THE UNIVERSITY OF TEXAS AT AUSTIN

Statistical Modeling II Project

Juliette Franqueville
May 6, 2022

1 Introduction

The COVID-19 pandemic resulted in stay-at-home orders across the world. People spent more time at home and in outdoor areas and less time at work and restaurants. This project explores the relationship between the changes in mobility patterns of the population and the number of incidents reported by fire departments in the United States in 2020. Several studies have explored the effect of COVID-19 on fire incidents. For example, Suzuki and Manzello [1] analyzed the effect of stay-at-home orders on cooking fires in major cities. Koester and Greatbatch [2] investigated the impact of the onset of the pandemic on fire and search and rescue incident frequency. This report uses NFORS [3] analytics data provided by the International Public Safety Data Institute and 2020 Google Mobility data [4]. The names of the fire departments used in the analysis were hidden to preserve anonymity.

2 Data Formatting

The raw NFORS and Google datasets were formatted to obtain:

- weekly averages for number of incident percentage change from baseline for 2020 for each fire department
- weekly mobility (for each mobility type) averages corresponding to the county of each NFORS department. The Google data was already expressed as a change from baseline.

The Google data columns of interest were date, FIPS code (which corresponds to unique county) and percentage change from baseline for workplace, residential, retail, parks, and grocery/pharmacy mobility. First, the Google mobility data were grouped by FIPS code. The data with FIPS codes corresponding to those of the counties of the fire departments in the NFORS data were kept.

The raw Google mobility data is reported as a percentage change from baseline. The baseline used is the median value for each day of the week over the January 3rd - February 6th 2020 period. To get a weekly average, a 7-day rolling average was used and the value for all Mondays was reported as the weekly average.

The columns of interest in the NFORS data were date, fire department, and number of incidents reported for each date. Note that all types of incidents (fire, car crashes, and EMS) were pooled together in the incident counts. For each fire department, gaps in data (where more than two consecutive days had zero incidents, which is very unlikely) were removed. Then, outliers were removed by fitting a negative binomial distribution to the incident counts of each department using moment matching and removing points outside the 95 % confidence interval.

An incident baseline was calculated from the 2019 NFORS data and applied to the 2020 data. The baseline was calculated by taking the median of the number of incidents per day of the week per month for each department. Accounting for month as well as day of the week ensured that the seasonality in fire incidents was accounted for. Then, the percentage change for each day in 2020 was calculated from this baseline. Note that some departments were missing significant amounts of data and did not allow for calculating a baseline for each month and

each day of the week in 2019. For missing baseline data points, the corresponding changes from baseline in 2020 could not be calculated. Some department had no data for 2019, so they were not used in the analysis, As with the Google data, a weekly average for incident percentage change from baseline for each department was calculated.

The resulting data format an $X_i \in \mathbb{R}^{n_i \times p}$ matrix for each department, where p is the number of mobility types (workplace, retail, etc), i is the department, and n_i is the number of weekly averages for each department. Note that ideally, each department would have had $n_i = 52$ for the 52 weeks of 2020, but the Google mobility data only begins in February 2020 and some departments had missing data. Each department also had a $y_i \in \mathbb{R}^{n_i}$ vector of incident percentage change from baseline corresponding to the X_i matrices. Additionally, each department had a $t_i \in \mathbb{R}^{n_i}$ vector containing the day of the year that the observations are for (days all correspond to Mondays because the weekly averages were evaluated on Mondays).

3 Data Exploration

The various mobility types were highly correlated. Figure 1 shows the correlation matrix for the mobility types for all departments. The correlation matrix is not surprising, for example workplace mobility is inversely correlated with residential mobility. Parks, however, is not highly correlated with the other mobility types.

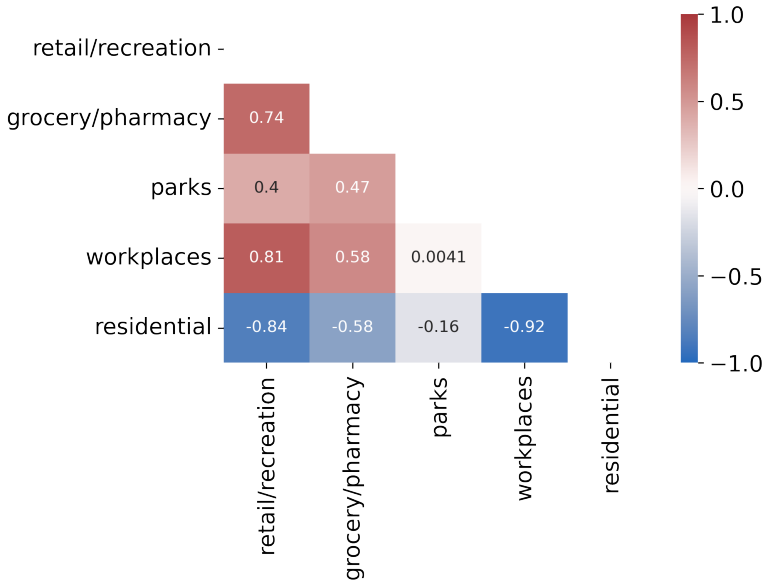

Figure 1: Correlation Matrix.

To avoid multicolinearity issues in the model, principal component analysis was performed and the first two components were kept. The components were calculated by first subtracting the mean of each mobility type from the corresponding mobility values. Then, the covariance matrix for the centered data was calculated. Then, the eigenvalues (or the explained variance corresponding to each component) and the eigenvectors (components) of the covariance

matrix were calculated. The two components corresponding to the two highest eigenvalues were kept. The first two components explained 99% of the variance. The first component pointed in the direction of "parks" and the second pointed in the direction of "retail/recreation", "grocery/pharmacy", "workplaces" and away from "residential"(see Table 1). The first component can easily be interpreted as park mobility, and the second as increasing "city" mobility and decreasing residential mobility. The original data were transformed using the chosen components.

	retail/recreation	**grocery/pharmacy**	**parks**	**workplaces**	**residential**
first component	0.108	0.080	0.991	0.008	-0.018
second component	0.646	0.311	-0.106	0.639	-0.258

Table 1: First two PCA components.

The next step was to determine whether linear regression could be used on the incident data, that is:

$$y_i \sim N(X_i\beta_i, \sigma_i^2 I)$$

Where i is the fire department, X_i is the design matrix including an intercept and transformed mobility data, y_i is the response vector, $\beta_i \in \mathbb{R}^p$, and I is the $n_i \times n_i$ identity matrix. OLS was used separately on each department's data. As shown in Figure 2, the Q-Q plots show that the residuals were approximately normally distributed; therefore no further transformations of the data were necessary.

Another concern was that the residuals (as plotted in Figure 2 - although some departments showed worse correlation than the 9 plotted) may be auto-correlated, in which case assuming that the covariance matrix is diagonal is not an appropriate assumption. Figure 3 shows ACF plots for all departments. Some departments (for example, departments 12, 13, 25, 29) had multiple correlation coefficients outside the 95 % confidence internal. For some departments, auto-correlation was not an issue.

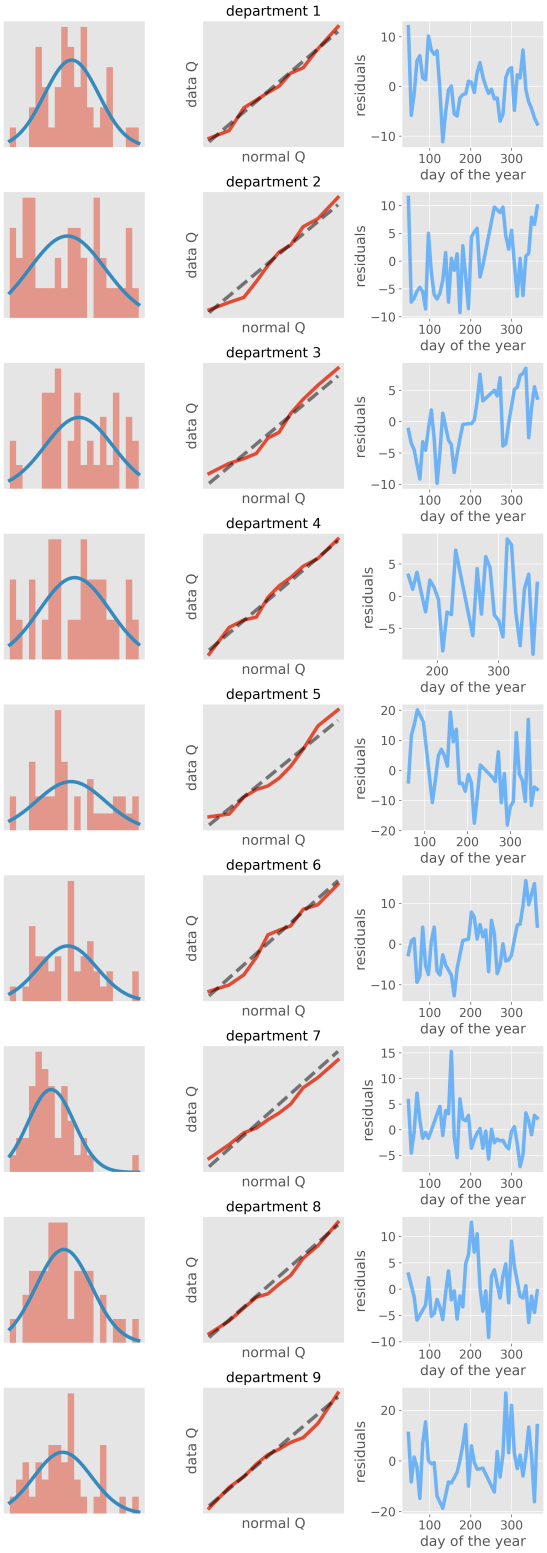

Figure 2: Plots showing residuals with fitted normal with same moments as the residuals (left), Q-Q plot (middle), residuals against time (right). Only the first 9 departments are shown.

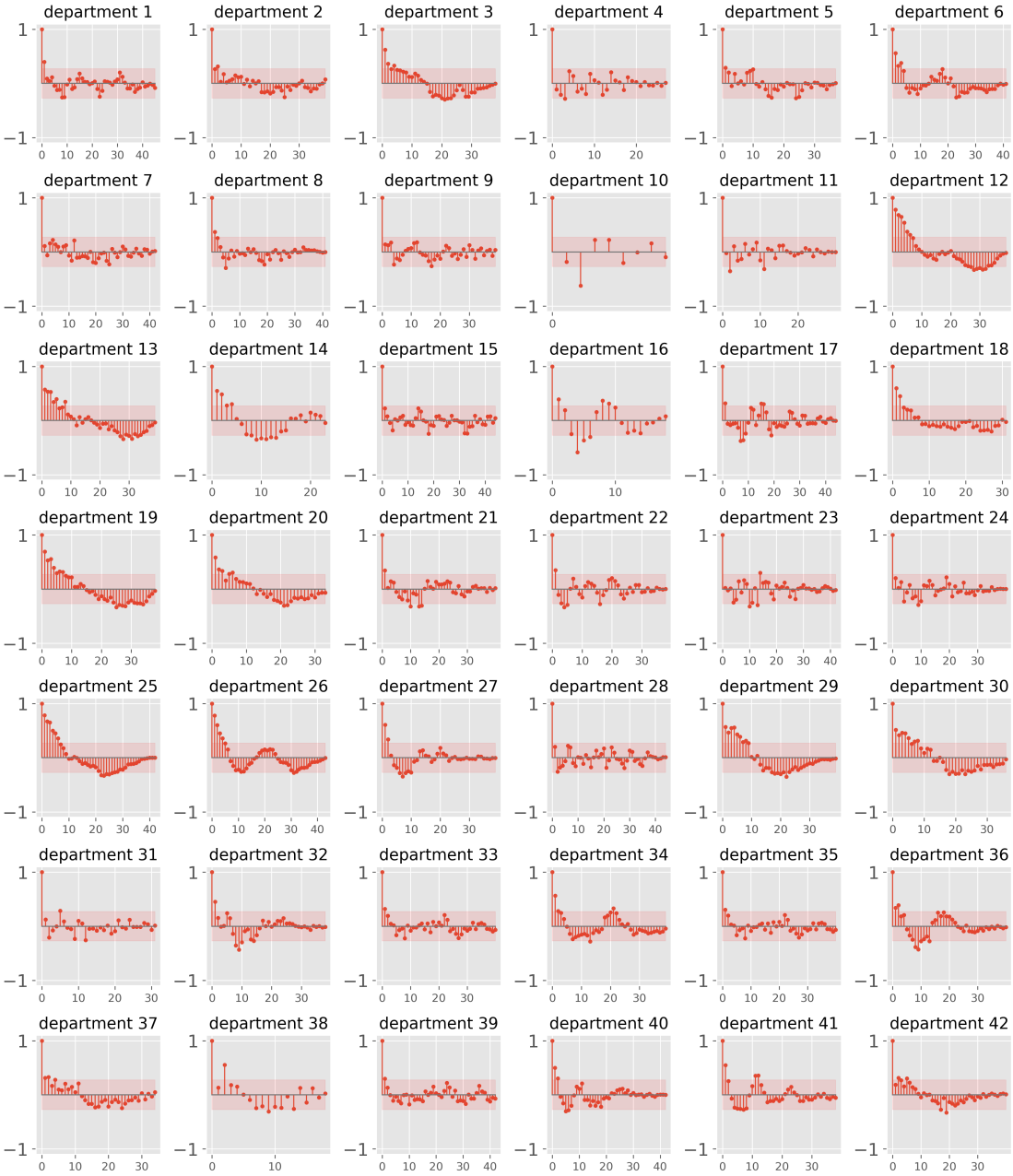

Figure 3: ACF plots for all departments for the residuals using linear regression on each department separately.

4 Models

4.1 Simple Hierarchical Model

As a first step before considering the autocorrelation in the residuals shown above, a simple partially pooled linear regression model was used:

$$
\begin{aligned}
y_i &= X_i \beta_i + \epsilon_i \\
\epsilon_i &\sim N(0, \sigma_i^2 I) \\
\beta_i &\sim N(\mu, \Sigma)
\end{aligned}
$$

The following priors were used:

$$
\begin{aligned}
\sigma_i^2 &\propto \frac{1}{\sigma_i^2} \\
\mu &\propto c \\
\Sigma_{jj} &\sim IG\left(\frac{1}{2}, \frac{1}{2}\right)
\end{aligned}
$$

The conditionals needed for Gibbs sampling were

$$
\begin{aligned}
P(\mu|\beta_i, \Sigma) &\propto \prod_{i=1}^{N} P(\beta_i|\mu, \Sigma) P(\mu) \\
&\propto \prod_{i=1}^{N} \exp\left\{-\frac{1}{2}(\beta_i - \mu)^T \Sigma^{-1} (\beta_i - \mu)\right\} \\
&\propto \exp\left\{-\frac{1}{2} \sum_{i=1}^{N} (\beta_i - \mu)^T \Sigma^{-1} (\beta_i - \mu)\right\} \\
&\propto \exp\left\{-\frac{1}{2} \sum_{i=1}^{N} \mu^T \Sigma^{-1} \mu - 2\mu^T \Sigma^{-1} \beta_i\right\} \\
&\propto \exp\left\{-\frac{1}{2}\left(N\mu^T \Sigma^{-1} \mu - 2\mu^T \Sigma^{-1} \sum_{i=1}^{N} \beta_i\right)\right\} \\
&\sim N_p(\mu^*, \Sigma^*)
\end{aligned}
$$

Where $\mu^* = \sum_{i=1}^{N} \beta_i / N$, $\Sigma^* = \Sigma / N$ and N is the number of departments.

$$
\begin{aligned}
P(\Sigma_{jj}|\beta_{ij},\mu_j) &\propto \prod_{i=1}^{N} P(\beta_{ij}|\mu_j,\Sigma_{jj})P(\Sigma_{jj}) \\
&\propto \prod_{i=1}^{N} \frac{1}{\Sigma_{jj}^{1/2}} \exp\left\{-\frac{1}{2\Sigma_{jj}}(\beta_{ij}-\mu_j)^2\right\} \Sigma_{jj}^{-1-\frac{1}{2}} \exp\left\{-\frac{1}{2\Sigma_{jj}}\right\} \\
&\propto \Sigma_{jj}^{-\frac{N+1}{2}-1} \exp\left\{-\frac{1}{2\Sigma_{jj}}\left[\sum_{i=1}^{N}(\beta_{ij}-\mu_j)^2+1\right]\right\} \\
&\sim IG\left(\frac{N+1}{2}, \frac{1}{2}\left[\sum_{i=1}^{N}(\beta_{ij}-\mu_j)^2+1\right]\right)
\end{aligned}
$$

$$
\begin{aligned}
P(\sigma_i^2|y_i,\beta_i) &\propto P(y_i|\sigma_i^2,\beta_i)P(\sigma_i^2) \\
&\propto \frac{1}{(\sigma_i^2)^{n_i/2}} \exp\left\{-\frac{1}{2}(y_i-X_i\beta_i)^T\sigma_i^{-2}(y_i-X_i\beta_i)\right\} \frac{1}{\sigma_i^2} \\
&\propto (\sigma_i^2)^{-\frac{n_i}{2}-1} \exp\left\{-\frac{\sigma_i^{-2}}{2}(y_i-X_i\beta_i)^T(y_i-X_i\beta_i)\right\} \\
&\sim IG\left(\frac{n_i}{2}, \frac{1}{2}(y_i-X_i\beta_i)^T(y_i-X_i\beta_i)\right)
\end{aligned}
$$

Where n_i is the number of observations in each department.

$$
\begin{aligned}
P(\beta_i|y_i,\sigma_i^2,\Sigma,\mu) &\propto P(y_i|\beta_i,\sigma_i^2)P(\beta_i|\Sigma,\mu) \\
&\propto \exp\left\{-\frac{1}{2}(y_i-X_i\beta_i)^T\sigma_i^{-2}(y_i-X_i\beta_i)\right\} \exp\left\{-\frac{1}{2}(\beta_i-\mu)^T\Sigma^{-1}(\beta_i-\mu)\right\} \\
&\propto \exp\left\{-\frac{1}{2}\left(-2\beta_i^T[\sigma_i^{-2}X_i^Ty_i+\Sigma^{-1}\mu]+\beta_i^T[\sigma_i^{-2}X_i^TX_i+\Sigma^{-1}]\beta_i\right)\right\} \\
&\sim N_p(\mu^*,\Sigma^*)
\end{aligned}
$$

Where $\Sigma^* = [\sigma_i^{-2}X_i^TX_i+\Sigma^{-1}]^{-1}$ $\mu^* = \Sigma^*[\sigma_i^{-2}X_i^Ty_i+\Sigma^{-1}\mu]$.

For learning purposes, the model was also implemented using PyMC. PyMC is a No U-Turn Sampler (NUTS), which is an extension of Hamiltonian Monte Carlo (HMC). HMC requires specifying a step size and number of steps. NUTS eliminates the need to set the number of steps and instead uses an algorithm that stops automatically when it starts to retrace its steps. Note that Jeffrey's prior is not an option in PyMC (although it could likely be implemented with extra work), so a uniform prior was used for σ_i^2.

4.2 Gaussian Process/Hierarchical Model

Next, the hierarchical model shown above was improved to account for temporal correlation. The improved model was structured as follows:

$$
\begin{aligned}
y_i &= f_i(t_i) + \epsilon_i \\
\epsilon_i &\sim N(0, \sigma_i^2 I) \\
f_i(t_i) &\sim GP(X_i\beta_i, C_i) \\
\beta_i &\sim N(\mu, \Sigma)
\end{aligned}
$$

Where i is the department number, t_i is the time vector for each department, X_i is the design matrix for each department, C_i is the GP covariance matrix which is a function of t_i, the bandwidth b_i, and τ_i^2. The following priors were used:

$$
\begin{aligned}
\sigma_i^2 &\propto \frac{1}{\sigma_i^2} \\
\mu &\propto c \\
\Sigma_{jj} &\sim IG\left(\frac{1}{2}, \frac{1}{2}\right) \\
b_i &\sim U(1, 150) \\
\tau^2 &\sim U(1, 100)
\end{aligned}
$$

The conditionals for Σ_{jj} and μ were the same as in the simple model. In addition, the following conditionals were needed:

$$
\begin{aligned}
P(\sigma_i^2 | y_i, f_i) &\propto P(y_i | \sigma_i^2, f_i) P(\sigma_i^2) \\
&\propto \frac{1}{(\sigma_i^2)^{n_i/2}} \exp\left\{-\frac{1}{2}(y_i - f_i)^T \sigma_i^{-2} (y_i - f_i)\right\} \frac{1}{\sigma_i^2} \\
&\propto (\sigma_i^2)^{-\frac{n_i}{2}-1} \exp\left\{-\frac{\sigma_i^{-2}}{2}(y_i - f_i)^T (y_i - f_i)\right\} \\
&\sim IG\left(\frac{n_i}{2}, \frac{1}{2}(y_i - f_i)^T (y_i - f_i)\right)
\end{aligned}
$$

Where n_i is the number of observations in each department.

$$
\begin{aligned}
P(\beta_i | f_i, C_i, \Sigma, \mu) &\propto P(f_i | \beta_i, C_i) P(\beta_i | \Sigma, \mu) \\
&\propto \exp\left\{-\frac{1}{2}(f_i - X_i\beta_i)^T C_i^{-1} (f_i - X_i\beta_i)\right\} \exp\left\{-\frac{1}{2}(\beta_i - \mu)^T \Sigma^{-1} (\beta_i - \mu)\right\} \\
&\propto \exp\left\{-\frac{1}{2}\left(-2\beta_i^T [X_i^T C_i^{-1} f_i + \Sigma^{-1}\mu] + \beta_i^T [X_i^T C_i^{-1} X_i + \Sigma^{-1}]\beta_i\right)\right\} \\
&\sim N_p(\mu^*, \Sigma^*)
\end{aligned}
$$

Where $\Sigma^* = [X_i^T C_i^{-1} X_i + \Sigma^{-1}]^{-1}$ $\mu^* = \Sigma^* [X_i^T C_i^{-1} f_i + \Sigma^{-1}\mu]$.

$$
\begin{aligned}
P(f_i|y_i,\sigma_i^2,C_i,\beta_i) &\propto P(y_i|f_i,\sigma_i^2)P(f_i|C_i,\beta_i) \\
&\propto \exp\left\{-\frac{1}{2}(y_i-f_i)^T\sigma_i^{-2}(y_i-f_i)\right\}\exp\left\{-\frac{1}{2}(f_i-X_i\beta_i)^T C_i^{-1}(f_i-X_i\beta_i)\right\} \\
&\propto \exp\left\{-\frac{1}{2}(-2f_i^T[\sigma_i^{-2}y_i+C_i^{-1}X_i\beta_i]+f_i^T[\sigma_i^{-2}I+C_i^{-1}]f_i)\right\} \\
&\sim N_{n_i}(\mu^*,\Sigma^*)
\end{aligned}
$$

Where $\Sigma^* = [\sigma_i^{-2}I + C_i^{-1}]^{-1}$ $\mu^* = \Sigma^*[\sigma_i^{-2}y_i + C_i^{-1}X_i\beta_i]$, and

$$P(b_i,\tau_i^2|f_i) \propto P(f_i|b_i,\tau_i^2,\beta_i)P(b_i)P(\tau_i^2)$$

The last conditional did not simplify to a known distribution. Random walk Metropolis-Hastings was used to jointly update b_i and τ_i^2 as shown below:

- Initialize b_i and τ_i^2 in a feasible region
- Draw $b_{i,t+1}$ and $\tau_{i,t+1}^2$ from separate normal distributions centered at $b_{i,t}$ and $\tau_{i,t}^2$ with given variances. If the proposed value is below 0, choose 0 instead.
- Calculate the logarithm of the ratio of the new posterior to the old posterior. Taking the logarithm helps with numerical stability.
- Draw a value from $U(0,1)$. If the ratio above is larger than the logarithm of the drawn value, accept the new parameters. Otherwise, keep the old parameters.

The variances for the proposal distributions were chosen by trial and error.

In order to calculate the GPs covariance matrices C_i, the squared exponential function was used:

$$C(x_1,x_2) = \tau_1^2\exp\left\{-\frac{1}{2}\left(\frac{x_2-x_1}{b}\right)^2\right\} + \tau_2^2\delta(x_1,x_2)$$

τ_2^2 was set to 10×10^{-5}. τ_1^2 is referred to as τ^2 throughout the report.

5 Results

5.1 Model Check

5.1.1 Simple Hierarchical Model with Gibbs Sampling and PyMC Implementations

Similar versions of the simple hierarchical model were implemented in class, so as expected, no issues were encountered in the implementation. The traces were well-mixed; as shown for example for μ and the diagonal elements of Σ in Figure 4. The Gibbs sampler was run for 10,000 iterations and the first 1,000 were discarded.

The Gibbs sampler was also compared to the PyMC results. The posteriors looked almost identical despite the different priors used, as shown in Figures 5 and 6.

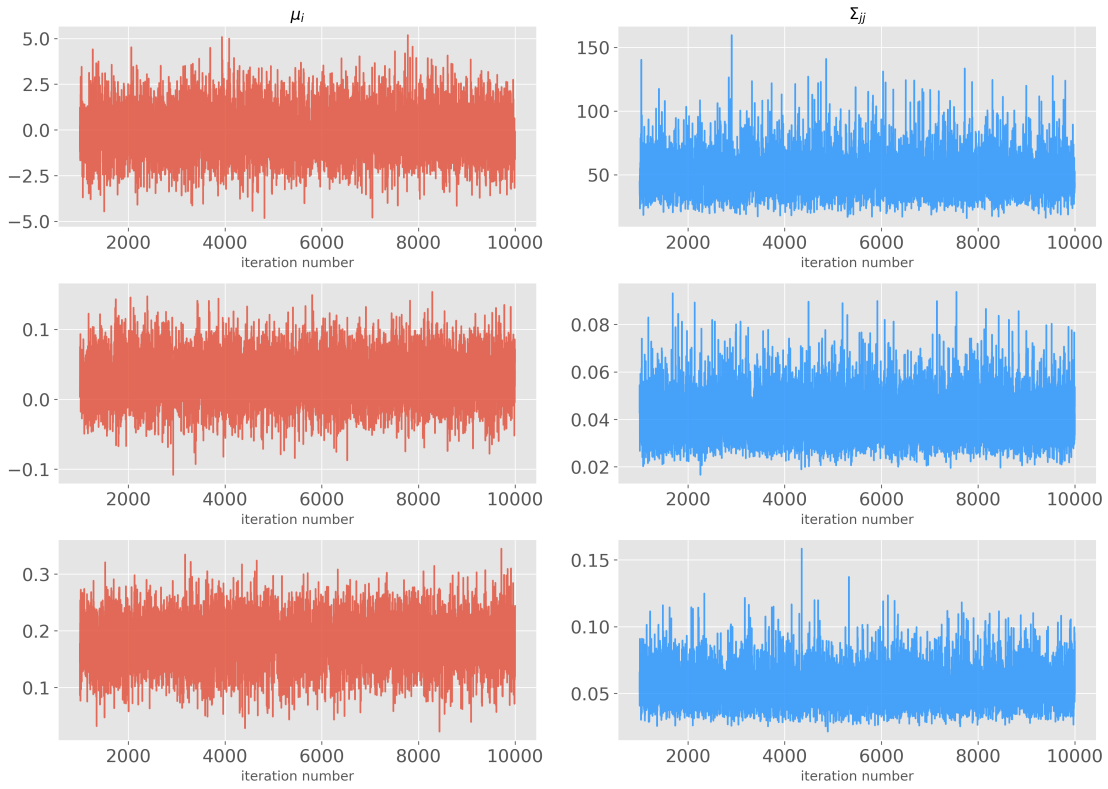

Figure 4: Traces for μ and the diagonal elements of Σ.

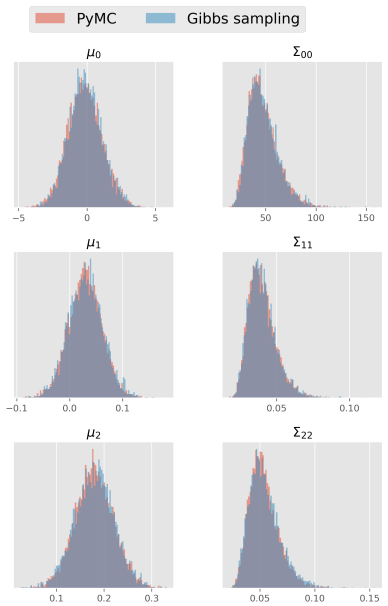

Figure 5: Posteriors for μ and Σ.

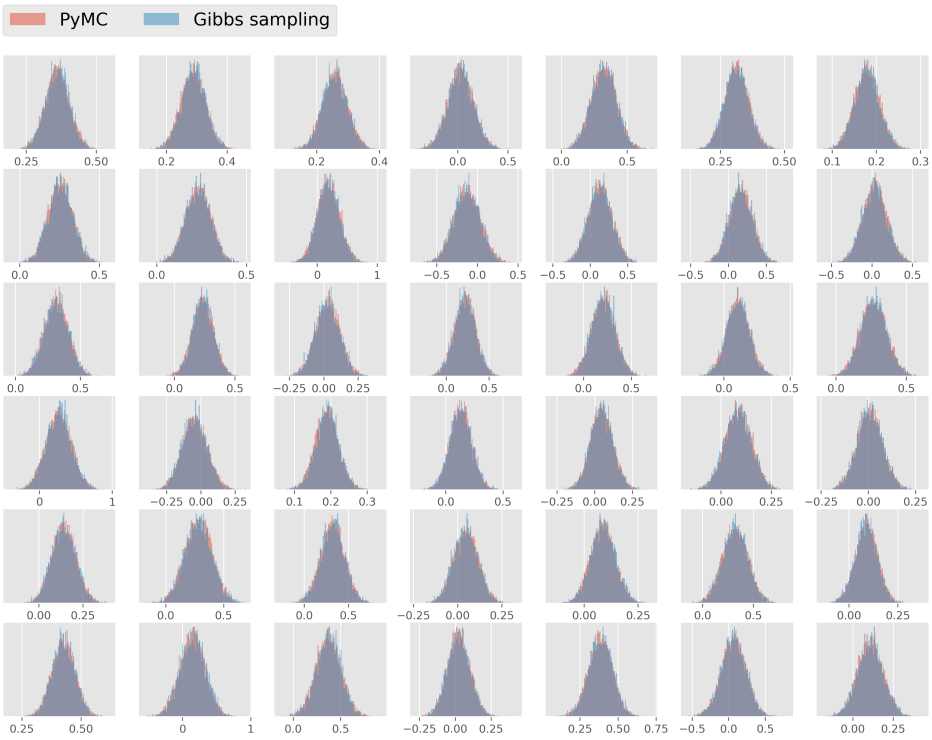

Figure 6: Posteriors for β_2.

5.1.2 Gaussian Process Model

The Gaussian process hierarchical model was run two separate times to verify that the traces converged to the same posterior distributions. However, issues were encountered with the model. The traces were not all well mixed, and some traces for the two chains did not converge to the same distribution. The acceptance rates for the bandwidths and τ_i^2 were around 53% (averaged across departments). Attempts were made to fix the issues encountered: several variances in the proposal distributions were tested and the model was run for as long as possible given time constraints (100,000 iterations). However, the issues remained. The issues discussed are illustrated in the plots below. Traces for one department (number 6) are shown .

Figure 7 shows the traces for the bandwidth for department number 6. The traces are relatively well-mixed; their Gelman-Rubin convergence diagnostic was around 1. Note that when calculating the diagnostic, the first 5,000 iterations were omitted. The chains for τ^2 looked more mixed; their Gelman-Rubin diagnostic was also around 1. The traces for β_0 were better, but those for β_1 and β_2 were most problematic. Unsurprisingly, their Gelman-Rubin diagnostics were not acceptable (> 1.1, the commonly used threshold). The traces for μ and the diagonal elements of Σ looked well-mixed, but did not seem to converge until about 80,000 iterations. The posteriors for these were also significantly different than those obtained with the simple Gibbs sampler, which was surprising and may further indicate that something was wrong with the GP model.

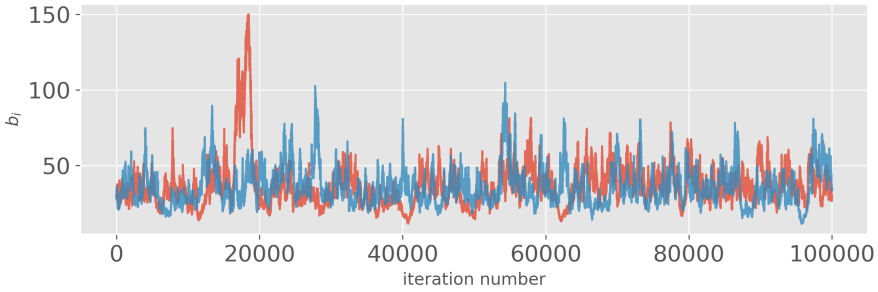

Figure 7: Traces for the bandwidth for department number 6.

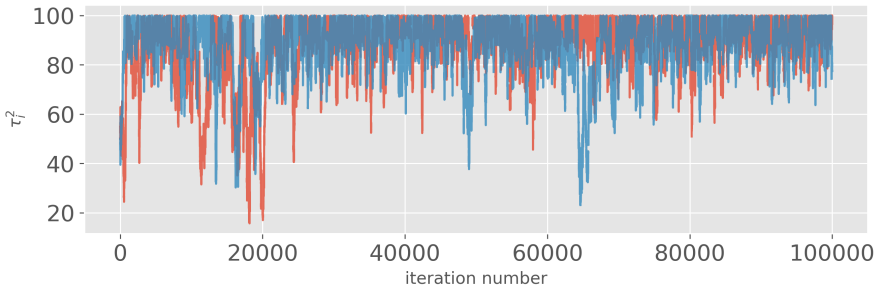

Figure 8: Traces for τ^2 for department number 6.

Some ideas to solve the issues encountered which were not attempted are:

- run the model for more iterations
- change proposal distributions again
- re-formulate the model

It is also possible that there was an issue with the code. Because of the issues with the traces, the interpretation of the results in the next section is based on the results from the Gibbs sampler without the GP model.

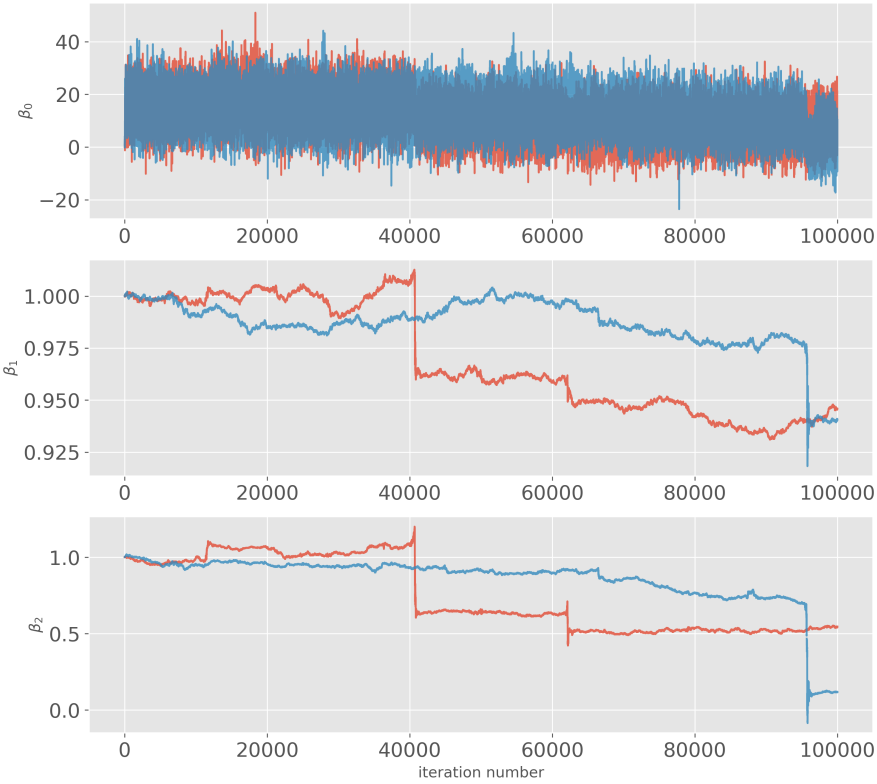

Figure 9: Traces for β for department number 6.

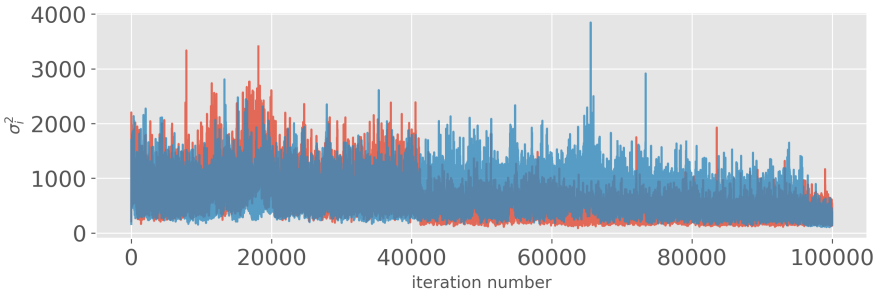

Figure 10: Traces for σ_i^2 for department number 6.

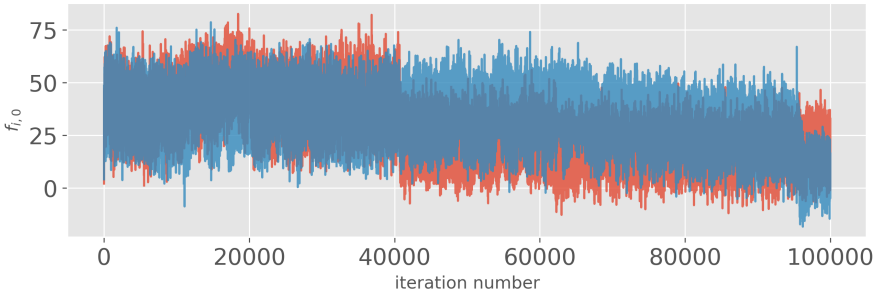

Figure 11: Traces for the first element of f for department number 6.

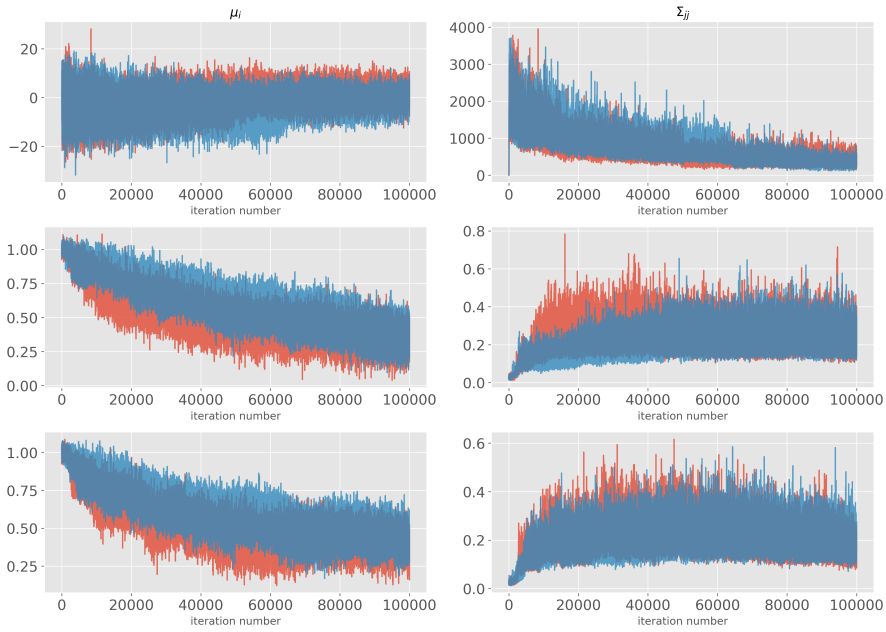

Figure 12: Traces for the μ and Σ for the GP model.

5.2 Interpretation

The first parameters of interest were the mean and diagonal terms for the covariance matrix for β. These are shown above in Figure 5; the simple Gibbs sampling posteriors are shown in blue. Table 2 shows the posterior means:

element	**0**	**1**	**2**
μ	-0.078	0.03	0.179
diagonal Σ	48.957	0.04	0.054

Table 2: posterior means for μ and the diagonal elements of Σ.

The first mean was negative and the other two positive. This means that on average, β_0 was negative and β_1 and β_2 were positive. This implies that on average, the incident counts in 2020 not explained by mobility were lower than in 2019. It also implies that an increase in park traffic resulted in an increase in incident compared to baseline, and that an increase in residential traffic/decrease in all other types of mobility resulted in a decrease in incident compared to baseline. It is interesting to see that the posterior for μ_2 only covers positive values, which is not the case for μ_1. The first diagonal elements of the covariance is large, meaning that there is a strong department effect for β_0. It is smaller for the two other β.

Figure 13 shows the posterior mean for the β vector for each department, and Figure 14 shows the corresponding box plots. As expected from the fact that the first diagonal element of Σ was large, the spread in β_0 across departments was significant. There are many variables which could explain a positive intercept: aging population, an unusually low baseline in 2019, decreasing population, covid EMS calls, growing population etc. A negative intercept may be explained by an unusually high baseline in 2019, decreasing population, etc.

The β_1 elements were mostly positive, which means that increased traffic in parks resulted in more incidents compared to baseline. This is intuitive given injury risks associated with going to parks, especially national/state parks and beaches. β_2 was also mostly positive, as expected. A couple coefficients were negative for both β_1 and β_2, which is difficult to explain intuitively. It is possible Covid cases is a "lurking" variable not accounted for by the model. Covid cases are likely inversely correlated with "city" mobility and may make the β_2 coefficients lower than they really are. To avoid this issue, one could add a Covid cases feature to the design matrix.

The relationship between the β coefficients and county population and state was explored, but no strong correlations were found.

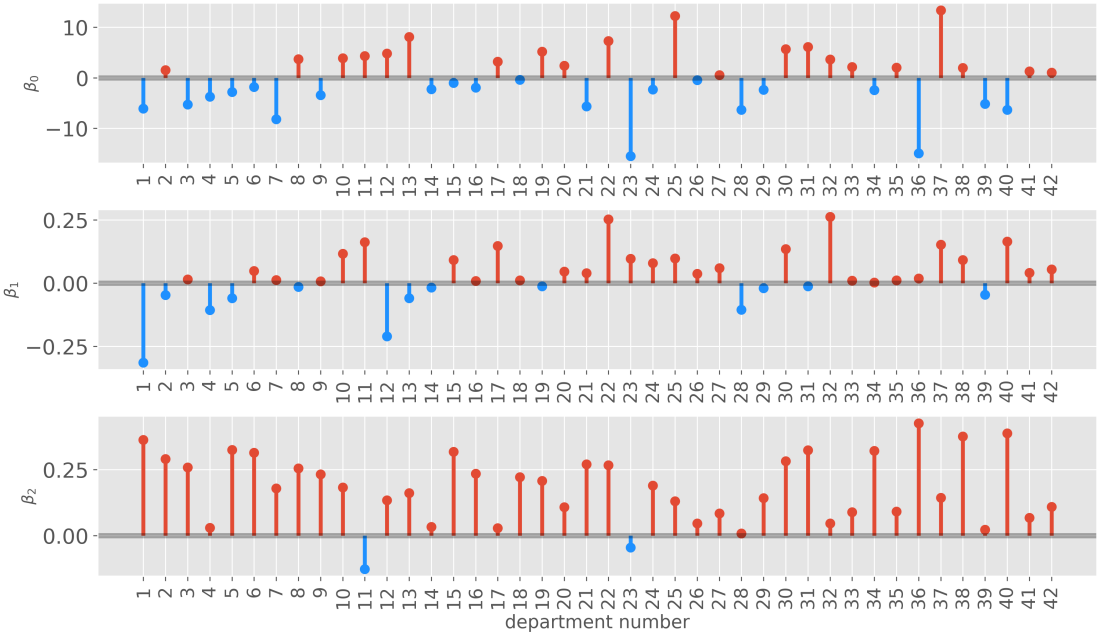

Figure 13: β means.

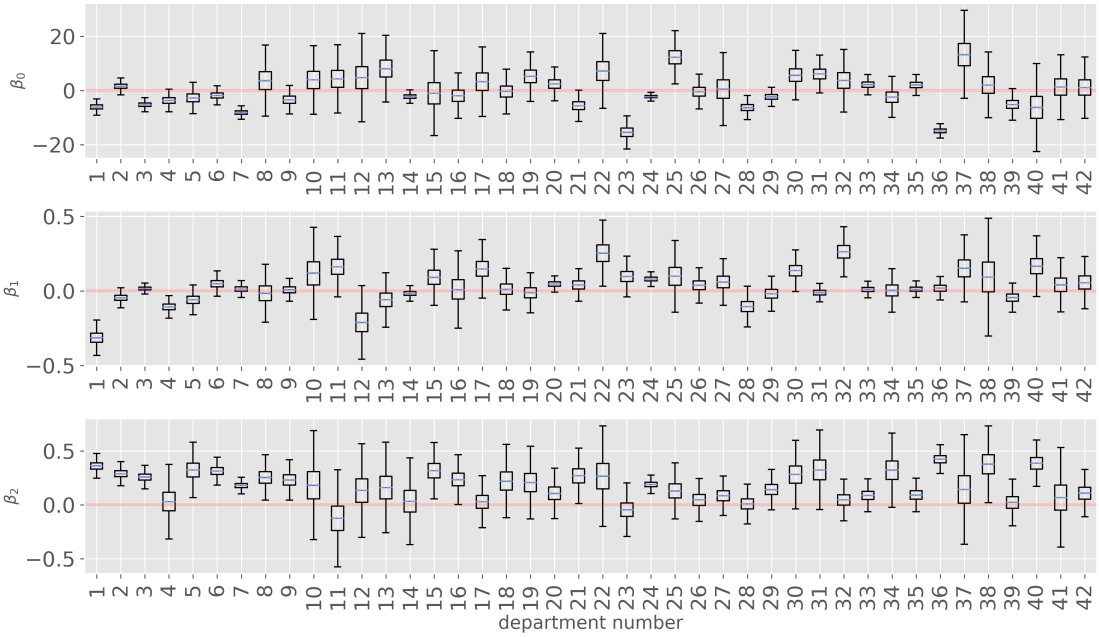

Figure 14: β boxplots.

Lastly, Figure 15 shows a visualization of the shrinkage of the β vectors. Shrinking means that the posterior mean for β is pulled towards the grand mean μ as opposed to being pulled towards the OLS value for β. High variance in the data, small diagonal values for Σ and few data points in departments' datasets can cause high shrinkage. Some departments showed

unexpected behavior. For example, β_1 for department 12 was shrunk from a positive to a negative value and its two other coefficients were significantly shrunk. This was caused by the fact that PCA was globally performed on the data of all departments (as opposed to for each department separately) . This caused issues for some departments for which the two components ended up being highly correlated. This explained the unexpected shrinkage behavior for some departments in this plot.

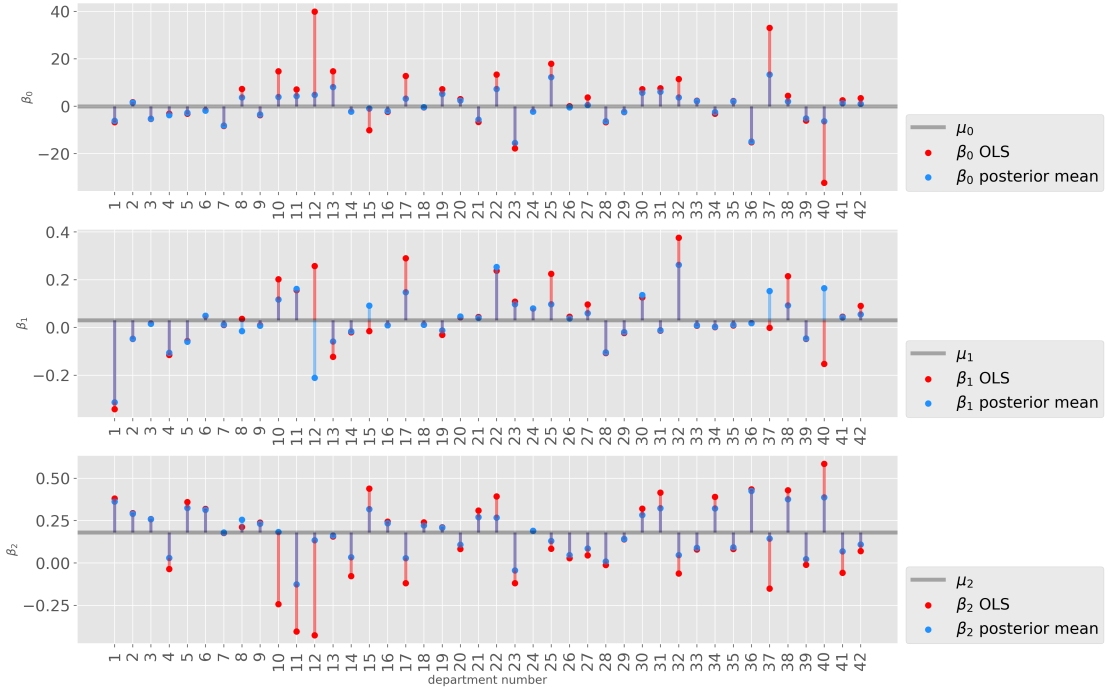

Figure 15: Visualization of shrinkage.

6 Conclusion and Future Work

The analysis above confirmed that on average across all departments, the decrease in "city" mobility over the pandemic caused a decrease in incidents. The increase in "park" mobility also caused an increase in incidents, though this was not the case for all departments. This information may help fire departments learn from this pandemic and better allocate resources in the future. The model could be improved significantly by including Covid cases as a variable, which would make it easier to interpret the β vectors. In this analysis, all incident types were pooled together. It may be useful to redo this analysis for each incident type, given that resource allocation is different for each incident type. Additionally, there were issues with the GP model; running it for longer, changing the proposal distributions or the priors on the covariance parameters may help the traces converge to a stationary distribution. It would also be interesting to implement the GP model in PyMC.

References

[1] S. Suzuki and S. L. Manzello, "The influence of covid-19 stay at home measures on fire statistics sampled from new york city, london, san francisco, and tokyo," *Fire Technology*, vol. 58, no. 2, pp. 679–688, 2022.

[2] R. J. Koester and I. Greatbatch, "Comparing the impact of covid-19 on search and rescue and fire emergency incident responses," *Journal of Search and Rescue*, vol. 4, no. 2, pp. 190–199, 2020.

[3] "Nfors." https://i-psdi.org/nfors-overview.html.

[4] "Google mobility data." https://www.google.com/covid19/mobility/.