Supplementary Materials
Crime in Philadelphia: Bayesian Clustering with Particle Optimization

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

S1 Proof of Proposition 1

In this Section 3.1 we state that we can find the set of L particles with largest posterior by finding a variational approximation of the tempered posterior Π_λ. Here we restate Proposition 1 and provide the proof.

Remember that we denote with $\Gamma_L = \{\boldsymbol{\gamma}^{(1)}, \ldots, \boldsymbol{\gamma}^{(L)}\}$ the set of L particles with largest posterior mass, with $q(\cdot \mid \Gamma, \mathbf{w})$ the discrete distribution that places probability w_ℓ on the particle $\boldsymbol{\gamma}_\ell$ and with $\mathcal{Q}_L$ the collection of all such distributions supported on at most L particles. Moreover, for each $\lambda > 0$, let π_λ be the mass function of the tempered marginal posterior Π_λ, where $\pi_\lambda(\boldsymbol{\gamma}) \propto \pi(\boldsymbol{\gamma} \mid \mathbf{y})^{\frac{1}{\lambda}}$.

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 $w_\ell^\star(\lambda) \propto \pi_\lambda(\boldsymbol{\gamma}_\ell^\star)$. 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. □

S2 Various hyper-parameter choices

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 overlapping clusters with equal variance σ^2_{cl} 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 $\hat{\sigma}^2_i$'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 $\hat{\sigma}^2_i$'s.

S3 Additional Synthetic Data Evaluation

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 S1 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.

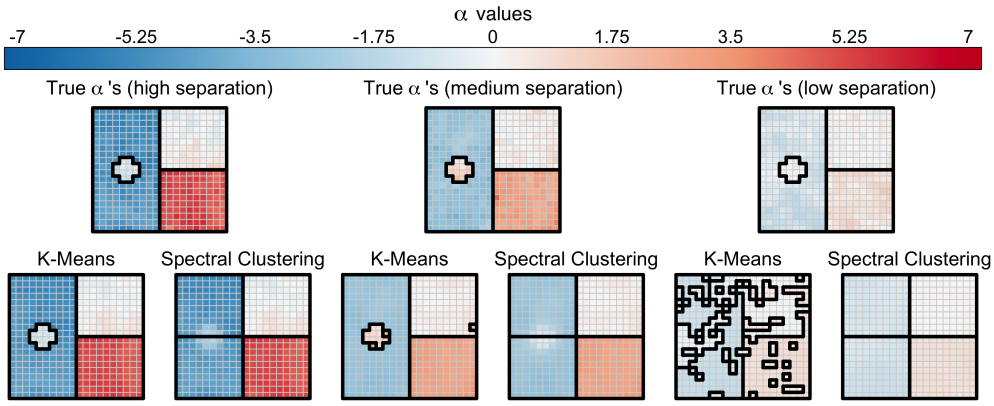

Figure S1: Partitions recovered by k-means and spectral clustering for three different cluster separation settings. The color of each tract corresponds to the estimated parameter value $\mathbb{E}[\alpha_i \mid \mathbf{y}, \boldsymbol{\gamma}]$.

S4 Additional Results for Clustering in Philadelphia

In figure S2 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).

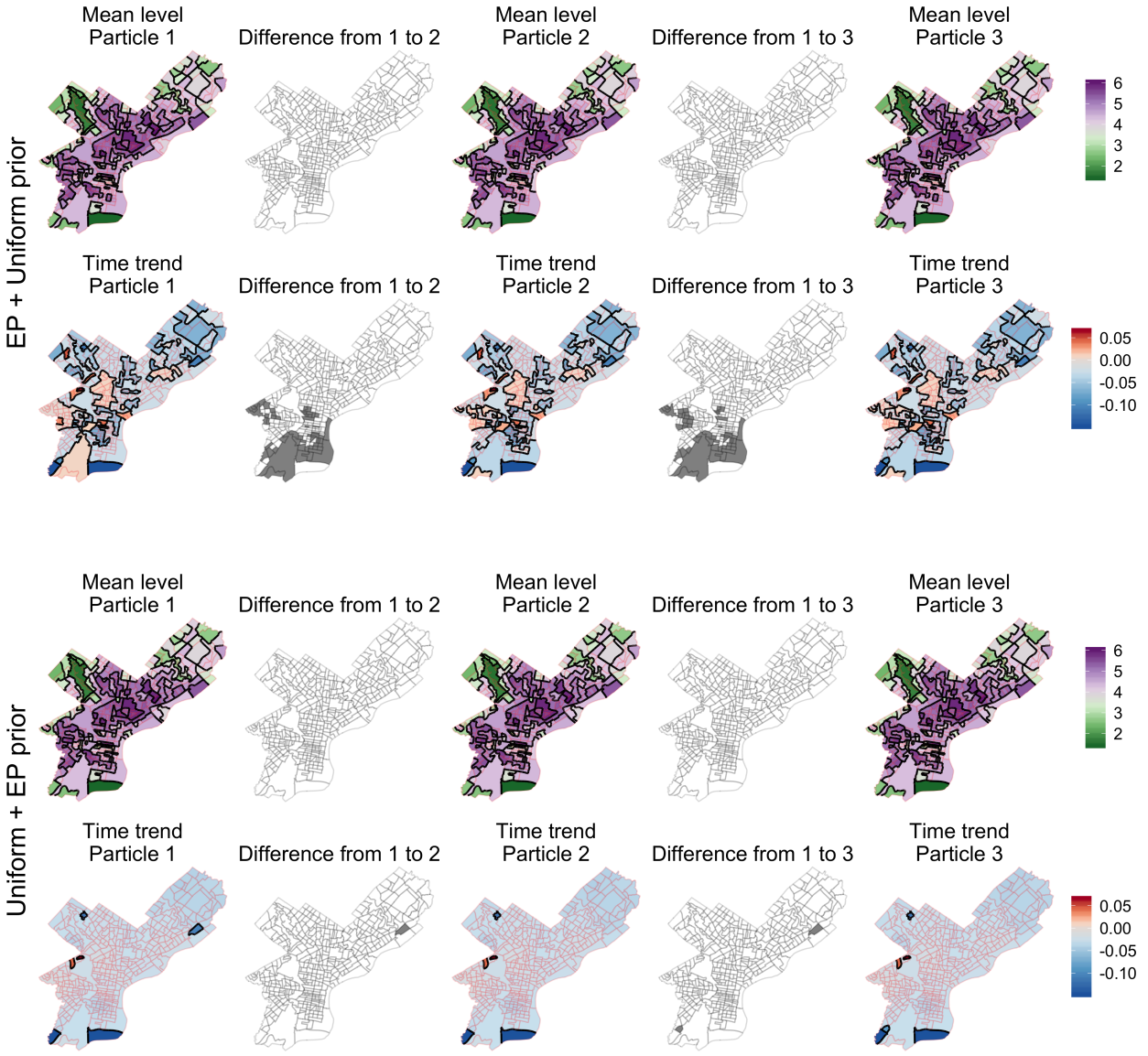

Figure S2: Colored plots: Top three models identified by our procedure. The thick borders represent the partition, and the color represents the posterior mean of the parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$. Black and white plots: transition from the model on the left to the model on 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 $\gamma^{(\alpha)}$.

S5 Derivation of Closed Form Expressions

S5.1 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

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

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.

S5.2 Two Partition Derivations

Recall from Section 2 that our full mode is:

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

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:

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

where $\Omega_{k,\mathrm{CAR}}^{(\alpha)} = \left(\Sigma_{k,\mathrm{CAR}}^{(\alpha)}\right)^{-1} = \rho(W_k^{(\alpha)})^* + (1-\rho)\mathbf{I}$; the second line follows from noticing that $\mathbf{1}$ is both a left and right eigenvector of $\Omega_{k,\mathrm{CAR}}^{(\alpha)}$ 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 $t = 1, \ldots T$) 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 $\sum_{t=1}^T x^2_{it}$; 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.