Collaborative Place Models
Supplement 2

Berk Kapicioglu
Foursquare Labs
berk.kapicioglu@gmail.com

David S. Rosenberg
YP Mobile Labs
david.davidr@gmail.com

Robert E. Schapire
Princeton University
schapire@cs.princeton.edu

Tony Jebara
Columbia University
jebara@cs.columbia.edu

1 Bayesian Gaussian Mixture Model (BGMM)

In this section, we provide a formal description of BGMM and describe its generative process. The model is depicted in Figure 1.

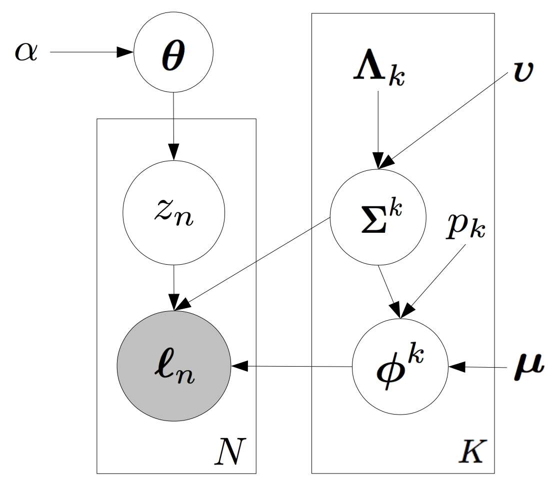

Figure 1: Graphical model representation of BGMM. The geographic coordinates, denoted by $\boldsymbol{\ell}$, are the only observed variables.

Let $\mathcal{N}$ denote the normal distribution and IW denote the inverse-Wishart distribution. Let $\text{Dirichlet}_K(\cdot)$ denote a symmetric Dirichlet if its parameter is a scalar and a general Dirichlet if its parameter is a vector. The generative process of BGMM is described in more formal terms below. Further technical details about the distributions we use in our model can be found in the appendix.

1. Draw a place distribution $\boldsymbol{\theta} \sim \text{Dirichlet}_K(\alpha)$.
2. For each place k,
   (a) Draw a place covariance $\boldsymbol{\Sigma}^k \sim IW(\boldsymbol{\Lambda}_k, v)$.
   (b) Draw a place mean $\boldsymbol{\phi}^k \sim \mathcal{N}\left(\boldsymbol{\mu}, \frac{\boldsymbol{\Sigma}^k}{p_k}\right)$.

3. For each observation index n,
   (a) Draw a place assignment $z_n \sim \text{Categorical}(\boldsymbol{\theta})$.
   (b) Draw a location $\boldsymbol{\ell}_n \sim \mathcal{N}\left(\boldsymbol{\phi}^{z_n}, \boldsymbol{\Sigma}^{z_n}\right)$.

2 Inference

In this section, we derive our inference algorithm, and we present our derivation in multiple steps. First, we use a strategy popularized by Griffiths and Steyvers [1], and derive a collapsed Gibbs sampler to sample from the posterior distribution of the categorical random variables conditioned on the observed geographic coordinates. Second, we derive the conditional likelihood of the posterior samples, which we use to determine the sampler's convergence. Finally, we derive formulas for approximating the posterior expectations of the non-categorical random variables conditioned on the posterior samples.

2.1 Collapsed Gibbs Sampler

In Lemma 1, we derive the collapsed Gibbs sampler for variable $\boldsymbol{z}$.

Given a vector $\boldsymbol{x}$ and an index k, let $\boldsymbol{x}_{-k}$ indicate all the entries of the vector excluding the one at index k. For Lemma 1, assume i denotes the index of the variable that will be sampled.

Lemma 1. *The unnormalized probability of z_i conditioned on the observed location data and remaining categorical variables is*

$$p\left(z_i = k \mid \boldsymbol{z}_{-i}, \boldsymbol{\ell}\right) \propto t_{\tilde{v}_k - 1}\left(\boldsymbol{\ell}_i \mid \tilde{\boldsymbol{\mu}}_k, \frac{\tilde{\boldsymbol{\Lambda}}^k\left(\tilde{p}_k + 1\right)}{\tilde{p}_k\left(\tilde{v}_k - 1\right)}\right)\left(\alpha + \tilde{m}^{k,\cdot}_{\cdot,\cdot}\right).$$

The parameters $\tilde{v}_k$, $\tilde{\boldsymbol{\mu}}_k$, $\tilde{\boldsymbol{\Lambda}}^k$, and $\tilde{p}_k$ are defined in the proof. t denotes the bivariate t-distribution and $\tilde{m}^{k,\cdot}_{\cdot,\cdot}$ denotes counts, both of which are defined in the appendix.

Proof. We decompose the probability into two components using Bayes' theorem:

$$\begin{aligned}
p\left(z_i = k \mid \boldsymbol{z}_{-i}, \boldsymbol{\ell}\right) &= p\left(\boldsymbol{\ell}_i \mid z_i = k, \boldsymbol{z}_{-i}, \boldsymbol{\ell}_{-i}\right) \\
&\quad \times \frac{p\left(z_i = k \mid \boldsymbol{z}_{-i}, \boldsymbol{\ell}_{-i}\right)}{p\left(\boldsymbol{\ell}_i \mid \boldsymbol{z}_{-i}, \boldsymbol{\ell}_{-i}\right)} \\
&= p\left(\boldsymbol{\ell}_i \mid z_i = k, \boldsymbol{z}_{-i}, \boldsymbol{\ell}_{-i}\right) \\
&\quad \times \frac{p\left(z_i = k \mid \boldsymbol{z}_{-i}\right)}{p\left(\boldsymbol{\ell}_i \mid \boldsymbol{z}_{-i}, \boldsymbol{\ell}_{-i}\right)} \\
&\propto p\left(\boldsymbol{\ell}_i \mid z_i = k, \boldsymbol{z}_{-i}, \boldsymbol{\ell}_{-i}\right) && (1) \\
&\quad \times p\left(z_i = k \mid \boldsymbol{z}_{-i}\right). && (2)
\end{aligned}$$

In the first part of the derivation, we operate on (1). We augment it with $\boldsymbol{\phi}$ and $\boldsymbol{\Sigma}$:

$$\begin{aligned}
p\left(\boldsymbol{\ell}_i \mid z_i = k, \boldsymbol{z}_{-i}, \boldsymbol{\ell}_{-i}\right) &= \int\int p\left(\boldsymbol{\ell}_i \mid z_i = k, \boldsymbol{z}_{-i}, \boldsymbol{\ell}_{-i}, \boldsymbol{\phi}^k, \boldsymbol{\Sigma}^k\right) \\
&\quad \times p\left(\boldsymbol{\phi}^k, \boldsymbol{\Sigma}^k \mid z_i = k, \boldsymbol{z}_{-i}, \boldsymbol{\ell}_{-i}\right) \mathrm{d}\boldsymbol{\phi}^k \, \mathrm{d}\boldsymbol{\Sigma}^k \\
&= \int\int p\left(\boldsymbol{\ell}_i \mid z_i = k, \boldsymbol{\phi}^k, \boldsymbol{\Sigma}^k\right) \\
&\quad \times p\left(\boldsymbol{\phi}^k, \boldsymbol{\Sigma}^k \mid z_i = k, \boldsymbol{z}_{-i}, \boldsymbol{\ell}_{-i}\right) \mathrm{d}\boldsymbol{\phi}^k \, \mathrm{d}\boldsymbol{\Sigma}^k \\
&= \int\int \mathcal{N}\left(\boldsymbol{\ell}_i \mid \boldsymbol{\phi}^k, \boldsymbol{\Sigma}^k\right) && (3) \\
&\quad \times p\left(\boldsymbol{\phi}^k, \boldsymbol{\Sigma}^k \mid z_i = k, \boldsymbol{z}_{-i}, \boldsymbol{\ell}_{-i}\right) \mathrm{d}\boldsymbol{\phi}^k \, \mathrm{d}\boldsymbol{\Sigma}^k. && (4)
\end{aligned}$$

We convert (4) into a more tractable form. Let $\tilde{M}^{k,\cdot}_{\cdot,\cdot}$ be a set of indices, which we define in the appendix, and let $\boldsymbol{\ell}_{\tilde{M}^{k,\cdot}_{\cdot,\cdot}}$ denote the subset of observations whose indices are in $\tilde{M}^{k,\cdot}_{\cdot,\cdot}$. In the derivation below, we treat all variables other than $\boldsymbol{\phi}^k$ and $\boldsymbol{\Sigma}^k$ as a constant:

$$
\begin{aligned}
p\left(\boldsymbol{\phi}^k, \boldsymbol{\Sigma}^k \mid z_i = k, \boldsymbol{z}_{-i}, \boldsymbol{\ell}_{-i}\right) &= p\left(\boldsymbol{\phi}^k, \boldsymbol{\Sigma}^k \mid z_i = k, \boldsymbol{z}_{-i}, \boldsymbol{\ell}_{\tilde{M}^{k,\cdot}_{\cdot,\cdot}}\right) \\
&= p\left(\boldsymbol{\phi}^k, \boldsymbol{\Sigma}^k \mid \boldsymbol{z}_{-i}, \boldsymbol{\ell}_{\tilde{M}^{k,\cdot}_{\cdot,\cdot}}\right) \\
&= p\left(\boldsymbol{\ell}_{\tilde{M}^{k,\cdot}_{\cdot,\cdot}} \mid \boldsymbol{\phi}^k, \boldsymbol{\Sigma}^k, \boldsymbol{z}_{-i}\right) \frac{p\left(\boldsymbol{\phi}^k, \boldsymbol{\Sigma}^k \mid \boldsymbol{z}_{-i}\right)}{p\left(\boldsymbol{\ell}_{\tilde{M}^{k,\cdot}_{\cdot,\cdot}} \mid \boldsymbol{z}_{-i}\right)} \\
&\propto p\left(\boldsymbol{\ell}_{\tilde{M}^{k,\cdot}_{\cdot,\cdot}} \mid \boldsymbol{\phi}^k, \boldsymbol{\Sigma}^k, \boldsymbol{z}_{-i}\right) p\left(\boldsymbol{\phi}^k, \boldsymbol{\Sigma}^k\right) \\
&= \left(\prod_{j \in \tilde{M}^{k,\cdot}_{\cdot,\cdot}} p\left(\boldsymbol{\ell}_j \mid \boldsymbol{\phi}^k, \boldsymbol{\Sigma}^k, \boldsymbol{z}_{-i}\right)\right) p\left(\boldsymbol{\phi}^k, \boldsymbol{\Sigma}^k\right) \\
&= \left(\prod_{j \in \tilde{M}^{k,\cdot}_{\cdot,\cdot}} p\left(\boldsymbol{\ell}_j \mid \boldsymbol{\phi}^k, \boldsymbol{\Sigma}^k, z_j\right)\right) p\left(\boldsymbol{\phi}^k, \boldsymbol{\Sigma}^k\right) \\
&= \left(\prod_{j \in \tilde{M}^{k,\cdot}_{\cdot,\cdot}} \mathcal{N}\left(\boldsymbol{\ell}_j \mid \boldsymbol{\phi}^k, \boldsymbol{\Sigma}^k\right)\right) \mathcal{N}\left(\boldsymbol{\phi}^k \mid \boldsymbol{\mu}, \frac{\boldsymbol{\Sigma}^k}{p_k}\right) \\
&\quad \times IW\left(\boldsymbol{\Sigma}^k \mid \boldsymbol{\Lambda}_k, v\right).
\end{aligned}
$$

Since the normal-inverse-Wishart distribution is the conjugate prior of the multivariate normal distribution, the posterior is also a normal-inverse-Wishart distribution,

$$
p\left(\boldsymbol{\phi}^k, \boldsymbol{\Sigma}^k \mid z_i = k, \boldsymbol{z}_{-i}, \boldsymbol{\ell}_{-i}\right) = \mathcal{N}\left(\boldsymbol{\phi}^k \mid \tilde{\boldsymbol{\mu}}_k, \frac{\boldsymbol{\Sigma}^k}{\tilde{p}_k}\right) IW\left(\boldsymbol{\Sigma}^k \mid \tilde{\boldsymbol{\Lambda}}^k, \tilde{v}_k\right), \tag{5}
$$

whose parameters are defined as

$$
\begin{aligned}
\tilde{p}_k &= p_k + \tilde{m}^{k,\cdot}_{\cdot,\cdot}, \\
\tilde{v}_k &= \nu + \tilde{m}^{k,\cdot}_{\cdot,\cdot}, \\
\tilde{\boldsymbol{\ell}}_k &= \frac{1}{\tilde{m}^{k,\cdot}_{\cdot,\cdot}} \sum_{j \in \tilde{M}^{k,\cdot}_{\cdot,\cdot}} \boldsymbol{\ell}_j, \\
\tilde{\boldsymbol{\mu}}_k &= \frac{p_k \boldsymbol{\mu} + \tilde{m}^{k,\cdot}_{\cdot,\cdot} \tilde{\boldsymbol{\ell}}_k}{\tilde{p}_k}, \\
\tilde{S}^k &= \sum_{j \in \tilde{M}^{k,\cdot}_{\cdot,\cdot}} \left(\boldsymbol{\ell}_j - \tilde{\boldsymbol{\ell}}_k\right) \left(\boldsymbol{\ell}_j - \tilde{\boldsymbol{\ell}}_k\right)^T, \\
\tilde{\boldsymbol{\Lambda}}^k &= \boldsymbol{\Lambda}_k + \tilde{S}^k + \frac{p_k \tilde{m}^{k,\cdot}_{\cdot,\cdot}}{p_k + \tilde{m}^{k,\cdot}_{\cdot,\cdot}} \left(\tilde{\boldsymbol{\ell}}_k - \boldsymbol{\mu}\right) \left(\tilde{\boldsymbol{\ell}}_k - \boldsymbol{\mu}\right)^T.
\end{aligned}
$$

The posterior parameters depicted above are derived based on the conjugacy properties of Gaussian distributions, as described in [2]. We rewrite (1) by combining (3), (4), and (5) to obtain

$$
\begin{aligned}
p\left(\boldsymbol{\ell}_i \mid z_i = k, \boldsymbol{z}_{-i}, \boldsymbol{\ell}_{-i}\right) &= \int\int \mathcal{N}\left(\boldsymbol{\ell}_i \mid \boldsymbol{\phi}^k, \boldsymbol{\Sigma}^k\right) \\
&\quad \times p\left(\boldsymbol{\phi}^k, \boldsymbol{\Sigma}^k \mid z_i = k, \boldsymbol{z}_{-i}, \boldsymbol{\ell}_{-i}\right) \,\mathrm{d}\boldsymbol{\phi}^k \,\mathrm{d}\boldsymbol{\Sigma}^k \\
&= \int\int \mathcal{N}\left(\boldsymbol{\ell}_i \mid \boldsymbol{\phi}^k, \boldsymbol{\Sigma}^k\right) \mathcal{N}\left(\boldsymbol{\phi}^k \mid \tilde{\boldsymbol{\mu}}_k, \frac{\boldsymbol{\Sigma}^k}{\tilde{p}_k}\right) \\
&\quad \times IW\left(\boldsymbol{\Sigma}^k \mid \tilde{\boldsymbol{\Lambda}}^k, \tilde{v}_k\right) \,\mathrm{d}\boldsymbol{\phi}^k \,\mathrm{d}\Sigma^k \\
&= t_{\tilde{v}_k - 1}\left(\boldsymbol{\ell}_i \mid \tilde{\boldsymbol{\mu}}_k, \frac{\tilde{\boldsymbol{\Lambda}}^k \left(\tilde{p}_k + 1\right)}{\tilde{p}_k \left(\tilde{v}_k - 1\right)}\right), \qquad (6)
\end{aligned}
$$

where t is the bivariate t-distribution. (6) is derived by applying Equation 258 from [2].

Now, we move onto the second part of the derivation. We operate on (2) and augment it with $\boldsymbol{\theta}$:

$$
\begin{aligned}
p\left(z_i = k \mid \boldsymbol{z}_{-i}\right) &= \int p\left(z_i = k \mid \boldsymbol{z}_{-i}, \boldsymbol{\theta}\right) p\left(\boldsymbol{\theta} \mid \boldsymbol{z}_{-i}\right) \,\mathrm{d}\boldsymbol{\theta} \\
&= \int p\left(z_i = k \mid \boldsymbol{\theta}\right) \qquad (7) \\
&\quad \times p\left(\boldsymbol{\theta} \mid \boldsymbol{z}_{-i}\right) \,\mathrm{d}\boldsymbol{\theta}. \qquad (8)
\end{aligned}
$$

We convert (8) into a more tractable form. As before, let $\tilde{M}_{\cdot,\cdot}^{\cdot,\cdot}$ be a set of indices, which we define in the appendix, and let $\boldsymbol{z}_{\tilde{M}_{\cdot,\cdot}^{\cdot,\cdot}}$ denote the subset of place assignments whose indices are in $\tilde{M}_{\cdot,\cdot}^{\cdot,\cdot}$. In the derivation below, we treat all variables other than $\boldsymbol{\theta}$ as a constant:

$$
\begin{aligned}
p\left(\boldsymbol{\theta} \mid \boldsymbol{z}_{-i}\right) &= p\left(\boldsymbol{\theta} \mid \boldsymbol{z}_{\tilde{M}_{\cdot,\cdot}^{\cdot,\cdot}}\right) \\
&= \frac{p\left(\boldsymbol{z}_{\tilde{M}_{\cdot,\cdot}^{\cdot,\cdot}} \mid \boldsymbol{\theta}\right) p\left(\boldsymbol{\theta}\right)}{p\left(\boldsymbol{z}_{\tilde{M}_{\cdot,\cdot}^{\cdot,\cdot}}\right)} \\
&\propto \prod_{j \in \tilde{M}_{\cdot,\cdot}^{\cdot,\cdot}} p\left(z_j \mid \boldsymbol{\theta}\right) p\left(\boldsymbol{\theta}\right) \\
&= \prod_{j \in \tilde{M}_{\cdot,\cdot}^{\cdot,\cdot}} \text{Categorical}\left(z_j \mid \boldsymbol{\theta}\right) \text{Dirichlet}_K\left(\boldsymbol{\theta} \mid \alpha\right) \\
\implies p\left(\boldsymbol{\theta} \mid \boldsymbol{z}_{-i}\right) &= \text{Dirichlet}_K\left(\boldsymbol{\theta} \mid \alpha + \tilde{m}_{\cdot,\cdot}^{1,\cdot}, \ldots, \alpha + \tilde{m}_{\cdot,\cdot}^{K,\cdot}\right), \qquad (9)
\end{aligned}
$$

where the last step follows because Dirichlet distribution is the conjugate prior of the categorical distribution. We rewrite (2) by combining (7), (8), and (9):

$$
\begin{aligned}
p\left(z_i = k \mid \boldsymbol{z}_{-i}\right) &= \int p\left(z_i = k \mid \boldsymbol{\theta}\right) p\left(\boldsymbol{\theta} \mid \boldsymbol{z}_{-i}\right) \,\mathrm{d}\boldsymbol{\theta} \\
&= \int \theta_k \text{Dirichlet}_K\left(\boldsymbol{\theta} \mid \alpha + \tilde{m}_{\cdot,\cdot}^{1,\cdot}, \ldots, \alpha + \tilde{m}_{\cdot,\cdot}^{K,\cdot}\right) \,\mathrm{d}\boldsymbol{\theta} \\
&= \frac{\alpha + \tilde{m}_{\cdot,\cdot}^{k,\cdot}}{K\alpha + \tilde{m}_{\cdot,\cdot}^{\cdot,\cdot}}. \qquad (10)
\end{aligned}
$$

The last step follows because it is the expected value of the Dirichlet distribution.

Finally, we combine (1), (2), (6), and (10) to obtain the unnormalized probability distribution:

$$
\begin{aligned}
p\left(z_i = k \mid \boldsymbol{z}_{-i}, \boldsymbol{\ell}\right) &\propto p\left(\boldsymbol{\ell}_i \mid z_i = k, \boldsymbol{z}_{-i}, \boldsymbol{\ell}_{-i}\right) \\
&\quad \times p\left(z_i = k \mid \boldsymbol{z}_{-i}\right) \\
&= t_{\tilde{v}_k - 1}\left(\boldsymbol{\ell}_i \mid \tilde{\boldsymbol{\mu}}_k, \frac{\tilde{\boldsymbol{\Lambda}}^k \left(\tilde{p}_k + 1\right)}{\tilde{p}_k \left(\tilde{v}_k - 1\right)}\right) \frac{\alpha + \tilde{m}_{\cdot,\cdot}^{k,\cdot}}{K\alpha + \tilde{m}_{\cdot,\cdot}^{\cdot,\cdot}}.
\end{aligned}
$$

□

2.2 Likelihoods

In this subsection, we derive the conditional likelihoods of the posterior samples conditioned on the observed geographical coordinates. We use these conditional likelihoods to determine the sampler's convergence.

We present the derivations in multiple lemmas and combine them in a theorem at the end of the subsection. Let Γ denote the gamma function.

Lemma 2. *The marginal probability of the categorical random variable $\boldsymbol{z}$ is*

$$p(\boldsymbol{z}) = \frac{\Gamma(K\alpha) \prod_{k=1}^{K} \Gamma\left(\alpha + m^{k,\cdot}_{\cdot,\cdot}\right)}{\Gamma(\alpha)^{K} \Gamma\left(K\alpha + m^{\cdot,\cdot}_{\cdot,\cdot}\right)},$$

where the counts $m^{k,\cdot}_{\cdot,\cdot}$ and $m^{\cdot,\cdot}_{\cdot,\cdot}$ are defined in the appendix.

Proof. Below, we will augment the marginal probability with $\boldsymbol{\theta}$, and then factorize it based on the conditional independence assumptions made by our model:

$$\begin{aligned} p(\boldsymbol{z}) &= \int p(\boldsymbol{z} \mid \boldsymbol{\theta})\, p(\boldsymbol{\theta})\, \mathrm{d}\boldsymbol{\theta} \\ &= \int p(\boldsymbol{\theta}) \left(\prod_{j} p(z_j \mid \boldsymbol{\theta}) \right) \mathrm{d}\boldsymbol{\theta} \\ &= \int \left(\mathrm{Dirichlet}_K(\boldsymbol{\theta} \mid \alpha) \prod_{j} \mathrm{Categorical}(z_j \mid \boldsymbol{\theta}) \right) \mathrm{d}\boldsymbol{\theta}. \end{aligned} \tag{11}$$

Now, we substitute the probabilities in (11) with Dirichlet and categorical distributions, which are defined in more detail in the appendix:

$$\begin{aligned} p(\boldsymbol{z}) &= \int \left(\mathrm{Dirichlet}_K(\boldsymbol{\theta} \mid \alpha) \prod_{j} \mathrm{Categorical}(z_j \mid \boldsymbol{\theta}) \right) \mathrm{d}\boldsymbol{\theta} \\ &= \int \left(\frac{1}{B(\alpha)} \prod_{k=1}^{K} \theta_k^{\alpha-1} \right) \left(\prod_{k=1}^{K} \theta_k^{m^{k,\cdot}_{\cdot,\cdot}} \right) \mathrm{d}\boldsymbol{\theta} \\ &= \int \left(\frac{1}{B(\alpha)} \prod_{k=1}^{K} \theta_k^{\alpha-1+m^{k,\cdot}_{\cdot,\cdot}} \right) \mathrm{d}\boldsymbol{\theta} \\ &= \frac{1}{B(\alpha)} B\left(\alpha + m^{1,\cdot}_{\cdot,\cdot}, \ldots, \alpha + m^{K,\cdot}_{\cdot,\cdot}\right) \\ &= \frac{\Gamma(K\alpha) \prod_{k=1}^{K} \Gamma\left(\alpha + m^{k,\cdot}_{\cdot,\cdot}\right)}{\Gamma(\alpha)^{K} \Gamma\left(K\alpha + m^{\cdot,\cdot}_{\cdot,\cdot}\right)}. \end{aligned}$$

□

For our final derivation, let Γ_2 denote the bivariate gamma function, and let $|\cdot|$ denote the determinant.

Lemma 3. *The conditional probability of the observed locations $\boldsymbol{\ell}$ conditioned on $\boldsymbol{z}$ is*

$$p(\boldsymbol{\ell} \mid \boldsymbol{z}) = \prod_{k=1}^{K} \frac{\Gamma_2\left(\frac{\hat{\upsilon}_k}{2}\right) |\boldsymbol{\Lambda}_k|^{\frac{\nu}{2}} p_k}{\pi^{m^{k,\cdot}_{\cdot,\cdot}} \Gamma_2\left(\frac{\nu}{2}\right) \left|\hat{\boldsymbol{\Lambda}}^k\right|^{\frac{\hat{\upsilon}_k}{2}} \hat{p}_k}.$$

The parameters $\hat{\upsilon}_k$, $\hat{\boldsymbol{\Lambda}}^k$, and $\hat{p}_k$ are defined in the proof, and the counts $m^{k,\cdot}_{\cdot,\cdot}$ are defined in the appendix.

Proof. We will factorize the probability using the conditional independence assumptions made by the model, and then simplify the resulting probabilities by integrating out the means and covariances associated with the place clusters:

$$
\begin{aligned}
p(\boldsymbol{\ell} \mid \boldsymbol{z}) &= \prod_{k=1}^{K} p\left(\boldsymbol{\ell}_{M^{k}_{\cdot,\cdot}} \mid \boldsymbol{z}\right) \\
&= \prod_{k=1}^{K} \int\int p\left(\boldsymbol{\ell}_{M^{k}_{\cdot,\cdot}} \mid \boldsymbol{z}, \boldsymbol{\phi}^k, \boldsymbol{\Sigma}^k\right) p\left(\boldsymbol{\phi}^k, \boldsymbol{\Sigma}^k\right) \mathrm{d}\boldsymbol{\phi}^k \,\mathrm{d}\boldsymbol{\Sigma}^k \\
&= \prod_{k=1}^{K} \int\int p\left(\boldsymbol{\phi}^k_u, \boldsymbol{\Sigma}^k_u\right) \prod_{j \in M^{k}_{\cdot,\cdot}} p\left(\boldsymbol{\ell}_j \mid z_j, \boldsymbol{\phi}^k, \boldsymbol{\Sigma}^k\right) \mathrm{d}\boldsymbol{\phi}^k \,\mathrm{d}\boldsymbol{\Sigma}^k \\
&= \prod_{k=1}^{K} \int\int \mathcal{N}\left(\boldsymbol{\phi}^k \mid \boldsymbol{\mu}, \frac{\boldsymbol{\Sigma}^k}{p_k}\right) IW\left(\boldsymbol{\Sigma}^k \mid \boldsymbol{\Lambda}_k, v\right) \qquad (12) \\
&\quad \times \prod_{j \in M^{k}_{\cdot,\cdot}} \mathcal{N}\left(\boldsymbol{\ell}_j \mid \boldsymbol{\phi}^k, \boldsymbol{\Sigma}^k\right) \mathrm{d}\boldsymbol{\phi}^k \,\mathrm{d}\boldsymbol{\Sigma}^k.
\end{aligned}
$$

We apply Equation 266 from [2], which describes the conjugacy properties of Gaussian distributions, to reformulate (12) into its final form:

$$
\begin{aligned}
p(\boldsymbol{\ell} \mid \boldsymbol{z}) &= \prod_{k=1}^{K} \int\int \mathcal{N}\left(\boldsymbol{\phi}^k \mid \boldsymbol{\mu}, \frac{\boldsymbol{\Sigma}^k}{p_k}\right) IW\left(\boldsymbol{\Sigma}^k \mid \boldsymbol{\Lambda}_k, v\right) \prod_{j \in M^{k}_{\cdot,\cdot}} \mathcal{N}\left(\boldsymbol{\ell}_j \mid \boldsymbol{\phi}^k, \boldsymbol{\Sigma}^k\right) \mathrm{d}\boldsymbol{\phi}^k \,\mathrm{d}\boldsymbol{\Sigma}^k \\
&= \prod_{k=1}^{K} \frac{\Gamma_2\left(\frac{\hat{v}_k}{2}\right) |\boldsymbol{\Lambda}_k|^{\frac{\nu}{2}} p_k}{\pi^{m^{k}_{\cdot,\cdot}} \Gamma_2\left(\frac{\nu}{2}\right) \left|\hat{\boldsymbol{\Lambda}}^k\right|^{\frac{\hat{v}_k}{2}} \hat{p}_k}.
\end{aligned}
$$

The definitions for $\hat{v}_k$, $\hat{\boldsymbol{\Lambda}}^k$, and $\hat{p}_k$ are provided in (5). □

Finally, we combine Lemmas 2 and 3 to provide the log-likelihood of the samples $\boldsymbol{z}$ conditioned on the observations $\boldsymbol{\ell}$.

Lemma 4. *The log-likelihood of the samples $\boldsymbol{z}$ conditioned on the observations $\boldsymbol{\ell}$ is*

$$
\begin{aligned}
\log p(\boldsymbol{z} \mid \boldsymbol{\ell}) = &\left(\sum_{k=1}^{K} \log \Gamma\left(\alpha + m^{k}_{\cdot,\cdot}\right)\right) \\
&+ \left(\sum_{k=1}^{K} \left(\log \Gamma_2\left(\frac{\hat{v}_k}{2}\right) - m^{k}_{\cdot,\cdot} \log \pi - \frac{\hat{v}_k}{2} \log\left|\hat{\boldsymbol{\Lambda}}^k\right| - \log \hat{p}_k\right)\right) + C,
\end{aligned}
$$

where C denotes the constant terms.

Proof. The result follows by multiplying the probabilities stated in Lemmas 2 and 3, and applying the logarithm function. □

2.3 Parameter estimation

In Subsection 2.1, we described a collapsed Gibbs sampler for sampling the posteriors of the categorical random variables. Below, Lemmas 5 and 6 show how these samples, denoted as $\boldsymbol{z}$, can be used to approximate the posterior expectations of $\boldsymbol{\theta}$, $\boldsymbol{\phi}$, and $\boldsymbol{\Sigma}$.

Lemma 5. *The expectation of $\boldsymbol{\theta}$ given the observed geographical coordinates and the posterior samples is*

$$\hat{\theta}_k = \mathbb{E}\left[\theta_k \mid \boldsymbol{z}, \boldsymbol{\ell}\right] = \frac{\alpha + m^{k}_{\cdot,\cdot}}{K\alpha + m^{\cdot}_{\cdot,\cdot}},$$

where the counts $m^{k}_{\cdot,\cdot}$ and $m^{\cdot}_{\cdot,\cdot}$ are defined in the appendix.

Proof.

$$\begin{aligned}
p(\boldsymbol{\theta} \mid \boldsymbol{z}, \boldsymbol{\ell}) &= p(\boldsymbol{\theta} \mid \boldsymbol{z}) \\
&= \frac{p(\boldsymbol{z} \mid \boldsymbol{\theta})\, p(\boldsymbol{\theta})}{p(\boldsymbol{z})} \\
&= \frac{\prod_j p(z_j \mid \boldsymbol{\theta})\, p(\boldsymbol{\theta})}{p(\boldsymbol{z})} \\
&\propto \text{Dirichlet}_K(\boldsymbol{\theta} \mid \alpha) \prod_j \text{Categorical}(z_j \mid \boldsymbol{\theta}) \\
&= \text{Dirichlet}_K\left(\boldsymbol{\theta} \mid \alpha + m_{\cdot,\cdot}^{1,\cdot}, \ldots, \alpha + m_{\cdot,\cdot}^{K,\cdot}\right) \\
\implies \hat{\theta}_k = \mathbb{E}[\theta_k \mid \boldsymbol{z}, \boldsymbol{\ell}] &= \frac{\alpha + m_{\cdot,\cdot}^{k,\cdot}}{K\alpha + m_{\cdot,\cdot}^{\cdot,\cdot}}.
\end{aligned}$$

□

Lemma 6. *The expectations of $\boldsymbol{\phi}$ and $\boldsymbol{\Sigma}$ given the observed geographical coordinates and the posterior samples is*

$$\hat{\boldsymbol{\phi}}_u^k = \mathbb{E}\left[\boldsymbol{\phi}_u^k \mid \boldsymbol{z}, \boldsymbol{\ell}\right] = \hat{\boldsymbol{\mu}}_k^u$$

and

$$\hat{\boldsymbol{\Sigma}}_u^k = \mathbb{E}\left[\boldsymbol{\Sigma}_u^k \mid \boldsymbol{z}, \boldsymbol{\ell}\right] = \frac{\hat{\boldsymbol{\Lambda}}_u^k}{\hat{v}_k^u - 3}.$$

Parameters $\hat{\boldsymbol{\mu}}_k^u$, $\hat{\boldsymbol{\Lambda}}_u^k$, and $\hat{v}_k^u$ are defined in the proof of Lemma 1.

Proof.

$$
\begin{aligned}
p\left(\boldsymbol{\phi}^k, \boldsymbol{\Sigma}^k \mid \boldsymbol{z}, \boldsymbol{\ell}\right) &= p\left(\boldsymbol{\phi}^k, \boldsymbol{\Sigma}^k \mid \boldsymbol{z}, \boldsymbol{\ell}\right) \\
&= p\left(\boldsymbol{\phi}^k, \boldsymbol{\Sigma}^k \mid \boldsymbol{z}, \boldsymbol{\ell}_{M^{k,\cdot}_{\cdot,\cdot}}\right) \\
&= \frac{p\left(\boldsymbol{\ell}_{M^{k,\cdot}_{\cdot,\cdot}} \mid \boldsymbol{\phi}^k, \boldsymbol{\Sigma}^k, \boldsymbol{z}\right) p\left(\boldsymbol{\phi}^k, \boldsymbol{\Sigma}^k \mid \boldsymbol{z}\right)}{p\left(\boldsymbol{\ell}_{M^{k,\cdot}_{\cdot,\cdot}} \mid \boldsymbol{z}\right)} \\
&= \frac{\prod_{j \in M^{k,\cdot}_{\cdot,\cdot}} p\left(\boldsymbol{\ell}_j \mid \boldsymbol{\phi}^k, \boldsymbol{\Sigma}^k, \boldsymbol{z}\right) p\left(\boldsymbol{\phi}^k, \boldsymbol{\Sigma}^k\right)}{p\left(\boldsymbol{\ell}_{M^{k,\cdot}_{\cdot,\cdot}} \mid \boldsymbol{z}\right)} \\
&= \frac{\mathcal{N}\left(\boldsymbol{\phi}^k \mid \boldsymbol{\mu}, \frac{\boldsymbol{\Sigma}^k}{p_k}\right) IW\left(\boldsymbol{\Sigma}^k \mid \boldsymbol{\Lambda}_k, v\right) \prod_{j \in M^{k,\cdot}_{\cdot,\cdot}} \mathcal{N}\left(\boldsymbol{\ell}_j \mid \boldsymbol{\phi}^k, \boldsymbol{\Sigma}^k\right)}{p\left(\boldsymbol{\ell}_{M^{k,\cdot}_{\cdot,\cdot}} \mid \boldsymbol{z}\right)} \\
&\propto \mathcal{N}\left(\boldsymbol{\phi}^k \mid \boldsymbol{\mu}, \frac{\boldsymbol{\Sigma}^k}{p_k}\right) IW\left(\boldsymbol{\Sigma}^k \mid \boldsymbol{\Lambda}_k, v\right) \prod_{j \in M^{k,\cdot}_{\cdot,\cdot}} \mathcal{N}\left(\boldsymbol{\ell}_j \mid \boldsymbol{\phi}^k, \boldsymbol{\Sigma}^k\right) \\
\implies p\left(\boldsymbol{\phi}^k, \boldsymbol{\Sigma}^k \mid \boldsymbol{z}, \boldsymbol{\ell}\right) &= \mathcal{N}\left(\boldsymbol{\phi}^k \mid \hat{\boldsymbol{\mu}}_k, \frac{\boldsymbol{\Sigma}^k}{\hat{p}_k}\right) IW\left(\boldsymbol{\Sigma}^k \mid \hat{\boldsymbol{\Lambda}}^k, \hat{v}_k\right) \\
\implies p\left(\boldsymbol{\phi}^k \mid \boldsymbol{z}, \boldsymbol{\ell}\right) &= t_{\hat{v}_k - 1}\left(\boldsymbol{\phi}^k \mid \hat{\boldsymbol{\mu}}_k, \frac{\hat{\boldsymbol{\Lambda}}^k}{\hat{p}_k\left(\hat{v}_k - 1\right)}\right) \\
\implies \hat{\boldsymbol{\phi}}^k = \mathbb{E}\left[\boldsymbol{\phi}^k \mid \boldsymbol{z}, \boldsymbol{\ell}\right] &= \hat{\boldsymbol{\mu}}_k \\
\implies p\left(\boldsymbol{\Sigma}^k \mid \boldsymbol{z}, \boldsymbol{\ell}\right) &= IW\left(\boldsymbol{\Sigma}^k_u \mid \hat{\boldsymbol{\Lambda}}^k, \hat{v}_k\right) \\
\implies \hat{\boldsymbol{\Sigma}}^k = \mathbb{E}\left[\boldsymbol{\Sigma}^k \mid \boldsymbol{z}, \boldsymbol{\ell}\right] &= \frac{\hat{\boldsymbol{\Lambda}}^k}{\hat{v}_k - 3}.
\end{aligned}
$$

□

3 Appendix

3.1 Miscellaneous notation

Throughout the paper, we use various notations to represent sets of indices and their cardinalities. Vectors $\boldsymbol{y}$ and $\boldsymbol{z}$ denote the component and place assignments in CPM, respectively. Each vector entry is identified by a tuple index (u, w, n), where $u \in \{1, \ldots, U\}$ is a user, $w \in \{1, \ldots, W\}$ is a weekhour, and $n \in \{1, \ldots, N_{u,w}\}$ is an iteration index.

For the subsequent notations, we assume that the random variables $\boldsymbol{y}$ and $\boldsymbol{z}$ are already sampled. We refer to a subset of indices using

$$M^{k_0, f_0}_{u_0, w_0} = \left\{(\dot{u}, \dot{w}, \dot{n}) \mid z_{\dot{u},\dot{w},\dot{n}} = k_0, y_{\dot{u},\dot{w},\dot{n}} = f_0, \dot{u} = u_0, \dot{w} = w_0\right\},$$

where u_0 denotes the user, w_0 denotes the weekhour, k_0 denotes the place, and f_0 denotes the component. If we want the subset of indices to be unrestricted with respect to a category, we use the placeholder "·". For example,

$$M^{\cdot, f_0}_{u_0, w_0} = \left\{(\dot{u}, \dot{w}, \dot{n}) \mid y_{\dot{u},\dot{w},\dot{n}} = f_0, \dot{u} = u_0, \dot{w} = w_0\right\}$$

has no constraints with respect to places.

Given a subset of indices denoted by M, the lowercase $m = |M|$ denotes its cardinality. For example, given a set of indices

$$M_{u_0,\cdot}^{\cdot,f_0} = \left\{ (\dot{u}, \dot{w}, \dot{n}) \mid z_{\dot{u},\dot{w},\dot{n}} = \cdot, y_{\dot{u},\dot{w},\dot{n}} = f_0, \dot{u} = u_0, \dot{w} = w_0 \right\},$$

its cardinality is

$$m_{u_0,\cdot}^{\cdot,f_0} = \left| M_{u_0,\cdot}^{\cdot,f_0} \right|.$$

For the collapsed Gibbs sampler, the sets of indices and cardinalities used in the derivations exclude the index that will be sampled. We use "$\sim$" to modify sets or cardinalities for this exclusion. Let (u, w, n) denote the index that will be sampled, then given an index set M, let $\tilde{M} = M - \{(u, w, n)\}$ represent the excluding set and let $\tilde{m} = \left| \tilde{M} \right|$ represent the corresponding cardinality. For example,

$$\tilde{M}_{u_0,\cdot}^{\cdot,f_0} = M_{u_0,\cdot}^{\cdot,f_0} - \{(u, w, n)\}$$

and

$$\tilde{m}_{u_0,\cdot}^{\cdot,f_0} = \left| \tilde{M}_{u_0,\cdot}^{\cdot,f_0} \right|.$$

In the proof of Lemma 1, parameters $\tilde{v}_k^u$, $\tilde{\boldsymbol{\mu}}_k^u$, $\tilde{\boldsymbol{\Lambda}}_u^k$, and $\tilde{p}_k^u$ are defined using cardinalities that exclude the current index (u, w, n). Similarly, in the proof of Lemma 6, parameters $\hat{\boldsymbol{\mu}}_k^u$, $\hat{\boldsymbol{\Lambda}}_u^k$, and $\hat{v}_k^u$ are defined like their wiggly versions, but the counts used in their definitions do not exclude the current index.

3.2 Probability distributions

Let Γ_2 denote a bivariate gamma function, defined as

$$\Gamma_2(a) = \pi^{\frac{1}{2}} \prod_{j=1}^{2} \Gamma\left(a + \frac{1-j}{2}\right).$$

Let $\nu > 1$ and let $\boldsymbol{\Lambda} \in \mathbb{R}^{2\times 2}$ be a positive definite scale matrix. The inverse-Wishart distribution, which is the conjugate prior to the multivariate normal distribution, is defined as

$$IW(\boldsymbol{\Sigma} \mid \boldsymbol{\Lambda}, \nu) = \frac{|\boldsymbol{\Lambda}|^{\frac{\nu}{2}}}{2^{\nu}\Gamma_2\left(\frac{\nu}{2}\right)} |\boldsymbol{\Sigma}|^{\frac{-\nu-3}{2}} \exp\left(-\frac{1}{2}\operatorname{tr}\left(\boldsymbol{\Lambda}\boldsymbol{\Sigma}^{-1}\right)\right).$$

Let $\boldsymbol{\Sigma} \in \mathbb{R}^{2\times 2}$ be a positive definite covariance matrix and let $\boldsymbol{\mu} \in \mathbb{R}^2$ denote a mean vector. The multivariate normal distribution is defined as

$$\mathcal{N}(\boldsymbol{\ell} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) = (2\pi)^{-1} |\boldsymbol{\Sigma}|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\boldsymbol{\ell} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{\ell} - \boldsymbol{\mu})\right).$$

Let $\nu > 1$ and let $\boldsymbol{\Sigma} \in \mathbb{R}^{2\times 2}$, then the 2-dimensional t-distribution is defined as

$$t_v(x \mid \boldsymbol{\mu}, \Sigma) = \frac{\Gamma\left(\frac{\nu}{2} + 1\right)}{\Gamma\left(\frac{\nu}{2}\right)} \frac{|\Sigma|^{-\frac{1}{2}}}{\nu\pi} \left(1 + \frac{1}{\nu}(x - \boldsymbol{\mu})^T \Sigma^{-1} (x - \boldsymbol{\mu})\right)^{-\frac{\nu}{2}-1}.$$

Let $K > 1$ be the number of categories and let $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K)$ be the concentration parameters, where $\alpha_k > 0$ for all $k \in \{1, \ldots, K\}$. Then, the K-dimensional Dirichlet distribution, which is the conjugate prior to the categorical distribution, is defined as

$$\text{Dirichlet}_K(\boldsymbol{x} \mid \boldsymbol{\alpha}) = \frac{1}{B(\boldsymbol{\alpha})} \prod_{k=1}^{K} x_k^{\alpha_k - 1},$$

where

$$B(\boldsymbol{\alpha}) = \frac{\prod_{k=1}^{K} \Gamma(\alpha_k)}{\Gamma\left(\sum_{k=1}^{K} \alpha_k\right)}.$$

We abuse the Dirichlet notation slightly and use it to define the K-dimensional *symmetric* Dirichlet distribution as well. Let $\beta > 0$ be a scalar concentration parameter. Then, the symmetric Dirichlet distribution is defined as

$$\text{Dirichlet}_K(\boldsymbol{x} \mid \beta) = \text{Dirichlet}_K(\boldsymbol{x} \mid \alpha_1, \ldots, \alpha_K),$$

where $\beta = \alpha_k$ for all $k \in \{1, \ldots, K\}$.

References

[1] Thomas L. Griffiths and Mark Steyvers. Finding scientific topics. *Proceedings of the National Academy of Sciences of the United States of America*, 101(Suppl 1):5228–5235, April 2004.

[2] Kevin Murphy. Conjugate bayesian analysis of the gaussian distribution. October 2007.