Competing Models: Inferring Exploration Patterns and Information Relevance via Bayesian Model Selection

Supplemental Material

Shayan Monadjemi, Roman Garnett, and Alvitta Ottley

A Bayesian Inference for Multivariate Gaussian

The multivariate Gaussian distribution is parameterized by a mean vector $\vec{\mu}$ and a covariance matrix Σ, representing the center and spread of the distribution respectively. Since the value of these parameters $(\vec{\mu},\Sigma)$ for a given set of observations $\mathscr{C}$ is unknown, we infer them thorough the Bayesian inference process which results in a closed form posterior predictive distribution function, $f(\vec{x} \mid \mathscr{C})$. In the equations below, we use Λ, the precision matrix, which is simply the the inverse of the covariance matrix, i.e. $\Lambda = \Sigma^{-1}$. Since both mean and precision matrix describing user interactions are unknown, we proceed with the Bayesian method and choose a **prior** across all possible values of $\vec{\mu}$ and Λ [2]:

$$\begin{aligned} p(\vec{\mu},\Lambda) &= \text{Normal-Wishart}\,(\vec{\mu},\Lambda \mid \vec{\mu}_0,\kappa,\nu,T) \\ &= \mathscr{N}\left(\vec{\mu} \mid \vec{\mu}_0,(\kappa\Lambda)^{-1}\right)\text{Wishart}_{\nu}\,(\Lambda \mid T) \end{aligned} \tag{1}$$

For any given value of parameters $\vec{\mu}$ and Λ, the **likelihood** of observing n interactions, $\mathscr{C} = \{\vec{c}_1,\vec{c}_2,...,\vec{c}_n\}$, is:

$$\begin{aligned} p(\mathscr{C} \mid \vec{\mu},\Lambda) &= p(\vec{c}_1 \mid \vec{\mu},\Lambda)p(\vec{c}_2 \mid \vec{\mu},\Lambda)...p(\vec{c}_n \mid \vec{\mu},\Lambda) \\ &= \mathscr{N}(\vec{c}_1;\vec{\mu},\Lambda)\mathscr{N}(\vec{c}_2;\vec{\mu},\Lambda)...\mathscr{N}(\vec{c}_n;\vec{\mu},\Lambda) \\ &= (2\pi)^{-nd/2}|\Lambda|^{n/2}\exp(-\frac{1}{2}\sum_{i=0}^{n}(\vec{c}_i-\vec{\mu})^T\Lambda(\vec{c}_i-\vec{\mu})) \end{aligned} \tag{2}$$

where clicks are assumed to be independent. After observing n clicks, $\mathscr{C}$, we update our belief over $\vec{\mu}$ and Λ to get the **posterior** distribution:

$$\begin{aligned} p(\vec{\mu},\Lambda|\mathscr{C}) &= p(\mathscr{C} \mid \vec{\mu},\Lambda)p(\vec{\mu},\Lambda) \\ &= \mathscr{N}(\vec{\mu}|\vec{\mu_n},(\kappa_n\Lambda)^{-1})\text{Wishart}_{\nu_n}(\Lambda,T_n) \end{aligned} \tag{3}$$

where:

$$\begin{aligned} \vec{\mu_n} &= \frac{\kappa\vec{\mu_0}+n\vec{\bar{c}}}{\kappa+n} \\ T_n &= T+S+\frac{\kappa n}{\kappa+n}(\vec{\mu_0}-\vec{\bar{c}})(\vec{\mu_0}-\vec{\bar{c}})^T \\ \nu_n &= \nu+n \\ \kappa_n &= \kappa+n \\ S &= \sum_{i=1}^{n}(\vec{c}_i-\vec{\bar{c}})(\vec{c}_i-\vec{\bar{c}})^T \end{aligned}$$

- *Shayan Monadjemi is with Washington University. monadjemi@wustl.edu.*
- *Roman Garnett is with Washington University. garnett@wustl.edu.*
- *Alvitta Ottley is with Washington University. alvitta@wustl.edu.*

Notice that our posterior distribution has the same form as our prior distribution. In Bayesian probability theory, we refer to these distributions as conjugate distributions. Now that we have an updated belief over the parameters of the Gaussian distribution which take into consideration user interactions, we want to integrate our belief about every parameter to come up with one **posterior predictive** distribution which is a multivariate Student's t-distribution conditioned on observed data [2]:

$$\begin{aligned} f_c(\vec{x}|\mathscr{C}) &= \int_{\vec{\mu}}\int_{\Sigma} p(\vec{x} \mid \vec{\mu},\Sigma)p(\vec{\mu},\Sigma \mid \mathscr{C})d\Sigma d\vec{\mu} \\ &= \text{St}_{\nu_n-d+1}\big(\mu_n,\frac{T_n(k_n+1)}{\kappa_n(\nu_n-d+1)}\big) \end{aligned} \tag{4}$$

The Normal-Wishart prior relies on four hyper-parameters: μ_0, T, κ, and ν. We set the measure of location (μ_0) and spread (T) based on the range of any given data set. Our best intuitive guess for an appropriate μ_0 is to be the midpoint of the domain. We chose the measure of spread T, to be the data co-variance divided by ν. κ is a measure of how sure we are about our prior, or what sample size of fake observations are explained by our prior. Higher values of κ make the posterior distribution slower in adapting to new observations, while lower values of κ make the prior less significant in our posterior. We set the value of κ to be 1, meaning we put less faith into our prior as we observe more user interactions. We let our posterior distribution be mainly influenced by the observations. Finally, ν is the degrees of freedom, representing the number of variables in our multivariate Gaussian distribution.

B Bayesian Inference for Categorical Distribution

The categorical model is used to explain the probability of discrete events occurring. For an attribute domain with K possible categories, the categorical model has a K-dimensional vector $\vec{\mu}$ which describes the probability of observing each of the K choices. Since the value of $\vec{\mu}$ is unknown, we choose a Dirichlet **prior** over all values of $\vec{\mu}$:

$$\begin{aligned} p(\vec{\mu} \mid \vec{\alpha}) &= \text{Dirichlet}(\vec{\mu} \mid \vec{\alpha}) \\ &= \frac{\Gamma(\sum_{k=1}^{K}\alpha_k)}{\Gamma(\alpha_1)\Gamma(\alpha_2)...\Gamma(\alpha_K)}\prod_{k=1}^{K}\mu_k^{\alpha_k-1} \end{aligned} \tag{5}$$

After observing N data points, $\mathscr{C}$, we know the observation counts for each of the K categories. We denote the observation count for the category i with m_i. The **likelihood** of such observation happening given a particular $\vec{\mu}$ is [1]:

$$p(\mathscr{C} \mid \vec{\mu}) = \binom{N}{m_1m_2...m_K}\prod_{k=1}^{K}\mu_k^{m_k} \tag{6}$$

Using the Bayes' rule, we update our belief over values of μ after we observe the user clicks $\mathscr{C}$ in order to get the **posterior** distribution [1]:

$$
\begin{aligned}
p(\vec{\mu} \mid \mathscr{C}, \vec{\alpha}) &= p(\mathscr{C} \mid \vec{\mu}) p(\vec{\mu} \mid \vec{\alpha}) \\
&= \text{Dirichlet}(\vec{\mu} \mid \vec{\alpha} + \vec{m}) \\
&= \frac{\Gamma\left(N + \sum_{k=1}^{K} \alpha_k\right)}{\Gamma(\alpha_1 + m_1)\Gamma(\alpha_2 + m_2)...\Gamma(\alpha_K + m_K)} \prod_{k=1}^{K} \mu_k^{\alpha_k + m_k - 1}
\end{aligned}
\tag{7}
$$

Notice that similar to the continuous case, our prior and posterior in the discrete scenario are also conjugate distributions. The Dirichlet prior for discrete attributes has only one hyper-parameter: $\vec{\alpha}$. For a categorical distribution with K categories, we set $\vec{\alpha}$ to be a vector of K equal small numbers (0.0001), which constitutes a uniform distribution across all possible categories. This choice of hyper-parameter is uninformative and puts almost no faith in the prior, as our number of fake observations in the prior distribution is very small. Finally, we integrate our updated belief over all possible values of $\vec{\mu}$ to get a closed-form **posterior predictive** for a given category k [3]:

$$
\begin{aligned}
f_d(k \mid \mathscr{C}, \vec{\alpha}) &= \int_{\vec{\mu}} p(k \mid \vec{\mu}) p(\vec{\mu} \mid \mathscr{C}, \vec{\alpha}) d\vec{\mu} \\
&= \frac{\alpha_k + m_k}{\sum_{i=1}^{K} (\alpha_i + m_i)}
\end{aligned}
\tag{8}
$$

where α_i is the pseudocount for category i (from hyperparameter $\vec{\alpha}$) and m_i is the observation count for category i.

C Example for Setting Hyper-parameters

Consider the example restaurant dataset from the paper, where each fictitious restaurant is a 3D vector of form $(latitude, longitude, type)$:

$$
\begin{aligned}
\mathscr{D} = \{&(0.35, 0.85, \text{Italian}), (0.8, 0.35, \text{Mexican}), (0.85, 0.1, \text{Persian}), \\
&(0.7, 0.3, \text{Italian}), (0.15, 0.75, \text{Mexican}), (0.1, 0.05, \text{Persian}), \\
&(0.9, 0.85, \text{Mexican})\}
\end{aligned}
$$

In this section, we give an example of setting hyper-parameters for this fictitious dataset.

Multivariate Gaussian Hyper-parameters The four hyper-parameters discussed in Appendix A are: $\vec{\mu}_0, T, \kappa, \nu$. We set the value of $\vec{\mu}_0$ to be the center of the data $\mathscr{D}$, with an added 0 to represent the mean of timesteps. As discussed in the paper, the added $time$ attribute is used to model the order of observations and predict future interaction.

$$
\vec{\mu}_0 = \begin{bmatrix} \mu_{0,latitude} \\ \mu_{0,longitude} \\ \mu_{0,time} \end{bmatrix} = \begin{bmatrix} \frac{0.9-0.1}{2} \\ \frac{0.85-0.05}{2} \\ 0 \end{bmatrix} = \begin{bmatrix} 0.4 \\ 0.4 \\ 0 \end{bmatrix}
$$

Next, we set the value of T to be the covariance of $latitude$ and $longitude$ dimensions of $\mathscr{D}$, with an added dimension for timesteps ($time$):

$$
T = \begin{bmatrix} 0.116 & -0.005 & 0 \\ -0.005 & 0.121 & 0 \\ 0 & 0 & 1 \end{bmatrix}
$$

Finally, we set $\kappa = 1$ (low faith in prior) and $\nu = 3$ (number of continuous dimensions including $time$).

Categorical Distribution Hyper-parameters In this example, we only have one categorical attribute ($type$) with three possible outcomes {Italian, Mexican, Persian}. We set the hyper-parameter $\vec{\alpha}$ to represent low pseudocount (hence low faith in prior):

$$
\vec{\alpha} = \begin{bmatrix} \alpha_{Italian} \\ \alpha_{Mexican} \\ \alpha_{Persian} \end{bmatrix} = \begin{bmatrix} 0.0001 \\ 0.0001 \\ 0.0001 \end{bmatrix}
$$

References

[1] C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

[2] K. P. Murphy. Conjugate Bayesian analysis of the Gaussian distribution. *Technical Report*, 2007.

[3] S. Tu. The Dirichlet-Multinomial and Dirichlet-Categorical models for Bayesian inference. *Computer Science Division, UC Berkeley*, 2014.