Bayesian Spatial Clustering of Crime in Philadelphia with Particle Optimization

Cecilia Balocchi, Sameer K. Deshpande, Edward I. George, Shane T. Jensen

This draft: 28 October 2019

Abstract

Bayesian hierarchical modeling is a natural way to study spatial variation in crime dynamics within a city at the neighborhood level, since it facilitates principled "sharing of information" between spatially adjacent neighborhoods. Typically, however, cities contain many physical and social boundaries that may manifest as spatial discontinuities in crime patterns. As a result, standard prior choices often yield overly-smooth parameter estimates, which can ultimately produce mis-calibrated forecasts. To prevent potential over-smoothing, we introduce a prior that first partitions the neighborhoods into several clusters and then encourages spatial smoothness within each cluster. In our prior, we allow the effect of different covariates to be partitioned independently. As a result, conventional stochastic search techniques are computationally prohibitive, as they must traverse a combinatorially vast product space of partitions. We introduce an ensemble optimization procedure that simultaneously identifies several high probability combinations of partitions by solving one optimization problem using a new local search strategy. We then use the identified combinations of partitions to estimate crime trends in Philadelphia between 2006 and 2017. On simulated and real data, our proposed method demonstrates good estimation and partition selection performance.

Keywords: Variational inference, Bayesian model averaging

1 Introduction

Accurate modeling of criminal behavior benefits many stakeholders: law enforcement officials can make more informed decisions when deploying resources to ensure public safety, urban planners can better understand how socio-economic factors and the built environment affect crime, and city officials can develop community programs and interventions to improve the overall quality of life in the city. In this paper, we study how crime has evolved in the city of Philadelphia between 2006 and 2017 with a focus on finding clusters of neighborhoods with similar crime dynamics.

For the first time in decades, Philadelphia is experiencing population growth and is currently rapidly evolving; this transformation makes it an interesting real-time case study for examining how crime evolves over time. While there has been an overall decrease in crime over the last decade, we can obtain a more nuanced understanding by examining the temporal trends at a local neighborhood level. Indeed, in examining initial estimates of neighborhood crime dynamics in Figure 1, it is clear that the broad negative trend is not uniform across the city and that in a small number of neighborhoods, crime has actually increased over the last decade. Nevertheless, we also see in Figure 1 that, with few notable exceptions, neighborhoods that are spatially adjacent tend to have similar estimates, suggesting that there is a high degree of spatial correlation in the neighborhood-level crime dynamics.

Bayesian hierarchical modeling is a very natural way to study crime at the neighborhood level as it allows us to "borrow strength" between spatially adjacent neighborhoods. In fact, Balocchi and Jensen (2019) have demonstrated that Bayesian models that encourage spatial shrinkage can yield more accurate predictions than models that do not introduce dependencies between parameters from adjacent neighborhoods. Following that work, we propose a model that extends Bernardinelli et al. (1995)'s linear model with spatially varying intercept (mean level of crime) and spatially varying slope (time trend).

Priors based on conditionally auto-regressive (CAR) models (Besag, 1974) are workhorses in the Bayesian spatial statistics literature that encourage shrinking the parameters for each neighborhood towards the average value of the parameters from adjacent neighborhoods. Though these models are an intuitive and popular way to "share information" between spatially adjacent regions, they can introduce a level of smoothness at odds with the realities of complex urban environments. In fact, while Figure 1 displays considerable spatial correlation, it also includes several sharp discontinuities. This is because cities often contain

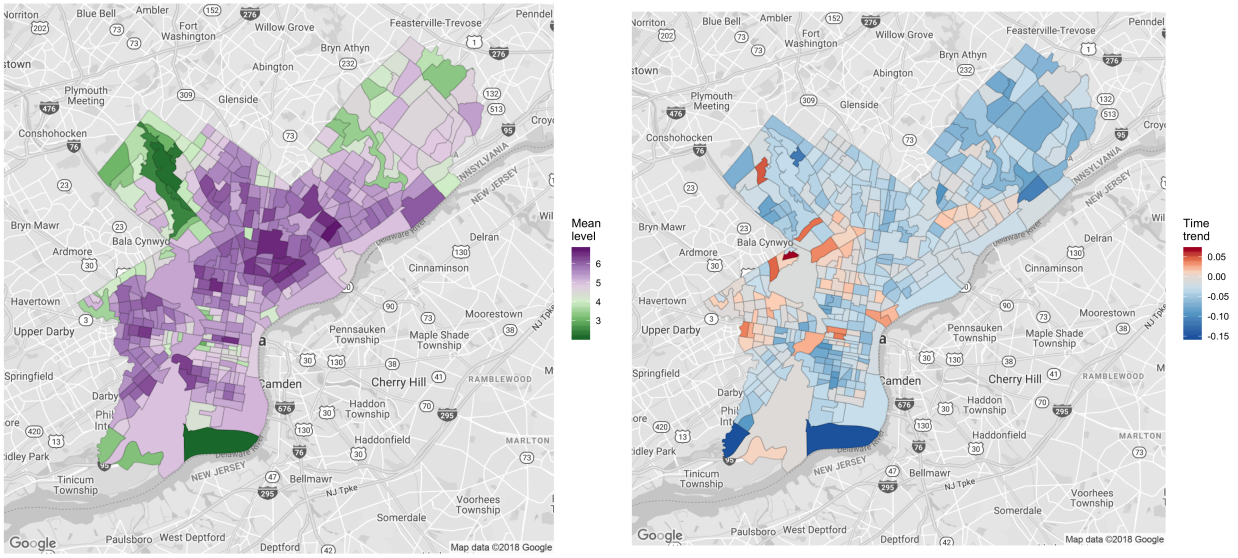

Figure 1: Visualization of the maximum likelihood estimates of the tract-level intercepts $\boldsymbol{\alpha}$ (left panel) and time-trends $\boldsymbol{\beta}$ (right panel) for the model defined in Section 2.1

boundaries within the physical or social environment that can manifest as sharp discontinuities in the modeled quantities. For example, major streets, parks or rivers may coincide with socioeconomic divisions that themselves may be associated with differences in crime patterns.

In the context of crime modeling, using a CAR prior that does not account for such discontinuities can potentially result in over-estimation (resp. under-estimation) of crime in low-crime (resp. high-crime) areas. While manually adjusting the CAR prior to prevent smoothing over these boundaries is conceptually simple, it presupposes knowledge about the location of these discontinuities, which is often not available. A far more elegant and agnostic approach is to use the data itself to identify the discontinuities.

There is a very rich literature on data-adaptive strategies for detecting discontinuities at the border between adjacent neighborhoods, also known as *wombling*. One approach to wombling involves first fitting a simple model that does not account for potential discontinuities and then identifying jumps in the fitted values (see, e.g., Boots (2001), Li et al. (2011), Banerjee et al. (2012), Lu and Carlin (2005), and Lee and Mitchell (2013)). Alternatively, many authors directly model uncertainty about which borders correspond to sharp discontinuities within larger Bayesian hierarchical models (see, e.g., Lee and Mitchell (2012), Lu et al.

(2007), and Balocchi and Jensen (2019)). By directly modeling uncertainty in discontinuity locations, these latter approaches introduce a large number of additional parameters to the model, one for each border between adjacent neighborhoods.

Rather than look for individual discontinuities between pairs of spatial units, we instead aim to identify *clusters* of neighborhoods that exhibit similar behavior within clusters but not necessarily between clusters; moreover, the boundaries between the clusters can be thought as continuous discontinuities. Besides allowing more flexible modeling and more accurate estimates, clustering aids with interpretation and dimensionality reduction of the complex phenomenon of crime change. In this paper, we propose a "CAR–within–clusters" model: we introduce two latent spatial partitions, for the mean levels of crime (intercepts) and the temporal trends (slopes), and specify independent CAR priors, with potentially different means, within each cluster. Like other spatial clustering approaches (see, e.g., Knorr-Held and Raßer (2000), Denison and Holmes (2001), Feng et al. (2016), and references therein), we treat parameters arising from different clusters independently *a priori*. However, unlike these other models, we do not require parameters within a cluster to be constant and instead allow them to vary smoothly. Our model is similar to the one in Anderson et al. (2017), which also allows for the mean and time trend to cluster differently, but we do not permit any spatial dependence between parameters from different clusters. Our approach therefore combines several positive aspects of clustering and wombling: we are able to find areas displaying different crime dynamics and simultaneously interpret borders between clusters as continuous boundaries corresponding to spatial discontinuities. We describe our data and introduce this model in Section 2.

We have two primary goals: (i) identify the two underlying spatial partitions and (ii) estimate the parameters and make predictions, taking into account the uncertainty about the partitions. These goals are complicated by the combinatorial vastness of the latent product space of spatial partitions, rendering typical stochastic search techniques computationally prohibitive. We instead focus on posterior optimization. However, rather than simply finding the *maximum a posteriori* (MAP) pair of partitions, we propose an extension of Ročková (2018)'s ensemble optimization framework that simultaneously identifies several pairs of partitions with high posterior probability by solving a *single* optimization problem. In Section 3, we show that solving this problem is formally equivalent to finding a particular variational approximation of the discrete posterior distribution of the pairs of partitions. We introduce a new local search strategy that, at a high level, involves running several greedy search trajectories that are made "mutually aware" with an entropy penalty. This penalty promotes

diversity among the trajectories by discouraging two search trajectories from visiting the same point in the latent discrete space. By identifying several high posterior probability pairs of partitions we can easily incorporate uncertainty about the latent clusterings into our estimation of the parameters and prediction, with a Bayesian Model Averaging (BMA; Raftery et al. (1997)) estimator. In Section 4, we illustrate our proposed methodology on simulated data before applying it to the Philadelphia data in Section 5. We conclude with a discussion of our results and an outline of potential directions for future work in Section 6.

2 Data, Model, and Direction

Our crime data comes from the Philadelphia Police Department, which publicly releases the location, time, and type of each reported crime[1]. Our analysis focuses on *violent* crimes, which include homicides, rapes, robberies, and aggravated assaults (FBI, 2011).

For the years between 2006 ($t = 0$) and 2017 ($t = 11$), let $c_{i,t}$ be the total number of violent crimes reported in tract i during year t. The city of Philadelphia is divided into $N = 384$ census tracts and we let $W = (w_{i,j})$ be a symmetric binary adjacency matrix where $w_{i,j} = 1$ if and only if census tracts i and j share a border.

The distribution of crime counts $c_{i,t}$ displays considerably skewness. Like Balocchi and Jensen (2019), rather than modeling $c_{i,t}$ directly, we work with an inverse hyperbolic sine transformation (Burbidge et al., 1988) of the violent crime counts:

$$y_{i,t} = \log\left(c_{i,t} + (c_{i,t}^2 + 1)^{1/2}\right) - \log(2).$$

This transformation is a close approximation of $\log(c_{i,t})$ which is well defined for neighborhoods that had a crime count of zero in certain years.

2.1 Model

To study the time-evolution of crime, we consider a simple linear regression model:

$$y_{i,t} = \alpha_i + \beta_i(t - \bar{t}) + \varepsilon_{i,t}; \quad \varepsilon_{i,t} \sim N(0, \sigma^2) \tag{1}$$

[1]The data is available for download at https://www.phillypolice.com/crime-maps-stats/index.html.

where time t has been centered, so that the parameters α_i and β_i represent respectively the mean level of crime and the trend over time of census tract i. When the number of time points is small or moderate, such models are typically employed as opposed to more complex non-linear models (Bernardinelli et al., 1995; Anderson et al., 2017).

We can find an initial estimate of α_i and β_i by treating the neighborhoods independently and computing the maximum likelihood estimates within each neighborhood. As seen in Figure 1, there is considerable spatial correlation between the MLEs. Thus, rather than fitting the model in (1) within each tract independently, we take a hierarchical Bayesian approach in order to "share strength" between census tracts. This involves specifying a prior distribution on the parameters $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_N)$ and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_N)$ and updating these priors with the observed data $\mathbf{y}$ to compute a posterior $\pi(\boldsymbol{\alpha}, \boldsymbol{\beta} \mid \mathbf{y})$. Because we expect the tract-specific parameters to display some spatial continuity, we use priors that explicitly introduce dependence between parameters from neighboring tracts.

Conditionally autoregressive (CAR) models are a particularly popular class of such priors and we use a version introduced in Leroux et al. (2000): given a binary adjacency matrix $W = (w_{i,j})$, we say the vector $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)$ follows a CAR model with grand mean $\overline{\theta}$ and variance scale τ^2 if and only if all of the full conditional distributions have the form

$$\theta_i \mid \boldsymbol{\theta}_{-i}, \overline{\theta}, \tau^2 \sim N\left(\frac{(1-\rho)\overline{\theta} + \rho \sum_j w_{i,j}\theta_j}{1 - \rho + \rho \sum_j w_{i,j}}, \frac{\tau^2}{1 - \rho + \rho \sum_j w_{i,j}}\right).$$

In this CAR model, the conditional mean of $\theta_i \mid \boldsymbol{\theta}_{-i}$ is a weighted average of the grand mean $\overline{\theta}$ and the average of the neighboring parameters. The degree to which θ_i is shrunk toward either of these targets is governed by a parameter ρ, which is typically set by the analyst, and the number of neighbors. These full conditionals uniquely determine the joint distribution $\boldsymbol{\theta} \sim N(\overline{\theta}\mathbf{1}_n, \tau^2\Sigma_{\text{CAR}})$ where

$$\Sigma_{\text{CAR}} = \begin{cases} [\rho W^\star + (1-\rho) I_n]^{-1} & \text{if } n \geq 2 \\ \frac{1}{1-\rho} & \text{if } n = 1 \end{cases},$$

$\mathbf{1}_n$ is the n-vector of ones and $W^\star$ is the unweighted graph Laplacian of the adjacency matrix W. For compactness, we will write $\boldsymbol{\theta} \mid \overline{\theta}, \tau^2 \sim \text{CAR}(\overline{\theta}, \tau^2, W)$.

Cities typically contain many natural and physical barriers like rivers and highways that

manifest in sharp spatial discontinuities. In the presence of these discontinuities, a naively specified CAR model can induce a level of spatial smoothness among the parameters at odds with the data. To guard against this possibility, we seek *clusters* of parameters that demonstrate considerable spatial continuity within but not between clusters. We introduce two latent partitions of $[N]$, $\gamma^{(\alpha)}$ and $\gamma^{(\beta)}$, where $\gamma^{(\cdot)} = \{S_1^{(\cdot)}, \ldots, S_{K^{(\cdot)}}^{(\cdot)}\}$. We refer to the sets $S_k^{(\cdot)}$ as *clusters* and restrict our attention only to those partitions whose clusters are spatially connected. We denote the set of all such partitions by $\mathcal{SP}$ and let $\boldsymbol{\gamma} := (\gamma^{(\alpha)}, \gamma^{(\beta)})$ be the pair of latent spatial partitions of interest. In what follows, we will simply refer to $\boldsymbol{\gamma}$ as a *particle*.

To simplify our presentation, we describe only the prior $\boldsymbol{\alpha} \mid \boldsymbol{\gamma}, \sigma^2$; we place an independent, analogous conditional prior on $\boldsymbol{\beta}$. Rather than placing a single CAR model on the entire collection $\boldsymbol{\alpha}$, conditional on $\gamma^{(\alpha)}$, we place independent CAR priors on the collections $\boldsymbol{\alpha}_k = \{\alpha_i : i \in S_k^{(\alpha)}\}$, so that the joint prior density $\pi(\boldsymbol{\alpha} \mid \gamma^{(\alpha)}, \sigma^2)$ factorizes over the collection of all clusters: $\pi(\boldsymbol{\alpha} \mid \gamma^{(\alpha)}, \sigma^2) = \prod_{k=1}^{K^{(\alpha)}} \pi(\boldsymbol{\alpha}_k \mid \sigma^2)$. To this end, we introduce a collection of grand cluster means $\overline{\boldsymbol{\alpha}} = \{\overline{\alpha}_1, \ldots, \overline{\alpha}_{K^{(\alpha)}}\}$ and model $\boldsymbol{\alpha}_k \mid \overline{\alpha}_k, \sigma^2 \sim \text{CAR}(\overline{\alpha}_k, a_1\sigma^2, W_k^{(\alpha)})$, where $W_k^{(\alpha)}$ is the sub-matrix of W whose rows and columns are indexed by the cluster $S_k^{(\alpha)}$. We further place independent $N(0, a_2\sigma^2)$ priors on the grand cluster means $\overline{\alpha}_k$ and place a fully-specific prior Π_γ on $\gamma^{(\alpha)}$. In Sections 4 and 5, we consider truncated uniform and Ewens-Pitman priors on the latent partitions, though the computational strategy introduced in Section 3 will work for general priors. We complete our hierarchical prior with an Inverse Gamma prior on the residual variance $\sigma^2 \sim \text{IG}\left(\frac{\nu_\sigma}{2}, \frac{\nu_\sigma \lambda_\sigma}{2}\right)$.

To summarize, our model is

$$
\begin{aligned}
\gamma^{(\alpha)}, \gamma^{(\beta)} &\overset{iid}{\sim} \Pi_\gamma \\
\sigma^2 &\sim \text{IG}\left(\frac{\nu_\sigma}{2}, \frac{\nu_\sigma \lambda_\sigma}{2}\right) \\
\overline{\alpha}_1, \ldots, \overline{\alpha}_{K_\alpha} \mid \gamma^{(\alpha)}, \sigma^2 &\overset{iid}{\sim} N(0, a_2\sigma^2) \\
\overline{\beta}_1, \ldots, \overline{\beta}_{K_\beta} \mid \gamma^{(\beta)}, \sigma^2 &\overset{iid}{\sim} N(0, b_2\sigma^2) \\
\boldsymbol{\alpha}_k \mid \overline{\alpha}_k, \sigma^2, \gamma^{(\alpha)} &\sim \text{CAR}(\overline{\alpha}_k, a_1\sigma^2, W_k^{(\alpha)}) \quad \text{for } k = 1, \ldots, K_\alpha \\
\boldsymbol{\beta}_{k'} \mid \overline{\beta}_{k'}, \sigma^2, \gamma^{(\beta)} &\sim \text{CAR}(\overline{\beta}_{k'}, b_1\sigma^2, W_{k'}^{(\beta)}) \quad \text{for } k' = 1, \ldots, K_\beta \\
y_{i,t} \mid \boldsymbol{\alpha}, \boldsymbol{\beta}, \sigma^2 &\sim N(\alpha_i + \beta_i(t - \overline{t}), \sigma^2)
\end{aligned}
\tag{2}
$$

The high degree of conditional conjugacy in (2) enables us to derive analytic expressions

for quantities such as the marginal likelihood $p(\mathbf{y} \mid \boldsymbol{\gamma})$ as well as the conditional posterior expectations $\mathbb{E}[\boldsymbol{\alpha}, \boldsymbol{\beta} \mid \boldsymbol{\gamma}, \mathbf{y}]$. The availability of these expressions will be crucial for the posterior exploration strategy we develop below.

Given the residual variance σ^2 and latent partitions $\gamma^{(\alpha)}$ and $\gamma^{(\beta)}$, parameters in different clusters are conditionally independent. In other words, our model falls with the class of conditional product partition models (PPMs) that have been widely used in Bayesian spatial statistics (see, e.g., Knorr-Held and Raßer (2000), Denison and Holmes (2001), and Feng et al. (2016)). Unlike these papers, however, we are interested in recovering two latent partitions, one each for the mean levels and time-trends within each census tract. Anderson et al. (2017) consider a model similar to ours but allow for spatial correlation between tracts in distinct clusters.

3 Posterior Exploration and Summarization

We have two simultaneous goals: (i) identify promising particles $\boldsymbol{\gamma}$ and (ii) estimate posterior expectations of the individual neighborhood-level parameters and make predictions; this second goal can generally be expressed in the form of $\mathbb{E}[f(\boldsymbol{\alpha}, \boldsymbol{\beta}) \mid \mathbf{y}]$ where f is an arbitrary functional. The combinatorial vastness of the space $\mathcal{SP}^2$, which contains all possible pairs of partitions, renders it impossible to enumerate all possible $\boldsymbol{\gamma}$ for even small values of N. As a result, we cannot compute posterior probability $\pi(\boldsymbol{\gamma} \mid \mathbf{y})$ exactly and it is tempting to resort to Markov Chain Monte Carlo (MCMC) simulations to approximate expectations $\mathbb{E}[f(\boldsymbol{\alpha}, \boldsymbol{\beta}) \mid \mathbf{y}]$. Naively, this might proceed in a Gibbs fashion, alternating between updating the two partitions in each $\boldsymbol{\gamma}$ and updating continuous parameters $(\boldsymbol{\alpha}, \boldsymbol{\beta}, \sigma^2)$, while holding the rest fixed. Unfortunately, because we must explore the vast space of pairs of partitions, such MCMC simulations may require a prohibitive amount of time to mix. To get around this difficulty, Anderson et al. (2017) arbitrarily restricted attention to partitions with no more than three to five clusters each. Even with such a restriction, which we will not impose, it is still quite difficult to distill the thousands of resulting draws of $\boldsymbol{\gamma}$ into a single point estimate and to quantify our particle selection uncertainty.

A popular alternative approach is posterior optimization, which usually focuses on identifying the *maximum a posteriori* (MAP) particle $\hat{\boldsymbol{\gamma}}_{MAP}$ or some other decision-theoretic optimal point estimate (see, e.g. Lau and Green, 2007). One then estimates the marginal expectation $\mathbb{E}[f(\boldsymbol{\alpha}, \boldsymbol{\beta}) \mid \mathbf{y}]$ with a "plug-in" estimator $\mathbb{E}[f(\boldsymbol{\alpha}, \boldsymbol{\beta}) \mid \mathbf{y}, \hat{\boldsymbol{\gamma}}_{MAP}]$. Though this procedure might

be substantially faster than MCMC, especially if the marginal likelihood $p(\mathbf{y} \mid \boldsymbol{\gamma})$ possesses certain ordering properties (Dahl, 2009), it completely eschews exploration of the uncertainty about $\boldsymbol{\gamma}$. As a result, the natural "plug-in" estimator $\mathbb{E}[f(\boldsymbol{\alpha}, \boldsymbol{\beta}) \mid \mathbf{y}, \hat{\boldsymbol{\gamma}}_{MAP}]$ may result in over-confident inference about the function f.

Notice, however, that this plug-in estimator may be viewed as a particular instantiation of Bayesian Model Averaging (BMA) (Raftery et al., 1997; Hoeting et al., 1999). At a very high-level, BMA aims to approximate the full marginal expectation

$$\mathbb{E}[f(\boldsymbol{\alpha}, \boldsymbol{\beta}) \mid \mathbf{y}] = \sum_{\boldsymbol{\gamma}} \pi(\boldsymbol{\gamma} \mid \mathbf{y}) \mathbb{E}[f(\boldsymbol{\alpha}, \boldsymbol{\beta}) \mid \mathbf{y}, \boldsymbol{\gamma}],$$

by first identifying some small subset Γ of models and then evaluating the more manageable sum

$$f_\Gamma = \sum_{\boldsymbol{\gamma} \in \Gamma} \pi_\Gamma(\boldsymbol{\gamma} \mid \mathbf{y}) \mathbb{E}[f(\boldsymbol{\alpha}, \boldsymbol{\beta}) \mid \mathbf{y}, \boldsymbol{\gamma}],$$

where π_Γ is the restriction of the posterior $\pi(\boldsymbol{\gamma} \mid \mathbf{y})$ to the set Γ.

Intuitively, the better the restricted posterior π_Γ approximates the full posterior $\pi(\boldsymbol{\gamma} \mid \mathbf{y})$, the closer f_Γ will be to the targeted marginal expectation $\mathbb{E}[f(\boldsymbol{\alpha}, \boldsymbol{\beta}) \mid \mathbf{y}]$. So rather than averaging over just the top $\boldsymbol{\gamma}$, a natural extension of the MAP plug-in may be to average over the top L $\boldsymbol{\gamma}$'s. Specifically if we let $\Gamma_L = \{\boldsymbol{\gamma}^{(1)}, \ldots, \boldsymbol{\gamma}^{(L)}\}$ be the L particles with largest posterior mass, we consider

$$f_L = \sum_{\ell=1}^{L} \tilde{\pi}(\boldsymbol{\gamma}^{(\ell)} | \mathbf{y}) \mathbb{E}[f(\boldsymbol{\alpha}, \boldsymbol{\beta}) \mid \boldsymbol{\gamma}^{(\ell)}, \mathbf{y}],$$

where $\tilde{\pi}(\cdot|\mathbf{y})$ is the truncation of $\pi(\boldsymbol{\gamma}|\mathbf{y})$ to Γ_L. In contrast to the MAP plug-in, $\mathbb{E}[f(\boldsymbol{\alpha}, \boldsymbol{\beta}) \mid \mathbf{y}, \hat{\boldsymbol{\gamma}}_{MAP}]$, f_L averages over more of the particle selection uncertainty and we might reasonably expect it to be a better approximation of the marginal posterior mean $\mathbb{E}[f(\boldsymbol{\alpha}, \boldsymbol{\beta}) \mid \mathbf{y}]$. Of course, in order to compute f_L exactly, we must have the L $\boldsymbol{\gamma}$'s with the most posterior in hand. Since we do not know these parameters *a priori*, we have to first estimate them using the observed data. In the next section, we introduce a generic strategy for identifying Γ_L based on approximation $\pi(\boldsymbol{\gamma} \mid \mathbf{y})$ without stochastic search.

3.1 A Variational Approximation

Before proceeding, we introduce a bit more notation. For any collection of L particles $\Gamma = \{\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_L\}$ and vector $\mathbf{w}$ in the L-dimensional simplex, let $q(\cdot \mid \Gamma, \mathbf{w})$ be the discrete distribution that places probability w_ℓ on the particle $\boldsymbol{\gamma}_\ell$. Following Ročková (2018), we will refer to the collection Γ as a *particle set* and $\mathbf{w}$ as *importance weights*. Let $\mathcal{Q}_L$ be the collection of all such distributions supported on at most L particles. Finally, for each $\lambda > 0$, let Π_λ be the tempered marginal posterior with mass function $\pi_\lambda(\boldsymbol{\gamma}) \propto \pi(\boldsymbol{\gamma} \mid \mathbf{y})^{\frac{1}{\lambda}}$. Note that the particles in Γ_L, which are the L particles with largest posterior mass, are also the L particles with largest tempered posterior mass for all λ. The following proposition provides the foundation for identifying this collection.

Proposition 1. *Suppose that $\pi(\boldsymbol{\gamma} \mid \mathbf{y})$ is supported on at least L distinct particles and that $\pi_\lambda(\boldsymbol{\gamma}) \neq \pi_\lambda(\boldsymbol{\gamma}')$ for $\boldsymbol{\gamma} \neq \boldsymbol{\gamma}'$. Let $q^\star_\lambda(\cdot|\Gamma^\star(\lambda), \mathbf{w}^\star(\lambda))$ be the distribution in $\mathcal{Q}_L$ that is closest to Π_λ in a Kullback-Leibler sense:*

$$q^\star_\lambda = \arg\min_{q \in \mathcal{Q}_L} \left\{ \sum_{\boldsymbol{\gamma}} q(\boldsymbol{\gamma}) \log \frac{q(\boldsymbol{\gamma})}{\pi_\lambda(\boldsymbol{\gamma})} \right\}.$$

Then $\Gamma^\star(\lambda) = \Gamma_L$ and for each $\ell = 1, \ldots, L$, $w^\star_\ell(\lambda) \propto \pi(\boldsymbol{\gamma}^{(\ell)}|\mathbf{y})^{\frac{1}{\lambda}}$

Proof. Denote the optimal particles $\Gamma^\star(\lambda) = \{\boldsymbol{\gamma}^\star_1, \ldots, \boldsymbol{\gamma}^\star_{L^\star}\}$. Straightforward calculus verifies that $w^\star_\ell(\lambda) \propto \pi_\lambda(\boldsymbol{\gamma}^\star_\ell)$. We thus compute

$$\mathrm{KL}(q^\star \parallel \pi_\lambda) = \sum_{\boldsymbol{\gamma}} q^\star(\boldsymbol{\gamma}) \log \frac{q^\star(\boldsymbol{\gamma})}{\pi_\lambda(\lambda)} = -\log \Pi_\lambda(\Gamma^\star(\ell))$$

Since Π_λ is supported on at least L models, we see from this computation that if $\Gamma^\star$ contained fewer than L particles, we could achieve a lower Kullback-Leibler divergence by adding another particle $\tilde{\boldsymbol{\gamma}}$ not currently in $\Gamma^\star$ that has positive Π_λ-probability to the particle set and updating the importance weights $\mathbf{w}$ accordingly.

Now if $\Gamma^\star$ contains L models but $\Gamma^\star(\lambda) \neq \Gamma_L$, we know $\Pi_\lambda(\Gamma^\star(\lambda)) < \Pi_\lambda(\Gamma_L)$. Thus, replacing $\Gamma^\star(\lambda)$ by Γ_L and adjusting the importances weights accordingly would also result in a lower Kullback-Liebler divergence. □

In other words, we can find Γ_L by finding a variational approximation of Π_λ *for any* $\lambda > 0$.

This is equivalent to solving

$$(\Gamma^\star(\lambda), \mathbf{w}^\star(\lambda)) = \arg\max_{(\Gamma,\mathbf{w})} \left\{ \sum_{\ell=1}^{L} w_\ell \left(\log p(\mathbf{y} \mid \boldsymbol{\gamma}_\ell) + \log \pi(\boldsymbol{\gamma}_\ell) \right) + \lambda H(\Gamma, \mathbf{w}) \right\}, \tag{3}$$

where $H(\Gamma, \mathbf{w}) = -\mathbb{E}_{q(\cdot|\Gamma,\mathbf{w})}[\log q(\cdot|\Gamma, \mathbf{w})]$ is the entropy of the approximating distribution $q(\cdot|\Gamma, \mathbf{w})$. Before proceeding, we stress that we are not finding a variational approximation of $\pi(\boldsymbol{\alpha}, \boldsymbol{\beta}, \sigma^2 \mid \mathbf{y})$, the marginal posterior distribution of the continuous parameters of interest. Rather, we seek the variational approximation of a *discrete* distribution within a class of discrete distributions with substantially smaller support.

We pause briefly to reflect on the two terms in Equation (3). The first term is, up to an additive constant depending only on $\mathbf{y}$, the $\mathbf{w}$-weighted average of the height of the log-posterior at each particle in the particle set Γ. This term is clearly maximized when all of the particles in Γ are equal to the MAP. On the other hand, the entropy $H(\Gamma, \mathbf{w})$ of the approximating distribution is maximized when all of the particles in Γ are distinct and each $w_\ell = L^{-1}$. The penalty term λ, which we may also view as an inverse temperature, balances these two opposing forces.

3.2 Particle optimization

Finding the global optimum of (3) exactly is practically impossible, given the enormous size of the set of all possible particle sets Γ. Instead, we deploy a coordinate ascent strategy, iteratively updating one of $\mathbf{w}$ and Γ, until we reach a stationary point. From Proposition 1, given the current value of the Γ, we can update $\mathbf{w}$ in closed form. Recall that each particle in Γ consists of two partitions, one for the intercepts $\boldsymbol{\alpha}$ and one for the slopes $\boldsymbol{\beta}$. Given $\mathbf{w}$, we update the particle set by sweeping over the $2L$ partitions once, updating each one sequentially while keeping the others fixed. Since $\mathcal{SP}$, the set of all spatial partitions, is quite vast, carrying out these single partition updates in a globally optimal fashion is also not feasible. Instead, we use local, greedy updates, in which we generate a manageable set of transitions and select the optimal transition within this set.

Ročková (2018) introduced essentially the same family of optimization problems to identify sparse high-dimensional linear regression models and described a similar general coordinate ascent strategy that iteratively updated $\mathbf{w}$ and Γ. In that work, $\boldsymbol{\gamma}$ was a binary vector indicating which variables to include in the model and the continuous parameters conditional on

$\boldsymbol{\gamma}$ were modeled with continuous spike-and-slab priors in the style of George and McCulloch (1993). To update each individual $\boldsymbol{\gamma}_\ell \in \Gamma$, Ročková (2018) restricted attention only to binary vectors which differed in one coordinate. While it is tempting to update each partition in our setting similarly by re-allocating a single neighborhood to a new or existing cluster, such a strategy is prone to lead to local entrapment.

Indeed, such one-neighborhood updates directly parallel conventional Gibbs samplers for Dirichlet process mixture models (i.e. Algorithms 1 – 8 in Neal (2000)). It is well-known (Celeux et al., 2000) that these samplers can mix very slowly, as their incremental nature make it virtually impossible to pass through regions of low probability between partitions that have similar probability but differ in the cluster assignment of multiple neighborhoods. In our optimization setting, such a restrictive search strategy results in premature termination at a sub-optimal ensemble Γ. Instead, a much more promising strategy for navigating the space of partitions is to allow multiple elements to be re-allocated at once (Jain and Neal, 2004). To this end, we consider both "fine" transitions, which re-allocate a single neighborhood to a new or existing cluster thereby enabling the creation of "islands", and "coarse" transitions, which re-allocate multiple neighborhoods simultaneously; see Figure 2.

We have two types of coarse transitions. The first exchanges multiple neighborhoods simultaneously across a border between adjacent clusters, while the second splits an existing cluster into several sub-clusters and merges some or all of the newly created sub-clusters with other existing clusters. Sometimes, removing a single neighborhood from a cluster leaves the resulting cluster disconnected. When this happens, we treat the resulting connected components as individual clusters.

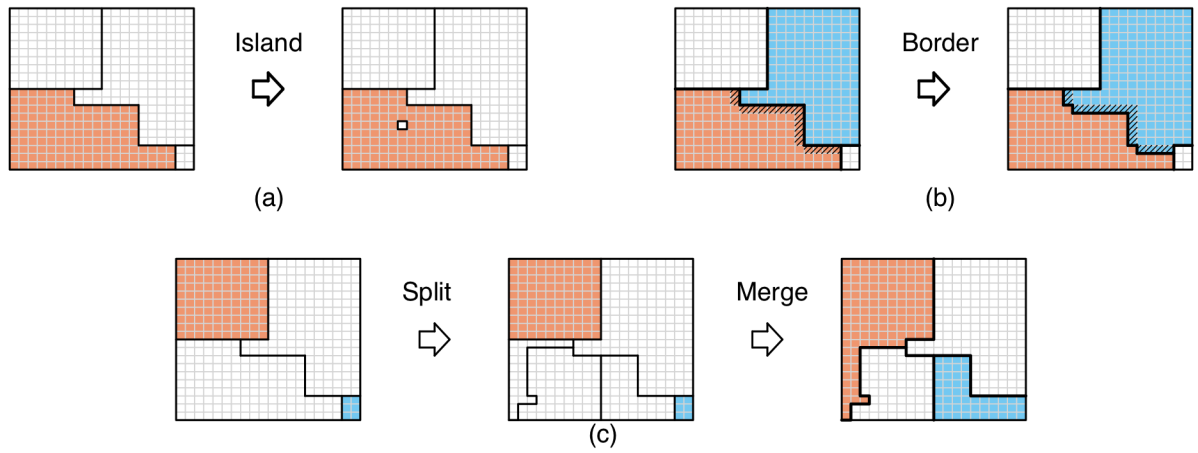

Figure 2: The three broad types of transitions that we consider. An "island" transition (a) removes a single neighborhood from an existing cluster (the lower left orange cluster) and creates a new singleton cluster. A "border" transition moves all neighborhoods at the interface of two adjacent clusters from one cluster to the other. In (b), the neighborhoods moved from the orange cluster to the blue cluster are shaded. The last type of transition (c) first splits an existing cluster (the left cluster in (c)) into multiple parts and then merges some or all of the new sub-clusters into already existing clusters.

In general, we do not attempt all possible coarse and fine transitions while updating a single partition. Indeed, there are $O(n)$ possible fine moves and if we allow each of K existing clusters to be split into up to K_{new} sub-clusters, there can be up to $O(K^2 + K \times K_{new}^K)$ possible coarse transitions. Rather than enumerating all of these transitions, we restrict our attention to a much smaller set using several heuristics outlined below. For brevity, we describe these heuristics only for transitions for the $\boldsymbol{\alpha}$-partition; we use exactly the same heuristics for the $\boldsymbol{\beta}$-partition.

The conditional conjugacy of our "CAR–within–cluster" model allows us to quickly compute $\mathbb{E}[\alpha_i \mid \boldsymbol{\gamma}, \mathbf{y}]$ and $\mathbb{E}[\overline{\alpha}_k \mid \gamma, \mathbf{y}]$. We use these conditional means as running estimates to propose transitions. For each cluster k, we can identify its nearest neighbor k', whose estimated grand cluster mean $\overline{\alpha}_{k'}$ is closest to the estimated grand cluster mean of cluster k, $\overline{\alpha}_k$. We then propose exchanging neighborhoods from k across the border between clusters k and k'. In this way, we only consider $O(K)$ coarse transitions of the first type. For coarse moves of the second type, which first split an existing cluster into many pieces, we cap the number of new sub-clusters at $K_{new} = 5$. To generate these sub-clusters, we run both k-means and spectral clustering on the running estimates of the α_i's within the cluster. We also propose

splits by removing the top or bottom 5% of these estimates.

Once we split a single cluster into many new sub-clusters, we can identify the nearest neighbor of each sub-cluster among the other existing clusters based on the estimated grand cluster means. We then propose a sequence of merges where a new sub-cluster is merged into its nearest neighbor only if all sub-clusters that are closer to their own nearest neighbors are also merged. For fine transitions, we initially only attempt to remove neighborhood i from its current cluster and move it to a new singleton if its estimated α_i is in the top or bottom 5% of the distribution of estimates within the cluster. Following these heuristics, we consider on the order of $N/10$ fine transitions and $O(K + K \times K_{new}^2)$ total coarse transitions while updating a single partition in our ensemble. During our coordinate ascent algorithm, if we find that none of these transitions are accepted, we then try all N fine moves. This last check ensures that our algorithm converges locally in the sense that no one-tract update to an individual partition will result in a higher objective. While these heuristics are somewhat arbitrary, we have found that they work quite well in practice.

4 Illustration

To illustrate the behavior of our proposed optimization procedure, we consider a simpler model of crime $y_{i,t} = \alpha_i + \sigma\varepsilon_{i,t}$ and we place a CAR–within–cluster prior over $\boldsymbol{\alpha}$. We simulate data on a 20×20 grid of census tracts partitioned into four clusters of sizes 12, 188, 100, and 100. Figure 3 show the four clusters in the true partition along with three of the different specifications of $\boldsymbol{\alpha}$.

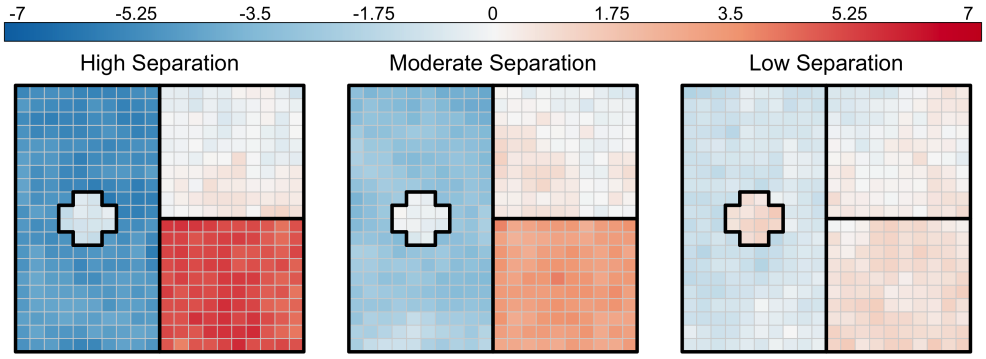

Figure 3: True data generating partition and three different settings for $\boldsymbol{\alpha}$. Going from left to right, the distances between the average of the α_i's within each cluster gets progressively smaller.

Figure 4 shows the top three partitions recovered when we run our procedure in each of the high, moderate, and low separation settings with two different entropy penalty parameters $\lambda = 1$ and $\lambda = 100$. We placed a truncated Ewens-Pitman prior on the latent partition so that the prior mass function on any partition $\gamma = \{S_1, \ldots, S_K\}$ is given by

$$\pi(\gamma) \propto \eta^K \prod_{k=1}^{K} (n_k - 1)! \times \mathbf{1}(\gamma \in \mathcal{SP}).$$

For this demonstration, we fixed $L = 10, \rho = 0.9$ and $\eta = 1$ and set the remaining hyper-parameters according to the heuristics detailed in Appendix A.

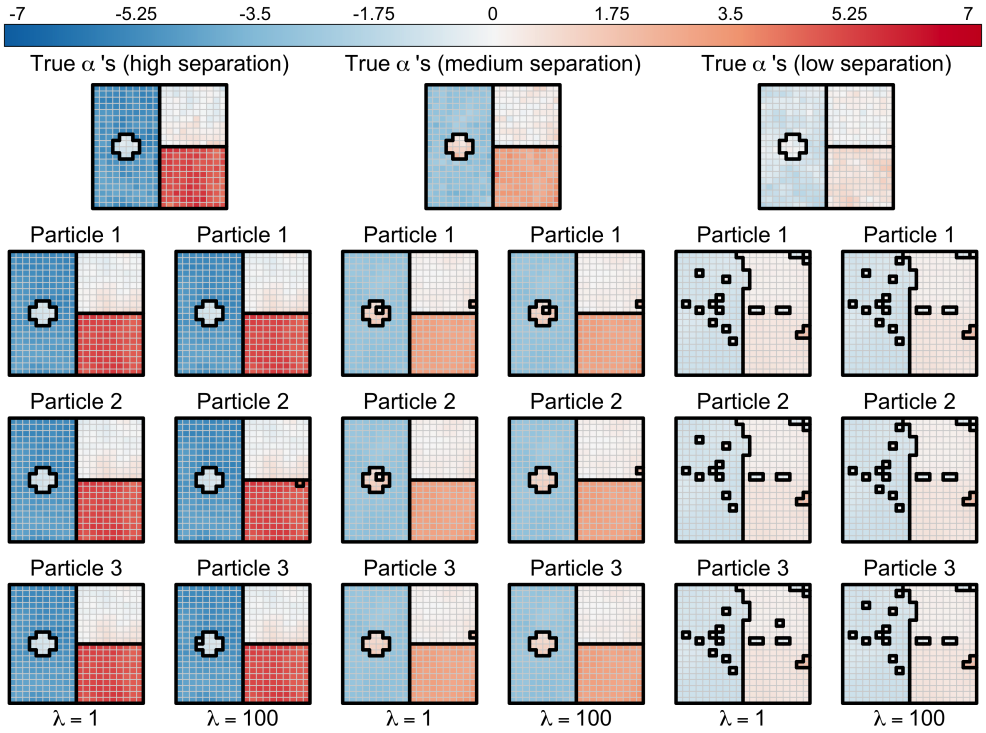

Figure 4: Top three partitions recovered by our particle optimization procedure across different specifications of $\boldsymbol{\alpha}$ and values of λ. The color of each square of the recovered particles corresponds to the value of the posterior mean $\mathbb{E}[\alpha_i \mid \mathbf{y}, \boldsymbol{\gamma}]$. Note, in the high separation setting with $\lambda = 1$, our final particle set contained 10 copies of the same partition.

It is reassuring to see that when the clusters are well-separated, our method identifies the true partition as the top particle for both values of λ and that when the clusters are only moderately separated, the top partitions identified are all quite close to the true partition that generated the data. On the other hand, when there is very little separation between the clusters, the partitions returned by our method are visually quite far from the truth. It turns out that these partitions had substantially more posterior probability than the true partition in this setting.

We know from Proposition 1 that the globally optimal particle set $\Gamma_L^\star$ must (i) contain exactly L particles and (ii) be identical for all values of λ. We see in Figure 4 that in each of the three settings, the top particles identified for $\lambda = 1$ and $\lambda = 100$ are different. In fact, in the high separation setting, all of the particles in our particle set collapsed to the true partition when $\lambda = 1$. Additionally, in the medium separation setting, the second partition identified when $\lambda = 1$ is not contained in the particle set obtained when $\lambda = 100$, despite having more

posterior probability than all but the top partition in the latter particle set. This behavior, which is at odds with what is expected from Proposition 1, highlights the local nature of our optimization algorithm.

Recall that the entropy term in Equation (3) attempts to offset any potential decrease in posterior probability that accompanies a transition away from a high probability particle set already present in the ensemble to a new particle. The fact that the particle set identified in the high separation setting with $\lambda = 1$ displays extreme redundancy – all of the particles collapsed to the same partition – suggests that by itself, this entropy term is not always sufficient to identify L distinct partitions.

This, in and of itself, is not altogether surprising: being bounded from above by $\log L$, the changes in entropy encountered by our algorithm are typically orders of magnitude smaller than changes in the $\mathbf{w}$-weighted (unnormalized) log-posterior. As we increase λ from 1 to 100, however, we find that our procedure recovers $L = 10$ distinct models. That being said, in all three settings, we find that some of the particles identified with one choice of λ may not be identified with the other choice of λ, despite having higher posterior probability than many of the particles found with the latter λ. This is again a direct artifact of the local, non-reversible, transitions that we consider. Typically, with larger values of λ, particles are encouraged to drift to regions of lower posterior probability more forcefully than with lower values of λ. Moreover, once in those regions, it is typically quite difficult for a particle to "double back" and return to a previously visited state with more posterior probability.

To assess the estimation and partition selection performance of our proposed method quantitatively, we computed the root mean square error (RMSE) of the proposed BMA estimator and the Rand index (Rand, 1971) between the top partition recovered and the true partition averaged over 20 simulated datasets for different choices of cluster separation. The Rand index is defined as the proportion of pairs of elements that are clustered together in both partitions, with values close to one indicating a high degree of similarity between the partitions. Figure 5 shows the average estimation and selection performance for our method run with $\lambda = 1$ along with the following four competitors: (i) the "1-Cluster" model that places all tracts into a single cluster, (ii) the "N-Clusters" model that places all tracts into singleton clusters, (iii) running k-means on the collection of MLE's $\hat{\alpha}_i = \overline{y}_{i,\cdot}$, and (iv) running spectral clustering on the tract averages. When running k-means and spectral clustering, we varied the number of clusters from one to ten. For k-means, we selected the number of clusters using the popular "elbow method" **?** and for spectral clustering, we found the number of

clusters which minimized the total within-cluster sum of squares. We then computed the conditional posterior expectation $\mathbb{E}[\boldsymbol{\alpha} \mid \mathbf{y}, \hat{\boldsymbol{\gamma}}]$ based on the partition $\hat{\boldsymbol{\gamma}}$ estimated from each of the k-means and spectral clustering procedures. Across our simulations, the estimation and partition selection performance of our method with $\lambda = 100$ is virtually identical to the performance with $\lambda = 1$.

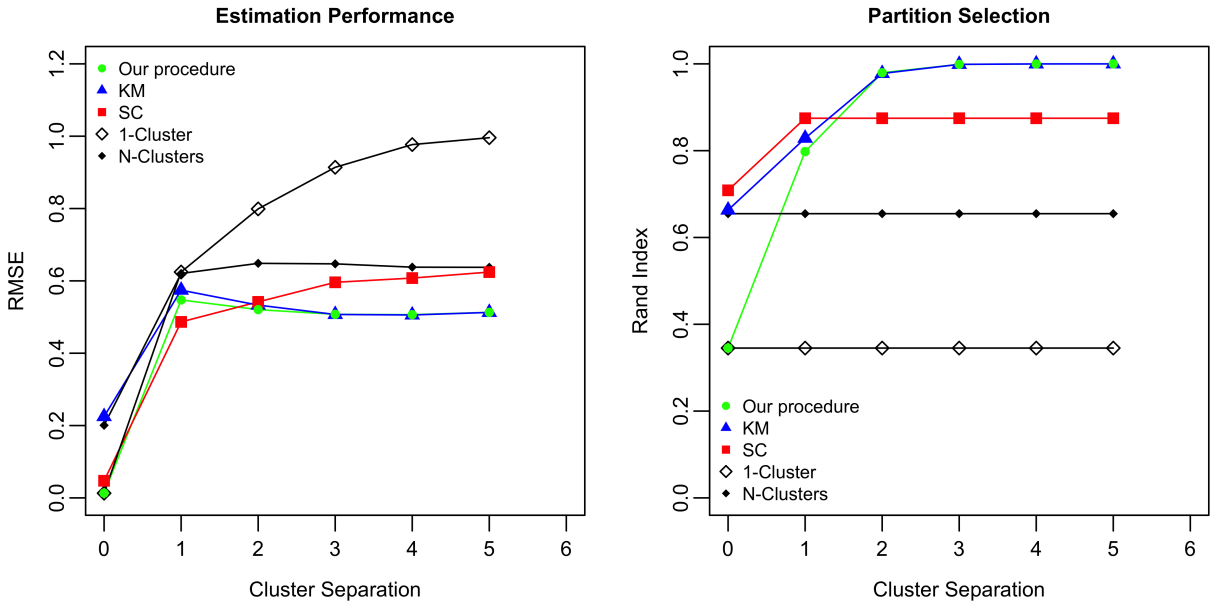

Figure 5: The estimation and partition selection performance, averaged over 20 Monte Carlo simulations, of our method run with $\lambda = 1$ and several competitors across a range of cluster separations.

Immediately we see that, in terms of estimation performance, our procedure is very similar to k-means for non-zero cluster separations. In a certain sense, this behavior is entirely expected when the clusters separation is high: the partition found by k-means in these settings was usually identical to or very close to the true partition, resulting in Rand indices very close to one. However, when the cluster separation is low, our proposed procedure, which identifies several high posterior probability partitions and averages over them, performs much better than k-means, which attempts only to identify a single partition with no reference to the posterior of interest. When there is in fact no separation between the cluster means, the top partition identified by our procedure was always equal to the partition that placed all tracts in a single cluster. In other words, when there truly was no difference between the cluster means, not only does the trivial "1-Cluster" partition have substantially higher posterior probability than other partitions but our particle optimization strategy is also able

to recover this partition reliably. This adaptation, in turn, results in excellent estimation performance in this setting.

Interestingly, our method outperforms spectral clustering, in terms of RMSE, except in one setting where the separation between clusters was low but non-zero. In fact, this was the same low separation setting from Figures 3 and 4. As seen in Figure 4, the partitions identified by our method are all quite different than the true partition. It turns out that in this setting, the partition identified by spectral clustering divided the tracts into four equally sized 10×10 grids; see Figure 9 in Appendix B. This partition is substantially closer to the true partition and it is therefore perhaps not surprising that spectral clustering achieved slightly better RMSE in this setting.

In Figure 5, we showed the RMSE for the full BMA estimator that averaged over all of the particles recovered by our method. Especially when the separation between clusters was very large, often the top partition identified had orders of magnitude more posterior probability than the other partitions identified. This raises a natural question: could we achieve somewhat better estimation performance by averaging over only a subset of the partitions identified by our method instead of averaging over all of them? In our experiments, we found that it was usually better to average over multiple partitions instead of focusing on the MAP plug-in. However, the RMSE was not monotonic in the number of particles averaged over. This is largely a by-product of simulation variability: a partition identified as the second-best particle in one simulation replication might be the third-best in another replication. We also found that the change in RMSE as we varied the number of particles averaged over was quite small, typically of order 10^{-4} or less.

5 Clustering Crime Dynamics in Philadelphia

As described in Section 2, we want to model the transformed number of violent crimes $y_{i,t}$ in neighborhood i at time t and study its change over time with a linear trend: $y_{i,t} = \alpha_i + \beta_i(t - \bar{t}) + \varepsilon_{i,t}$. Moreover we want cluster regions that display similar behavior in the mean level of crime α_i and in the trend over time β_i. For this reason, we consider the two latent partitions $\gamma^{(\alpha)}, \gamma^{(\beta)}$, and we consider a CAR-within-cluster prior for $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$; see Section 2.1 for details on the model.

For the analysis of crime in Philadelphia we consider two priors on the partitions $\gamma^{(\alpha)}, \gamma^{(\beta)}$:

the Ewens-Pitman with hyper-parameter $\eta = 5$ and the uniform distribution. Both of these priors were truncated to the set of spatially connected partitions $\mathcal{SP}$. The hyper-parameters a_1, b_1, a_2, b_2 were selected with the heuristic described in Appendix A.

To initialize our procedure we consider all pairs of partitions $(\gamma_i^{(\alpha)}, \gamma_j^{(\beta)})$ found by applying the k-means algorithm on the MLE for α_i's and β_i's, with k respectively equal to i and j, for all values of $i, j = 1, \ldots, \lfloor \log(N) \rfloor$. We then start our algorithm by randomly drawing L of these pairs (with replacement) where the probability of drawing any pair is proportional to its posterior probability. This initialization allows the algorithm to pursue several search directions simultaneously but also allows for some redundancy in initial particle set. In regions of high posterior probability, this redundancy allows multiple particles to search around a dominant mode, providing a measure of local uncertainty. For our analysis we set $L = 10$ and $\lambda = 100$ and display the top three particles found with the Ewens-Pitman prior in top panel Figure 6 and the top three particles found with the uniform prior in the bottom panel of Figure 6.

In each panel of Figure 6 we represent the top three particles as colored maps: the thick lines depict the borders between clusters, and the colors illustrate the conditional posterior mean of the α_i's (resp. β_i's) given the value of the particle. Between these colored maps we plot black and white maps that represent the differences between the identified partitions. Specifically, these "difference" plots highlight neighborhoods which are clustered differently in different particles, by shading such neighborhoods in grey. When the two partitions are equal, no greyed areas are shown.

Figure 6 reveals several key aspects: firstly, the difference in the particles is often found in only one of the two partitions. For example, when the uniform prior is used, there is only variation in $\gamma^{(\beta)}$ among the best three particles; when the Ewens-Pitman prior is used instead, there is only variation in $\gamma^{(\beta)}$ between the first and third particle, but both $\gamma^{(\alpha)}$ and $\gamma^{(\beta)}$ vary between the first and second particle. This difference shows that it is not an artifact of our algorithm, but a property of the posterior distribution.

A second aspect that can be derived from Figure 6 is that while often the difference between partitions consist of only one or two neighborhoods, the method also recovers larger differences: for example in the bottom panel of Figure 6, $\gamma_{(2)}^{(\beta)}$ differs from $\gamma_{(1)}^{(\beta)}$ for a large collection of census tracts in Northeast Philadelphia: the partition $\gamma_{(1)}^{(\beta)}$ estimates the trend in time to be strongly decreasing, while the $\gamma_{(2)}^{(\beta)}$ allows for more nuances, with some neighborhoods displaying mildly decreasing or almost increasing trends. Another example can

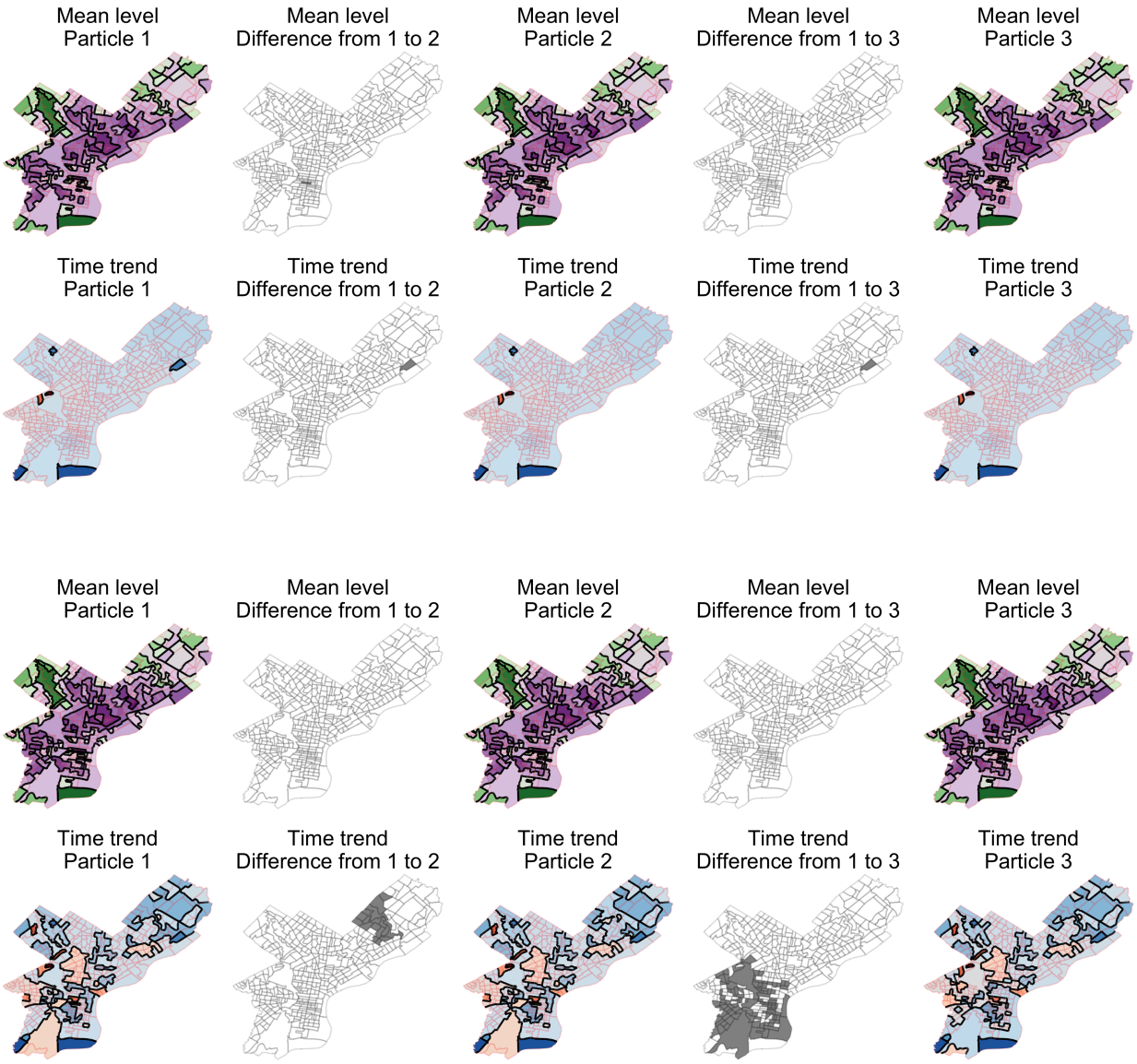

Figure 6: Colored plots: Top three models identified by our procedure. The thick borders represent the partition, and the color represents the posterior mean of the parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$. Black and white plots: transition from the model on the left to the model on the right. The greyed areas represent the neighborhoods whose cluster assignments change in the partitions on the sides. **Top:** Ewens-Pitman prior with $\eta = 5$. **Bottom:** Uniform prior on $\mathcal{SP}$.

be found in the difference between $\gamma^{(\beta)}_{(3)}$ and $\gamma^{(\beta)}_{(1)}$, where many neighborhoods in South and West Philadelphia are estimated to have quite a different trend over time: for example, the tract corresponding to the South part of the Schuylkill river is estimated to have a mildly increasing trend in $\boldsymbol{\gamma}_{(1)}$ and a decreasing trend in $\boldsymbol{\gamma}_{(3)}$. Smaller differences can be found in the top panel of Figure 6, where the particles recovered under the Ewens-Pitman prior are reported: the difference between $\gamma^{(\beta)}_{(1)}$ and $\gamma^{(\beta)}_{(2)}$ (and similarly, between $\gamma^{(\beta)}_{(2)}$ and $\gamma^{(\beta)}_{(3)}$) is only in one census tract in Northeast Philadelphia that is moved from a cluster with strongly increasing time trend to one with a mildly increasing trend.

Figure 6 also shows that the partitions recovered are sensitive to the prior choice: when the uniform prior on $\mathcal{SP}$ is used, many more clusters are found, compared to when the Ewens-Pitman prior is used. For the partition of the time trend $\boldsymbol{\gamma}^{(\beta)}$, the clusters found under the Ewens-Pitman correspond to areas of high or medium increase (two census tracts in West Philadelphia), of high decrease (four census tracts in South and North Philadelphia) and of mild decrease (the remaining of the city). Under the Uniform prior instead the clusters represent more nuanced behaviors: a small number of tracts characterized by high decrease, the majority of the city with moderate and mild decreases, some areas of mild and moderate and high increase. While the uniform prior seems to recover more intuitive results for the partition of $\boldsymbol{\beta}$, arguably too many clusters are found for the partitions of the mean level $\boldsymbol{\gamma}^{(\alpha)}$, with no substantial difference in the estimate of α_i between the different clusters. The results for $\gamma^{(\alpha)}$ recovered under the Ewens-Pitman are instead preferable for the increased interpretability.

To display a summary of the different models found together with the best partition we combine the difference plots between $\boldsymbol{\gamma}_{(1)}$ and all the remaining $\boldsymbol{\gamma}_{(i)}$ for $i = 2, \ldots, L$ into one plot. In the left panel of Figure 7 the models from the Ewens-Pitman prior are displayed, with the different regions of uncertainty highlighted in the top models compared to the MAP; the partition depicted with thicker border is the MAP, identified by $\gamma^{(\alpha)}_{(1)}$. The differences recovered consist only of a small number of census tracts. On the right panel of Figure 7 the same map is reported for the partition of the time trend $\boldsymbol{\beta}$ when the uniform prior on $\mathcal{SP}$ is used: we notice here that more differences are recovered compared to the ones displayed between the top three particles in Figure 6, covering most of the neighborhoods in South, West and North Philadelphia.

It is not surprising to notice that the partitions under the Ewens-Pitman prior vary to a lesser degree than the ones under the Uniform prior. In the former, the prior plays an

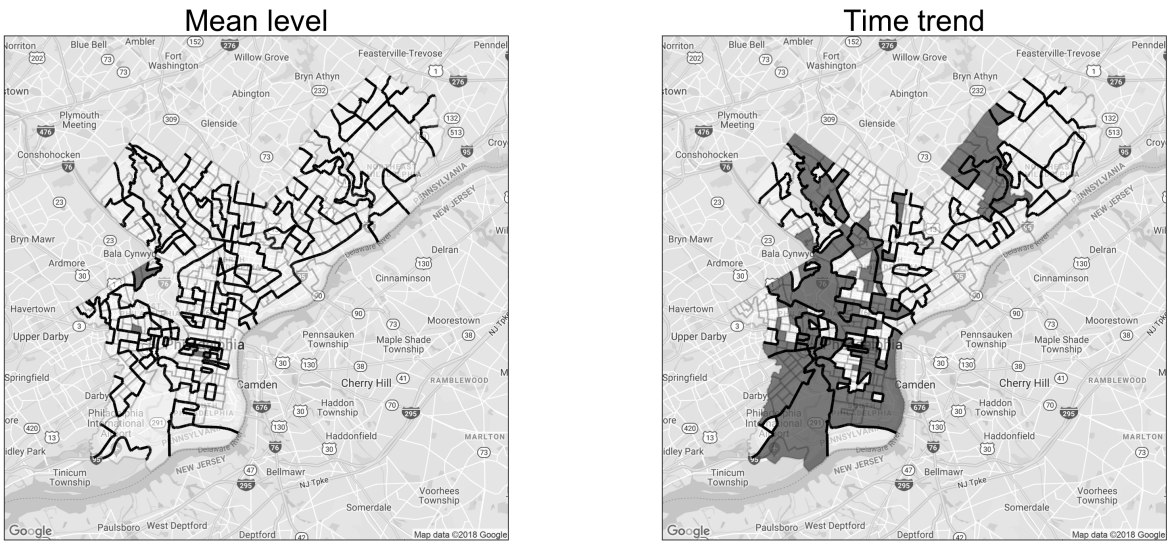

Figure 7: Partitions recovered in the top model (thick black lines) and differences with the remaining nine top models (grayed areas). Left panel: results for $\gamma^{(\alpha)}$ when using the Ewens-Pitman prior. Right panel: results for $\gamma^{(\beta)}$ when using the uniform prior on $\mathcal{SP}$.

important role in the posterior and only small changes in the partition are recovered because they don't decrease the prior excessively. Under the Uniform prior instead changes in the partitions only affect the posterior through the likelihood, so the model can recover very different partitions with similarly large posterior.

Besides analyzing the partitions recovered, we consider the predictive accuracy of this model, under the different priors for partitions, the Ewens-Pitman and the Uniform over $\mathcal{SP}$. Additionally, we consider two "hybrid" prior distributions: we consider the model where the prior for $\gamma^{(\alpha)}$ is the Ewens-Pitman and the prior for $\gamma^{(\beta)}$ is the Uniform, and we consider the opposite, where the prior for $\gamma^{(\alpha)}$ is the Uniform and the prior for $\gamma^{(\beta)}$ is the Ewens-Pitman. The top particles recovered under these two models are reported in Figure 8 and the full representation of the top three particles is given in Appendix C.

In Table 1 we report the out-of-sample prediction RMSE: using the data from 2006 to 2017 we predict the levels of crime in each neighborhood for 2018. For each of the models considered, we report the error for the predictions recovered both using the estimate of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ from the top particle ("Top Particle" column) and averaging the estimates from all the particles in a Bayesian Model Averaging fashion ("BMA" column). We compare our models with the performance of the model where no shrinkage or clustering are used, where the estimates for the coefficients are independently recovered with maximum likelihood estimation.

	Top particle	BMA
MLE	0.2340	-
EP-EP prior	0.2568	0.2560
Unif-Unif prior	0.2327	0.2325
EP-Unif prior	0.2339	0.2319
Unif-EP prior	0.2546	0.2539

Table 1: Out-of-sample prediction errors using different combinations of priors for the partitions $\gamma^{(\alpha)}$ and $\gamma^{(\beta)}$.

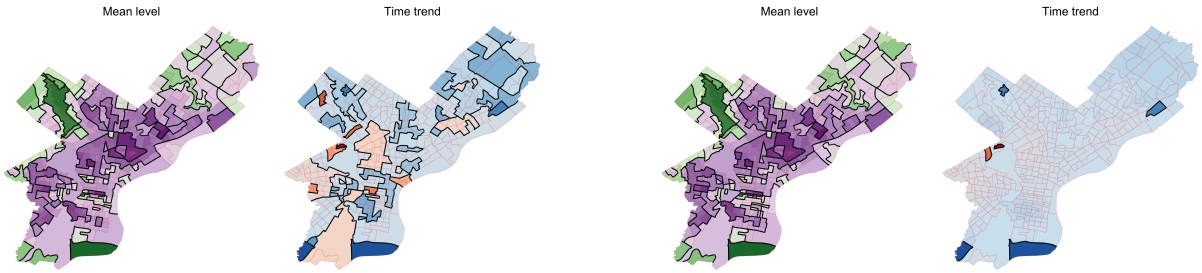

Figure 8: Partitions recovered in the top model (thick black lines) and estimated coefficients (colors). Left panel: results recovered under the EP prior on $\gamma^{(\alpha)}$ and uniform prior on $\gamma^{(\beta)}$. Right panel: results recovered under the uniform prior on $\gamma^{(\alpha)}$ and EP prior on $\gamma^{(\beta)}$.

As Table 1 shows, the best predictive performance is recovered under the models where a Uniform prior is specified on $\gamma^{(\beta)}$. The explanation is quite intuitive and comes from the fact that under the Ewens-Pitman prior a lot of neighborhood are inaccurately estimated to have a decreasing trend, and this affects the predictions in a negative way.

6 Discussion

Accurate estimation of the change in crime over time is a critical first step towards better understanding of public safety in large urban environments. An especially important challenge to such estimation is the potential presence of sharp discontinuities, which may be smoothed over by naive spatial shrinkage procedures. Focusing on the city of Philadelphia, we introduced a Bayesian hierachical model that naturally identifies these discontinuities by partitioning the city into several clusters of neighborhoods and introduces spatial smoothness within but not between clusters. In particular, we focused on recovering two latent spatial partitions, one for the mean-level of crime over the twelve year period 2006 – 2017 and one for the time-trend.

Rather than use computationally prohibitive stochastic search, we instead sought to identify the pairs of partitions with highest posterior probability simultaneously by solving a single optimization problem. We showed that optimizing the proposed objective function is formally equivalent to finding a particular variational objective and introduced a local search strategy for solving this problem. While our primary focus has been on crime in the city of Philadelphia, our ensemble optimization framework is decidedly more general and there are a number of areas of future development, which we discuss below.

The results of our applied analysis were quite sensitive to the choice of prior placed on the underlying spatial partition. To wit, when we deployed Ewens-Pitman priors, the recovered partitions of time trend clustered nearly all of the neighborhoods into a single cluster. However, with a uniform prior on $\gamma^{(\beta)}$, we obtained a much richer cluster structure. It would be interesting to constructing an objective prior for spatial partitions along the lines of Casella et al. (2004).

In our application to Philadelphia, we found that γ_β, the partition based on the time-trends, contained very clusters, especially when we placed a truncated Ewens-Pitman prior on this partition. Looking back at Figure 1, this is not especially surprising – most of the census tracts had very similar maximum likelihood estimates of β_i. This raises an interesting question, however: in order to explain the variation in crime changes at the census tract level adequately, do we need a distinct time-trend β_i for each census tract? To explore this possibility, we can replace the CAR–within–cluster model on $\boldsymbol{\beta} \mid \boldsymbol{\gamma}$ with a "constant–within–clusters" model in which a single time-trend is shared across all of the tracts within each cluster. Incorporating this modification is relatively straightforward within our framework, especially if we place a conditionally conjugate normal prior on the cluster-specific time trend. Although this modification makes the computation of the marginal data likelihood $p(\mathbf{y} \mid \boldsymbol{\gamma})$ somewhat more involved and requiress some slight modifications of the heuristics used to split clusters, the basic search strategy remains the same.

Similarly, while it may be sufficient to consider a linear temporal model of crime when there are relatively few time points (Bernardinelli et al., 1995; Anderson et al., 2017), with more observations per census tract, it is reasonable to consider more flexible models. For instance, we could model $y_{i,t} \sim N(f_i(\mathbf{x}_{i,t}, \sigma^2)$ and place Gaussian process priors over the f_i's within each cluster. Such an elaboration retains conditional conjugacy and we can still use our ensemble optimization strategy to identify clusters with high posterior probability, though computing the marginal likelihood $p(\mathbf{y} \mid \boldsymbol{\gamma})$ is somewhat more involved. It is more difficult

to deploy our ensemble optimization strategy directly to non-conjugate models for which the marginal likelihood $p(\mathbf{y} \mid \boldsymbol{\gamma})$ is difficult to compute in closed-form. While it may sometimes be possible to use an EM-algorithm like Ročková (2018), this is not always be possible for more complicated models. One very natural idea would be to estimate the marginal likelihood with a Laplace approximation.

In nearly all of our empirical examples, we have found that once the particle set navigates near a dominant mode, it tends to remain in the vicinity. This is especially pronounced when we used a Ewens-Pitman prior – all of the identified partitions were rather similar to one another. In fact, we found that once the particle set was near a dominant posterior mode, nearly all of the coarse transitions that moved multiple tracts had substantially less posterior mass than the fine transitions that moved a single tract. This behavior is largely an artifact of the Ewens-Pitman prior, which tends to discourage splitting clusters into two non-singleton sub-clusters and leaving them un-merged with other existing clusters. Put another way, once all of the particles are near a dominant mode, proposed transitions that split a cluster into two non-singleton clusters were almost never accepted as any increase in the log-likelihood was drowned out by the decrease in log-prior. Fine transitions, on the other hand, were overwhelmingly accepted once the ensemble was in the vicinity of the dominant mode, as both the log-prior and log-likelihood are relatively insensitive to single-index updates. As a result, once the ensemble navigated near a dominant mode, it tends to remain in the vicinity so that all of the models are close to one another in some metric. In such a case, it is not immediately obvious whether the posterior truly concentrates around a single dominant mode or if there are other pockets of substantial posterior mass that are far away.

Unfortunately, the entropy term in Equation 3 may provide insufficient repulsion between the particles to probe this latter possibility. Operationally, it discourages redundancy in the ensemble by penalizing exact equality between particles but does not penalize placing a particlel in the vicinity of another model that is already present in the particle set. One way around this potential weakness is to augment the optimization objective in (3) with an additional penalty term that directly penalizes the pairwise distance between particles in the particle set. In doing so, however, we lose the guarantee of optimality afforded by Proposition 1.

References

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

Balocchi, C. and Jensen, S. T. (2019). Spatial modeling of trends in crime over time in Philadelphia. arXiv:1901.08117.

Banerjee, S., Carlin, B. P., Li, P., and McBean, A. M. (2012). Bayesian areal wombling using false discovery rates. *Statistics and its Interface*, 5(2):149–158.

Bernardinelli, L., Clayton, D., Pascutto, C., Montomoli, C., Ghislandi, M., and Songini, M. (1995). Bayesian analysis of space—time variation in disease risk. *Statistics in medicine*, 14(21-22):2433–2443.

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

Boots, B. (2001). Using local statistics for boundary characterization. *GeoJournal*, 53(4):339–345.

Burbidge, J. B., Magee, L., and Robb, A. L. (1988). Alternative transformations to handle extreme values of the dependent variable. *Journal of the American Statistical Association*, 83(401):123–127.

Casella, G., Moreno, E., and Girón, F. J. (2004). Cluster analysis, model selection, and prior distributions on models. *Bayesian Analysis*, 9(3):613 – 685.

Celeux, G., Hurn, M., and Robert, C. P. (2000). Computational and inferential difficulties with mixture posterior distributions. *Journal of the American Statistical Association*, 95(451):957 – 970.

Dahl, D. B. (2009). Modal clustering in a class of product partition models. *Bayesian Analysis*, 4(2):243 – 264.

Denison, D. and Holmes, C. (2001). Bayesian partitioning for estimating disease risk. *Biometrics*, 57(1):143–149.

FBI (2011). Uniform Crime Reporting program, definitions. https://ucr.fbi.gov/crime-in-the-u.s/2011/crime-in-the-u.s.-2011/offense-definitions. Accessed: 2016-09-15.

Feng, W., Lim, C. Y., Maiti, T., and Zhang, Z. (2016). Spatial regression and estimation of disease risks: A clustering-based approach. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 9(6):417–434.

George, E. I. and McCulloch, R. E. (1993). Variable selection via Gibbs sampling. *Journal of the American Statistical Association*, 88(423):881 – 889.

Hoeting, J. A., Madigan, D., Raferty, A. E., and Volinsky, C. T. (1999). Bayesian model averaging: a tutorial. *Statistical Science*, 14(4):382 – 417.

Jain, S. and Neal, R. (2004). A split-merge Markov chain Monte Carlo procedure for the Dirichlet Process mixture model. *Journal of Computational and Graphical Statistics*, 13(1):158 – 182.

Knorr-Held, L. and Raßer, G. (2000). Bayesian detection of clusters and discontinuities in disease maps. *Biometrics*, 56(1):13–21.

Lau, J. W. and Green, P. J. (2007). Bayesian model-based clustering procedures. *Journal of Computational and Graphical Statistics*, 16(3):526–558.

Lee, D. and Mitchell, R. (2012). Boundary detection in disease mapping studies. *Biostatistics*, 13(3):415–426.

Lee, D. and Mitchell, R. (2013). Locally adaptive spatial smoothing using conditional auto-regressive models. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 62(4):593–608.

Leroux, B. G., Lei, X., and Breslow, N. (2000). Estimation of disease rates in small areas: a new mixed model for spatial dependence. In *Statistical models in epidemiology, the environment, and clinical trials*, pages 179–191. Springer.

Li, P., Banerjee, S., and McBean, A. M. (2011). Mining boundary effects in areally referenced spatial data using the bayesian information criterion. *Geoinformatica*, 15(3):435–454.

Lu, H. and Carlin, B. P. (2005). Bayesian areal wombling for geographical boundary analysis. *Geographical Analysis*, 37(3):265–285.

Lu, H., Reilly, C. S., Banerjee, S., and Carlin, B. P. (2007). Bayesian areal wombling via adjacency modeling. *Environmental and Ecological Statistics*, 14(4):433–452.

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

Raftery, A. E., Madigan, D., and Hoeting, J. A. (1997). Bayesian model averaging for linear regression models. *Journal of the American Statistical Association*, 92(437):179 – 191.

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

Ročková, V. (2018). Particle EM for variable selection. *Journal of the American Statistical Association*, 113(524):1684 – 1697.

A Hyper-parameter choice

The main model described in Section 2 depends on several hyper-parameters, which need to be fixed by the practitioner: the parameters for the prior for σ (ν_σ and λ_σ) and the multiplicative constants to specify within and between cluster variance (a_1, a_2, b_1 and b_2). We will now describe the heuristic used to specify such values.

Let us consider each neighborhood separately and fit a simple linear regression model in each one: let $\hat{\alpha}_i$ and $\hat{\beta}_i$ be the least square estimates and $\hat{\sigma}_i^2$ be the estimated residual variance for neighborhood i. Since these estimates do not incorporate any prior information or sharing of information, we can think of them as an approximation of α_i, β_i given the partition with N clusters γ_N; in fact under such configuration the coefficients are exchangeable and the only shrinkage induced is through the common variance parameter. Given this, one heuristic desideratum is that the marginal prior on $\boldsymbol{\alpha} \mid \gamma = \gamma_N$ should assign substantial probability to range of the $\hat{\alpha}_i$. Specifically, we will make sure that this conditional prior places 95% of its probability over the range of the $\hat{\alpha}_i$'s. Since $\boldsymbol{\alpha} \mid \gamma = \gamma_N \sim N(0, \sigma^2(a_1/(1-\rho) + a_2)I_n)$, we constrain a_1 and a_2 so that

$$\frac{a_1}{1-\rho} + a_2 = \frac{\max_i |\hat{\alpha}_i|^2}{4\hat{\sigma^2}}.$$

In order to determine each of a_1 and a_2, we need a second constraint. To this end, consider the highly stylized setting in which we have K over-lapping clusters with equal variance σ_{cl}^2 whose means are equally spaced at distance $2\sigma_{cl}$. The idea of this second heuristic is to match such a stylized description to the observe distribution of $\hat{\alpha}_i$. In essence, this involves covering the range of $\hat{\alpha}_i$ with $K+1$ "chunks" of length $2\sigma_{\text{cl}}$. While the exact value of σ_{cl} is unknown, we have found it useful to approximate it $a_1\sigma^2/(1-\rho)$. This approximation tends to produce smaller values of a_1, which in turn encourages a relatively small number of clusters.

With these two constraints we find:

$$a_1 = \frac{(\max(\hat{\alpha}_i) - \min(\hat{\alpha}_i))^2}{4(K+1)^2\hat{\sigma}^2/(1-\rho)}$$
$$a_2 = \frac{\max_i |\hat{\alpha}_i|^2}{4\hat{\sigma^2}} - \frac{a_1}{1-\rho}.$$

Similarly for the $\hat{\beta}_i$'s we find:

$$b_1 = \frac{(\max(\hat{\beta}_i) - \min(\hat{\beta}_i))^2}{4(K+1)^2\hat{\sigma}^2/(1-\rho)}$$
$$b_2 = \frac{\max_i |\hat{\beta}_i|^2}{4\hat{\sigma^2}} - \frac{b_1}{1-\rho}.$$

In order to operationalize these heuristics, we must specify an initial guess at K. We have found in our experiments, setting $K = \lfloor \log N \rfloor$ works quite well. It, moreover, accords with the general behavior of the Ewens-Pitman prior.

Finally, to specify the prior for σ^2 we can use the collection of $\hat{\sigma}_i^2$'s: by matching mean and variance, we can recover $\nu_\sigma = 2\frac{m^2}{v} + 4$ and $\lambda_\sigma = m(1 - \frac{2}{\nu_\sigma})$, where m and v are the empirical mean and variance of the $\hat{\sigma}_i^2$'s.

B Additional Simulation Results

In Section 4, we generated several synthetic datasets based on a 20 grid of census tracts partitioned into four clusters of size 12, 188, 100, and 100, as seen in Figure 3. Within each cluster, we drew the α_i's from a CAR model centered at a specified cluster mean with $\rho = 0.95$ and variance scale 0.2. Across the different specifications of cluster means, we always fixed the cluster mean of the 12-tract "cross" and the 100 tract square in the upper right corner to be zero. We then fixed the mean of the 188-tract cluster on the left hand side to be $-\Delta$ and the mean of the 100-tract cluster in the lower right corner to be Δ. We generated datasets for each of $\Delta = 0, 1, \ldots, 5$. The high, medium, and low separation settings in Figure 3 and 4 correspond to $\Delta = 5, 3$, and 1, respectively.

In Section 4, we compared the partition selection performance of our method to that of k-means and spectral clustering. Figure 9 shows the estimated partitions from k-means and spectral clustering on the same dataset used to generate Figure 4. Across these datasets, the optimal number of clusters for k-means was always three, according to the "elbow method." However, because k-means does not implicitly account for our spatial connectedness constraints, we post-processed the recovered partition by treating disconnected parts of clusters identified by k-means as their own separate clusters.

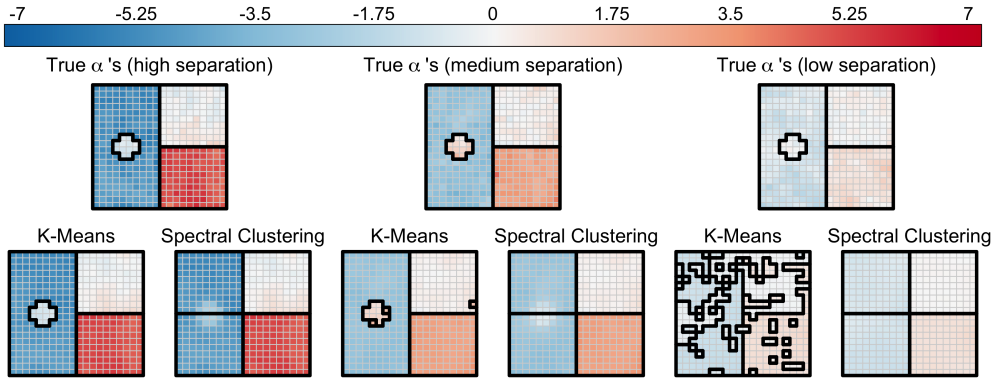

Figure 9: Partitions recovered by k-means and spectral clustering for three different cluster separation settings. The color of each tract corresponds to the estimated parameter value $\mathbb{E}[\alpha_i \mid \mathbf{y}, \boldsymbol{\gamma}]$.

C Additional Results for Clustering in Philadelphia

In figure 10 we represent the best three particles recovered by the models where the priors are specified as Ewens-Pitman prior with $\eta = 5$ for $\gamma^{(\alpha)}$ and Uniform on $\mathcal{SP}$ for $\gamma^{(\beta)}$ (top panel) and Uniform prior on $\mathcal{SP}$ for $\gamma^{(\beta)}$ and Ewens-Pitman prior with $\eta = 5$ for $\gamma^{(\alpha)}$ (bottom panel).

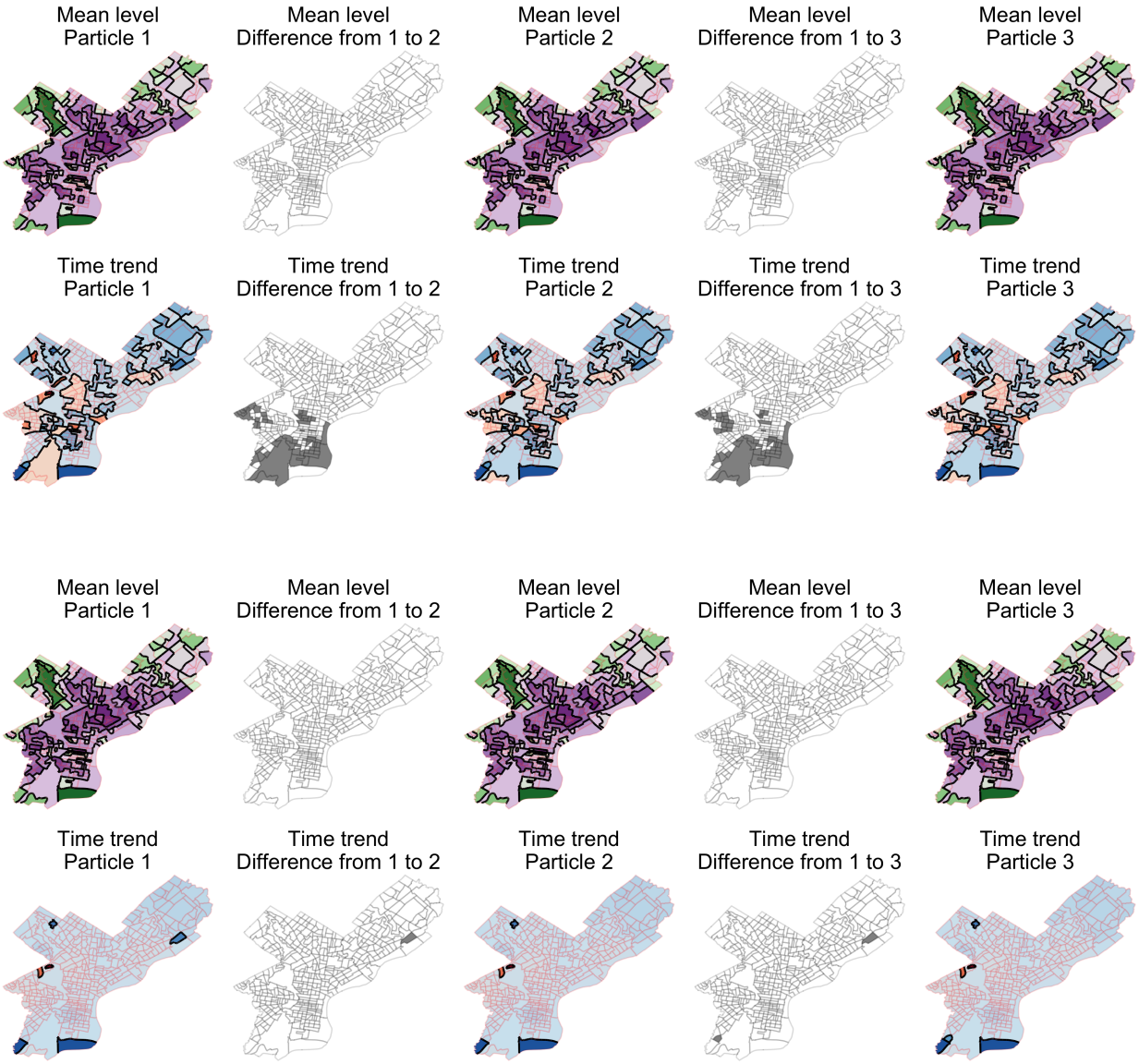

Figure 10: Colored plots: Top three models identified by our procedure. The thick borders represent the partition, and the color represents the posterior mean of the parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$. Black and white plots: transition from the model on the left to the model on the right. The greyed areas represent the neighborhoods whose cluster assignments change in the partitions on the sides. **Top:** Ewens-Pitman prior with $\eta = 5$ for $\gamma^{(\alpha)}$ and Uniform on $\mathcal{SP}$ for $\gamma^{(\beta)}$. **Bottom:** Uniform prior on $\mathcal{SP}$ for $\gamma^{(\beta)}$ and Ewens-Pitman prior with $\eta = 5$ for $\gamma^{(\alpha)}$.

D Derivation of Closed Form Expressions

D.1 Two Partition Derivations

Recall from Section 2.1 that our full mode is:

$$\begin{aligned}
\gamma^{(\alpha)}, \gamma^{(\beta)} &\sim \mathrm{EP}(\eta; \mathcal{SP}) \\
\sigma^2 &\sim \mathrm{IG}\left(\frac{\nu_\sigma}{2}, \frac{\nu_\sigma \lambda_\sigma}{2}\right) \\
(\overline{\alpha}_k)_k &\overset{iid}{\sim} N(0, a_2\sigma^2) \\
(\overline{\beta}_{k'})_{k'} &\overset{iid}{\sim} N(0, b_2\sigma^2) \\
(\boldsymbol{\alpha}_k)_k &\overset{ind}{\sim} \mathrm{CAR}(\overline{\alpha}_k, a_1\sigma^2, W_k^{(\alpha)}) \\
(\boldsymbol{\beta}_{k'})_{k'} &\overset{ind}{\sim} \mathrm{CAR}(\overline{\beta}_{k'}, b_1\sigma^2, W_{k'}^{(\beta)}) \\
(y_{i,t})_{i,t} &\overset{ind}{\sim} \mathrm{N}(\alpha_i + \beta_i(t - \bar{t}), \sigma^2)
\end{aligned}$$

We exploit the conditional conjugacy present in this model in several places. First, we have closed form expressions for the conditional posterior means $\mathbb{E}[\boldsymbol{\alpha} \mid \mathbf{y}, \boldsymbol{\gamma}]$ and $\mathbb{E}[\boldsymbol{\beta} \mid \mathbf{y}, \boldsymbol{\gamma}]$, which we use in our particle optimization procedure to propose new transitions. Second, we can compute the marginal likelihood $p(\mathbf{y} \mid \boldsymbol{\gamma})$ in closed form, which we use to evaluate the optimization objective and pick between multiple transitions. Below, we carefully derive these closed form expressions, noting that in several places, we can avoid potentially expensive matrix inversions. In particular, the choice to center the time variable, thereby ensuring an orthogonal design matrix within each neighborhood, facilitates rapid likelihood evaluations.

Distribution of $\boldsymbol{\alpha}_k$ Let us first consider the vector of parameters $\boldsymbol{\alpha}_k$ in cluster $S_k^{(\alpha)}$ given σ^2: by marginalizing the distribution of the grand cluster mean $\overline{\alpha}_k$, we find that its distribution is a multivariate normal with covariance matrix $\sigma^2\Sigma_k^{(\alpha)}$, where $\Sigma_k^{(\alpha)} = a_1\Sigma_{k,\mathrm{CAR}}^{(\alpha)} + a_2\mathbf{1}\mathbf{1}^\top = a_1\left[\rho(W_k^{(\alpha)})^* + (1-\rho)\mathbf{I}\right]^{-1} + a_2\mathbf{1}\mathbf{1}^\top$. Note that its precision matrix can be computed

using Woodbury's formula without having to invert any matrix:

$$(\Sigma_k^{(\alpha)})^{-1} = a_1^{-1}\Omega_{k,\text{CAR}}^{(\alpha)} - a_1^{-1}\Omega_{k,\text{CAR}}^{(\alpha)}\mathbf{1}\left(a_1^{-1}\mathbf{1}^\top\Omega_{k,\text{CAR}}^{(\alpha)}\mathbf{1} + a_2^{-1}\right)^{-1}\mathbf{1}^\top a_1^{-1}\Omega_{k,\text{CAR}}^{(\alpha)} =$$
$$= a_1^{-1}\Omega_{k,\text{CAR}}^{(\alpha)} - \frac{a_1^{-2}(1-\rho)^2}{a_1^{-1}n_k(1-\rho)+a_2^{-1}}\mathbf{1}\mathbf{1}^\top$$

where $\Omega_{k,\text{CAR}}^{(\alpha)} = \left(\Sigma_{k,\text{CAR}}^{(\alpha)}\right)^{-1} = \rho(W_k^{(\alpha)})^* + (1-\rho)\mathbf{I}$; the second line follows from noticing that $\mathbf{1}$ is both a left and right eigenvector of $\Omega_{k,\text{CAR}}^{(\alpha)}$ with eigenvalue $1-\rho$. Similarly this holds for the distribution of $\boldsymbol{\beta}_{k'}$.

Distribution of $\boldsymbol{\alpha}$ Next, we can write the distribution of the whole vector $\boldsymbol{\alpha}$ given σ^2 and $\gamma^{(\alpha)}$: by combining the distributions of the cluster specific parameters $\boldsymbol{\alpha}_k$'s, and using the independence between different clusters, we find that the distrubution of $\boldsymbol{\alpha}$ given σ^2 and $\gamma^{(\alpha)}$ is a multivariate normal with mean zero and covariance matrix that can be found by combining the $\Sigma_k^{(\alpha)}$'s. Because of the independence between clusters, *there exists an ordering of the indices of* $\boldsymbol{\alpha}$ so that the covariance matrix of $\boldsymbol{\alpha}|\gamma_\alpha, \sigma^2$ has a block-diagonal structure. We denote such permutation of the indices with $\pi^{(\alpha)}$, and it can be constructed by mapping the first n_1 elements to the indices in the first cluster ($\{\pi^{(\alpha)}(1), \ldots, \pi^{(\alpha)}(n_1)\} = S_1^{(\alpha)}$), the following n_2 elements to the indices in the second cluster ($\{\pi^{(\alpha)}(n_1+1), \ldots, \pi^{(\alpha)}(n_1+n_2)\} = S_2^{(\alpha)}$), and so on. With such ordering, the kth diagonal block of the covariance matrix is $\sigma^2\Sigma_k^{(\alpha)}$. Similarly, we can find a (potentially different) permutation $\pi^{(\beta)}$ for $\boldsymbol{\beta}$ and derive the distribution of $\boldsymbol{\beta}_\pi|\sigma^2, \gamma^{(\beta)}$.

Notation To describe the distributions of interest we can represent our model in the form of a unique linear model, by combining all the observations in a vector Y, combining the reodered coefficients in a unique vector $\boldsymbol{\theta} = (\boldsymbol{\alpha}_\pi, \boldsymbol{\beta}_\pi)$ and appropriately constructing the covariate matrix X. In the next paragraphs we will provide with the details on how we constructed such vectors and matrix.

To build the column vector Y we stack the vectors $\mathbf{y}_i$ with $i = 1, \ldots, N$: Y is a vector of length $N \cdot T$ and each block of T rows corresponds to a particular neighborhood; in particular, the $((i-1)T+t)$th entry of Y corresponds to $y_{i,t}$.

The vector of coefficients $\boldsymbol{\theta}$ is found by concatenating the reordered $\boldsymbol{\alpha}_\pi$ and $\boldsymbol{\beta}_\pi$: for $i = 1, \ldots, N$, elements $\theta_i = \alpha_{\pi^{(\alpha)}(i)}$ and $\theta_{N+i} = \beta_{\pi^{(\beta)}(i)}$.

The matrix of covariates X then has dimensions $NT \times 2N$; each block of T rows corresponds to a neighborhood and each column corresponds to an element of $\boldsymbol{\theta}$: the first N columns correspond to the elements of $\boldsymbol{\alpha}_\pi$ and the second N columns to $\boldsymbol{\beta}_\pi$. The rows of X corresponding to neighborhood i (rows $(i-1)T+t$ with $t = 1, \dots T$) have an element equal to 1 in the $(\pi^{(\alpha)})^{-1}(i)$th column, an element equal to $x_{it} = t - \bar{t}$ in the $(N + (\pi^{(\beta)})^{-1}(i))$th column, and zero elsewhere. With such construction, the $(i-1)T+t$ row of the equation $Y = X\boldsymbol{\theta}$ corresponds to $y_{i,t} = \theta_{(\pi^{(\alpha)})^{-1}(i)} + x_{it}\theta_{N+(\pi^{(\beta)})^{-1}(i)} = \alpha_i + (t-\bar{t})\beta_i$.

Marginal likelihood $Y|\gamma^{(\alpha)}, \gamma^{(\beta)}$ To recover the marginal likelihood $p(Y|\gamma^{(\alpha)}, \gamma^{(\beta)})$ we compute

$$\begin{aligned}
&\int \left[\int p(Y|\boldsymbol{\alpha}, \boldsymbol{\beta}, \sigma^2) p(\boldsymbol{\alpha}|\gamma^{(\alpha)}, \sigma^2) p(\boldsymbol{\beta}|\gamma^{(\beta)}, \sigma^2) d\boldsymbol{\alpha} d\boldsymbol{\beta} \right] p(\sigma^2) d\sigma^2 = \\
&= \int \left[\int p(Y|\boldsymbol{\alpha}_\pi, \boldsymbol{\beta}_\pi, \sigma^2) p(\boldsymbol{\alpha}_\pi|\gamma^{(\alpha)}, \sigma^2) p(\boldsymbol{\beta}_\pi|\gamma^{(\beta)}, \sigma^2) d\boldsymbol{\alpha}_\pi d\boldsymbol{\beta}_\pi \right] p(\sigma^2) d\sigma^2 = \\
&= \int \left[\int p(Y|\boldsymbol{\theta}, \sigma^2) p(\boldsymbol{\theta}|\gamma^{(\alpha)}, \gamma^{(\beta)}, \sigma^2) d\boldsymbol{\theta} \right] p(\sigma^2) d\sigma^2.
\end{aligned}$$

Let us first compute $p(Y|\sigma^2, \gamma^{(\alpha)}, \gamma^{(\beta)}) = \int p(Y|\boldsymbol{\theta}, \sigma^2) p(\boldsymbol{\theta}|\gamma^{(\alpha)}, \gamma^{(\beta)}, \sigma^2) d\boldsymbol{\theta}$. Using the notation for linear regression we can write $p(Y|\boldsymbol{\theta}, \sigma^2) = N(X\boldsymbol{\theta}, \sigma^2 \mathbf{I})$. The prior for $\boldsymbol{\theta}$ is a normal distribution with mean zero and block covariance matrix Σ_θ: the first $n \times n$ block corresponds to the covariance matrix of $\boldsymbol{\alpha}$ and the second to the one for $\boldsymbol{\beta}$.

By integrating out $\boldsymbol{\theta}$, $p(Y|\gamma^{(\alpha)}, \gamma^{(\beta)}, \sigma^2) = N(\mathbf{0}, \sigma^2 \Sigma_Y)$ where $\Sigma_Y = \mathbf{I} + X\Sigma_\theta X^\top$. Its precision matrix can be computed using Woodbury's formula again: $\Sigma_Y^{-1} = \mathbf{I} - X(\Sigma_\theta^{-1} + X^\top X)^{-1} X^\top$. Note that $X^\top X$ is a diagonal matrix, and we derive its form at the end of this chapter.

The marginal likelihood can now be derived by integrating out σ^2:

$$\begin{aligned}
p(Y|\gamma^{(\alpha)}, \gamma^{(\beta)}) &= \int p(Y|\sigma^2, \gamma^{(\alpha)}, \gamma^{(\beta)}) p(\sigma^2) d\sigma^2 = \\
&= \pi^{-nT/2} \det(\Sigma_Y)^{-1/2} \frac{(\nu_\sigma \lambda_\sigma / 2)^{\nu_\sigma/2}}{\Gamma(\frac{\nu_\sigma}{2})} \int (\sigma^2)^{-\frac{NT+\nu_\sigma}{2}-1} e^{-\frac{Y^\top \Sigma_Y^{-1} Y + \nu_\sigma \lambda_\sigma}{2\sigma^2}} d\sigma^2 = \\
&= \pi^{-nT/2} \det(\Sigma_Y)^{-1/2} \frac{\Gamma(\frac{NT+\nu_\sigma}{2})}{\Gamma(\frac{\nu_\sigma}{2})} \left(\frac{\nu_\sigma \lambda_\sigma}{2} \right)^{\nu_\sigma/2} \left(\frac{\nu_\sigma \lambda_\sigma + Y^\top \Sigma_Y^{-1} Y}{2} \right)^{-(NT+\nu_\sigma)/2} = \\
&= \pi^{-nT/2} \det(\Sigma_Y)^{-1/2} \frac{\Gamma(\frac{NT+\nu_\sigma}{2})}{\Gamma(\frac{\nu_\sigma}{2})} \left(\frac{\nu_\sigma \lambda_\sigma}{2} \right)^{-NT/2} \left(1 + \frac{Y^\top \Sigma_Y^{-1} Y}{\nu_\sigma \lambda_\sigma} \right)^{-(NT+\nu_\sigma)/2}.
\end{aligned}$$

Note that if $\lambda_\sigma = 1$, this is multivariate t-distribution with ν_σ degrees of freedom.

For this we need to compute the quadratic form

$$Y^\top \Sigma_Y^{-1} Y = Y^\top Y - Y^\top X(\Sigma_\theta^{-1} + X^\top X)^{-1} X^\top Y.$$

Because of the block diagonal structure of $\Sigma_\theta^{-1} + X^\top X$ we can write this as a sum over the clusters of the two partitions. Consider the column vector $X^\top Y$ of length $2N$: the first N elements correspond to the summary statistics related to the $\alpha_{\pi(i)}$'s and we will denote the ones corresponding to cluster $S_k^{(\alpha)}$ with $(X^\top Y)_k^{(\alpha)}$, while the second N elements are for the β_i's and we denote with $(X^\top Y)_{k'}^{(\beta)}$ the ones for cluster $S_{k'}^{(\beta)}$. Now we can write

$$\begin{aligned} Y^\top X(\Sigma_\theta^{-1} + X^\top X)^{-1} X^\top Y = &\sum_{k=1}^{K^{(\alpha)}} (X^\top Y)_k^{(\alpha)\top} ((\Sigma_k^{(\alpha)})^{-1} + T\mathbf{I})^{-1} (X^\top Y)_k^{(\alpha)} \\ &+ \sum_{k'=1}^{K^{(\beta)}} (X^\top Y)_{k'}^{(\beta)\top} ((\Sigma_{k'}^{(\beta)})^{-1} + \textstyle\sum x_t^2 \mathbf{I})^{-1} (X^\top Y)_{k'}^{(\beta)} \end{aligned}$$

where $(\Sigma_k^{(\alpha)})^{-1} + T\mathbf{I}$ is the diagonal blocks of $\Sigma_\theta^{-1} + X^\top X$ corresponding to cluster $S_k^{(\alpha)}$ and $(\Sigma_{k'}^{(\beta)})^{-1} + \sum x_t^2 \mathbf{I}$ corresponds to $S_{k'}^{(\beta)}$; each of them can be inverted using methods for symmetric positive definite matrices.

To compute the marginal likelihood we are left we calculating the determinant of Σ_Y, where we can use the reciprocal of the determinant of its inverse

$$\det(\Sigma_Y^{-1}) = \det(\mathbf{I} - X(\Sigma_\theta^{-1} + X^\top X)^{-1} X^\top) = \det(\mathbf{I} - (\Sigma_\theta^{-1} + X^\top X)^{-1} X^\top X)$$

where the last equality is given by Sylvester's formula, and allows us to compute the determinant of a smaller dimensional matrix. Moreover, because of its block diagonal structure, we can compute the determinant block-wise.

Posterior mean of $\boldsymbol{\alpha}, \boldsymbol{\beta}$ The calculations for the posterior mean of $\boldsymbol{\alpha}, \boldsymbol{\beta}$ are very similar: using the same notation and the results for linear regression, we can find

$$\mathbb{E}\left[\boldsymbol{\theta} | Y, \gamma^{(\alpha)}, \gamma^{(\beta)}, \sigma^{-1}\right] = \left(X^\top X + \Sigma_\theta^{-1}\right)^{-1} X^\top Y$$

and since this does not depend on σ^2, it coincides with $\mathbb{E}\left[\boldsymbol{\theta}|Y, \gamma^{(\alpha)}, \gamma^{(\beta)}\right]$. Because of the block diagonal structure of the matrices involved, we can compute the estimate of the parameter for each cluster independently. Moreover, note that the inverse of $X^\top X + \Sigma_\theta^{-1}$ is computed in the likelihood calculation, so it can be stored and does not need to be computed two times.

Derivation of $X^\top X$ Since in our formulation the covariates are orthogonal, i.e. $\sum_{t=1}^{T} x_{it} = 0$ for all i, $X^\top X$ is a diagonal matrix. Note that column $X_{(\pi^{(\alpha)})^{-1}(i')}$ contains T 1's in rows $t + (i'-1)\times$T and zeros elsewhere; similarly column $X_{N+(\pi^{(\beta)})^{-1}(i')}$ contains elements $(x_{i't})$ in rows $t + (i'-1) \times T$ and zero's elsewhere. Thus, when we compute $(X^\top X)_{ij}$ we consider the cross product of columns X_i and X_j. Depending on the value of i and j, we have the following cases:

- if $i = j \leq N$, then $(X^\top X)_{ij} = T$,
- if $i = j \geq N$, then $(X^\top X)_{ij} = \sum_t x^2_{\pi^{(\beta)}(j-N),t}$,
- if $i \leq N$ and $j = N + i$, then $(X^\top X)_{ij} = \sum_t x_{\pi^{(\beta)}(i),t} = 0$,
- if $j \leq N$ and $i = N + j$, then $(X^\top X)_{ij} = \sum_t x_{\pi^{(\beta)}(j),t} = 0$,
- for any other i, j, $(X^\top X)_{ij} = 0$.

Thus the matrix $X^\top X$ is a diagonal matrix: the first $n \times n$ diagonal block is $T\mathbf{I}$, and the second diagonal block is a diagonal matrix whose entries are $\sum_{t=1}^{T} x_{it}^2$; when we have fixed design, $x_{it} = x_t = t - \bar{t}$, then $\sum_{t=1}^{T} x_{it}^2 = \sum_{t=1}^{T} (t - \bar{t})^2$ is constant, so the second diagonal block is $\sum x_{it}^2 \mathbf{I}$. Because of the orthogonality of the covariates, the upper-right and lower-left blocks are zero matrices, since $\sum_{t=1}^{T} x_{it} = 0$.

Note on cluster-wise update of calculations. In our greedy search when we perform a move only one or two clusters in only one partition is changed: in a *split* move for $\gamma^{(\cdot)}$, a cluster is divided into two sub-clusters, and the original cluster replaced by the first, while the second creates an additional cluster; in a *merge* move, one of two clusters is deleted and the other is replaced to the merge of the two original clusters. In each case, we need to update the value of the marginal likelihood, of the prior for $\gamma^{(\cdot)}$ and of the estimate of the parameters.

Because of the block structure given by orthogonality of covariates and by the reordering

of the parameters, changing the structure of some clusters does not affect the parameter estimates for other clusters that are not involved in the move. This implies that updates for updates to $S_k^{(\alpha)}$ do not affect the parameter estimates $\boldsymbol{\alpha}_h$ for $h \neq k$ or $\boldsymbol{\beta}_{k'}$ for any k'. Similarly, since the quadratic form $Y^\top \Sigma_Y^{-1} Y$ can be written as sum of cluster-specific quadratic forms, we can update only the quadratic form of the clusters affected and we can compute the determinant of the blocks of Σ_Y corresponding to the modified clusters.

This allows us to invert matrices that scale like the size of the clusters, reducing the computational costs dramatically.

D.2 One Partition Derivations

In Section 4, we considered a simpler model, in which we ignored the time trend and only focused on clustering the intercepts. That model was:

$$\begin{aligned}
\boldsymbol{\gamma} = \{S_1, \ldots, S_K\} &\sim \mathcal{P}_{\boldsymbol{\gamma}} \\
\sigma^2 &\sim \text{Inv. Gamma}\left(\frac{\nu_\sigma}{2}, \frac{\nu_\sigma \lambda_\sigma}{2}\right) \\
\overline{\alpha}_k | \sigma^2 &\sim N(0, a_2 \sigma^2) \;\; \text{for each } k = 1, \ldots, K \\
\boldsymbol{\alpha}_{S_k} | \overline{\alpha}_k, \sigma^2 &\sim N_{n_k}(\overline{\alpha} \mathbf{1}_{n_k}, a_1 \sigma^2 \Sigma_k^{(\alpha)}) \;\; \text{for each } k = 1, \ldots, K \\
y_{i,t} | \alpha_i, \sigma^2 &\sim N(\alpha_i, \sigma^2) \;\; \text{for each } i = 1, \ldots, N, \text{ and } t = 1, \ldots, T
\end{aligned}$$

For the sake of completeness, we derive the corresponding marginal likelihood $p(\mathbf{y} \mid \boldsymbol{\gamma})$ and conditional expectation $\mathbb{E}[\boldsymbol{\alpha} \mid \boldsymbol{\gamma}, \mathbf{y}]$ for this simpler setting.

Now observe

$$\begin{aligned}
p(\mathbf{y} | \boldsymbol{\alpha}, \sigma^2, \boldsymbol{\gamma}) &\propto \prod_{k=1}^{K} \prod_{i \in S_k} \left(\sigma^2\right)^{-\frac{T}{2}} \exp\left\{ -\frac{T(\overline{y}_i - \alpha_i)^2 + (T-1)s_i^2}{2\sigma^2} \right\} \\
&\propto \left(\sigma^2\right)^{-\frac{N(T-1)}{2}} \exp\left\{ -\frac{(T-1)\sum_{i=1}^{N} s_i^2}{2\sigma^2} \right\} \prod_{k=1}^{K} p(\overline{\mathbf{y}}_{S_k} | \boldsymbol{\alpha}_{S_k}, \sigma^2, \boldsymbol{\gamma})
\end{aligned}$$

where $\overline{\mathbf{y}}_{S_k}|\boldsymbol{\alpha}_{S_k}, \sigma^2, \boldsymbol{\gamma} \sim N_{n_k}(\boldsymbol{\alpha}_{S_k}, T^{-1}\sigma^2 I_{n_k})$. From here, we conclude

$$p(\overline{y}|\sigma^2, \boldsymbol{\gamma}) \propto \left(\sigma^2\right)^{-\frac{N(T-1)}{2}} \exp\left\{-\frac{(T-1)\sum_{i=1}^N s_i^2}{2\sigma^2}\right\} \prod_{k=1}^K p(\overline{\mathbf{y}}_{S_k}|\sigma^2, \boldsymbol{\gamma})$$

To derive $p(\overline{\mathbf{y}}_{S_k}|\sigma^2, \boldsymbol{\gamma})$, we first note that marginally

$$\boldsymbol{\alpha}_{S_k}|\sigma^2 \sim N_{n_k}(0 \cdot \mathbf{1}_{n_k}, \sigma^2[a_1\Sigma_k^{(\alpha)} + a_2\mathbf{1}_{n_k}\mathbf{1}_{n_k}^\top]).$$

Now marginalizing out $\boldsymbol{\alpha}_{S_k}$ we have

$$\overline{\mathbf{y}}_{S_k}|\sigma^2, \boldsymbol{\gamma} \sim N_{n_k}\left(0\mathbf{1}_{n_k}, \sigma^2\left[a_1\Sigma_k^{(\alpha)} + a_2\mathbf{1}_{n_k}\mathbf{1}_{n_k}^\top + T^{-1}I_{n_k}\right]\right)$$

Hence

$$\begin{aligned} p(\mathbf{y}|\sigma^2, \boldsymbol{\gamma}) \propto \left(\sigma^2\right)^{-\frac{N(T-1)}{2}} \exp\left\{-\frac{(T-1)\sum_{i=1}^N s_i^2}{2\sigma^2}\right\} \\ \times \prod_{k=1}^K \left(\sigma^2\right)^{-\frac{n_k}{2}} |\Omega_k^{(y)}|^{\frac{1}{2}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{k=1}^K \overline{\mathbf{y}}_k^\top \Omega_k^{(y)} \overline{\mathbf{y}}_k\right\} \end{aligned}$$

where $\Omega_k^{(y)} = [a_1\Sigma_k^{(\alpha)} + a_2\mathbf{1}_{n_k}\mathbf{1}_{n_k}^\top + T^{-1}I_{n_k}]^{-1}$.

Marginalizing out σ^2, we conclude

$$p(\overline{\mathbf{y}}|\boldsymbol{\gamma}) = C(N, \nu_\sigma, \lambda_\sigma) \times \left(\prod_{k=1}^K |\Omega_k^{(y)}|\right)^{\frac{1}{2}} \times \left[\frac{\nu_\sigma\lambda_\sigma}{2} + \frac{1}{2}\sum_{k=1}^K \overline{\mathbf{y}}_k^\top \Omega_k^{(\alpha)} \overline{\mathbf{y}}_k + \frac{(T-1)}{2}\sum_{i=1}^N s_i^2\right]^{-\frac{\nu_\sigma + NT}{2}}$$

We further compute

$$p(\overline{y}_{S_k}, \boldsymbol{\alpha}_{S_k}|\sigma^2, \boldsymbol{\gamma}) \propto \exp\left\{-\frac{1}{2\sigma^2}\left[\boldsymbol{\alpha}_{S_k}^\top V^{-1}\boldsymbol{\alpha}_{S_k} - 2\boldsymbol{\alpha}_{S_k}^\top T\overline{\mathbf{y}}_{S_k}\right]\right\},$$

where $V^{-1} = \left[TI_{n_k} + \left(a_1\Sigma_k^{(\alpha)} + a_2\mathbf{1}_{n_k}\mathbf{1}_{n_k}^\top\right)^{-1}\right]$. From here, we immediate conclude that

$$\mathbb{E}[\boldsymbol{\alpha}_{S_k}|\overline{\mathbf{y}}_{S_k}, \boldsymbol{\gamma}] = T \times V\overline{\mathbf{y}}_{S_k}.$$

Finally, note that

$$\begin{aligned} p(\overline{\alpha}_k, \boldsymbol{\alpha}_{S_k}, \overline{\mathbf{y}}_{S_k} | \sigma^2, \boldsymbol{\gamma}) \propto & \exp\left\{ -\frac{1}{2\sigma^2} \left[(\overline{\mathbf{y}} - \boldsymbol{\alpha}_{S_k})^\top T \, (\overline{\mathbf{y}} - \boldsymbol{\alpha}_{S_k})^\top \right] \right\} \\ & \times \exp\left\{ -\frac{1}{2\sigma^2} \left[(\boldsymbol{\alpha}_{S_k} - \overline{\alpha}_k \mathbf{1}_{n_k})^\top a_1^{-1} \Omega_k^{(\alpha)} \, (\boldsymbol{\alpha}_{S_k} - \overline{\alpha}_k \mathbf{1}_{n_k}) \right] \right\} \\ & \times \exp\left\{ -\frac{1}{2\sigma^2} \overline{\alpha}_k^2 a_2^{-1} \right\} \end{aligned}$$

Therefore,

$$p(\overline{\alpha}_k | \boldsymbol{\alpha}_{S_k}, \overline{\mathbf{y}}, \sigma^2, \boldsymbol{\gamma}) \propto \exp\left\{ -\frac{1}{2\sigma^2} \left[\overline{\alpha}_k^2 \left(a_2^{-1} + a_1^{-1} \mathbf{1}_{n_k}^\top \Omega_k^{(\alpha)} \mathbf{1}_{n_k} \right) - 2\overline{\alpha}_k a_1^{-1} \mathbf{1}_{n_k}^\top \Omega_k^{(\alpha)} \boldsymbol{\alpha}_{S_k} \right] \right\}$$

By the Woodbury identity, we compute

$$\begin{aligned} \left[a_1 \Sigma_k^{(\alpha)} + a_2 \mathbf{1}_{n_k} \mathbf{1}_{n_k}^\top \right]^{-1} &= a_1^{-1} \Omega_k^{(\alpha)} - a_1^{-1} \Omega_k^{(\alpha)} \mathbf{1}_{n_k} \left[a_2^{-1} + a_1^{-1} \mathbf{1}_{n_k}^\top \Omega_k^{(\alpha)} \mathbf{1}_{n_k} \right]^{-1} \mathbf{1}_{n_k}^\top \Omega_k^{(\alpha)} a_1^{-1} \\ &= a_1^{-1} \Omega_k^{(\alpha)} - a_1^{-2} (1-\rho)^2 \times \left[a_2^{-1} + a_1^{-1} (1-\rho) n_k \right]^{-1} \times \mathbf{1}_{n_k} \mathbf{1}_{n_k}^\top \end{aligned}$$

So the posterior conditional mean of $\overline{\alpha}_k$ is given by

$$\mathbb{E}[\overline{\alpha}_k | \boldsymbol{\alpha}_{S_k}, \mathbf{y}_{S_k}, \boldsymbol{\gamma}] = \frac{a_1^{-1} \mathbf{1}^\top \Omega_k^{(\alpha)} \boldsymbol{\alpha}_{S_k}}{a_2^{-1} + a_1^{-1} \mathbf{1}_{n_k}^\top \Omega_k^{(\alpha)} \mathbf{1}_{n_k}^\top} = \frac{a_1^{-1} (1-\rho) \mathbf{1}_{n_k}^\top \boldsymbol{\alpha}_{S_k}}{a_2^{-1} + a_1^{-1} n_k (1-\rho)}$$

Note: observe that as $a_2 \to \infty$ (i.e. as we allow the variability of the cluster means to increase), this conditional expectation converges to the $n_k^{-1} \mathbf{1}^\top \boldsymbol{\alpha}_{S_k}$, the arithmetic mean of the parameters within each block-group.