Appendix A: Proof of proposition 1

Proof. By the property of Kronecker products, we only need to show that $\mathbf{R}_{\theta_t}$, $\mathbf{R}_B$, and $\mathbf{R}_{\theta_y}$ are all positive definite. First, since $\mathbf{R}_{\theta_t}$ is the exponential of a symmetric matrix, it is positive definite.

Second, denote the basis matrix as $\mathbf{B}$ where

$$\mathbf{B} = \begin{pmatrix} B_1(1) & B_1(2) & \cdots & B_1(T) \\ B_2(1) & B_2(2) & \cdots & B_2(T) \\ \vdots & \vdots & \ddots & \vdots \\ B_K(1) & B_K(2) & \cdots & B_K(T) \end{pmatrix}.$$

Then $\mathbf{R}_B$ can be written as $\mathbf{R}_B = \mathbf{B}'\mathbf{B}$. Since $\mathbf{B}$ is a basis matrix (which is full rank), we can conclude that $\mathbf{R}_B$ is positive definite since for any non-zero vector $\boldsymbol{x}$,

$$\boldsymbol{x}'\boldsymbol{R}_B\boldsymbol{x} = \boldsymbol{x}'\mathbf{B}'\mathbf{B}\boldsymbol{x} = (\mathbf{B}\boldsymbol{x})'(\mathbf{B}\boldsymbol{x}) > \mathbf{0}.$$

Third, for $\mathbf{R}_{\theta_y}$, denote $\mathbf{R}_{\theta_y} = \boldsymbol{\sigma} \circ \mathbf{P}$, where $P_{j_1,j_2} = \exp\left\{-\theta_y\left|\boldsymbol{y}_{j_1} - \boldsymbol{y}_{j_2}\right|^2\right\}$ is the (j_1, j_2) element of $\mathbf{P}$ and the operation "$\circ$" is the entrywise matrix product. Suppose $\boldsymbol{\sigma}$ is a block diagonal binary matrix with the form after permutation

$$\boldsymbol{\sigma} = \begin{pmatrix} \boldsymbol{\sigma}_1 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\sigma}_2 & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \boldsymbol{\sigma}_n \end{pmatrix}.$$

Then a corresponding permutation of $\mathbf{R}_{\theta_y}$ can be written as

$$\begin{pmatrix} \mathrm{P}_1 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathrm{P}_2 & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathrm{P}_n \end{pmatrix},$$

where P_j $(j = 1, \cdots, n)$ is a diagonal submatrix of $\mathbf{P}$ with the same dimension as $\boldsymbol{\sigma}_j$. It can be observed that each submatrix P_j is an exponential of a symmetric matrix and is thus positive definite. Therefore, $\mathbf{R}_{\theta_y}$ is also positive definite. Hence, we have proved proposition 1.

Appendix B: Metropolis-Hastings algorithm in the E-step.

Metropolis-Hasting algorithms for sampling $\boldsymbol{u}_i$:

Steps:

Choose $\boldsymbol{u}_i^{(0)}$ as initial value, and let $v \leftarrow 1$;

While $v < m$ do

Draw a candidate $\boldsymbol{u}_i^*$ from $g(\boldsymbol{u}_i^* | \boldsymbol{u}_i^{(v-1)}, \boldsymbol{\Theta}^{(s)})$;

$\alpha \leftarrow \min\left(\frac{f(\boldsymbol{u}_i^*)/g(\boldsymbol{u}_i^*|\boldsymbol{u}_i^{(v-1)},\boldsymbol{\Theta}^{(s)})}{f(\boldsymbol{u}_i^{(v-1)})/g(\boldsymbol{u}_i^{(v-1)}|\boldsymbol{u}_i^*,\boldsymbol{\Theta}^{(s)})}, 1\right)$;

Draw a uniform random number w on $[0, 1]$;

If $w \leq \alpha$,

$\boldsymbol{u}_i \leftarrow \boldsymbol{u}_i \cup \boldsymbol{u}_i^{(v)}$;

$v = v + 1$;

End

End

Return $\left\{\boldsymbol{u}_i^{(1)}, \boldsymbol{u}_i^{(2)}, \cdots, \boldsymbol{u}_i^{(m)}\right\}$.

Here $g(\boldsymbol{u}^* | \boldsymbol{u}, \boldsymbol{\Theta}^{(s)})$ is the density function of the proposal distribution; and $f(\boldsymbol{u}^*) = p(\boldsymbol{u}, \boldsymbol{N} | \boldsymbol{\Theta}^{(s)})$.

Chib et al. (1998) discussed possible options to select a proposal distribution to reduce the burn-in time of the Metropolis-Hastings algorithm. It is advised in the literature to use a tailored normal distribution (Chib et al., 1998, Wu et al., 2018) that targets the modal value $\widehat{\boldsymbol{u}}_i = arg\max_{\boldsymbol{u}} \ln p(\boldsymbol{u}, \boldsymbol{N}_i | \boldsymbol{\Theta}^{(s)})$, where the distribution parameters can be derived as $N(\boldsymbol{\kappa}_i, c\boldsymbol{H}_i)$ and

$$\boldsymbol{\kappa}_i = \left[diag(\boldsymbol{N}_i) + \boldsymbol{\Sigma}^{(s)^{-1}}\right]^{-1} \left(diag(\boldsymbol{N}_i) \log \boldsymbol{N}_i + \boldsymbol{\Sigma}^{(s)^{-1}} \boldsymbol{\mu}^{(s)}\right),$$

$$\boldsymbol{H}_i = (diag(e^{\boldsymbol{\kappa}_i}) + \boldsymbol{\Sigma}^{-1})^{-1}.$$

Here c is a tuning parameter and $\boldsymbol{H}_i$ can be replaced by its nearest positive semidefinite matrix if it is not positive semidefinite (Wu et al., 2018).

Appendix C: Algorithm to determine $\boldsymbol{\sigma}$ in the M-step.

Algorithm to determine $\boldsymbol{\sigma}$ in the M-step:

Input: estimated covariance matrix $\widehat{\boldsymbol{\Sigma}}$, estimated parameters at the previous iteration $\theta_y{}^{(s)}, \tau^{(s)}$.

Output: estimation of $\boldsymbol{\sigma}^{(s+1)}$ at iteration $s+1$.

Steps:

Get a direct estimation $\boldsymbol{\pi}$ based on $\widehat{\boldsymbol{\Sigma}}$ where

$$\pi_{i,j} = \frac{1}{T} \sum_{l=1}^{T} \left(\widehat{\Sigma}_{T(i-1)+l, T(j-1)+l} / \exp\left\{-\theta_y^{(s)} \left|\boldsymbol{y}_i - \boldsymbol{y}_j\right|^2\right\} \left(\sum_{k=1}^{K} B_k(l)^2 + \left(\tau^{(s)}\right)^2 \right) \right);$$

Initialize $\boldsymbol{\pi}^{(1)} \leftarrow \boldsymbol{\pi}$, $\boldsymbol{\pi}^{(2)} \leftarrow \boldsymbol{\pi}$, $\epsilon \leftarrow$ average$\left(\boldsymbol{\pi}^{(1)} \backslash diag\left(\boldsymbol{\pi}^{(1)}\right)\right)$, list $\boldsymbol{L} \leftarrow \emptyset$, and scalar $q \leftarrow 1$, where "\" represents the set operation of difference.

Determine the index of the largest element of $\boldsymbol{\pi}^{(2)}$: $(i_1, i_2) \leftarrow \arg\max\left(\boldsymbol{\pi}^{(2)}\right)$, and let $\pi_{i_1,i_2}^{(2)} \leftarrow -\infty$, $\pi_{i_2,i_1}^{(2)} \leftarrow -\infty$, $L_q \leftarrow \{i_1, i_2\}$, $q \leftarrow q+1$.

Determine the value and index of the next largest element: $m \leftarrow \max\left(\boldsymbol{\pi}^{(2)}\right)$, $(i_1, i_2) \leftarrow \arg\max\left(\boldsymbol{\pi}^{(2)}\right)$, and let $\pi_{i_1,i_2}^{(2)} \leftarrow -\infty$, $\pi_{i_2,i_1}^{(2)} \leftarrow -\infty$.

While $m > \epsilon$,

Let flag $= 0$;

For $l = 1, \cdots, q-1$,

If $i_1 \in L_l$,

If average $\left(\boldsymbol{\pi}^{(1)}_{L_l \cup \{i_2\}, L_l \cup \{i_2\}}\right) > \epsilon$,

Let $L_l \leftarrow L_l \cup \{i_2\}$, $\boldsymbol{\pi}^{(2)}_{L_l \cup \{i_2\}, L_l \cup \{i_2\}} \leftarrow -\infty$, flag $= 1$;

End

Else if $i_2 \in L_l$

If average $\left(\boldsymbol{\pi}^{(1)}_{L_l \cup \{i_1\}, L_l \cup \{i_1\}}\right) > \epsilon$,

Let $L_l \leftarrow L_l \cup \{i_1\}$, $\boldsymbol{\pi}^{(2)}_{L_l \cup \{i_1\}, L_l \cup \{i_1\}} \leftarrow -\infty$, flag $= 1$;

End

End

End

If flag $= 0$,

$L_q \leftarrow \{i_1, i_2\}$, $q \leftarrow q + 1$.

End

$m \leftarrow \max\left(\boldsymbol{\pi}^{(2)}\right)$, $(i_1, i_2) \leftarrow \arg\max\left(\boldsymbol{\pi}^{(2)}\right)$, and let $\pi^{(2)}_{i_1, i_2} \leftarrow -\infty$, $\pi^{(2)}_{i_2, i_1} \leftarrow -\infty$

End

Let $\boldsymbol{\sigma}^{(s+1)} \leftarrow \mathbf{I}_{\boldsymbol{n}}$ ($n \times n$ identity matrix).

For $l = 1, \cdots, q-1$,

Let $\boldsymbol{\sigma}^{(s+1)}_{L_l, L_l} = 1$;

End

End

Return $\boldsymbol{\sigma}^{(s+1)}$

Please note that $\boldsymbol{\pi}$ is a direct estimation of $\boldsymbol{\sigma}$ based on the traffic network information and the estimation of other parameters in the previous step. Then we estimate the clusters of routes denoted as the lists $\{L_1, L_2, \cdots, L_{q-1}\}$, where the routes in the same cluster are highly correlated. After the evaluation of clusters, we set the submatrices corresponding to all clusters as all-one matrices. In this way, it is naturally guaranteed that the output $\boldsymbol{\sigma}^{(s+1)}$ is equivalent to a block diagonal matrix with regard to permutation. Since this algorithm is based on thresholding of the estimator $\pi_{i,j}$, we establish a proposition that shows the consistency of this estimator.

Proposition 2. For $\epsilon \in (0,1)$, the thresholding estimator $\Phi_\epsilon(\pi_{i,j})$ is a consistent estimator of σ_{ij}, where $\pi_{i,j} = \frac{1}{T}\sum_{l=1}^{T}\left(\hat{\Sigma}_{T(i-1)+l,T(j-1)+l}/\exp\left\{-\theta_y\left|\boldsymbol{y}_i - \boldsymbol{y}_j\right|^2\right\}(\sum_{k=1}^{K} B_k(l)^2 + \tau^2)\right)$, and $\Phi_\epsilon(x) = \begin{cases} 0, & x < \epsilon \\ 1, & x \ge \epsilon \end{cases}$.

Proof. According to the property of the MLE for covariance matrix, $\hat{\Sigma}_{T(i-1)+l,T(j-1)+l}$ converges to $\Sigma_{T(i-1)+l,T(j-1)+l}$ in probability as sample size $m \to \infty$. Note that $\sigma_{i,j}\exp\left\{-\theta_y\left|\boldsymbol{y}_i - \boldsymbol{y}_j\right|^2\right\}(\sum_{k=1}^{K} B_k(l)^2 + \tau^2) = \Sigma_{T(i-1)+l,T(j-1)+l}$, for $l = 1, 2, \cdots, T$. We can then derive that

$$\left|\pi_{ij} - \sigma_{ij}\right| = \frac{1}{T}\left|\sum_{l=1}^{T}\left(\left(\hat{\Sigma}_{T(i-1)+l,T(j-1)+l} - \Sigma_{T(i-1)+l,T(j-1)+l}\right)/\exp\left\{-\theta_y\left|\boldsymbol{y}_i - \boldsymbol{y}_j\right|^2\right\}\left(\sum_{k=1}^{K} B_k(l)^2 + \tau^2\right)\right)\right|$$

$$\le \frac{1}{T\exp\left\{-\theta_y\left|\boldsymbol{y}_i - \boldsymbol{y}_j\right|^2\right\}(B_m + \tau^2)}\sum_{l=1}^{T}\left|\hat{\Sigma}_{T(i-1)+l,T(j-1)+l} - \Sigma_{T(i-1)+l,T(j-1)+l}\right|$$

where $B_m = \max_l \sum_{k=1}^{K} B_k(l)^2$. According to the convergence of the sample covariance, for any $\delta > 0$, there exists a large enough M_1 such that when $m > M_1$, $\left|\hat{\Sigma}_{T(i-1)+l,T(j-1)+l} - \Sigma_{T(i-1)+l,T(j-1)+l}\right| \leq \exp\left\{-\theta_y \left|\boldsymbol{y}_i - \boldsymbol{y}_j\right|^2\right\}(B_m + \tau^2)\delta$ holds. Then

$$\left|\pi_{ij} - \sigma_{ij}\right| \leq \frac{1}{T \exp\left\{-\theta_y \left|\boldsymbol{y}_i - \boldsymbol{y}_j\right|^2\right\}(B_m + \tau^2)} T \exp\left\{-\theta_y \left|\boldsymbol{y}_i - \boldsymbol{y}_j\right|^2\right\}(B_m + \tau^2)\delta = \delta.$$

For the thresholding estimator $\Phi_\epsilon(\pi_{i,j})$, it can be easily observed that $\Phi_\epsilon(\pi_{i,j}) = \sigma_{ij}$ when $\left|\pi_{ij} - \sigma_{ij}\right| < \min(\epsilon, 1-\epsilon)$. By the same argument above, we can show there exists a large enough M_2 such that when $m > M_2$, $\left|\pi_{ij} - \sigma_{ij}\right| < \min(\epsilon, 1-\epsilon)$. Therefore, we have demonstrated that when $m > M_2$, $\left|\Phi_\epsilon(\pi_{i,j}) - \sigma_{ij}\right| = 0$. This shows the consistency of the thresholding estimator $\Phi_\epsilon(\pi_{i,j})$.

Appendix D: First-order descent method to estimate the parameters in the M-step.

Recall that the optimization problem is formulated as follows:

$$\left(\theta_y^{(s+1)}, \theta_t^{(s+1)}, \tau^{(s+1)}\right) = \arg\min_{\theta_y,\theta_t,\tau} \left(\left\|\hat{\boldsymbol{\Sigma}}^{\boldsymbol{\Theta}} - \hat{\boldsymbol{\Sigma}}\right\|_F^2\right) = \arg\min_{\theta_y,\theta_t,\tau} \left(\sum_{i=1}^{JT} \sum_{j=1}^{JT} \left(\hat{\Sigma}_{i,j}^{\boldsymbol{\Theta}} - \hat{\Sigma}_{i,j}\right)^2\right).$$

The objective function can be further derived as

$$\mathcal{F} = \left\|\hat{\boldsymbol{\Sigma}}^{\boldsymbol{\Theta}} - \hat{\boldsymbol{\Sigma}}\right\|_F^2 = \sum_{i=1}^{JT} \sum_{j=1}^{JT} \left(\hat{\Sigma}_{i,j}^{\boldsymbol{\Theta}} - \hat{\Sigma}_{i,j}\right)^2$$

$$= \sum_{i=1}^{J} \sum_{j=1}^{J} \sum_{l=1}^{T} \sum_{p=1}^{T} \left[\sigma_{i,j} \exp\left\{-\theta_y \left|\boldsymbol{y}_i - \boldsymbol{y}_j\right|^2\right\} \left(\sum_{k=1}^{K} B_k(l) B_k(p) + \tau^2 \exp\left\{-\theta_t \left|t_l - t_p\right|\right\}\right) - \hat{\Sigma}_{T(i-1)+l,T(j-1)+p}\right]^2$$

$$= \sum_{i=1}^{J} \sum_{j=1}^{J} \sum_{l=1}^{T} \sum_{p=1}^{T} g(i,j,l,p)^2,$$

where $g(i,j,l,p) = \sigma_{i,j} \exp\left\{-\theta_y \left|\boldsymbol{y}_i - \boldsymbol{y}_j\right|^2\right\} \left(\sum_{k=1}^{K} B_k(l) B_k(p) + \tau^2 \exp\{-\theta_t |t_l - t_p|\}\right) - \widehat{\boldsymbol{\Sigma}}_{T(i-1)+l,T(j-1)+p}$. Then the derivative of the objective function with regard to the parameters $\nabla \mathcal{F}(\theta_y, \theta_t, \tau) = \left(\frac{\partial \mathcal{F}}{\partial \theta_y}, \frac{\partial \mathcal{F}}{\partial \theta_t}, \frac{\partial \mathcal{F}}{\partial \tau}\right)$ can be derived as

$$\frac{\partial \mathcal{F}}{\partial \theta_y} = \sum_{i=1}^{J}\sum_{j=1}^{J}\sum_{l=1}^{T}\sum_{p=1}^{T} \frac{\partial g^2(i,j,l,p)}{\partial \theta_y} = -2 \sum_{i=1}^{J}\sum_{j=1}^{J}\sum_{l=1}^{T}\sum_{p=1}^{T} g\left(g + \widehat{\boldsymbol{\Sigma}}_{n(i-1)+l,T(j-1)+p}\right)\left|\boldsymbol{y}_i - \boldsymbol{y}_j\right|^2,$$

$$\frac{\partial \mathcal{F}}{\partial \theta_t} = \sum_{i=1}^{J}\sum_{j=1}^{J}\sum_{l=1}^{T}\sum_{p=1}^{T} \frac{\partial g^2(i,j,l,p)}{\partial \theta_t}$$

$$= -2 \sum_{i=1}^{J}\sum_{j=1}^{J}\sum_{l=1}^{T}\sum_{p=1}^{T} \frac{\tau^2 g\left(g + \widehat{\boldsymbol{\Sigma}}_{n(i-1)+l,T(j-1)+p}\right) \exp\{-\theta_t |t_l - t_p|\} |t_l - t_p|}{\sum_{k=1}^{K} B_k(l) B_k(p) + \tau^2 \exp\{-\theta_t |t_l - t_p|\}},$$

$$\frac{\partial \mathcal{F}}{\partial \tau} = \sum_{i=1}^{J}\sum_{j=1}^{J}\sum_{l=1}^{T}\sum_{p=1}^{T} \frac{\partial g^2(i,j,l,p)}{\partial \tau} = 4 \sum_{i=1}^{J}\sum_{j=1}^{J}\sum_{l=1}^{T}\sum_{p=1}^{T} g \sigma_{i,j} \exp\left\{-\theta_y \left|\boldsymbol{y}_i - \boldsymbol{y}_j\right|^2\right\} \exp\{-\theta_t |t_l - t_p|\} \tau.$$

Then the gradient descent method searches the optimal parameters iteratively, and at step $v + 1$,

$$\left(\theta_y^{(v+1)}, \theta_t^{(v+1)}, \tau^{(v+1)}\right) = \left(\theta_y^{(v)}, \theta_t^{(v)}, \tau^{(v)}\right) - \alpha \cdot \nabla \mathcal{F}\left(\theta_y^{(v)}, \theta_t^{(v)}, \tau^{(v)}\right).$$

Here, α is the step length, the value of which is determined by a backtracking line search method.

References

Chib, S., Greenberg, E., and Winkelmann, R. (1998), "Posterior Simulation and Bayes Factors in Panel Count Data Models," *Journal of Econometrics*, 86, 33–54.

Wu, H., Deng, X., and Ramakrishnan, N. (2018), "Sparse Estimation of Multivariate Poisson Log-Normal Models From Count Data," *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 11, 66–77.