A DERIVATION OF FULL CONDITIONALS

For each term $\boldsymbol{\beta}_r$ in $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_k)$, the full conditional distribution is:

$$\begin{aligned} f(\boldsymbol{\beta}_r \mid -) &\propto \mathrm{MLG}(\mathbf{0}_p, \boldsymbol{V}, \boldsymbol{\alpha}, \boldsymbol{\kappa}) \prod_{z_i=r} \mathrm{Poisson}(\exp\left(\boldsymbol{X}(\boldsymbol{s}_i)\boldsymbol{\beta}(s_{z_i})\right) \\ &\propto \exp(\boldsymbol{\alpha}'\boldsymbol{V}^{-1}\boldsymbol{\beta}_r - \boldsymbol{\kappa}' \exp(\boldsymbol{V}^{-1}\boldsymbol{\beta}_r)) \prod_{z_i=r} \exp\left(\boldsymbol{X}(\boldsymbol{s}_i)\boldsymbol{\beta}(s_{z_i})\right)^{y(\boldsymbol{s}_i)} \exp(-\exp\left(\boldsymbol{X}(\boldsymbol{s}_i)\boldsymbol{\beta}(s_{z_i})\right)) \\ &\propto \exp\left(\boldsymbol{\alpha}'V^{-1}\boldsymbol{\beta}_r + \sum_{z_i=r} y(\boldsymbol{s}_i)\boldsymbol{X}(\boldsymbol{s}_i)\boldsymbol{\beta}(s_{z_i})\right) \\ &\exp\left(-\boldsymbol{\kappa}' \exp(\boldsymbol{V}^{-1}\boldsymbol{\beta}_r) - \sum_{z_i=r} I_{(z_i=r)} \exp(\boldsymbol{X}(\boldsymbol{s}_i)\boldsymbol{\beta}(s_{z_i}))\right) \\ &\propto \exp\left[(\alpha, \sum_{z_i=r} y(\boldsymbol{s}_i))' \begin{bmatrix} \boldsymbol{V}^{-1} \\ \boldsymbol{X}(\boldsymbol{s}_i) \end{bmatrix} \beta_r\right] \exp\left[-(\kappa, \sum_{z_i=r} I_{(z_i=r)})' \exp(\begin{bmatrix} \boldsymbol{V}^{-1} \\ \boldsymbol{X}(\boldsymbol{s}_i) \end{bmatrix} \beta_r)\right] \end{aligned} \tag{23}$$

This implies that $f(\boldsymbol{\beta}_r \mid -) \sim \mathrm{cMLG}(\boldsymbol{H}_\beta, \boldsymbol{\alpha}_\beta, \boldsymbol{\kappa}_\beta)$.

For each term z_i in $z = (z_i, \ldots, z_n)$, the full conditional distribution is:

$$P(z_i = c \mid z_1, \ldots, z_{i-1}) \propto \begin{cases} P(z_i = c \mid z_{-i}) d\mathrm{Poisson}(y(\boldsymbol{s}_i), \exp(\boldsymbol{X}(\boldsymbol{s}_i)\boldsymbol{\beta}_r)), & \text{at table labeled } c \\ \frac{V_n(|C_{-i}|+1)}{V_n(|C_{-i}|)} \gamma m(y(\boldsymbol{s}_i)), & \text{if } c \text{ is a new table} \end{cases}.$$

where

$$\begin{aligned} m(y(\boldsymbol{s}_i)) &= \int \mathrm{MLG}(\mathbf{0}_p, \boldsymbol{V}, \boldsymbol{\alpha}, \boldsymbol{\kappa}) \mathrm{Poisson}(y(\boldsymbol{s}_i) \mid \boldsymbol{\beta}_r) d\boldsymbol{\beta}_r \\ &\propto \int \frac{1}{det(\boldsymbol{V}\boldsymbol{V}')^{\frac{1}{2}}} \left(\prod_{i=1}^{p} \frac{\kappa_i^{\alpha_i}}{\Gamma(\alpha_i)}\right) \exp(\boldsymbol{\alpha}'\boldsymbol{V}^{-1}\boldsymbol{\beta}_r - \ +\kappa' \exp(\boldsymbol{V}^{-1}\boldsymbol{\beta}_r)) \\ &\exp\left[\boldsymbol{X}(\boldsymbol{s}_i)\boldsymbol{\beta}_r\right]^{y(\boldsymbol{s}_i)} \exp\left[-\exp(\boldsymbol{X}(\boldsymbol{s}_i)\boldsymbol{\beta}_r)\right] \\ &= \frac{1}{det(\boldsymbol{V}\boldsymbol{V}')^{\frac{1}{2}}} \left(\prod_{i=1}^{p} \frac{\kappa_i^{\alpha_i}}{\Gamma(\alpha_i)}\right) \\ &\int \exp\left[(\alpha, \sum_{z_i=r} y(\boldsymbol{s}_i))' \begin{bmatrix} \boldsymbol{V}^{-1} \\ \boldsymbol{X}(\boldsymbol{s}_i) \end{bmatrix} \boldsymbol{\beta}_r\right] \exp\left[-(\kappa, \sum_{z_i=r} I_{(z_i=r)})' \exp(\begin{bmatrix} \boldsymbol{V}^{-1} \\ \boldsymbol{X}(\boldsymbol{s}_i) \end{bmatrix} \boldsymbol{\beta}_r)\right] \\ &= \frac{1}{det(\boldsymbol{V}\boldsymbol{V}')^{\frac{1}{2}}} \left(\prod_{i=1}^{p} \frac{\kappa_i^{\alpha_i}}{\Gamma(\alpha_i)}\right) \frac{1}{M_1} \end{aligned}$$

and

$$M_1 = det(\begin{bmatrix} \boldsymbol{H}_\beta & Q_2 \end{bmatrix}) \left(\prod_{i=1}^{n+p} \frac{\kappa_i^{\alpha_i}}{\Gamma(\alpha_i)}\right) \frac{1}{\int f(y(\boldsymbol{s}_i) \mid \mathbf{0}_{n+p}, \boldsymbol{V} = \begin{bmatrix} \boldsymbol{H}_\beta \ Q_2 \end{bmatrix}^{-1}, \boldsymbol{\alpha}, \boldsymbol{\kappa})}$$

and "det" is a short hand as determinant of a matrix.

B ADDITIONAL COMPARISON FOR SIMULATION (STATE OF GEORGIA)

We present additional comparison for simulation section (State of Georgia). We compare our proposed method to LGP and CAR in two cluster design. In Figure 4, the values above zero indicate that our method has higher LPML than comparator. The results shown that we have a better result for both comparator.

C ADDITIONAL SIMULATION FOR DIFFERENT SPATIAL GRAPH (STATE OF MISSISSIPPI)

We provide another simulation design with different spatial graph. This additional analysis is based on the spatial structure of the state of Mississippi, which contains 82 counties. We consider a different spatial cluster designs shown in Figure 5. This design consists of two disjoint parts located in the top and bottom parts of Mississippi.

Two different scenarios are considered. The first scenario does not take into account spatial random effects, while in the second scenario, spatial random effects are included for each design. The spatial random effects are assumed to follow a multivariate normal distribution with a mean zero and exponential covariogram. Based on the estimated number of clusters and Rand Index (RI), the clustering performance

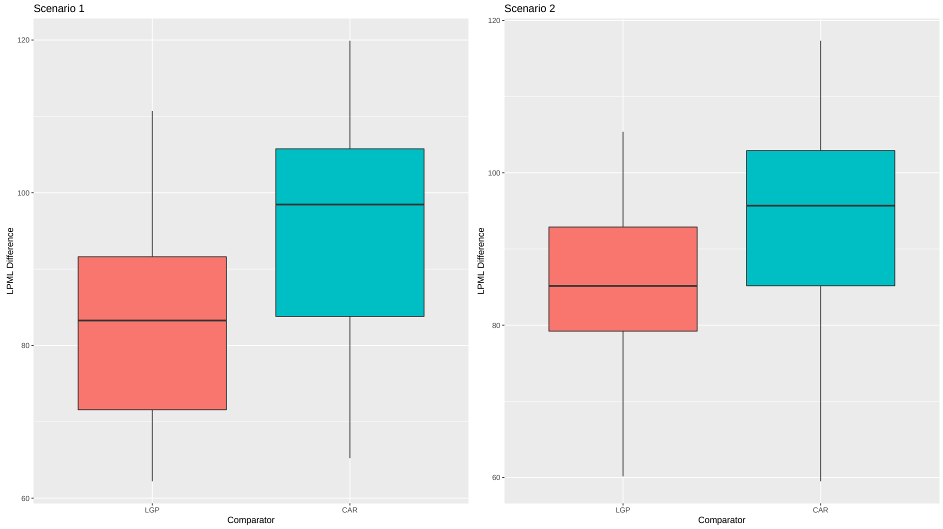

Figure 4: Additional Comparison for Two Cluster Simulation (State of Georgia).

is evaluated. Each replicate is also used to calculate the final number of clusters estimated. A total of 50 sets of data are generated under different scenarios. We run 3000 iterations of the MCMC chain and burn-in the first 1000 for each replicate.

The results of the comparison of LPML, Rand index, and estimation of the number of clusters for each design can be found in Table 5. Our proposed method outperforms vanilla MFM with respect to model fitness and clustering, as demonstrated by the LPML values and Rand index. Additional comparison to LGP and CAR also presented. In Figure 6, the values above zero indicate that our method has higher LPML than comparator. The results shown that we have a better result for both comparator.

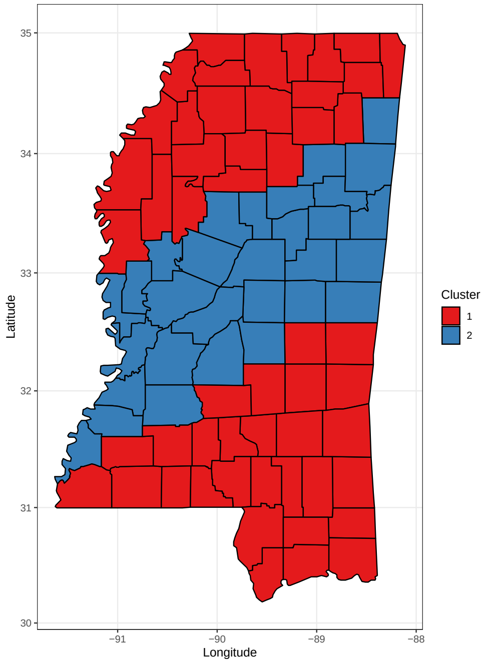

Figure 5: Simulation design with two cluster assignments. (State of Mississippi)

D COMPARISON WITH [23] AND [1]

Both [23] and [1] perform clustered coefficients separately for different coefficients, see Figures 2,4 in [23] and Figures 1,3 in[1], while our approach performs an entire clustered coefficients, where all the coefficients share the same cluster configuration. For example, see Figure 3

Table 5: Simulation Results including LPML, Rand Index (RI), and number of true cluster cover rate (CR) by MRF-MFM (optimal) model and MFM model. We provide mean and standard deviation for both LPML and RI.

Method	Scenario	LPML	RI	CR	Scenario	LPML	RI	CR
Optimal	1	-295.79 (9.29)	0.9954 (0.0179)	100%	2	-291.76 (10.93)	0.9966 (0.0135)	100%
MFM		-819.39 (407.15)	0.9901 (0.0257)	98%		-727.47 (272.14)	0.9901 (0.0269)	96%

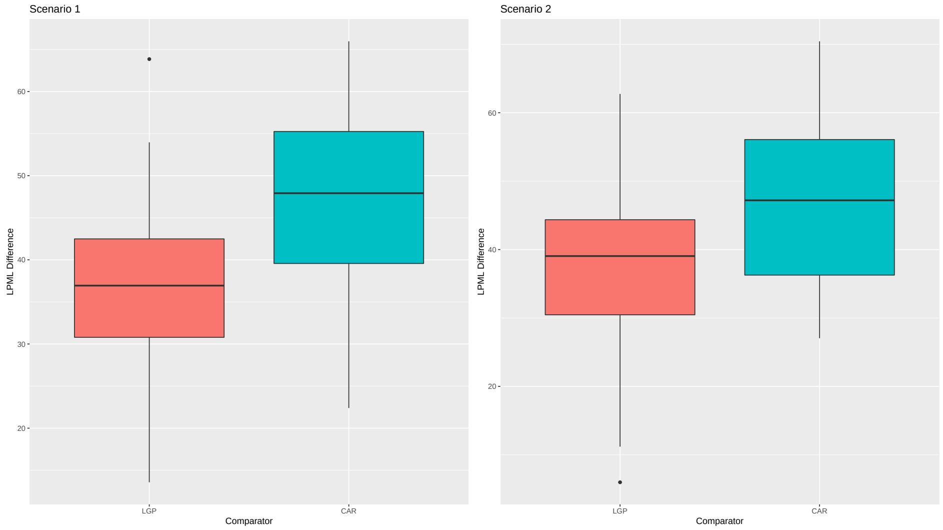

Figure 6: Additional Comparison for Two Cluster Simulation (State of Mississippi).

where our approach generates one clustered configuration on the county map despite three coefficients being estimated. In particular, the correct specification of single partition vs. different partitions for a model is very important in practice. Here we consider a single partition for different coefficients for the following reasons:

- Methodology: A single partition for different coefficients is more interpretable as it automatically produces a cluster assignment of the spatial objects.
- Simulation: When the true data generation process is based on a single cluster assignment, a single partition with different coefficients performs better than having separate partitions as the latter has more variance in estimations and therefore overfits the model. To demonstrate this, we implemented [23] approach using R package *genlasso* to report the comparison of AMSE for two clusters design without spatial random effect. Our estimation error is around 0.085 according to Table 2 while the estimation error of [23]'s approach is 26.48. The reproducible code fusedlasso.R is provided in the supplement. The reasons that estimation errors have such a huge gap are because: (1). Our model is correctly specified with the Poisson likelihood while for *genlasso* the likelihood is Gaussian; (2). Our model is correctly specified with the single cluster assignment.
- Theoretical details: Single clusters for different coefficients are less frequent in the literature, perhaps because they are mathematically more complex. Take penalized regression, [23], as an example, to induce separate clusters, a simple penalty on the aggregated coefficients is sufficient, corresponding to the fused lasso. However, some group-wise penalizations should be considered to induce a common cluster assignment among different coefficients, corresponding to a more complex group-wise type of regularizations.
- Application: for our state of Georgia's premature death data, the coefficients of interest, which are PM 2.5 and the food environment index, are both variables primarily influenced by environmental factors. Therefore, claiming that their effects cluster in homogenous groups makes sense.

E PROOF OF THE THEOREM 2.1

By Bayes' theorem, we have:

$$\Pi\left(\boldsymbol{\beta}_i|\boldsymbol{\beta}_{-i}\right) \propto \Pi\left(\boldsymbol{\beta}_1,\ldots,\boldsymbol{\beta}_{n_0}\right) = P\left(\boldsymbol{\beta}_1,\ldots,\boldsymbol{\beta}_{n_0}\right) M\left(\boldsymbol{\beta}_1,\ldots,\boldsymbol{\beta}_{n_0}\right) \propto P\left(\boldsymbol{\beta}_i|\boldsymbol{\beta}_{-i}\right) M\left(\boldsymbol{\beta}_i|\boldsymbol{\beta}_{-i}\right). \tag{24}$$

As shown in [29], by conditioning on the different possible situations of the cluster for the new observations, we have

$$P\left(\boldsymbol{\beta}_i|\boldsymbol{\beta}_{-i}\right) \propto \frac{V_{n_0}(t+1)\gamma}{V_{n_0}(t)}P(\boldsymbol{\beta}_i) + \sum_{i=1}^{t}\left(n_i+\gamma\right)\delta_{\boldsymbol{\beta}_i^*}. \tag{25}$$

Let $\partial(i) := \{j : (i, j) \in E\}$. When considering the full conditional distribution

$$M\left(\boldsymbol{\beta}_i|\boldsymbol{\beta}_{-i}\right) \propto \exp\left(H_{i|-i}(\boldsymbol{\beta}_i|\boldsymbol{\beta}_{-i})\right), \tag{26}$$

where $H_{i|-i}(\boldsymbol{\beta}_i|\boldsymbol{\beta}_{-i})$ only depends on $H_{ij}(\boldsymbol{\beta}_i\boldsymbol{\beta}_j)$ for $(i, j) \in E$. Note that

$$H_{i|-i}\left(\boldsymbol{\beta}_i|\boldsymbol{\beta}_{-i}\right) = 0 \quad \text{if } S_i \notin S_{\partial(i)} \;\;, \tag{27}$$

where $\boldsymbol{s}_i$ specifies the cluster that $\boldsymbol{\beta}_i$ belongs to. With the property in equation (27) and the assumption that P is continuous, $\exp\left(H_{i|-i}\left(\boldsymbol{\beta}_i|\boldsymbol{\beta}_{-i}\right)\right) = 1$ almost surely for $\boldsymbol{\beta}_i \sim P$. Then given any measurable function f for $P(\boldsymbol{\beta})$ and any subset A for the domain of $\boldsymbol{\beta}_i$,

$$\begin{aligned}\int_A f\left(\boldsymbol{\beta}_i\right) M\left(\boldsymbol{\beta}_i|\boldsymbol{\beta}_{-i}\right) P\left(\boldsymbol{\beta}_i\right) d\boldsymbol{\beta}_i = &\int_A f\left(\boldsymbol{\beta}_i\right)\frac{1}{Z_{H'}}\exp\left(H_{i|-i}\left(\boldsymbol{\beta}_i|\boldsymbol{\beta}_{-i}\right)\right) P\left(\boldsymbol{\beta}_i\right) d\boldsymbol{\beta}_i \\ &= \int_A f\left(\boldsymbol{\beta}_i\right)\frac{1}{Z_{H'}} P\left(\boldsymbol{\beta}_i\right) d\boldsymbol{\beta}_i,\end{aligned} \tag{28}$$

where the constant $Z_{H'}$ only depends on the constant part of $M\left(\boldsymbol{\beta}_i|\boldsymbol{\beta}_{-i}\right)$. Hence, the full conditional of Π can be derived

$$\Pi\left(\boldsymbol{\beta}_i|\boldsymbol{\beta}_{-i}\right) \propto \frac{V_{n_0}(t+1)\gamma}{V_{n_0}(t)}P(\boldsymbol{\beta}_i) + \sum_{i=1}^{t}\exp\left(H_{i|-i}\left(\boldsymbol{\beta}_i|\boldsymbol{\beta}_{-i}\right)\right)\left(n_i+\gamma\right)\delta_{\boldsymbol{\beta}_i^*}. \tag{29}$$

F PROOF OF THE THEOREM 3.1

Proposition F.1. *If the data generating process follows equation* (6) *with H replaced by the hierarchical distribution in equation* (21), *then we have*

$$\begin{aligned} p(C) = V_n(t)\prod_{c\in C}\gamma^{(|c|)}, \quad p(C \mid k) = \frac{k_{(t)}}{(\gamma k)^{(n)}}\prod_{c\in C}\gamma^{(|c|)}, \\ p(K = t \mid T = t) = \frac{t_{(t)}}{V_n(t)(rt)^{(n)}}p_K(t) \to 1, \quad C \perp K \mid T, \end{aligned} \tag{30}$$

where $t = |C|$ is the number of clusters while T is the corresponding random variable of t and $V_n(t)$ is defined in equation (4).

The proof of this proposition directly follows from [29], since all conclusions only involves on C, K and T, while the i.i.d assumption on $\boldsymbol{\beta}$ is not used.

Lemma F.2. *Suppose the data generating process in Proposition F.1, such that the distribution is correctly specified. Given the cluster configuration C, the data $\boldsymbol{y}$ and the number of components K are independent.*

Remark F.3. As with MFM, we generalize the same result to exchangeable cases. Since the dependence between $\boldsymbol{y}$ is totally decided by $\boldsymbol{\beta}$, when $\boldsymbol{\beta}$ are exchangeable, all the $\boldsymbol{\beta}$ play the same role in generating $\boldsymbol{y}$. When $\boldsymbol{\beta}$ are marginalized, the cluster configuration C covers the same information with the number of components K and the latent labels $\boldsymbol{z}$.

F.1 Proof of the Lemma F.2

Proof. We show that the conditional independence among data $\boldsymbol{y}$ and the number of components $K = k$ given the cluster configuration C still holds when all $\boldsymbol{\beta}$ are exchangeable.
Let $E_i = \{j : z_j = i\}$, based on the definition of E_i and $\boldsymbol{z}$, we have

$$p(\boldsymbol{y}|\boldsymbol{\beta}, z, k) = \prod_{i=1}^{k}\prod_{j\in E_i}p(y_j|\boldsymbol{\beta}_i) = \prod_{i=1}^{t}\prod_{j\in E_i}p(y_j|\boldsymbol{\beta}_i^*), \tag{31}$$

where $\boldsymbol{\beta}_i^*$, $i = 1, 2, \ldots, t$ are the distinct values of $\boldsymbol{\beta}_{1:k}$ decided by $\boldsymbol{z}$ and $\boldsymbol{y}$, and $\boldsymbol{\beta}_{1:k} = (\boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_k)^\top$. Given $\boldsymbol{z}$, the transformation from variable $\boldsymbol{\beta}_{1:k}$ to $\boldsymbol{\beta}_{1:t}^*$, is totally decided, so when marginalizing the unused $\boldsymbol{\beta}_{(t+1):k}^*$, given any function $g(\boldsymbol{\beta}_{1:t}^*)$, we have the identity

$$\int_{\Theta^k} g(\boldsymbol{\beta}_{1:t}^*)p(\boldsymbol{\beta}|z, k)\,(d\boldsymbol{\beta}) = \int_{\Theta^t} g(\boldsymbol{\beta}_{1:t}^*)p(\boldsymbol{\beta}_{1:t}^*)d\boldsymbol{\beta}^*. \tag{32}$$

Note that $\boldsymbol{\beta}^*_{1:t}$ are exchangeable based on assumption, then the density after marginalizing $\boldsymbol{\beta}$ can be seen

$$\begin{aligned}
p(\boldsymbol{y}|\boldsymbol{z},k) &= \int_{\Theta^k} p(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{z},k)p(\boldsymbol{\beta}|\boldsymbol{z},k)d\boldsymbol{\beta} = \int_{\Theta^k}\prod_{i=1}^{k}\prod_{j\in E_i} p(y_j|\boldsymbol{\beta}_i)p(\boldsymbol{\beta}|\boldsymbol{z},k)\,(d\boldsymbol{\beta}) \\
&= \int_{\Theta^t}\prod_{i=1}^{t}\prod_{j\in E_i} p(y_j|\boldsymbol{\beta}^*_i)p(\boldsymbol{\beta}^*_{1:t})d\boldsymbol{\beta}^* \\
&\overset{(i)}{=} \int_{\Theta^t}\prod_{i=1}^{t} p(\boldsymbol{y}_{E_i}|\boldsymbol{\beta}^*_i)\int\prod_{i=1}^{t} p(\boldsymbol{\beta}^*_i|\boldsymbol{\theta})dF(\boldsymbol{\theta})d\boldsymbol{\beta}^* \\
&\overset{(ii)}{=} \int\int_{\Theta^t}\prod_{i=1}^{t}\left[p(\boldsymbol{y}_{E_i}|\boldsymbol{\beta}^*_i)p(\boldsymbol{\beta}^*_i|\boldsymbol{\theta})\right]d\boldsymbol{\beta}^* dF(\boldsymbol{\theta}) \\
&\overset{(iii)}{=} \int\prod_{i=1}^{t} m_i(\boldsymbol{y}_{E_i},\boldsymbol{\theta})dF(\boldsymbol{\theta}),
\end{aligned} \tag{33}$$

where $m_i(\boldsymbol{y}_{E_i},\boldsymbol{\theta})$ is a function only depends on $\boldsymbol{y}_{E_i}$ and $\boldsymbol{\theta}$. In addition, (i) directly follows from de Finetti's Theorem; for (ii), we apply the Fubini's theorem; (iii) is because the expression depends only on $\boldsymbol{z}, k$ through $C = C(\boldsymbol{z})$ since there is no correspondence between E_i and $\boldsymbol{\beta}^*_i$ after integrating out $\boldsymbol{\beta}^*$. From the last expression, we can see $p(\boldsymbol{y}|\boldsymbol{z},k)$ can be represented as a function of $C, \boldsymbol{y}$, which implies that $\boldsymbol{y}$ and K are conditional independent given the cluster configuration C. □

Based on the fourth line in Proposition 1 and Lemma F.2, we have $C \perp K \mid T$ and $\boldsymbol{y} \perp K \mid C$. Then we have

$$p(\boldsymbol{y}|t,k) = \sum_{C:|C|=t} p(\boldsymbol{y}|C,t,k)p(C|t,k) = \sum_{C:|C|=t} p(\boldsymbol{y}|C,t)p(C|t) = p(\boldsymbol{y}|t), \tag{34}$$

which implies $\boldsymbol{y} \perp K \mid T$. Then for any $n \geq k$,

$$p\left(K=k\mid\boldsymbol{y}\right) = \sum_{t=1}^{k} p\left(K=k\mid T=t,\boldsymbol{y}\right)p\left(T=t\mid\boldsymbol{y}\right) = \sum_{t=1}^{k} p\left(K=k\mid T=t\right)p\left(T=t\mid\boldsymbol{y}\right). \tag{35}$$

In addition, $p(K=t|T=t) = 1/V_n(t) \longrightarrow 1$ as $n \longrightarrow \infty$ based on the third equation in Proposition 1. Thus

$$p\left(K=k\mid\boldsymbol{y}\right) \to \sum_{t=1}^{k} I(k=t)p(T=t\mid\boldsymbol{y}) = p(T=t\mid\boldsymbol{y}). \tag{36}$$