Learning Community-based Preferences via Dirichlet Process Mixtures of Gaussian Processes

1 Approximation of the Community Utilities

The posterior of the utility (latent) functions $\mathbf{f}$ given all the preferences is:

$$p(\mathbf{f}|\mathbf{c},\mathcal{D},\boldsymbol{\theta}) = \frac{1}{Z}p(\mathcal{D}|\mathbf{f},\mathbf{c},\boldsymbol{\theta})p(\mathbf{f}|\mathbf{c}) \tag{1}$$

Since the likelihood is factorized we can take advantage of sequential approximation methods such as *Expectation Propagation* (EP). EP approximates the posterior $p(\mathbf{f}|\mathbf{c},\mathcal{D},\boldsymbol{\theta})$ by a tractable distribution $q(\mathbf{f}|\mathbf{c})$. EP assumes that each likelihood term $p(\mathbf{x}_i \succ \mathbf{x}_j|f_i^{\mathbf{u}}, f_j^{\mathbf{u}}, \mathbf{c_u} = c, \mathbf{K_x})$ can be approximated by a distribution $q(f_i^{\mathbf{u}}, f_j^{\mathbf{u}}|\mathbf{c_u} = c, \mathbf{K_x})$ such that the approximated posterior $q(\mathbf{f}|\mathbf{c})$ factorizes over $q(f_i^{\mathbf{u}}, f_j^{\mathbf{u}}|\mathbf{c_u})$. Then EP iteratively approximates each $q(f_i^{\mathbf{u}}, f_j^{\mathbf{u}}|\mathbf{c_u})$ in turn by dividing it out from the approximated posterior $q(\mathbf{f}|\mathbf{c})$ (obtaining the cavity distribution), multiplying in the true likelihood $p(\mathbf{x}_i \succ \mathbf{x}_j|f_i^{\mathbf{u}}, f_j^{\mathbf{u}}, \mathbf{c_u}, \mathbf{K_x})$, and projecting the result back to its factorized form by matching its moments to an updated $q(f_i^{\mathbf{u}}, f_j^{\mathbf{u}}|\mathbf{c_u})$.

This overall procedure is motivated by the aim to minimize the KL−divergence between the true posterior $p(\mathbf{f}|\mathcal{D},\mathbf{c},\mathbf{K})$ and its approximation $q(\mathbf{f}|\mathbf{c})$.

In the preference learning case we detailed earlier, we can approximate the posterior with a Gaussian:

$$\begin{aligned} q(\mathbf{f}|\mathbf{c}) &= \prod_c \frac{1}{\tilde{Z}^c} \prod_{\mathbf{u}\in U} p(\mathbf{f}|\mathbf{c_u} = c) \prod_{\{\{i,j\}|\mathbf{x}_i \succ \mathbf{x}_j \in \mathcal{D}^{\mathbf{u}}\}} q(f_i^{\mathbf{u}}, f_j^{\mathbf{u}}|\mathbf{c_u} = c) \\ &= \mathcal{N}(\mathbf{f}; \boldsymbol{\mu}^c, \boldsymbol{\Sigma}^c). \end{aligned} \tag{2}$$

where $\boldsymbol{\mu}^c$ and $\boldsymbol{\Sigma}^c$ denote the mean and covariance of the Gaussian distribution for the community user $\mathbf{u}$ belongs to corresponding to $\boldsymbol{\theta}_{\mathbf{c_u}}$. We are interested in locally approximating each likelihood term as:

$$\begin{aligned} p(\mathbf{x}_i \succ \mathbf{x}_j|f_i^{\mathbf{u}}, f_j^{\mathbf{u}}, \mathbf{c_u}) &\approx q(f_i^{\mathbf{u}}, f_j^{\mathbf{u}}|\mathbf{c_u}) \\ &= \tilde{Z}_{i,j}^{\mathbf{u}} \mathcal{N}(f_i^{\mathbf{u}}, f_j^{\mathbf{u}}; \tilde{\boldsymbol{\mu}}^c{}_{\mathbf{u},[i,j]}, \tilde{\boldsymbol{\Sigma}}^c{}_{\mathbf{u},[i,j]}), \end{aligned} \tag{3}$$

where $\mathcal{N}(f_i^{\mathbf{u}}, f_j^{\mathbf{u}}; \tilde{\boldsymbol{\mu}}^c{}_{\mathbf{u},[i,j]}, \tilde{\boldsymbol{\Sigma}}^c{}_{\mathbf{u},[i,j]})$ denotes the local two-dimensional Gaussian over $[f_i^{\mathbf{u}}, f_j^{\mathbf{u}}]^\top$ with mean $\tilde{\boldsymbol{\mu}}^c{}_{\mathbf{u},[i,j]}$ and covariance $\tilde{\boldsymbol{\Sigma}}^c{}_{\mathbf{u},[i,j]}$ corresponding to items i and j.

Hence we can approximate the posterior in Equation 2 with the following parameters:

$$\boldsymbol{\mu}^{c}{}_{\mathbf{u},[i,j]} = \boldsymbol{\Sigma}^{c}{}_{\mathbf{u},[i,j]} \tilde{\boldsymbol{\Sigma}}^{c\,-1}_{\mathbf{u},[i,j]} \tilde{\boldsymbol{\mu}}^{c}{}_{\mathbf{u},[i,j]} \tag{4}$$

$$\boldsymbol{\Sigma}^{c\,-1}_{\mathbf{u},[i,j]} = (\mathbf{K}^{c\,-1}_{\mathbf{u},[i,j]} + \tilde{\boldsymbol{\Sigma}}^{c\,-1}_{\mathbf{u},[i,j]}). \tag{5}$$

This means that in order to determine the parameters of our approximate posterior, we need to compute estimates of the local parameters $\tilde{\boldsymbol{\mu}}^{c}$ and $\tilde{\boldsymbol{\Sigma}}^{c}$. To show these updates, we need to define additional distributions: (a) the *cavity* distribution which we will denote with the backslash symbol "$\backslash$" and (b) the *unnormalized marginal posterior*, which we will denote with the hat symbol "$\hat{}$".

Here we only show how to compute the parameters necessary to estimate the posterior. We iterate through the following steps:

1. Update the cavity distribution: The cavity distribution $q_{\backslash}(f_i^{\mathbf{u}}, f_j^{\mathbf{u}}|\mathbf{c_u} = c)$ results from multiplying the prior by all the local approximate likelihood terms except $q(f_i^{\mathbf{u}}, f_j^{\mathbf{u}}|\mathbf{c_u} = c)$ and marginalizing all latent dimensions except $f_i^{\mathbf{u}}$ and $f_j^{\mathbf{u}}$. This is done in practice simply by removing the current approximate likelihood term from the approximate posterior. Hence we obtain:

$$q_{\backslash}(f_i^{\mathbf{u}}, f_j^{\mathbf{u}}|\mathbf{c_u}) = \mathcal{N}(f_i^{\mathbf{u}}, f_j^{\mathbf{u}}; \boldsymbol{\mu}^{c}_{\backslash \mathbf{u},[i,j]}, \boldsymbol{\Sigma}^{c}_{\backslash \mathbf{u},[i,j]}) \tag{6}$$

$$\boldsymbol{\mu}^{c}_{\backslash \mathbf{u},[i,j]} = \boldsymbol{\Sigma}^{c}_{\backslash \mathbf{u},[i,j]} (\boldsymbol{\Sigma}^{c}_{\mathbf{u},[i,j]}{}^{-1} \boldsymbol{\mu}^{c}_{\mathbf{u},[i,j]} - \tilde{\boldsymbol{\Sigma}}^{c\,-1}_{\mathbf{u},[i,j]} \tilde{\boldsymbol{\mu}}^{c}{}_{\mathbf{u},[i,j]}) \tag{7}$$

$$\boldsymbol{\Sigma}^{c}_{\backslash \mathbf{u},[i,j]} = (\boldsymbol{\Sigma}^{c\,-1}_{\mathbf{u},[i,j]} - \tilde{\boldsymbol{\Sigma}}^{c\,-1}_{\mathbf{u},[i,j]})^{-1}. \tag{8}$$

2. Update the unnormalized marginal posterior: This results from finding the unnormalized Gaussian that best approximates the product of the cavity distribution and the exact likelihood:

$$\hat{q}(f_i^{\mathbf{u}}, f_j^{\mathbf{u}}|\mathbf{c_u}) \approx p(\mathbf{x}_i \succ \mathbf{x}_j|f_i^{\mathbf{u}}, f_j^{\mathbf{u}}|\mathbf{c_u}) q_{\backslash}(f_i^{\mathbf{u}}, f_j^{\mathbf{u}}|\mathbf{c_u} = c) \tag{9}$$

$$\hat{q}(f_i^{\mathbf{u}}, f_j^{\mathbf{u}}|\mathbf{c_u}) = \hat{Z}^{-1} \mathcal{N}(f_i^{\mathbf{u}}, f_j^{\mathbf{u}}; \hat{\boldsymbol{\mu}}^{c}_{\mathbf{u},[i,j]}, \hat{\boldsymbol{\Sigma}}^{c}_{\mathbf{u},[i,j]}) \quad \text{with} \tag{10}$$

$$\hat{Z} = \Phi(r_{i,j}) \tag{11}$$

$$\hat{\boldsymbol{\mu}}^{c}_{\mathbf{u},[i,j]} = \boldsymbol{\mu}^{c}_{\backslash \mathbf{u},[i,j]} + \boldsymbol{\Sigma}^{c}_{\backslash \mathbf{u},[i,j]} \mathbf{w}_{\mathbf{u},[i,j]} \tag{12}$$

$$\hat{\boldsymbol{\Sigma}}^{c}_{\mathbf{u},[i,j]} = \boldsymbol{\Sigma}^{c}_{\backslash \mathbf{u},[i,j]} - \boldsymbol{\Sigma}^{c}_{\backslash \mathbf{u},[i,j]} (\mathbf{w}_{\mathbf{u},[i,j]} \mathbf{w}^{\top}_{\mathbf{u},[i,j]} \hat{r}_{i,j} \mathbf{w}_{\mathbf{u},[i,j]} \mathbf{1_1}^{\top}) \boldsymbol{\Sigma}^{c}_{\backslash \mathbf{u},[i,j]}, \tag{13}$$

where

$$\mathbf{w}_{\mathbf{u},[i,j]} = \frac{\mathcal{N}(r_{i,j})}{\Phi(r_{i,j})(\alpha^2 + \mathrm{tr}(\boldsymbol{\Sigma}^{c}_{\backslash \mathbf{u},[i,j]} \mathbf{1_2}))} \mathbf{1_1},$$

$$r_{i,j} = \frac{\boldsymbol{\mu}^{c}_{\backslash \mathbf{u},[i,j]} \mathbf{1_1}}{\alpha^2 + \mathrm{tr}(\boldsymbol{\Sigma}^{c}_{\backslash \mathbf{u},[i,j]} \mathbf{1_2})}$$

$$\hat{r}_{i,j} = \frac{r_{i,j}}{\alpha^2 + \mathrm{tr}(\boldsymbol{\Sigma}^{c}_{\backslash \mathbf{u},[i,j]} \mathbf{1_2})}$$

$$\text{and} \quad \mathbf{1_1} = \begin{bmatrix} 1 \\ -1 \end{bmatrix}, \qquad \mathbf{1_2} = \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}. \tag{14}$$

3. Update the local factor approximation: by performing moment matching, we can calculate the corresponding parameters in $q(f_i^{\mathbf{u}}, f_j^{\mathbf{u}}|\mathbf{c_u} = c)$ as:

$$\begin{aligned}\tilde{\boldsymbol{\mu}}^c_{\mathbf{u},[i,j]} &= \tilde{\boldsymbol{\Sigma}}^c{}_{\mathbf{u},[i,j]}(\hat{\boldsymbol{\Sigma}}^c{}^{-1}_{\mathbf{u},[i,j]}\hat{\boldsymbol{\mu}}^c{}_{\mathbf{u},[i,j]} - \boldsymbol{\Sigma}^c{}^{-1}_{\setminus\mathbf{u},[i,j]}\boldsymbol{\mu}^c{}_{\setminus\mathbf{u},[i,j]}) \\ \tilde{\boldsymbol{\Sigma}}^c_{\mathbf{u},[i,j]} &= (\hat{\boldsymbol{\Sigma}}^c{}^{-1}_{\mathbf{u},[i,j]} - \boldsymbol{\Sigma}^c{}^{-1}_{\setminus\mathbf{u},[i,j]})^{-1}. \end{aligned} \quad (15)$$

At each iteration once we have local factor parameters $\tilde{\boldsymbol{\mu}}^c$ and $\tilde{\boldsymbol{\Sigma}}^c$, we can compute the parameters of the full posterior approximation using 3. We iterate through all the factors and update the local approximations sequentially.

2 Inferring Community Membership

In our Gibbs sampler, given $\mathbf{f}$, we now wish to sample $\mathbf{c}$ – the community memberships for all users. Assuming that our blocked Gibbs sampler has already provided us with a sample of $\mathbf{f}$ for some fixed $\mathbf{c}^*$ sampled on the previous iteration, we now wish to sample each new $\mathbf{c_u}$ in turn for the current iteration provided that we can define $p(\mathbf{c_u}|\mathbf{c}_{\setminus\mathbf{u}}, \mathbf{f}, \mathcal{D}, \lambda, \alpha)$.

While we could sample $\mathbf{f}$ from $p(\mathbf{f}|\mathbf{c}^*, \mathcal{D}, \lambda, \alpha)$ to compute $p(\mathbf{c_u}|\mathbf{c}_{\setminus\mathbf{u}}, \mathbf{f}, \mathcal{D}, \lambda, \alpha)$, this seems inefficient given that we can derive the full posterior $p(\mathbf{f}|\mathbf{c}^*, \mathcal{D}, \lambda, \alpha)$ in closed-form given our previously discussed Gaussian Process inference machinery. So instead we propose to compute $\mathbb{E}_{p(\mathbf{f}|\mathbf{c}^*,\mathcal{D},\lambda,\alpha)}[p(\mathbf{c_u}|\mathbf{c}_{\setminus\mathbf{u}}, \mathbf{f}, \mathcal{D}, \lambda, \alpha)]$.[1]

Now we derive an efficiently computable closed-form for sampling $\mathbf{c_u}$ where we abbreviate the previous expectation to the shorter form $\mathbb{E}_{\mathbf{f}|\mathbf{c}^*}[p(\mathbf{c_u}|\mathbf{c}_{\setminus\mathbf{u}}, \mathbf{f}, \mathcal{D}, \lambda, \alpha)]$:

$$\mathbb{E}_{\mathbf{f}|\mathbf{c}^*}[p(\mathbf{c_u}|\mathbf{c}_{\setminus\mathbf{u}}, \mathbf{f}, \mathcal{D}, \lambda, \alpha)] \propto \mathbb{E}_{\mathbf{f}|\mathbf{c}^*}\left[\int p(\underbrace{\mathbf{c_u}, \mathbf{c}_{\setminus\mathbf{u}}}_{\mathbf{c}}, \mathbf{f}, \boldsymbol{\pi}|\mathcal{D}, \lambda, \alpha)d\boldsymbol{\pi}\right] \quad (16)$$

$$= \int \mathbb{E}_{\mathbf{f}|\mathbf{c}^*}\left[\prod_{\mathbf{u}\in U}\underbrace{\left[\prod_{(i,j)\in\mathcal{D}^{\mathbf{u}}} p(\mathbf{x}_i \succ \mathbf{x}_j|f_i^{c_\mathbf{u}}, f_j^{c_\mathbf{u}}, \alpha)\right]}_{p(\mathcal{D}^{\mathbf{u}}|\mathbf{f},\mathbf{c_u},\alpha)} p(\mathbf{c_u}|\boldsymbol{\pi})\right] p(\boldsymbol{\pi}|\lambda)d\boldsymbol{\pi} \quad (17)$$

$$\propto \mathbb{E}_{\mathbf{f}|\mathbf{c}^*}[p(\mathcal{D}^{\mathbf{u}}|\mathbf{f}, \mathbf{c_u}, \alpha)] \underbrace{\int p(\mathbf{c}|\boldsymbol{\pi})p(\boldsymbol{\pi}|\lambda)d\boldsymbol{\pi}}_{p(\mathbf{c}|\lambda)\,\propto\, p(\mathbf{c_u}|\mathbf{c}_{\setminus\mathbf{u}},\lambda)} \quad (18)$$

$$\propto \left[\int \underbrace{p(\mathcal{D}^{\mathbf{u}}|\mathbf{f}, \mathbf{c_u}, \alpha)}_{\text{Likelihood}} \underbrace{p(\mathbf{f}|\mathbf{c}^*, \mathcal{D}, \lambda, \alpha)}_{\text{Gaussian Process}}]d\mathbf{f}\right] \underbrace{p(\mathbf{c_u}|\mathbf{c}_{\setminus\mathbf{u}}, \lambda)}_{\text{Dirichlet Process}} \quad (19)$$

In this derivation, we note the proportionalities can be introduced anytime a rewrite induces a constant normalizer that is independent of $\mathbf{c_u}$ since this can

[1] Of course, one can always sample $\mathbf{f}$ and avoid this expectation if preferred, but we conjecture that using the expectation will induce a lower-variance Gibbs sampling process with faster convergence.

be absorbed into the global normalizer. Since $\mathbf{c_u}$ is discrete, the normalization can be easily computed on demand when required.

In (16), we rewrote the conditional in terms of the full joint since the normalizer required to obtain the LHS is $p(\mathbf{c}_{\setminus \mathbf{u}}, \mathbf{f}, \boldsymbol{\pi} | \mathcal{D}, \lambda, \alpha)$, which is independent of $\mathbf{c_u}$.

In (17), we expanded the full joint into its definition from our graphical model in Figure 1 of the main paper.

In (18), we removed the product for $\mathbf{u}' \in U \setminus \{\mathbf{u}\}$ since the likelihood of each user $\mathbf{u}'$'s preferences only depend on $\mathbf{c}_{\mathbf{u}'}$ and hence these likelihoods are a constant w.r.t. $\mathbf{c_u}$ for $\mathbf{u} \neq \mathbf{u}'$. We also note that the expectation over $\mathbf{f}$ only applies to terms involving $\mathbf{f}$ and likewise the integral over $\boldsymbol{\pi}$ only applies to terms involving $\boldsymbol{\pi}$.

In (19), we expanded the definition of the expectation and replaced $p(\mathbf{c}|\lambda)$ with $p(\mathbf{c_u}|\mathbf{c}_{\setminus \mathbf{u}}, \lambda)$ since the latter only has a normalizer $p(\mathbf{c}_{\setminus \mathbf{u}}|\lambda)$ which is independent of $\mathbf{c_u}$.

Thus in (19) we arrive at a closed-form computation that is straightforward to compute. In the square brackets, we need only use our GP posterior $f|\mathbf{c}^*$ to compute the product of the probabilities of each of user $\mathbf{u}$'s preferences $\mathbf{x}_i \succ \mathbf{x}_j \in \mathcal{D}^{\mathbf{u}}$. This is defined in Sections 3.2–3.4 in the paper and leaves us only to compute $p(\mathbf{c_u} = c|\mathbf{c}_{\setminus \mathbf{u}}, \lambda)$ as is standard in Gibbs sampling for Dirichlet processes:

1. If c is an *active community* ($\exists \mathbf{c_u} \in \mathbf{c}_{\setminus \mathbf{u}}$ s.t. $\mathbf{c_u} = c$), then

$$p(\mathbf{c_u} = c|\mathbf{c}_{\setminus \mathbf{u}}, \lambda) = \frac{\sum_{\mathbf{u}' \neq \mathbf{u}} \mathbb{I}[\mathbf{c}_{\mathbf{u}'} = c]}{N - 1 + \lambda} \tag{20}$$

   where N is the number of non-empty communities.

2. Else c is a *new community* so

$$p(\mathbf{c_u} = c|\mathbf{c}_{\setminus \mathbf{u}}, \lambda) = \frac{\lambda}{N - 1 + \lambda} \tag{21}$$

Hence all quantities required to sample $\mathbf{c_u}$ have now been defined permitting sampling of each $\mathbf{c_u}$ in turn to complete the community process sampling portion of the Gibbs sampling inference for our model. And the result is intuitive: a user $\mathbf{u}$ is more likely to join a community which provides a higher likelihood on its preference data $D^{\mathbf{u}}$. Additionally, this sampling process displays the well-known "rich-get-richer" effect of Dirichlet Processes since communities with more members have a higher probability of being selected.

In practice, classes $c \in C$ that have no assignment in $\mathbf{c}$ after all $\mathbf{c}$ have been sampled can be "garbage collected" meaning that in practice the size of actively maintained communities (for which Gaussian Process utilities must be learned) can never exceed the size of $|\mathbf{c}|$, i.e., the number of users $|U|$. While this number may still be large, tuning the concentration $\lambda \to 0$ encourages fewer (sparse) communities with a larger number of members.