Biometrika (2016), **1**, 1, *pp.* 1–25
Printed in Great Britain Advance Access publication on 31 July 2016

Supplementary material for: A semiparametric extension of the stochastic block model for longitudinal networks

BY C. MATIAS, T. REBAFKA AND F. VILLERS

*Sorbonne Université, Université Paris Diderot, Centre National de la Recherche Scientifique,
Laboratoire de Probabilités, Statistique et Modélisation,
75005 Paris, France.*

catherine.matias@upmc.fr tabea.rebafka@upmc.fr fanny.villers@upmc.fr

All the references are from the main manuscript, except for those appearing as S-xx that are within this document.

S.1. IDENTIFIABILITY PROOFS

Proof of Proposition 1. The proof follows ideas similar to those of Theorem 12 in Allman et al. [2011]. For notational convenience this proof is presented in the undirected setup where the set of intensities is $\alpha = \{\alpha^{(q,l)} : q, l = 1, \dots, Q;\ q \leq l\}$. The directed case is treated in the same way.

To explain the general idea of the proof we start by considering the distribution of one marginal process $N_{i,j}$. This is a Cox process directed by the random measure

$$A_{i,j} \sim \sum_{q=1}^{Q} \sum_{l=1}^{Q} \pi_q \pi_l \delta_{A^{(q,l)}}.$$

For any $q \leq l$, we use $A^{(q,l)}$ for the measure on $[0, T]$ defined by $A^{(q,l)}(I) = \int_I \alpha^{(q,l)}(u) du$ for all measurable $I \subset [0, T]$. We recall that δ_u is the Dirac mass at point u. It is known that the mapping of probability laws of random measures into laws of Cox processes directed by them is a bijection, see for example Proposition 6.2.II in Daley and Vere-Jones [2003]. In other words the distribution of $N_{i,j}$ uniquely determines the finite measure

$$\sum_{q=1}^{Q} \sum_{l=1}^{Q} \pi_q \pi_l \delta_{A^{(q,l)}},$$

on the set of measures on $[0, T]$. According to Assumption 1 the intensities $\alpha^{(q,l)}$ are distinct. Hence, the corresponding measures $A^{(q,l)}$ are all different and we may recover from the distribution of our counting process $N_{i,j}$ the set of values $\{(\pi_q^2, A^{(q,q)}) : q = 1, \dots, Q\} \cup \{(2\pi_q\pi_l, A^{(q,l)}) : q, l = 1, \dots, Q;\ q < l\}$ or equivalently the set $\{(\pi_q^2, \alpha^{(q,q)}) : q = 1, \dots, Q\} \cup \{(2\pi_q\pi_l, \alpha^{(q,l)});\ q, l = 1, \dots, Q;\ q < l\}$. In particular we recover the functions $\alpha^{(q,l)}$ almost everywhere on $[0, T]$ up to a permutation on the pairs of groups (q, l). However for the recovery up to a permutation in $\mathfrak{S}_Q$ it is necessary to consider higher-order marginals.

We now fix three distinct integers i, j, k in $\{1, \dots, n\}$ and consider the trivariate counting process $(N_{i,j}, N_{i,k}, N_{j,k})$. In the same way, these are Cox processes directed by the triplet of random measures $(A_{i,j}, A_{i,k}, A_{j,k})$ such that

$$(A_{i,j}, A_{i,k}, A_{j,k}) \sim \sum_{q=1}^{Q} \sum_{l=1}^{Q} \sum_{m=1}^{Q} \pi_q \pi_l \pi_m \delta_{(A^{(q,l)}, A^{(q,m)}, A^{(l,m)})}.$$

We write this distribution in such a way that distinct components appear only once

$$\sum_{q=1}^{Q}\pi_q^3\delta_{(A^{(q,q)},A^{(q,q)},A^{(q,q)})} + \sum_{q=1}^{Q}\sum_{\substack{l=1\\ l\neq q}}^{Q}\pi_q^2\pi_l\Big\{\delta_{(A^{(q,q)},A^{(q,l)},A^{(q,l)})}+\delta_{(A^{(q,l)},A^{(q,q)},A^{(q,l)})}+\delta_{(A^{(q,l)},A^{(q,l)},A^{(q,q)})}\Big\} + \sum_{q=1}^{Q}\sum_{\substack{l=1\\ l\neq q}}^{Q}\sum_{\substack{m=1\\ m\neq q,l}}^{Q}\pi_q\pi_l\pi_m\delta_{(A^{(q,l)},A^{(q,m)},A^{(l,m)})}. \tag{S.1}$$

Using the same reasoning we identify the triplets of values $\{(A^{(q,l)},A^{(q,m)},A^{(l,m)}) : q,l,m=1,\dots,Q\}$ up to a permutation on the triplets (q,l,m). Among these, the only values with three identical components are $\{(A^{(q,q)};A^{(q,q)};A^{(q,q)}) : q=1,\dots,Q\}$ and thus the measures $\{A^{(q,q)} : q=1,\dots,Q\}$ are identifiable up to a permutation in $\mathfrak{S}_Q$. Going back to (S.1) and looking for the Dirac terms at points that have two identical components, namely of the form $(A^{(q,q)},A^{(q,l)},A^{(q,l)})$ and two other similar terms with permuted components, we can now identify the set of measures

$$\{(A^{(q,q)},\{A^{(q,l)} : l=1,\dots,Q;\ l\neq q\}) : q=1,\dots,Q\}.$$

This is equivalent to saying that we identify the measures $\{A^{(q,l)} : q,l=1,\dots,Q;\ q\leq l\}$ up to a permutation in $\mathfrak{S}_Q$. Obviously this also identifies the corresponding intensities $\{\alpha^{(q,l)} : q,l=1,\dots,Q;\ q\leq l\}$ almost everywhere on $[0,T]$ up to a permutation in $\mathfrak{S}_Q$. To finish the proof we need to identify the proportions π_q. As we identified the components $\{A^{(q,q)} : q=1,\dots,Q\}$, we recover from (S.1) the set of values $\{\pi_q^3 : q=1,\dots,Q\}$ up to the same permutation as on the $A^{(q,q)}$'s. This concludes the proof. □

Proof of Proposition 2. The setup considered here is undirected. We follow some of the arguments already appearing in the proof of Proposition 1. Let A^{in} and A^{out} denote the measures whose intensities are α^{in} and α^{out}, respectively. The univariate process $N_{i,j}$ is a Cox process directed by the random measure $A_{i,j}$ that is distributed as

$$A_{i,j}\sim(\sum_{q=1}^{Q}\pi_q^2)\delta_{A^{\text{in}}}+(\sum_{q=1}^{Q}\sum_{\substack{l=1\\ l\neq q}}^{Q}\pi_q\pi_l)\delta_{A^{\text{out}}}.$$

Thus the measures A^{in} and A^{out} are identifiable from the distribution of $N_{i,j}$, but only up to a permutation. Similarly to the previous proof we rather consider the trivariate Cox processes $(N_{i,j},N_{i,k},N_{j,k})$ directed by the random measures $(A_{i,j},A_{i,k},A_{j,k})$ whose distribution in the affiliation case has now five atoms

$$\Big(\sum_{q=1}^{Q}\pi_q^3\Big)\delta_{(A^{\text{in}},A^{\text{in}},A^{\text{in}})}+\Big(\sum_{q=1}^{Q}\sum_{\substack{l=1\\ l\neq q}}^{Q}\pi_q^2\pi_l\Big)\delta_{(A^{\text{in}},A^{\text{out}},A^{\text{out}})}+\Big(\sum_{q=1}^{Q}\sum_{\substack{l=1\\ l\neq q}}^{Q}\pi_q^2\pi_l\Big)\delta_{(A^{\text{out}},A^{\text{in}},A^{\text{out}})} + \Big(\sum_{q=1}^{Q}\sum_{\substack{l=1\\ l\neq q}}^{Q}\pi_q^2\pi_l\Big)\delta_{(A^{\text{out}},A^{\text{out}},A^{\text{in}})}+\Big(\sum_{q=1}^{Q}\sum_{\substack{l=1\\ l\neq q}}^{Q}\sum_{\substack{m=1\\ m\neq q,l}}^{Q}\pi_q\pi_l\pi_m\Big)\delta_{(A^{\text{out}},A^{\text{out}},A^{\text{out}})}.$$

As previously, these five components are identifiable up to a permutation on $\mathfrak{S}_5$. Now it is easy to identify the three components for which two marginals have same parameters and the third one has a different parameter. Thus we recover exactly the measures A^{in} and A^{out}. This also identifies the corresponding intensities α^{in} and α^{out} almost everywhere on $[0,T]$.

Now identification of the proportions $\{\pi_q : q=1,\dots,Q\}$ follows an argument already used in the proof of Theorem 13 in Allman et al. [2011] that we recall here for completeness. From the trivariate distribution of $(N_{i,j}, N_{i,k}, N_{j,k})$ and the already recovered values A^{in} and A^{out}, we identify the proportion $\sum_{q=1}^{Q} \pi_q^3$. Similarly for any $n \geq 1$, by considering the multivariate distribution of $(N_{i,j})_{(i,j)\in\mathcal{R}}$, we can identify the Dirac mass at point $(A^{\text{in}}, \dots, A^{\text{in}})$ and thus its weight which is equal to $\sum_{q=1}^{Q} \pi_q^n$. By the Newton identities the values $\{\sum_{q=1}^{Q} \pi_q^n : n = 1, \dots, Q\}$ determine the values of elementary symmetric polynomials $\{\sigma_n(\pi_1, \dots, \pi_Q) : n = 1, \dots, Q\}$. These, in turn, are up to sign the coefficients of the monic polynomial whose roots with multiplicities are precisely $\{\pi_q : q = 1, \dots, Q\}$. Thus the proportion parameters are recovered up to a permutation. □

S.2. Variational E-step: Proof of Proposition 3

For the Kullback-Leibler divergence we compute

$$\begin{aligned}\text{KL}\{\text{pr}_\tau(\cdot \mid \mathcal{O}) \| \text{pr}_\theta(\cdot \mid \mathcal{O})\} &= E_\tau \left\{ \log \frac{\text{pr}_\tau(\mathcal{Z} \mid \mathcal{O})}{\text{pr}_\theta(\mathcal{Z} \mid \mathcal{O})} \mid \mathcal{O} \right\} = E_\tau \left\{ \log \frac{\text{pr}_\tau(\mathcal{Z} \mid \mathcal{O})\text{pr}_\theta(\mathcal{O})}{\mathcal{L}(\mathcal{O}, \mathcal{Z} \mid \theta)} \mid \mathcal{O} \right\} \\ &= \sum_{i=1}^{n} E_\tau \left(\log \tau^{i,Z_i} \mid \mathcal{O} \right) + \log \text{pr}_\theta(\mathcal{O}) - E_\tau \{ \log \mathcal{L}(\mathcal{O}, \mathcal{Z} \mid \theta) \mid \mathcal{O} \}.\end{aligned}$$

The complete-data log-likelihood $\log \mathcal{L}(\mathcal{O}, \mathcal{Z} \mid \theta)$ is

$$-\sum_{q=1}^{Q} \sum_{l=1}^{Q} Y_{\mathcal{Z}}^{(q,l)} A^{(q,l)}(T) + \sum_{q=1}^{Q} \sum_{l=1}^{Q} \sum_{m=1}^{M} Z_m^{(q,l)} \log \left\{ \alpha^{(q,l)}(t_m) \right\} + \sum_{i=1}^{n} \sum_{q=1}^{Q} Z^{i,q} \log \pi_q,$$

where $Y_{\mathcal{Z}}^{(q,l)}$ and $Z_m^{(q,l)}$ have been introduced in Equations (1) and (3), respectively. Now note that $E_\tau(Z^{i,q} \mid \mathcal{O}) = \text{pr}_\tau(Z^{i,q} = 1 \mid \mathcal{O}) = \text{pr}_\tau(Z_i = q \mid \mathcal{O}) = \tau^{i,q}$. Moreover by the factorised form of pr_τ, for every $i \neq j$ we have

$$E_\tau(Z^{i,q} Z^{j,l} \mid \mathcal{O}) = E_\tau(Z^{i,q} \mid \mathcal{O}) E_\tau(Z^{j,l} \mid \mathcal{O}) = \tau^{i,q} \tau^{j,l}.$$

The quantity $Y^{(q,l)}$ is thus equal to $E_\tau(Y_{\mathcal{Z}}^{(q,l)} \mid \mathcal{O})$, namely the variational approximation of the mean number of dyads with latent groups (q, l). Similarly $\tau_m^{(q,l)}$ equals $E_\tau(Z_m^{(q,l)} \mid \mathcal{O})$, the variational approximation of the probability that observation (t_m, i_m, j_m) corresponds to a dyad with latent groups (q, l). It follows that

$$\hat{\tau} = \arg\min_{\tau \in \mathcal{T}} \text{KL}\{\text{pr}_\tau(\cdot \mid \mathcal{O}) \| \text{pr}_\theta(\cdot \mid \mathcal{O})\} = \arg\max_{\tau \in \mathcal{T}} J(\theta, \tau),$$

where $J(\theta, \tau)$ is

$$-\sum_{q=1}^{Q} \sum_{l=1}^{Q} Y^{(q,l)} A^{(q,l)}(T) + \sum_{q=1}^{Q} \sum_{l=1}^{Q} \sum_{m=1}^{M} \tau_m^{(q,l)} \log \left\{ \alpha^{(q,l)}(t_m) \right\} + \sum_{i=1}^{n} \sum_{q=1}^{Q} \tau^{i,q} \log \left(\frac{\pi_q}{\tau^{i,q}} \right). \qquad \text{(S.2)}$$

The variational E-step consists in maximizing J with respect to the $\tau^{i,q}$'s which are constrained to satisfy $\sum_{q=1}^{Q} \tau^{i,q} = 1$ for all i. In other words we maximize

$$M(\tau, \gamma) = J(\theta, \tau) + \sum_{i=1}^{n} \gamma_i \left(\sum_{q=1}^{Q} \tau^{i,q} - 1 \right),$$

with Lagrange multipliers γ_i. The partial derivatives are

$$\frac{\partial}{\partial \tau^{i,q}} M(\tau,\gamma) = -\sum_{l=1}^{Q}\sum_{j\neq i} \tau^{j,l}\left\{A^{(q,l)}(T) + A^{(l,q)}(T)\right\} + \sum_{l=1}^{Q}\sum_{m=1}^{M} \mathbb{1}_{\{i_m=i\}}\tau^{j_m,l}\log\left\{\alpha^{(q,l)}(t_m)\right\}$$
$$+ \sum_{l=1}^{Q}\sum_{m=1}^{M} \mathbb{1}_{\{j_m=i\}}\tau^{i_m,l}\log\left\{\alpha^{(l,q)}(t_m)\right\} + \log\left(\frac{\pi_q}{\tau^{i,q}}\right) - 1 + \gamma_i,$$
$$\frac{\partial}{\partial \gamma_i} M(\tau,\gamma) = \sum_{q=1}^{Q} \tau^{i,q} - 1.$$

The partial derivatives are null if and only if $\sum_{q=1}^{Q} \tau^{i,q} = 1$ and the $\tau^{i,q}$'s satisfy the fixed point equations (6), with $\exp(\gamma_i - 1)$ being the normalizing constant.

S.3. DETAILS ON THE ALGORITHM

Initialisation is a crucial point for any clustering method. Our variational expectation-maximization algorithm starts with a classification of the nodes and iterates an M-step followed by a variational E-step. We apply the algorithm on multiple initial classifications of the nodes based on various aggregations of the data: on the whole time interval and on sub-intervals. Sub-intervals are obtained through regular dyadic partitions of $[0,T]$, this is parameter `l.part` in the R code, and a k-means algorithm applied on the rows of the adjacency matrices of each of the aggregated datasets provides starting points for our algorithm. To obtain further starting values we use perturbations of the different k-means classifications: a given percentage of the total number of individuals is picked at random, this is parameter `perc.perturb`, and their group memberships are shuffled. The perturbation procedure can be applied several times, this is parameter `n.perturb`. The algorithm returns as final result the run that achieves the largest value of criterion J.

There is no theoretical result on the existence of a solution to the fixed point equation (6). The iterations for the fixed point equation are initialized with the value of τ obtained at the previous variational E-step. In practice, convergence is fast and we stop the fixed-point iterations either when convergence is achieved $|\tau^{[s]} - \tau^{[s-1]}| < \varepsilon = 10^{-6}$ or when the maximal number of iterations is attained, here `fix.iter`=10.

As the variational expectation-maximization algorithm aims at maximizing J defined in (S.2), the algorithm is stopped when the increase of J is less than a given threshold, here $\varepsilon = 10^{-6}$, that is when

$$\left|\frac{J(\theta^{[s+1]},\tau^{[s+1]}) - J(\theta^{[s]},\tau^{[s]})}{J(\theta^{[s]},\tau^{[s]})}\right| < \varepsilon,$$

or when the maximal number of iterations has been attained, here `nb.iter`=50.

S.4. ADDITIONAL TABLES AND FIGURES

Figure S.1 shows the intensities used in Scenario 1 from the synthetic experiments to assess the clustering performances of our method.

Table S.1 gives the risks $\text{RISK}(q,l)$ and standard deviations of the histogram and the kernel versions of our method as well as oracle quantities obtained with known group labels in Scenario 2 when $n = 20$. This is the analogue of Table 1 in the main manuscript where $n = 50$.

Figure S.2 shows boxplots of the adjusted Rand index obtained from the synthetic experiments from Scenario 2. They are computed over 1000 datasets with different numbers n of individuals and for the two estimation methods: histogram and kernel.

Figure S.3 shows the model selection results on the number of groups based on the integrated classification likelihood in Scenario 2 with $n = 20$: the left panel shows the frequency of the selected values $\hat{Q}$ over the 1000 datasets; the right panel shows boxplots of the adjusted Rand index between the estimated classification with 3 groups and the true latent structure as a function of the number of groups selected by the integrated classification likelihood criterion. On this right panel, the adjusted Rand index of the classification with three groups is rather low when the criterion does not select the correct number Q, indicating that the algorithm has failed in the classification task and probably only a local maximum of the criterion J has been found.

Turning to the London bike-sharing system dataset, Figure S.4 shows the temporal profiles of the 2 stations in the smallest cluster for day 1. One can see that these are outgoing stations around 8am and incoming stations between 5 and 7pm. Figure S.5 shows the highest intensities estimated by our model between these 6 clusters, all other intensities are almost null. The most important interactions occur from the smallest cluster that is number 4 to cluster 5 'City of London' cluster, in the morning and conversely from cluster 5 to cluster 4 at the end of the day.

Concerning the analysis of the Enron dataset we provide here additional tables and figures for $Q = 4$ groups. Table S.2 gives the size and composition of the four groups. For a part of the persons in the dataset the position at Enron is not available. This is the reason why the total size of the group sometimes exceeds the sum of the number of managers and employees in the group.

Figure S.6 gives the logarithm of the mean values of the estimated intensities $\alpha^{(q,l)}$. The lack of symmetry of the matrix indicates that communication is far from being symmetric and that the use of the directed model is appropriate.

Finally, Figure S.7 shows the estimated intensities and associated bootstrap confidence intervals. The bootstrap intervals are obtained by parametric bootstrap. More precisely, every bootstrap sample contains the same number of individuals, here $n = 147$ and for every individual i the group membership Z_i^* is drawn from the multinomial distribution $\mathcal{M}(1, \hat{\pi})$ where $\hat{\pi}$ is the vector of group probabilities estimated from the data. Then for every pair of individuals (i, j) realizations from a Poisson process with intensity $\hat{\alpha}^{(Z_i^*, Z_j^*)}$ are simulated, where $\{\hat{\alpha}^{(q,l)}\}_{q,l=1,\dots,Q}$ denote the estimated intensities. Finally, bootstrap confidence intervals are obtained by the percentile method.

Here the bootstrap intervals suffer from the fact that some of the group probabilities $\hat{\pi}_k$ are very low, implying that the probability that a bootstrap sample contains empty groups is relatively high (0.15). That is, about 15% of the bootstrap samples do not provide any information on some of the intensities $\alpha^{(q,l)}$, implying that the associated estimators are completely erroneous. The groups that are the most concerned by the problem are group 2 and 3.

S.5. Additional example: Primary school temporal network dataset

To understand contacts between children at school and to quantify the transmission opportunities of respiratory infections, data on face-to-face interactions in a French primary school were collected. The dataset is presented in detail in Stehlé et al. [2011] and available online [SocioPatterns, 2015]. Children are aged from 6 to 12 years and the school is composed of five grades, each of them comprising two classes, for a total of 10 classes, denoted by $1A, 1B, \dots, 5A, 5B$. Each class has an assigned teacher and an assigned room. The school day runs from 8.30am to 4.30pm, with a lunch break from 12pm to 2pm and two breaks of 20-25 min in the morning and in the afternoon. Lunch is served in a common canteen and a shared playground is located outside the main building. As the playground and the canteen do not have enough capacity to host all pupils at a time, only two or three classes have breaks together, and lunch is served in two turns. The dataset contains $125, 773$ face to face contacts among $n = 242$ individuals, 232 children and 10 teachers, observed during two days. We applied our procedure in the undirected setup with histograms based on a regular dyadic partitions with maximum size 256, corresponding to $d_{\max} = 8$.

The integrated classification likelihood criterion achieves its maximum with $\hat{Q} = 17$ latent groups. Figure S.8 shows the clustering of the n individuals into the 17 groups, where children from different classes are represented with different colors. Some groups correspond exactly or almost exactly to classes: for example, group 9 consists almost perfectly of class 1A, whereas other classes are split into several groups;

for example class 1B is split into groups 1 and 16. Moreover, one group which is group 6, corresponds to the entire class 4B with, in addition, pupils coming from almost all other classes. Teachers never form a particular group apart, but they are generally in the cluster of their assigned class.

The highest intensities are the intra-group intensities. As groups mainly correspond to classes, this highlights that most contacts involve individuals of the same class and that the dataset is structured into communities, i.e. groups of highly connected individuals and with few inter-groups interactions. Figure S.9 shows the estimated intra-group intensities for each group with at least 3 individuals. Peaks of interactions are observed during the two breaks in the morning and in the afternoon. At lunch time interactions between children vary from the first to the second day and are less important than during the breaks when they play together. We also observe periods with no interaction at all. For example, the estimated intensity of group 9 corresponding to class 1A, is null between 3:30am and 4am suggesting that some particular school activity like sports takes place during which contacts were not observable for technical reasons. The group number 6 composed with the entire class 4B and others pupils clearly appears as the group with the lowest intra-group intensity. This means that this cluster gathers the individuals having less interactions with others. Class 4B also appears as the class having the least intra-class interactions in Stehlé et al. [2011].

Concerning inter-group connections most of the estimated intensities for groups (q, l) with $q \neq l$ can be considered as null, except for some that we discuss now. First, as our procedure splits some children of the same class into separate groups, the inter-group interactions associated with these clusters correspond in fact to intra-class interactions. For example, class 1B is split into group 1 with 18 pupils and group 16 with 7 pupils. The estimated inter-group intensity shows that those two groups interact. Our clustering has formed two separate groups because group 1 has more intra-group interactions than the other, see Figure S.9.

Second, intensities between groups made of children of the same grade are significant, suggesting that children mostly interact with children of the same age, see e.g. Figure S.10 that shows the case $(q, l) = (5, 13)$. Those interactions are observed during the two breaks in the morning and in the afternoon as well as at lunch time.

Third, the estimated intensities suggest particular behaviour of some pupils. Consider for example class 2B, except the two pupils assigned to group 6, which is separated into group 12 with 21 pupils, group 11 with 2 pupils and group 17 with only one pupil. The estimated intensities in Figure S.11 suggest first that the two children in group 11 have very strong interaction with the pupil in group 17: notice the different y-scale used in the Figure; and second that those interactions do not occur during the lunch time.

Similar results are obtained for classes 2A and 5B. This means that our procedure detects subgroups of pupils with a specific behaviour.

S.6. THE SPARSE SETUP: THEORY

We consider an extended setup where some of the processes $N_{i,j}$ may have a null intensity. We thus introduce additional latent variables $U_{i,j} \in \{0, 1\}$, $((i, j) \in \mathcal{R})$ that conditional on the Z_i's are independent Bernoulli with $\beta_{q,l}$ being the parameter of the distribution of $U_{i,j}$ conditional on $Z^{i,q}Z^{j,l} = 1$. We keep the global conditional independence assumption by imposing that conditional on $(Z_i, U_{i,j})_{(i,j)\in\mathcal{R}}$ the counting processes $(N_{i,j})_{(i,j)\in\mathcal{R}}$ are independent. Then conditional on $U_{i,j}, Z_i, Z_j$, the counting process $N_{i,j}$ is an inhomogeneous Poisson process with intensity

$$U_{i,j}\alpha^{Z_i,Z_j} = \sum_{q,l=1}^{Q} U_{ij} Z^{i,q} Z^{j,l} \alpha^{(q,l)}, \quad ((i, j) \in \mathcal{R}).$$

In this way the additional latent variable $U_{i,j}$ accounts for sparsity in the interaction processes; in each pair of groups (q, l), there is now a proportion $1 - \beta_{q,l}$ of dyads $(i, j) \in \mathcal{R}$ that do not interact, so that the corresponding process $N_{i,j}$ is almost surely 0. As such these non interacting dyads will not tend to decrease the estimate of the common intensity $\alpha^{(q,l)}$. Moreover clustering in this model should give different groups, less driven by the absence of interactions.

We let $\mathcal{U} = (U_{i,j})_{(i,j)\in\mathcal{R}}$ and the parameter value is $\theta = (\pi, \beta, \alpha)$.

Identifiability may be proved under the same assumptions, requiring moreover that none of the intensities $\alpha^{(q,l)}$ is itself equal to zero. We discuss this in the undirected case, similarly to the identifiability proof of the main model. Indeed $N_{i,j}$ is now a counting process directed by the random measure $A_{i,j}$ whose distribution is

$$A_{i,j} \sim \sum_{q=1}^{Q}\sum_{l=1}^{Q} \pi_q \pi_l \{\beta_{q,l}\delta_{A^{(q,l)}} + (1-\beta_{q,l})\delta_0\}.$$

Fixing three distinct integers i, j, k in $\{1, \ldots, n\}$ and considering the trivariate counting process $(N_{i,j}, N_{i,k}, N_{j,k})$ we end up with the distribution of the triplet of random measures $(A_{i,j}, A_{i,k}, A_{j,k})$. There the expressions become more cumbersome but the very same reasoning may be applied to identify the measures $\{A^{(q,l)} : q, l = 1, \ldots, Q;\ q \leq l\}$ up to a permutation in $\mathfrak{S}_Q$. Concerning identification of π and β, we obtain the set of weights $\{\pi_q^3\beta_{q,q}^3 : q = 1, \ldots, Q\}$ that is attached to the Q components corresponding to Dirac masses at points of the form $(A^{(q,q)}, A^{(q,q)}, A^{(q,q)})$. Moreover we also obtain the set of weights $\{\pi_q^3\beta_{q,q}^2(1-\beta_{q,q}) : q = 1, \ldots, Q\}$ attached to Dirac masses at points $(A^{(q,q)}, A^{(q,q)}, 0)$. As the $A^{(q,q)}$'s are unique we can match the value $\pi_q^3\beta_{q,q}^3$ with $\pi_q^3\beta_{q,q}^2(1-\beta_{q,q})$ and thus obtain (through a simple ratio) $\beta_{q,q}$ and also π_q. In other words the sets $\{\beta_{q,q} : q = 1, \ldots, Q\}_q$ and $\{\pi_q : q = 1, \ldots, Q\}_q$ are identifiable. Finally we may look at the weights $\pi_q^2\pi_l\beta_{q,q}\beta_{q,l}^2$ associated with Dirac masses at points of the form $(A^{(q,q)}, A^{(q,l)}, A^{(q,l)})$ with $q \neq l$. As the cumulative intensities $\{A^{(q,l)} : q, l = 1, \ldots, Q;\ q \leq l\}$ have been identified up to a permutation in $\mathfrak{S}_Q$, together with the pair of $(\pi_q, \beta_{q,q})$'s for the same permutation, we obtain the values $\{\beta_{q,l} : q \neq l\}$.

Let us turn to inference of this model. To fix the notation we use the directed setup but similar equations may be derived in the undirected case. The complete-data likelihood is

$$\begin{aligned}
\mathcal{L}_{\text{sparse}}(\mathcal{O}, \mathcal{Z}, \mathcal{U} \mid \theta) =& \mathcal{L}(\mathcal{O} \mid \mathcal{Z}, \mathcal{U}, \theta) \times \mathcal{L}(\mathcal{U} \mid \mathcal{Z}, \theta) \times \mathcal{L}(\mathcal{Z} \mid \theta) \\
=& \exp\left\{ -\sum_{(i,j)\in\mathcal{R}} U_{i,j} A^{(Z_i,Z_j)}(T) \right\} \times \prod_{m=1}^{M} \alpha^{(Z_{i_m}, Z_{j_m})}(t_m) \\
& \times \prod_{(i,j)\in\mathcal{R}} \prod_{q=1}^{Q}\prod_{l=1}^{Q} \left\{ \beta_{q,l}^{U_{i,j}} (1-\beta_{q,l})^{1-U_{i,j}} \right\}^{Z^{i,q}Z^{j,l}} \times \prod_{i=1}^{n}\prod_{q=1}^{Q} \pi_q^{Z^{i,q}}.
\end{aligned}$$

The true conditional distribution of the latent variables $(\mathcal{Z}, \mathcal{U})$ given the observations writes

$$\mathrm{pr}_\theta(\mathcal{Z}, \mathcal{U} \mid \mathcal{O}) = \mathrm{pr}_\theta(\mathcal{Z} \mid \mathcal{O})\mathrm{pr}_\theta(\mathcal{U} \mid \mathcal{Z}, \mathcal{O}) = \mathrm{pr}_\theta(\mathcal{Z} \mid \mathcal{O}) \prod_{(i,j)\in\mathcal{R}} \mathrm{pr}_\theta(U_{i,j} \mid Z_i, Z_j, N_{i,j}).$$

A main difference with the previous setting is that now this conditional distribution has two parts: the one concerning $\mathcal{U}$ has a factorised form and can thus be computed exactly, while the part concerning $\mathcal{Z}$ still has an intricate dependence structure and we rely on a variational approximation to deal with it. We thus introduce a new conditional factorised distribution $\tilde{\mathrm{pr}}_{\tau,\theta}(\cdot \mid \mathcal{O})$ on the variables $\mathcal{Z}, \mathcal{U}$ that depends on the observations $\mathcal{O}$ and is defined as

$$\begin{aligned}
& \tilde{\mathrm{pr}}_{\tau,\theta}\left(\mathcal{Z} = (q_1, \ldots, q_n), \mathcal{U} = (u_{i,j})_{(i,j)\in\mathcal{R}} \mid \mathcal{O}\right) \\
&= \prod_{i=1}^{n} \mathrm{pr}_\tau(Z_i = q_i \mid \mathcal{O}) \prod_{(i,j)\in\mathcal{R}} \mathrm{pr}_\theta(U_{i,j} = u_{i,j} \mid Z_i = q_i, Z_j = q_j, N_{i,j}) \\
&= \prod_{i=1}^{n} \tau^{i,q_i} \prod_{(i,j)\in\mathcal{R}} \mathrm{pr}_\theta(U_{i,j} = u_{i,j} \mid Z_i = q_i, Z_j = q_j, N_{i,j}),
\end{aligned}$$

for any $(q_1,\dots,q_n)\in\{1,\dots,Q\}^n$ and $(u_{i,j})_{(i,j)\in\mathcal{R}}\in\{0,1\}^r$. As it does not depend on θ, we let $\tilde{\text{pr}}_\tau(\mathcal{Z}\mid\mathcal{O})$ denote the marginal distribution on $\mathcal{Z}$ of the distribution $\tilde{\text{pr}}_{\tau,\theta}(\cdot\mid\mathcal{O})$ and $\tilde{E}_\tau(\cdot\mid\mathcal{O})$ the corresponding expectation. Moreover, the true conditional distribution of $U_{i,j}$ is given by

$$\text{pr}_\theta(U_{i,j}=1\mid Z_i=q,Z_j=l,N_{i,j})=1\{N_{i,j}(T)>0\}+\rho_\theta(q,l)1\{N_{i,j}(T)=0\}:=\rho_\theta(i,j,q,l),$$
$$\text{where } \rho_\theta(q,l)=\frac{\beta_{q,l}\exp\{-A^{(q,l)}(T)\}}{1-\beta_{q,l}+\beta_{q,l}\exp\{-A^{(q,l)}(T)\}}. \tag{S.3}$$

Indeed, whenever we observe an interaction event between (i,j), namely $N_{i,j}(T)>0$, we know that $U_{ij}=1$ almost surely. Otherwise ($N_{i,j}(T)=0$), we either have a null intensity process or a non-null intensity process with zero observations. Note that the parameters $\rho_\theta(q,l)$, or equivalently the $\rho_\theta(i,j,q,l)$, are not additional variational parameters; these are just functions of the original parameter θ. Finally we have

$$\tilde{\text{pr}}_{\tau,\theta}(\mathcal{Z},\mathcal{U}\mid\mathcal{O})=\Big\{\prod_{i=1}^n\tilde{\text{pr}}_\tau(Z_i\mid\mathcal{O})\Big\}\times\prod_{(i,j)\in\mathcal{R}}\rho_\theta(i,j,Z_i,Z_j)^{U_{i,j}}\{1-\rho_\theta(i,j,Z_i,Z_j)\}^{1-U_{i,j}}.$$

Let us now derive our variational approximation. Denoting by $\tilde{E}_{\tau,\theta}(\cdot\mid\mathcal{O})$ the expectation under the distribution $\tilde{\text{pr}}_{\tau,\theta}(\cdot\mid\mathcal{O})$ on $(\mathcal{Z},\mathcal{U})$ and by $\theta^{[s]}$ the current parameter value, we write as usual

$$\begin{aligned}&\log\text{pr}_{\theta^{[s]}}(\mathcal{O})\\&=\tilde{E}_{\tau,\theta}\{\log\text{pr}_{\theta^{[s]}}(\mathcal{O},\mathcal{Z},\mathcal{U})\mid\mathcal{O}\}-\tilde{E}_{\tau,\theta}\{\log\text{pr}_{\theta^{[s]}}(\mathcal{Z},\mathcal{U}\mid\mathcal{O})\mid\mathcal{O}\}\\&=\tilde{E}_{\tau,\theta}\{\log\text{pr}_{\theta^{[s]}}(\mathcal{O},\mathcal{Z},\mathcal{U})\mid\mathcal{O}\}+\mathcal{H}\{\tilde{\text{pr}}_{\tau,\theta}(\mathcal{Z},\mathcal{U}\mid\mathcal{O})\}+\text{KL}\{\tilde{\text{pr}}_{\tau,\theta}(\mathcal{Z},\mathcal{U}\mid\mathcal{O})\|\text{pr}_{\theta^{[s]}}(\mathcal{Z},\mathcal{U}\mid\mathcal{O})\}.\end{aligned}$$

As a consequence, we introduce a new criterion $\tilde{J}(\tau,\theta;\theta^{[s]})$ that is a lower bound on the log-likelihood $\log\text{pr}_{\theta^{[s]}}(\mathcal{O})$ and defined as

$$\begin{aligned}\tilde{J}(\tau,\theta;\theta^{[s]})=&\tilde{E}_{\tau,\theta}\{\log\text{pr}_{\theta^{[s]}}(\mathcal{O},\mathcal{Z},\mathcal{U})\mid\mathcal{O}\}+\mathcal{H}\{\tilde{\text{pr}}_{\tau,\theta}(\mathcal{Z},\mathcal{U}\mid\mathcal{O})\}\\=&-\sum_{(i,j)\in\mathcal{R}}\sum_{q=1}^Q\sum_{l=1}^Q\tau^{i,q}\tau^{j,l}\rho_\theta(i,j,q,l)(A^{[s]})^{(q,l)}(T)\\&+\sum_{q=1}^Q\sum_{l=1}^Q\sum_{m=1}^M\tau^{i_m,q}\tau^{j_m,l}\log\Big\{(\alpha^{[s]})^{(q,l)}(t_m)\Big\}\\&+\sum_{(i,j)\in\mathcal{R}}\sum_{q=1}^Q\sum_{l=1}^Q\tau^{i,q}\tau^{j,l}\Big[\rho_\theta(i,j,q,l)\log\beta^{[s]}_{q,l}+\{1-\rho_\theta(i,j,q,l)\}\log(1-\beta^{[s]}_{q,l})\Big]\\&+\sum_{i=1}^n\sum_{q=1}^Q\tau^{i,q}\log\left(\frac{\pi^{[s]}_q}{\tau^{i,q}}\right)-\sum_{(i,j)\in\mathcal{R}}\sum_{q=1}^Q\sum_{l=1}^Q\tau^{i,q}\tau^{j,l}\psi\{\rho_\theta(i,j,q,l)\},\end{aligned}$$

where $\psi(\rho)=\rho\log\rho+(1-\rho)\log(1-\rho)$ is the entropy of the Bernoulli distribution with parameter ρ. Using the definition of $\rho_\theta(i,j,q,l)$ and $\psi(1)=0$, the last term in the right-hand side simplifies to

$$\sum_{(i,j)\in\mathcal{R}}\sum_{q=1}^Q\sum_{l=1}^Q\tau^{i,q}\tau^{j,l}\psi\{\rho_\theta(i,j,q,l)\}=\sum_{q=1}^Q\sum_{l=1}^Q\psi\{\rho_\theta(q,l)\}\sum_{(i,j)\in\mathcal{R}}\tau^{i,q}\tau^{j,l}1\{N_{i,j}(T)=0\}.$$

The variational E-step consists in maximizing $\tilde{J}(\tau,\theta;\theta^{[s]})$ with respect to (τ,θ). This is equivalent to choosing the variational distribution $\tilde{\text{pr}}_{\tau,\theta}$ that minimises the Kullback-Leibler divergence $\text{KL}\{\tilde{\text{pr}}_{\tau,\theta}(\mathcal{Z},\mathcal{U}\mid\mathcal{O})\|\text{pr}_{\theta^{[s]}}(\mathcal{Z},\mathcal{U}\mid\mathcal{O})\}$. The solution in θ is naturally obtained for $\theta=\theta^{[s]}$.

We need to choose the variational parameter τ that maximizes $\tilde{J}(\tau, \theta^{[s]}; \theta^{[s]})$. Similarly to the non sparse setup we obtain that $\tau^{[s]}$ satisfies a fixed point equation in τ,

$$\tau^{i,q} \propto \pi_q^{[s]} \exp\{\tilde{D}_{iq}(\tau, \theta^{[s]})\}, \quad (i = 1, \ldots, n;\ q = 1, \ldots, Q), \tag{S.4}$$

where

$$\begin{aligned}
\tilde{D}_{iq}(\tau, \theta) = &- \sum_{l=1}^{Q} \sum_{\substack{j=1 \\ j \neq i}}^{n} \tau^{j,l} \Big\{ \rho_\theta(i,j,q,l) A^{(q,l)}(T) + \rho_\theta(j,i,l,q) A^{(l,q)}(T) \Big\} \\
&- \sum_{l=1}^{Q} \psi\{\rho_\theta(q,l)\} \sum_{\substack{j=1 \\ j \neq i}}^{n} \tau^{j,l} 1\{N_{i,j}(T) = 0\} - \sum_{l=1}^{Q} \psi\{\rho_\theta(l,q)\} \sum_{\substack{j=1 \\ j \neq i}}^{n} \tau^{j,l} 1\{N_{j,i}(T) = 0\} \\
&+ \sum_{l=1}^{Q} \sum_{m=1}^{M} \Big[\mathbb{1}_{\{i_m = i\}} \tau^{j_m,l} \log\Big\{\alpha^{(q,l)}(t_m)\Big\} + \mathbb{1}_{\{j_m = i\}} \tau^{i_m,l} \log\Big\{\alpha^{(l,q)}(t_m)\Big\} \Big] \\
&+ \sum_{l=1}^{Q} \sum_{\substack{j=1 \\ j \neq i}}^{n} \tau^{j,l} [\rho_\theta(i,j,q,l) \log \beta_{q,l} + \{1 - \rho_\theta(i,j,q,l)\} \log(1 - \beta_{q,l}) \\
&\qquad + \rho_\theta(j,i,l,q) \log \beta_{l,q} + \{1 - \rho_\theta(j,i,l,q)\} \log(1 - \beta_{l,q})].
\end{aligned}$$

The M-step consists in maximizing $\tilde{J}(\tau^{[s]}, \theta^{[s]}; \theta)$ with respect to θ. It is again divided into two parts, treating the finite-dimensional parameter (π, β) differently than the infinite dimensional one α. We thus first maximize $\tilde{J}(\tau^{[s]}, \theta^{[s]}; \pi, \beta, \alpha)$ with respect to (π, β) using the current parameter value $\alpha = \alpha^{[s]}$. The solution with respect to π is the same as in the non sparse case and given in (7). Now optimization with respect to β leads to (denoting $\rho^{[s]} = \rho_{\theta^{[s]}}$),

$$\beta_{q,l}^{[s+1]} = \frac{\sum_{(i,j) \in \mathcal{R}} (\tau^{[s]})^{i,q} (\tau^{[s]})^{j,l} \rho^{[s]}(i,j,q,l)}{\sum_{(i,j) \in \mathcal{R}} (\tau^{[s]})^{i,q} (\tau^{[s]})^{j,l}}, \quad q, l = 1, \ldots, Q. \tag{S.5}$$

Then estimation of the intensities $\alpha^{(q,l)}$ is done exactly as previously, except that we replace the variational process $N^{(q,l)}$ by $\tilde{N}^{(q,l)} = \sum_{(i,j) \in \mathcal{R}} \rho^{[s]}(i,j,q,l) (\tau^{[s]})^{i,q} (\tau^{[s]})^{j,l} N_{i,j}$.

Finally we start from an initial value of the clusters $\mathcal{Z}$, see Section S.3, that we treat as probabilities $\{(\tau^{i,q})_{1 \le q \le Q}; 1 \le i \le n\}$. Then we initialise the sparsity parameters $\beta_{q,l}$ and mean intensities $A^{(q,l)}(T)$ with

$$\beta_{q,l} = \frac{\sum_{(i,j) \in \mathcal{R}} Z^{i,q} Z^{j,l} 1\{N_{i,j}(T) > 0\}}{\sum_{(i,j) \in \mathcal{R}} Z^{i,q} Z^{j,l}}, \qquad A^{(q,l)}(T) = \frac{\sum_{(i,j) \in \mathcal{R}} Z^{i,q} Z^{j,l} N_{i,j}(T)}{\sum_{(i,j) \in \mathcal{R}} Z^{i,q} Z^{j,l} 1\{N_{i,j}(T) > 0\}}.$$

This enables to initialise $\rho(i,j,q,l)$ with (S.3). After these initialisations, we are ready to iterate the following steps. At iteration $s \ge 1$ we do

- M-step: Update $\pi^{[s+1]}$ via (7) with $\tau^{[s]}$; Update $\beta^{[s+1]}$ via (S.5) with $\tau^{[s]}$, $\rho^{[s]}$; Update $\alpha^{[s+1]}$ either via Equation (8) for histogram method or (9) for kernel method, using the process $\tilde{N}^{(q,l)}$ and variational parameters $\rho^{[s]}$ and $\tau^{[s]}$.
- VE-step: Update the values $\rho^{[s+1]}$ via (S.3) with $\beta^{[s+1]}$ and $A^{[s+1]}(T)$ derived from $\alpha^{[s+1]}$; Update $\tau^{[s+1]}$ as the solution to the fixed point equation (S.4) relying on the current values $\pi^{[s+1]}, \rho^{[s+1]}, \alpha^{[s+1]}, \beta^{[s+1]}$.

The integrated classification likelihood criterion becomes

$$\mathrm{ICL}_{\text{sparse}}(Q) = \log \mathbb{P}_{\hat{\theta}(Q)}\{\mathcal{O}, \hat{\tau}(Q), \hat{\rho}(Q)\} - \frac{1}{2}(Q-1)\log n - \frac{1}{2}\log r\Big(Q^2 + \sum_{q=1}^{Q}\sum_{l=1}^{Q} 2^{\hat{d}(q,l)}\Big).$$

S.7. The sparse setup: examples

We first discuss the results of the sparse analysis on the London bike-sharing system dataset. In this dataset only 7% of pairs of bike stations have at least one interaction. The main model ignores that fact and this impacts the results as groups are mainly driven by these absences of interactions. For instance the clusters obtained are mostly geographic, revealing absences of connections between distant bikes stations We explore whether one can decipher different structure with our sparse setup. In the following we focus on day 1 as similar results were obtained for day 2.

First our sparse integrated classification likelihood criterion selects only $\hat{Q} = 2$ groups, compared to $\hat{Q} = 6$ in the non sparse case. Geographic locations of the bike stations and the resulting clusters are represented on a city map thanks to the OpenStreetMap project, see Figure S.12. There is one group containing a central part of the city while the remaining stations form a large peripheral cluster. From the estimated intensities in Figure S.13 we see that the second group, i.e. the central geographical group, has large intra-group intensity with three modes: one in the morning around 8:30 am, one at lunch around 1pm and the last at the end of the day at 5:50pm. Group 1, the peripheral one, mostly consists in leaving stations in the morning, see mode in the estimated intensity for $(q, l) = (1, 2)$ around 8:20am, and in arriving stations at the end of the day with a mode in the estimated intensity for $(q, l) = (2, 1)$ at 5:50pm. Intra-group interactions in group 1 have a much lower intensity. On this dataset the sparse setup appears as a complementary model that may shed some different light on the data.

We also analysed the Enron corpus with the sparse model as 91% of the pairs of individuals do not exchange any email during the observation time. The sparse integrated classification likelihood criterion chooses $\hat{Q} = 10$ as the optimal number of groups, which is smaller than in the non sparse model where no optimum has been found in the range of Q from 1 to 20. As in the non sparse model the algorithm identifies one large group with 125 members, while the other nine groups contain at most five individuals. The adjusted Rand index of the clustering in the non sparse model with $Q = 4$ and in the sparse case equals 0.51 with $Q = 10$ and 0.52 when $Q = 4$, which means that there are substantial differences between the clusterings in the two models. Figure S.14 shows the estimated values of the connectivity probabilities $\beta_{q,l}$. Most of these probabilities are significantly lower than 1 justifying the application of the sparse model to these data. A consequence of low connectivity probabilities $\beta_{q,l}$ is that the estimated intensities are more elevated than in the non sparse case which can be observed in Figure S.15 in comparison to the intensity values obtained in the non sparse model, see Figure S.6. We can also compare the estimated intensities in the sparse model with $Q = 4$, Figure S.16, with those in the non sparse case. Again we see that the intensities in the sparse setup are much more elevated. Moreover, the form of the intensities involving two small groups, i.e. $(q, l) \in \{2, 3, 4\}^2$, are all quite different in the two models.

We conclude that as in the bike-sharing example the results in the sparse model differ much from those in the non sparse case. The sparse model tends to select a smaller number of groups which makes interpretation of results easier. As many real datasets are sparse in the sense that only a small percentage of individuals effectively interact with another the sparse model seems to be particularly adapted to real data and provides the possibility of further insights on the data.

We also analysed the primary school dataset with the sparse model. In this dataset 28% of pairs of individuals have at least one interaction. Our sparse integrated classification likelihood criterion selects $\hat{Q} = 13$ groups, which is smaller than in the non sparse model. The clustering in the non sparse model and in the sparse model are quite close, with some groups being the same. The main difference between the two clusterings concerns the group composed in the non sparse model of class 4B with additional pupils coming from almost all other classes. This group was characterized by the lowest intra-group intensity. In the sparse model, class 4B is separated into two groups: 6 pupils are gathered with class 4A to form one group, that is group 1, whereas the 17 remaining pupils are gathered with class 1A and some pupils coming from almost all other classes, composing group 3. Looking at the estimated intensities, we see that group 3 has a low intra-group intensity during the lunch time contrary to group 1, see Figure S.17. Moreover the estimated intra-connectivity probability for groups 1 and 3 are given by $\hat{\beta}_{1,1} = 0.84$ and $\hat{\beta}_{3,3} = 0.29$. Therefore group 3 is composed of individuals which only a few proportion interacts, and characterized by very few interactions during the lunch time. On this dataset with the non sparse and sparse models we mainly recover the same clustering based on communities, but the sparse model also exhibits particular temporal profile of some individuals.

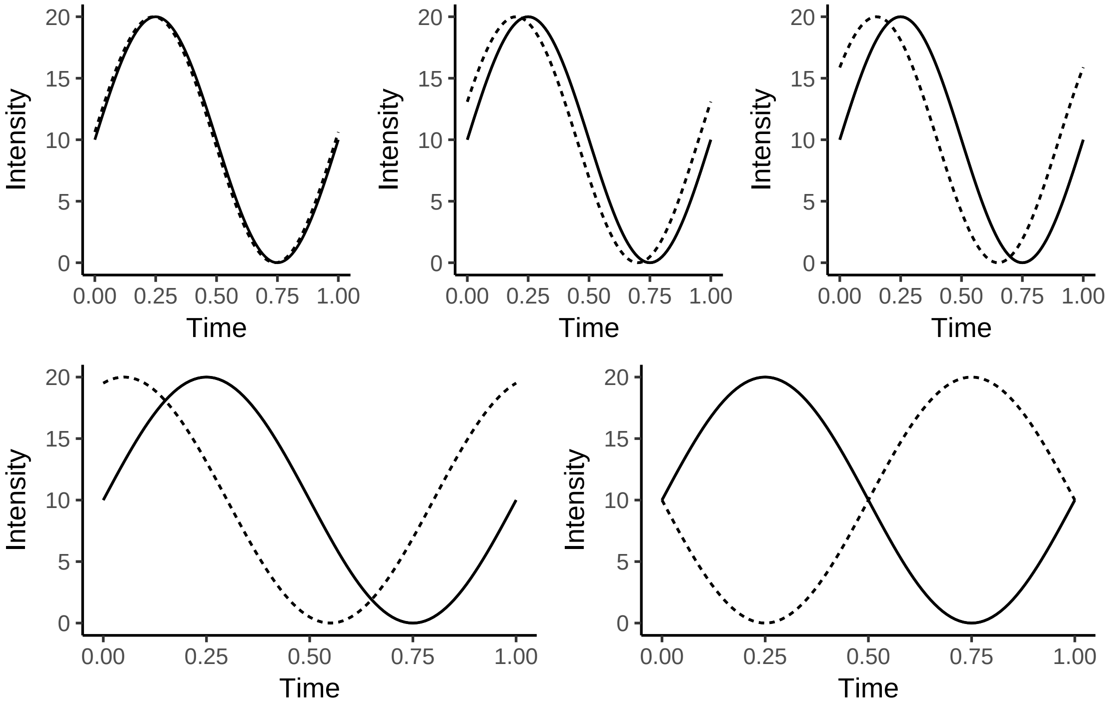


Fig. S.1: Intensities in the synthetic experiments from Scenario 1. Each picture represents the intra-group intensity α^{in} in bold line and the inter-group intensity α^{out} in dashed line with different shift parameter values $\varphi \in \{0.01, 0.05, 0.1, 0.2, 0.5\}$.

Table S.1: Mean number of events and risks with standard deviations (sd) in scenario 2 with $n = 20$. Histogram (Hist) and kernel (Ker) estimators are compared with their oracle counterparts (Or.Hist, Or.Ker). All values associated with the risks are multiplied by 100.

Groups (q, l)	Nb.events	Hist (sd)	Or.Hist (sd)	Ker (sd)	Or.Ker (sd)
$(1, 1)$	84	136 (92)	50 (49)	215 (83)	113 (55)
$(1, 2)$	146	177 (146)	98 (27)	270 (107)	194 (23)
$(1, 3)$	86	211 (160)	78 (20)	178 (143)	43 (18)
$(2, 2)$	32	136 (72)	108 (29)	139 (109)	71 (41)
$(2, 3)$	130	265 (72)	217 (28)	238 (78)	182 (22)
$(3, 3)$	48	173 (61)	158 (47)	171 (111)	85 (43)

Table S.2: Enron: Total size and group composition with $Q = 4$ groups, some people's positions are unknown.

	total	managers	employees
group 1	127	62	36
group 2	4	0	3
group 3	2	1	1
group 4	14	12	1

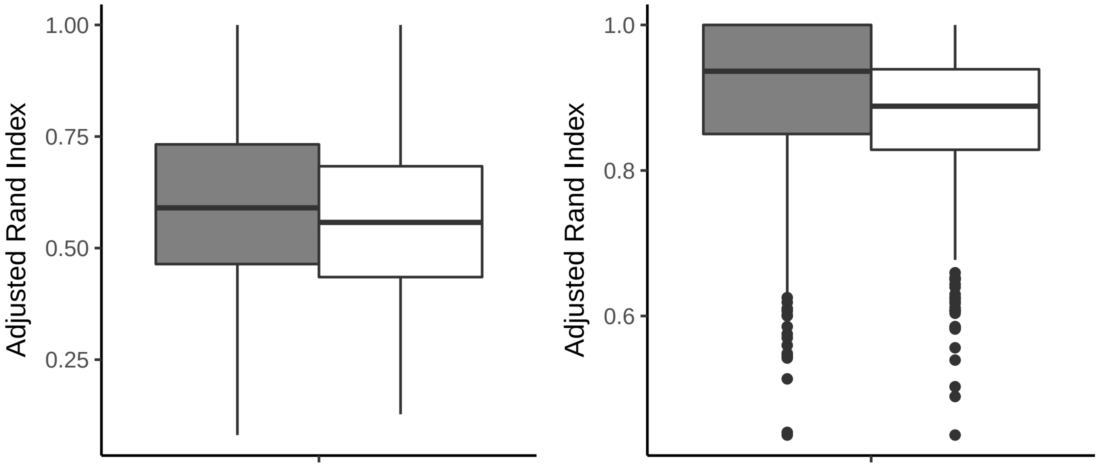


Fig. S.2: Boxplots of the adjusted Rand index in the synthetic experiments for Scenario 2, for the histogram estimator (darkgrey) and the kernel estimator (white). Left panel: $n = 20$, right panel: $n = 50$.

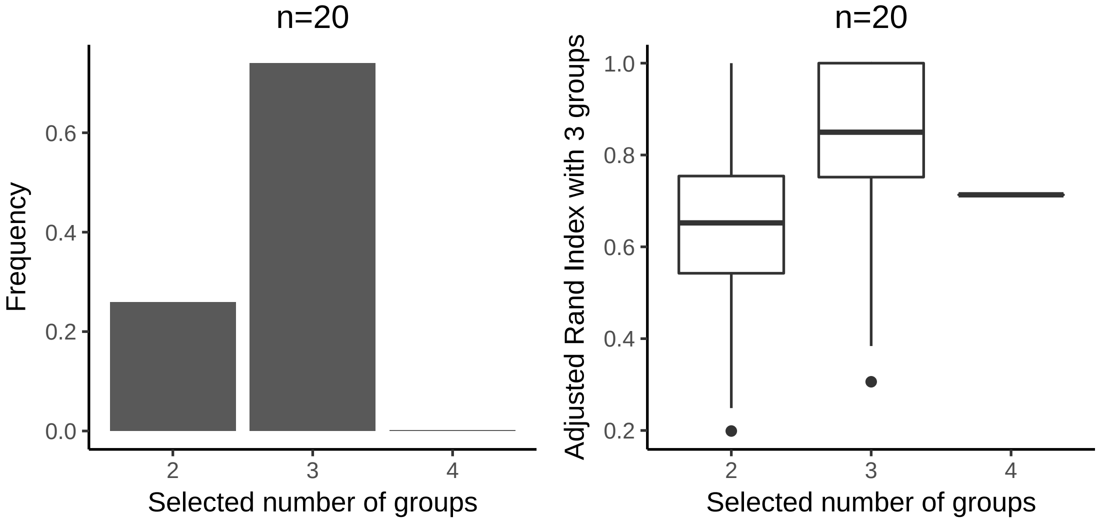


Fig. S.3: Selection of the number of latent groups via the integrated classification likelihood criterion in Scenario 2 with $n = 20$. Left panel: frequencies of selected number of groups. Right panel: adjusted Rand index comparing the classification into three groups and true classification as a function of the number of groups selected by ICL.

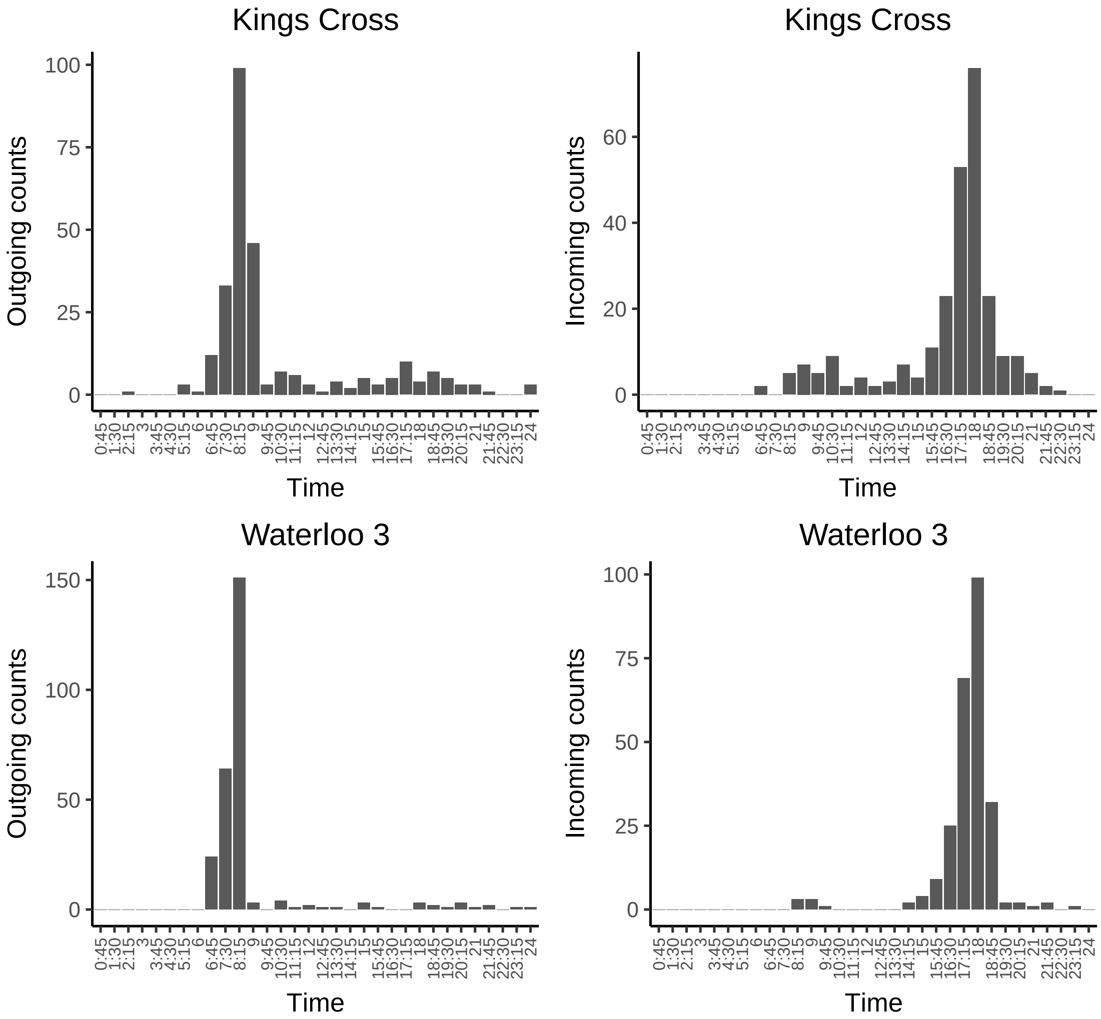


Fig. S.4: London bike-sharing system: Barplots of outgoing counts $N_{i\cdot}(\cdot)$ on the left, and incoming counts $N_{\cdot i}(\cdot)$ on the right, for the two stations i in the smallest cluster as top row and bottom row, respectively: representation of volumes of connections to all other stations during day 1, with time on the x-axis.

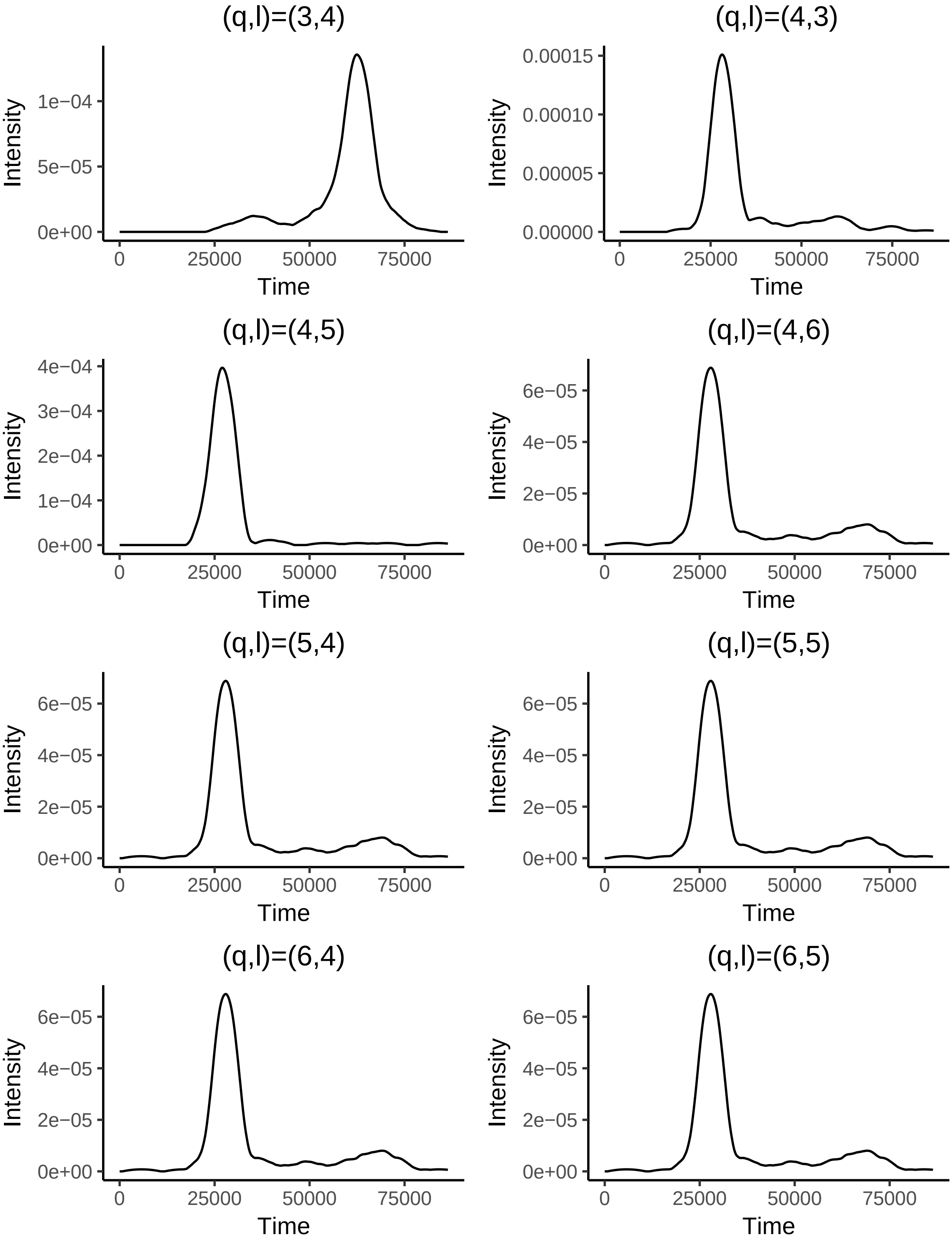


Fig. S.5: London bike-sharing system: estimated non almost null intensities for day 1; the x-axis gives the time in seconds.

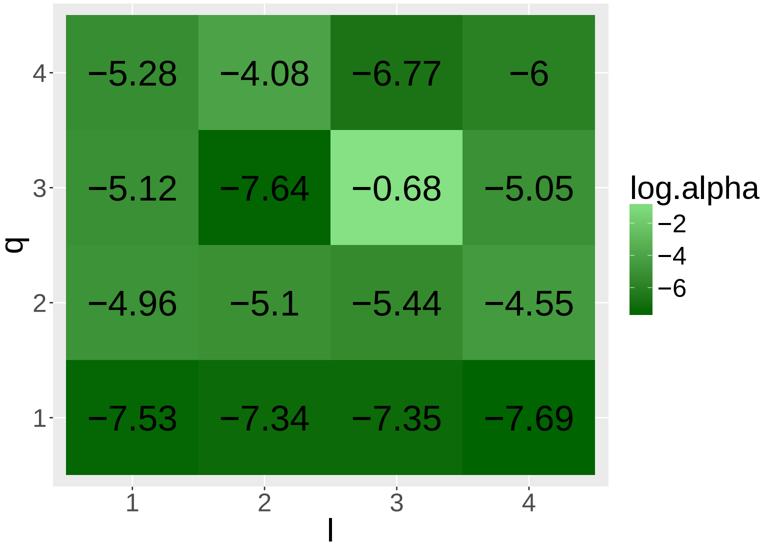


Fig. S.6: Enron: Logarithm of the mean values of the estimated intensities $\alpha^{(q,l)}$ with $Q = 4$ groups.

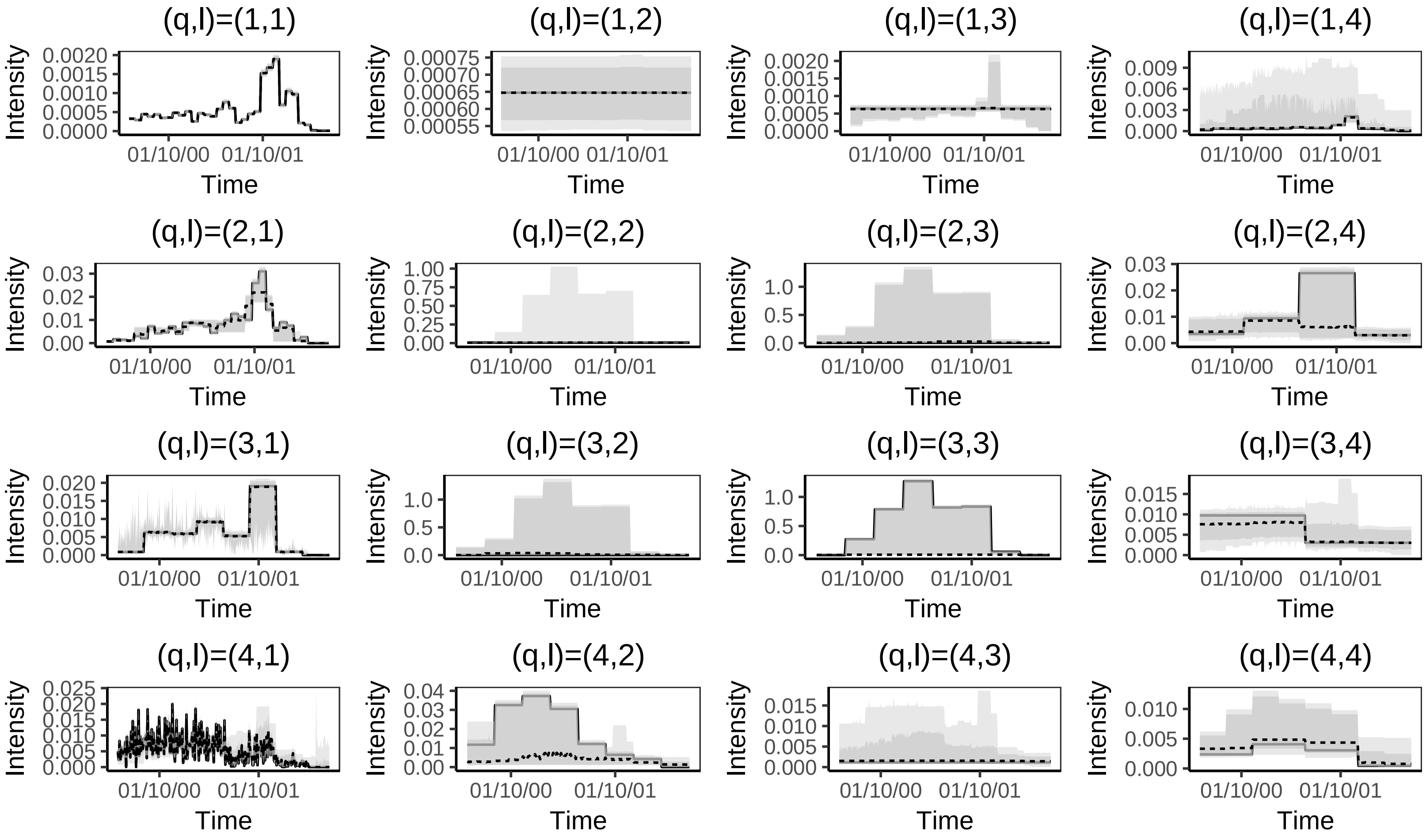


Fig. S.7: Enron: Estimated intensities $\hat{\alpha}^{(q,l)}$ with $Q = 4$ groups with bootstrap confidence intervals with confidence level 90% in lightgrey and 80% in darkgrey, and the median bootstrap values at 80% level in dotted lines.

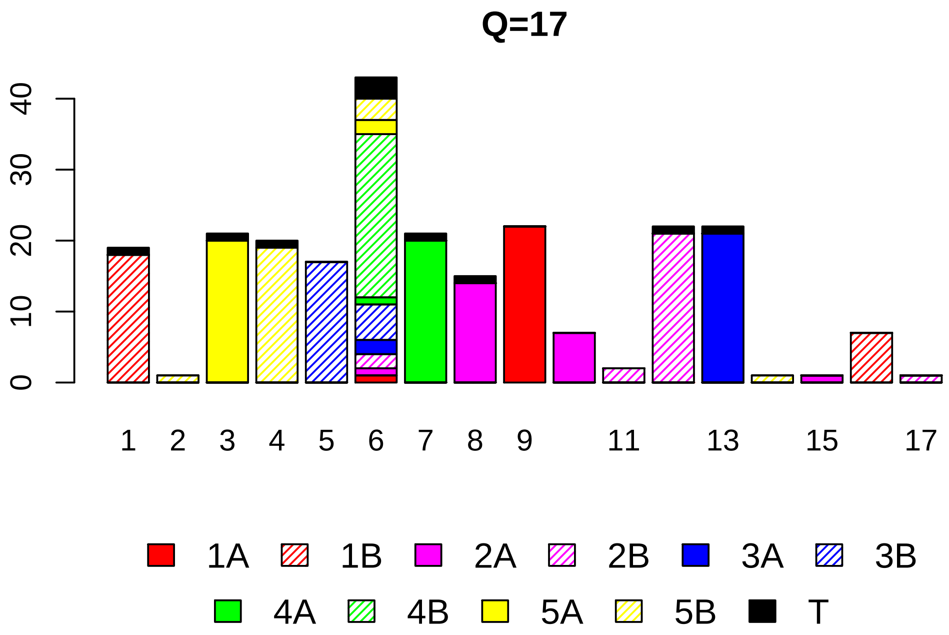


Fig. S.8: Primary school: clustering of the 242 individuals into $Q = 17$ groups. Vertical bars represent the Q clusters. Colours indicate the grades and the teachers, plain and hatching distinguish the two classes in the same grade.

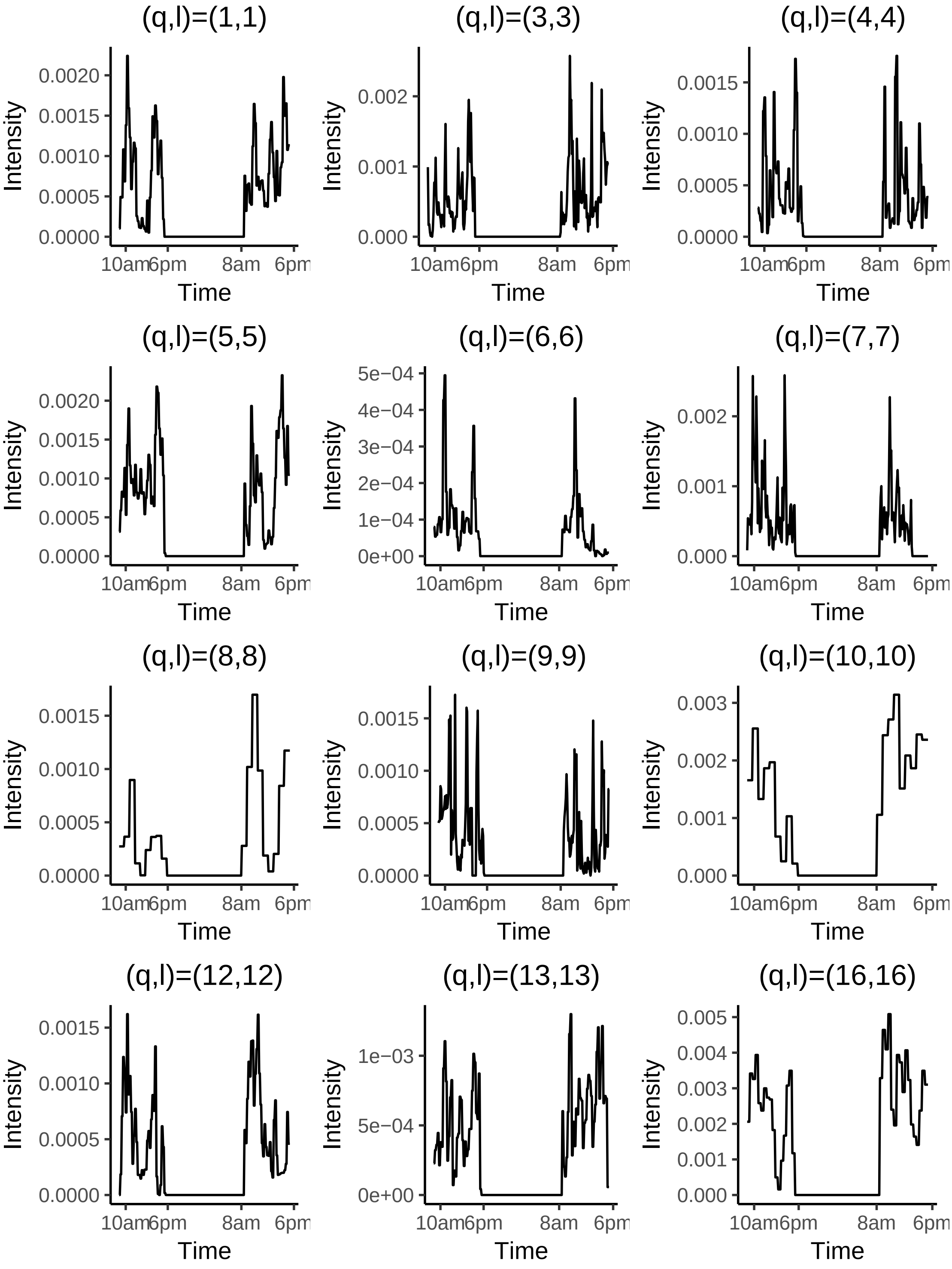


Fig. S.9: Primary school: Estimated intra-group intensities

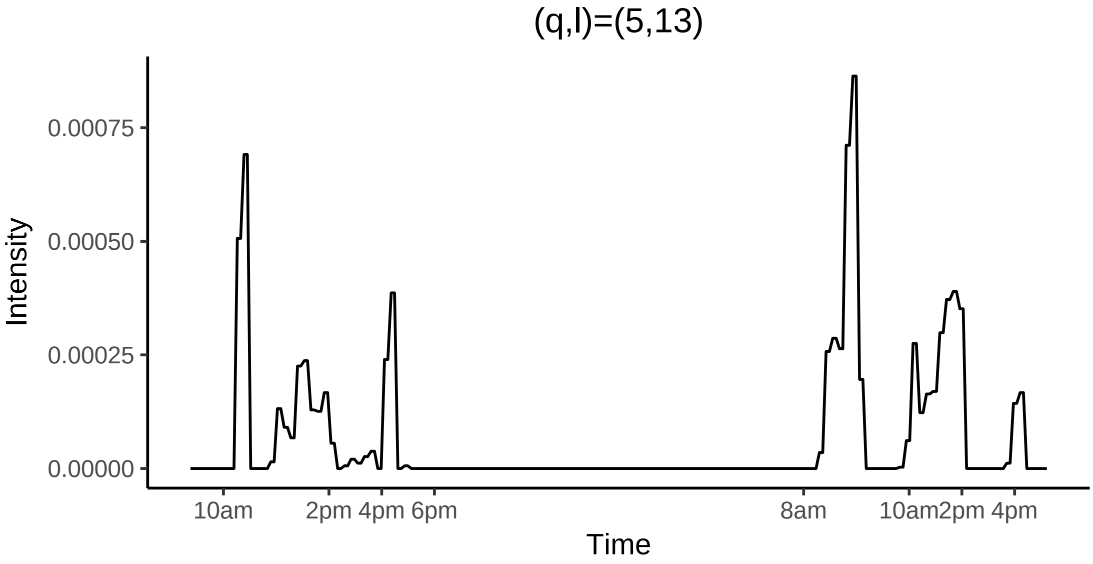


Fig. S.10: Primary school: Estimated inter-group intensity between two classes of the same grade: classes $3A$ that is group 13 and $3B$ that is group 5.

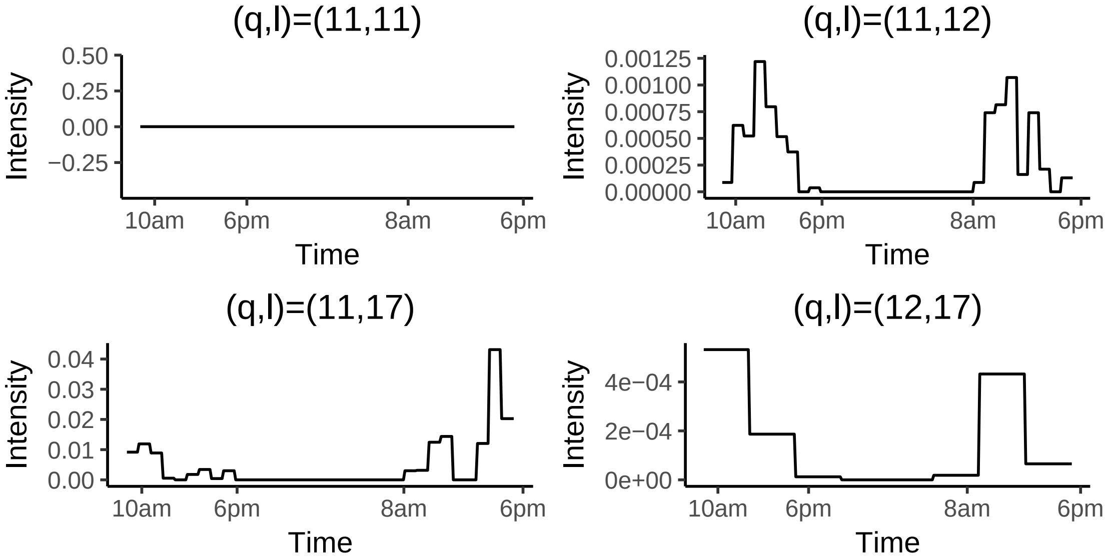


Fig. S.11: Primary school: Estimated intensities. Example of class 2B split into group 12 with 21 pupils, group 11 with 2 pupils, and group 17 with only one pupil.

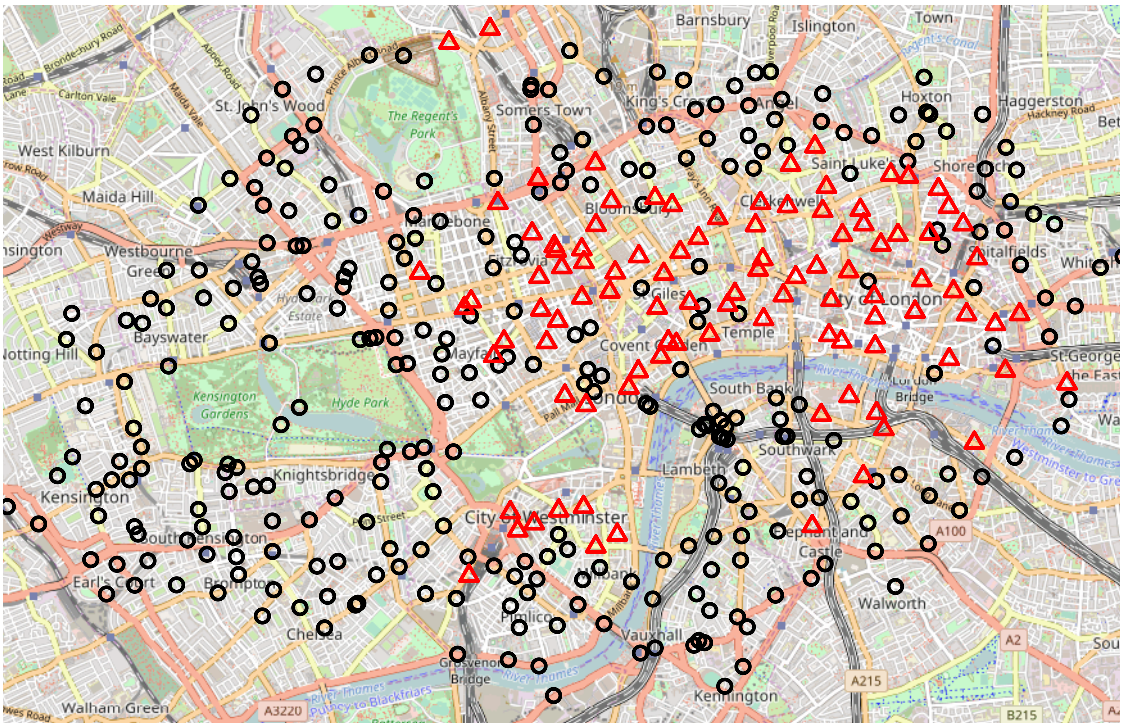


Fig. S.12: London bike-sharing system: Geographic positions of the stations and clustering into two clusters, represented by different colors and symbols, obtained from the sparse model for day 1.

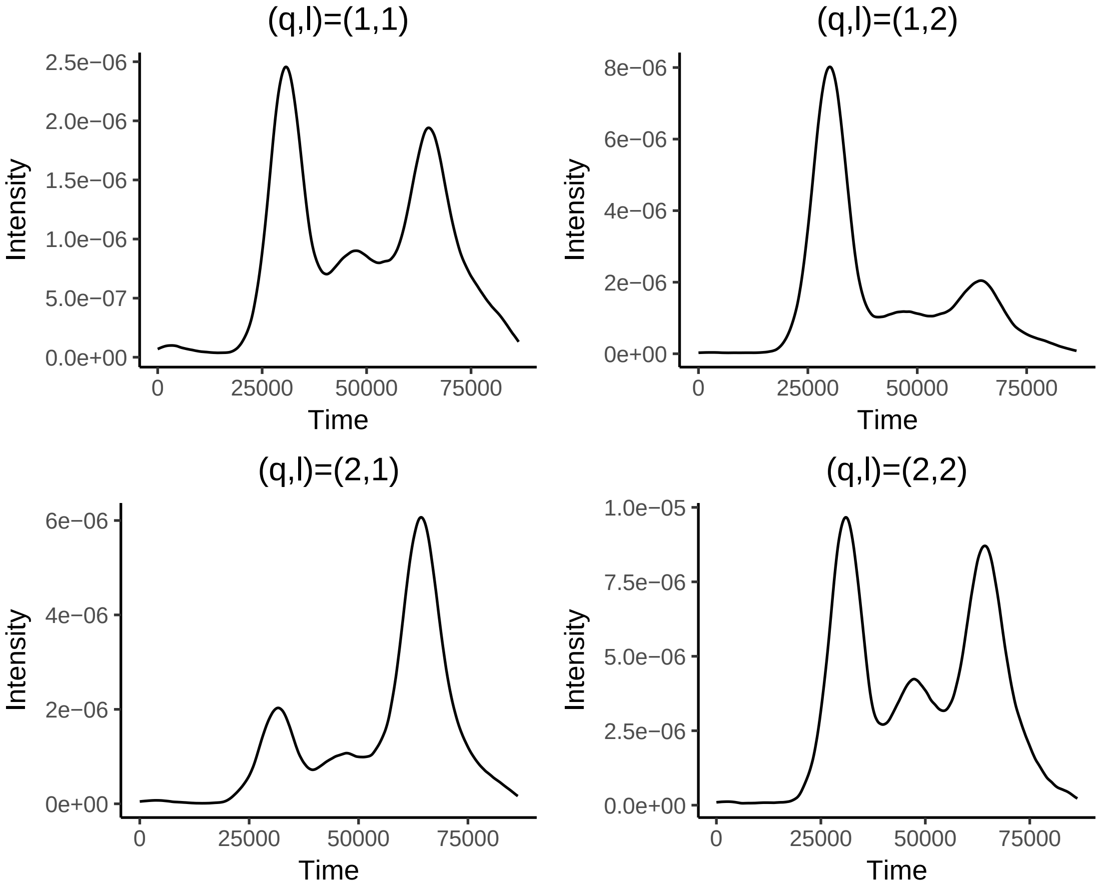
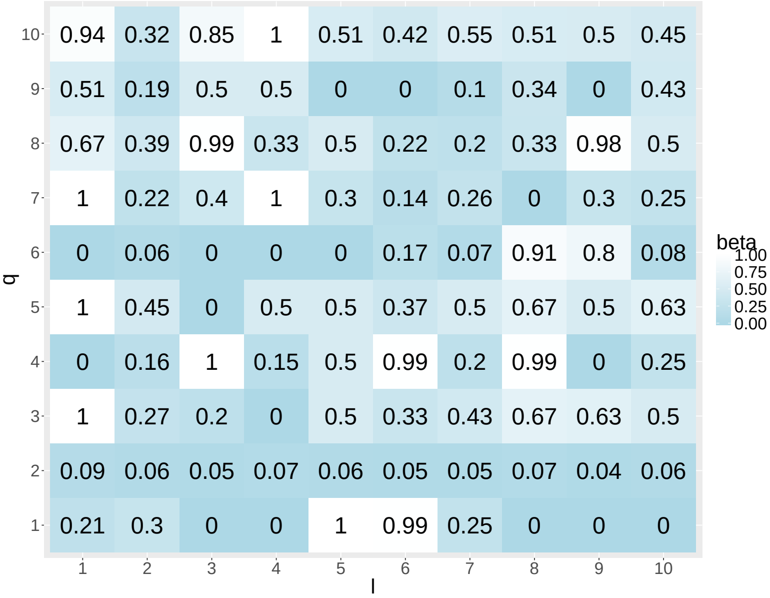
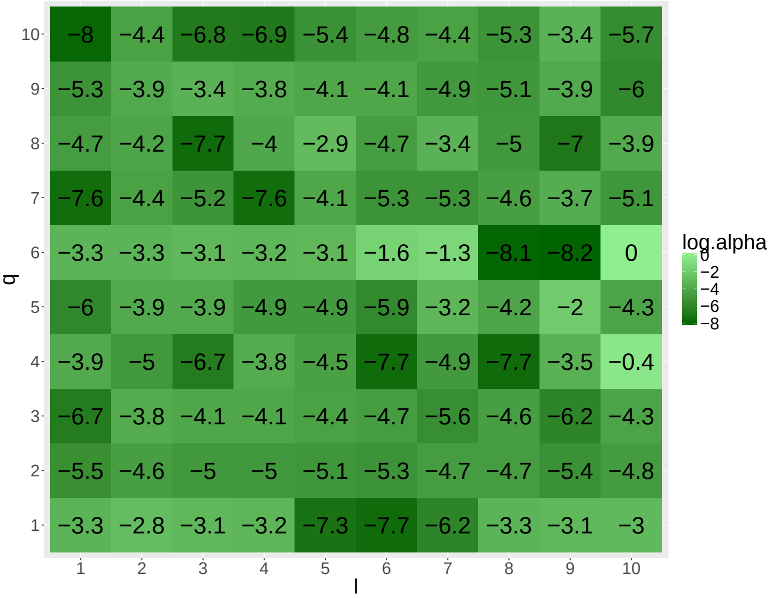


Fig. S.13: London bike-sharing system: estimated intensities from the sparse model for day 1, time on the x-axis is in seconds.



q \ l	1	2	3	4	5	6	7	8	9	10
10	0.94	0.32	0.85	1	0.51	0.42	0.55	0.51	0.5	0.45
9	0.51	0.19	0.5	0.5	0	0	0.1	0.34	0	0.43
8	0.67	0.39	0.99	0.33	0.5	0.22	0.2	0.33	0.98	0.5
7	1	0.22	0.4	1	0.3	0.14	0.26	0	0.3	0.25
6	0	0.06	0	0	0	0.17	0.07	0.91	0.8	0.08
5	1	0.45	0	0.5	0.5	0.37	0.5	0.67	0.5	0.63
4	0	0.16	1	0.15	0.5	0.99	0.2	0.99	0	0.25
3	1	0.27	0.2	0	0.5	0.33	0.43	0.67	0.63	0.5
2	0.09	0.06	0.05	0.07	0.06	0.05	0.05	0.07	0.04	0.06
1	0.21	0.3	0	0	1	0.99	0.25	0	0	0

Fig. S.14: Enron: Estimated connectivity probabilities $\beta_{q,l}$ in the sparse model with $Q = 10$ groups.



q \ l	1	2	3	4	5	6	7	8	9	10
10	−8	−4.4	−6.8	−6.9	−5.4	−4.8	−4.4	−5.3	−3.4	−5.7
9	−5.3	−3.9	−3.4	−3.8	−4.1	−4.1	−4.9	−5.1	−3.9	−6
8	−4.7	−4.2	−7.7	−4	−2.9	−4.7	−3.4	−5	−7	−3.9
7	−7.6	−4.4	−5.2	−7.6	−4.1	−5.3	−5.3	−4.6	−3.7	−5.1
6	−3.3	−3.3	−3.1	−3.2	−3.1	−1.6	−1.3	−8.1	−8.2	0
5	−6	−3.9	−3.9	−4.9	−4.9	−5.9	−3.2	−4.2	−2	−4.3
4	−3.9	−5	−6.7	−3.8	−4.5	−7.7	−4.9	−7.7	−3.5	−0.4
3	−6.7	−3.8	−4.1	−4.1	−4.4	−4.7	−5.6	−4.6	−6.2	−4.3
2	−5.5	−4.6	−5	−5	−5.1	−5.3	−4.7	−4.7	−5.4	−4.8
1	−3.3	−2.8	−3.1	−3.2	−7.3	−7.7	−6.2	−3.3	−3.1	−3

Fig. S.15: Enron: Logarithm of the mean values of the estimated intensities $\alpha^{(q,l)}$ in the sparse model with $Q = 10$ groups.

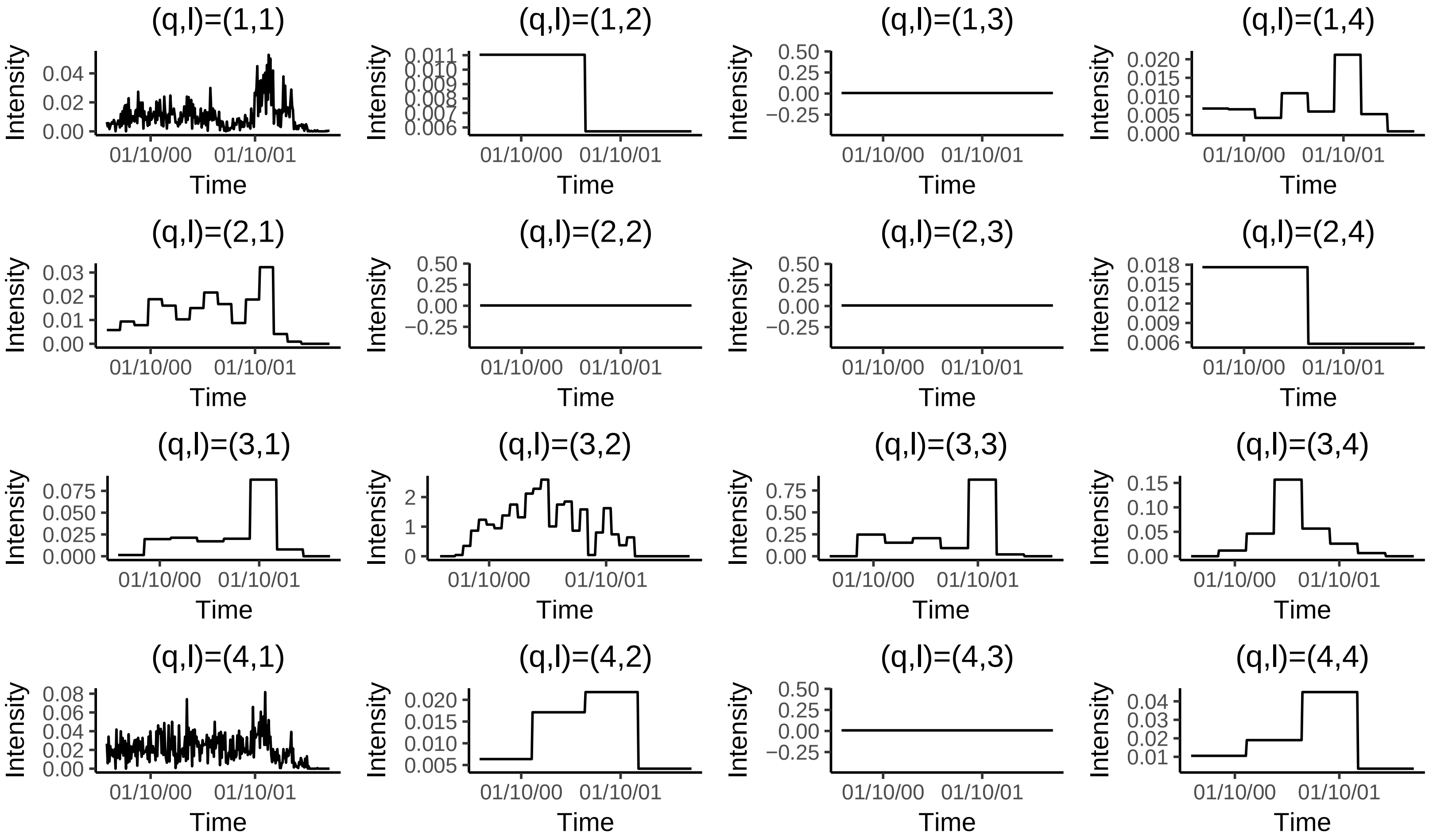


Fig. S.16: Enron: Estimated intensities $\hat{\alpha}^{(q,l)}$ in the sparse model with $Q = 4$ groups.

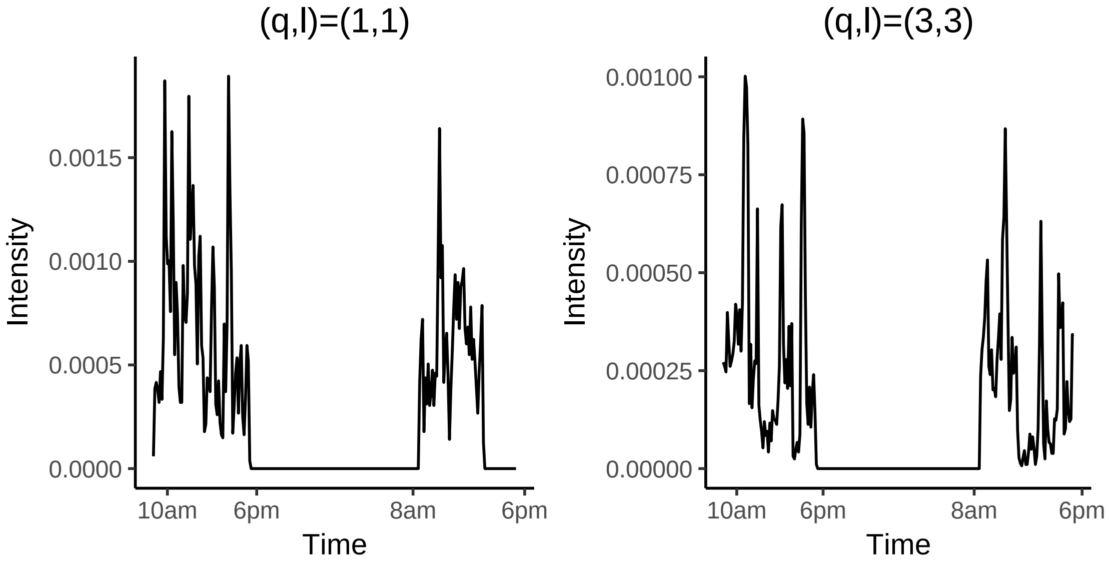


Fig. S.17: Primary school: Estimated intensities in the sparse model. Example of group 1 composed of class 4A and 6 pupils of class 4B and group 3 composed of the entire class 1A, with 17 pupils of class 4B and pupils from almost all other classes.

REFERENCES

E. Allman, C. Matias, and J. Rhodes. Parameters identifiability in a class of random graph mixture models. *J. Stat. Plan. Inference*, 141:1719–1736, 2011.

D. J. Daley and D. Vere-Jones. *An introduction to the theory of point processes. Vol. I.* Probability and its Applications (New York). Springer-Verlag, New York, second edition, 2003. Elementary theory and methods.

SocioPatterns. http://www.sociopatterns.org/, 2015.

J. Stehlé, N. Voirin, A. Barrat, C. Cattuto, L. Isella, J.-F. Pinton, and et al. High-resolution measurements of face-to-face contact patterns in a primary school. *PLoS ONE*, 6(8):e23176, 2011.

[*Received April* 2012. *Revised October* 2015]