$, where $X_1(\boldsymbol{s}_i), X_2(\boldsymbol{s}_i) \overset{\text{ind}}{\sim} \text{Unif}(1, 2)$, $i = 1, \ldots, n$, $(\beta_{11}, \beta_{21}) = (0.5, 0.5)$, $(\beta_{12}, \beta_{22}) = (1, 1)$, $(\beta_{13}, \beta_{23}) = (1.5, 1.5)$.

4. $y(\boldsymbol{s}_i) \sim \text{Poisson}(X_1(\boldsymbol{s}_i)\beta_{1z_i} + X_2(\boldsymbol{s}_i)\beta_{2z_i} + w(\boldsymbol{s}_i))$, where $X_1(\boldsymbol{s}_i), X_2(\boldsymbol{s}_i) \overset{\text{ind}}{\sim} \text{Unif}(1, 2)$, $i = 1, \ldots, n$, $(\beta_{11}, \beta_{21}) = (0.5, 0.5)$, $(\beta_{12}, \beta_{22}) = (1, 1)$, $(\beta_{13}, \beta_{23}) = (1.5, 1.5)$. $\omega \sim \text{N}(0, \sigma^2_\omega H(\phi))$, where $H(\phi) = \exp(-\phi\|s_i - s_j\|)$, we set $\sigma^2_\omega = 0.3$ and $\phi = 0.05$.

I Additional Comparison for Simulation (State of Georgia)

We present additional comparison for simulation section (State of Georgia). We compare our proposed method to LGP and CAR in two cluster design. In Figure 4, the values above zero indicate that our method has higher LPML than comparator. The results shown that we have a better result for both comparator.

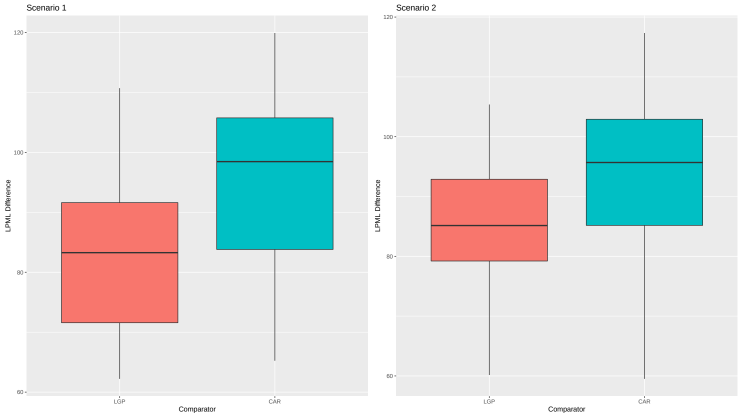


Figure 4: Additional Comparison for Two Cluster Simulation (State of Georgia).

J Additional Simulation for Different Spatial Graph (State of Mississippi)

We provide another simulation design with different spatial graph. This additional analysis is based on the spatial structure of the state of Mississippi, which contains 82 counties. We consider a different spatial cluster designs shown in Figure 5. This design consists of two disjoint parts located in the top and bottom parts of Mississippi.

Two different scenarios are considered. The first scenario does not take into account spatial random effects, while in the second scenario, spatial random effects are included for each design. The spatial random effects are assumed to follow a multivariate normal distribution with a mean zero and exponential covariogram. Based on the estimated number of clusters and Rand Index (RI), the clustering performance is evaluated. Each replicate is also used to calculate the final number of clusters estimated. A total of 50 sets of data are generated under different scenarios. We run 3000 iterations of the MCMC chain and burn-in the first 1000 for each replicate.

The results of the comparison of LPML, Rand index, and estimation of the number of clusters for each design can be found in Table 5. Our proposed method outperforms vanilla MFM with respect to model fitness and clustering, as demonstrated by the LPML values and Rand index. Additional comparison to LGP and CAR also presented. In Figure 6, the values above zero indicate that our method has higher LPML than comparator. The results shown that we have a better result for both comparator.

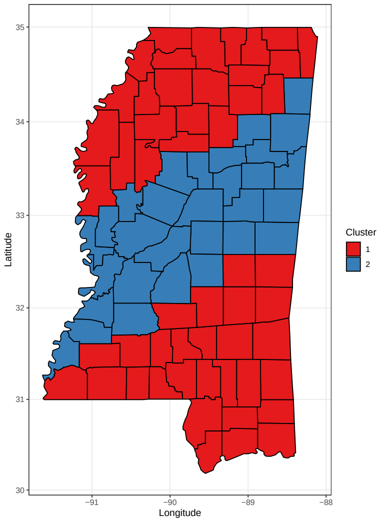


Figure 5: Simulation design with two cluster assignments. (State of Mississippi)

Table 5: Simulation Results including LPML, Rand Index (RI), and number of true cluster cover rate (CR) by MRF-MFM (optimal) model and MFM model. We provide mean and standard deviation for both LPML and RI.

Method	Scenario	LPML	RI	CR	Scenario	LPML	RI	CR
Optimal	1	-295.79	0.9954	100%	2	-291.76	0.9966	100%
		(9.29)	(0.0179)			(10.93)	(0.0135)	
MFM		-819.39	0.9901	98%		-727.47	0.9901	96%
		(407.15)	(0.0257)			(272.14)	(0.0269)	

References

Bernardo, J. M. and Smith, A. F. *Bayesian Theory*, volume 405. John Wiley & Sons, 2009.

Blackwell, D., MacQueen, J. B., et al. Ferguson distributions via pólya urn schemes. *The Annals of Statistics*, 1(2):353–355, 1973.

Bradley, J. R., Holan, S. H., Wikle, C. K., et al. Computationally efficient multivariate spatio-temporal models for high-dimensional count-valued data (with discussion). *Bayesian Analysis*, 13 (1):253–310, 2018.

Cai, D., Campbell, T., and Broderick, T. Finite mixture models do not reliably learn the number of components. In *International Conference on Machine Learning*, pp. 1158–1169. PMLR, 2021.

Dahl, D. B. Model-based clustering for expression data via a Dirichlet process mixture model. *Bayesian Inference for Gene Expression and Proteomics*, 4:201–218, 2006.

De Finetti, B. Funzione caratteristica di un fenomeno aleatorio. In *Atti del Congresso Internazionale dei Matematici: Bologna del 3 al 10 de settembre di 1928*, pp. 179–190, 1929.

Ferguson, T. S. A Bayesian analysis of some nonparametric problems. *The Annals of Statistics*, 1(2): 209–230, 1973.

Ibrahim, J. G., Chen, M.-H., and Sinha, D. *Bayesian Survival Analysis*. Springer Science & Business Media, 2013.

Miller, J. W. and Harrison, M. T. A simple example of Dirichlet process mixture inconsistency for the number of components. In *Advances in Neural Information Processing Systems*, pp. 199–206, 2013.

Miller, J. W. and Harrison, M. T. Mixture models with a prior on the number of components. *Journal of the American Statistical Association*, 113(521):340–356, 2018.

Orbanz, P. and Buhmann, J. M. Nonparametric Bayesian image segmentation. *International Journal of Computer Vision*, 77(1-3):25–45, 2008.

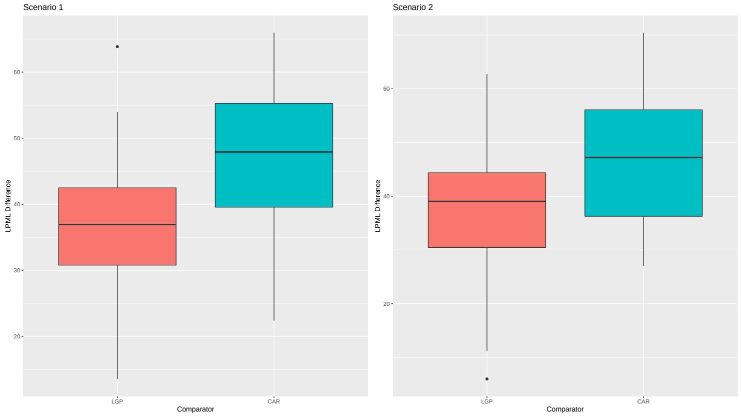


Figure 6: Additional Comparison for Two Cluster Simulation (State of Mississippi).

Rand, W. M. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, 66(336):846–850, 1971.

Wade, S., Ghahramani, Z., et al. Bayesian cluster analysis: Point estimation and credible balls (with discussion). *Bayesian Analysis*, 13(2):559–626, 2018.