Review of Joint Clustering with Correlated Variables

Saran Ahluwalia, Minji Kim and Hannah Park

Abstract

A succinct description of the authors' methods, a simulation study with replication of two independent experiments and implications of the study is provided. Fundamental derivations for both prior and posterior distributions used in the authors' Bayesian approach for estimation is included. Finally, a review of the of the Dirichlet process is included in order to provide context for this distribution's utility.

Keywords: MCMC, Bayesian Estimation, Gibbs Sampling, Inverse Wishart, Conjugate Priors, Dirichlet Process

1 Motivation for Introducing Joint Clustering

Cluster analysis is a technique used to classify a set of objects into relative groups called clusters. The goal is to form clusters in such a way that objects within the same cluster are similar to one another but are dissimilar to those in other clusters. To study patterns of wheal sizes in reaction to different allergens at different ages, authors of this paper were interested in forming clusters such that subjects within the same cluster were dissimilar to those in other clusters in the following ways: (1) their reaction to certain allergens was different from their reaction to other allergens and (2) their reaction to certain allergens was different from the reaction of other subjects in other groups to the same allergens. Classical methods for cluster analysis, however, generally only satisfy one of the two conditions; they focus on clustering subjects or variables (e.g. allergens) but not both. Focusing on one criterion can cause lack of homogeneity in the other, and though methods, such as bi-clustering, exist that can address both the interdependence between phenotypes and the homogeneity in subjects, they fail to explain how certain external variables, such as time, affect the formation of patterns. The ability to explain such external variable effects is crucial in this study, which explores changes in allergic sensitization with age.

In order to form clusters that satisfy the two stated conditions and explain the time effect, a novel method, called joint clustering, is proposed by the authors. This probabilistic clustering method considers both the correlation between variables and the interrelationship between variables and subjects by introducing an indicator variable for variable cluster assignment and applying a Dirichlet process mixture model to cluster the subjects. It also takes into account the relationship between the variables and the covariates by evaluating their associations using a semi-parametric model via penalized splines. Details of the joint clustering method are further explained below.

2 Setting and Methods Overview

2.1 Clustering of Subjects and Clustering of Variable Clusters

First, we briefly review the setting that the authors use for the procedure to clustering the variables. M clusters and K variables are considered for the clustering procedure. The authors use a a vector function to model the relationship between variables and covariates of interest. More specifically they use penalized splines as described in Eilers and Marx (1).

P-splines possess properties that are particularly efficacious for modeling relationships. This is primarily due to the inherited properties from B-spline basis functions. For brevity, we outline two:

1. **Boundary effects** P-splines show no boundary effects. In other words, the spreading of a fitted curve or density outside of the (physical) domain of the data, generally accompanied by bending toward zero (2).

2. **P-splines fit polynomial data exactly.** Given data (x_i, y_i), if the y_i are a polynomial in x of degree k, then B-splines of degree k or higher will fit the data exactly. In practical applications, this is problematic as this leads to over-fitting.

The subjects within each variable cluster are further grouped such that each group reflects a different relationship between variables and covariates of interest.

The authors propose an indicator matrix in order to determine the subject clusters within each variable cluster - primarily through grid search. To do so, the authors propose the Dirichlet process parameterized by a base distribution that describes the base association between the response variable and the cluster variable m and subject i.

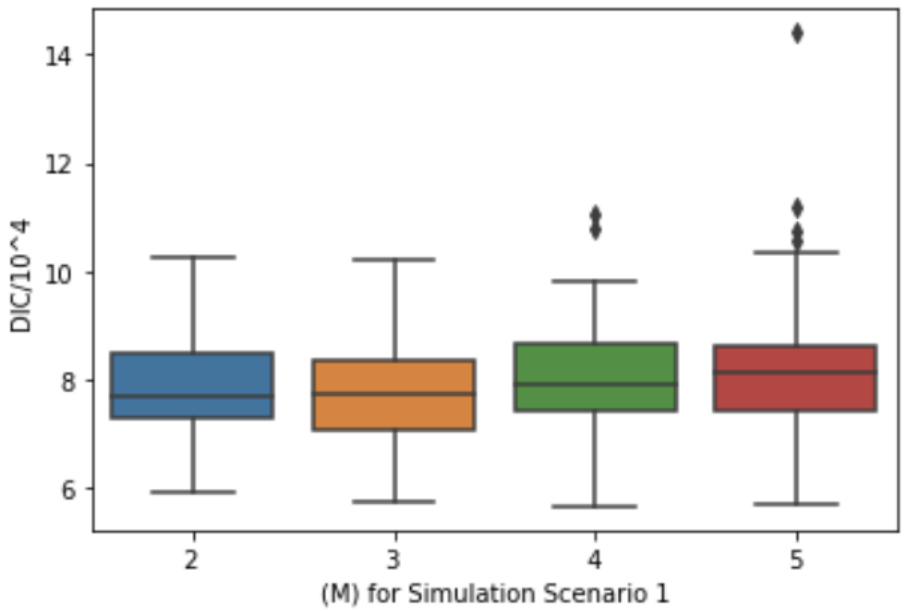
2.2 Constructing Mixture Models from the Dirichlet Process

In the Appendix Section an overview of the Dirichlet process (DP) and Gaussian Mixture is provided that describes the joint clustering approach by combining sections 2.1 and 2.2. For the purpose of this report, the most salient methods are elaborated on in order to provide greater detail on the DP as an invaluable tool for modeling hierarchical relationships.

Furthermore, a very brief review of the computational procedure is elaborated on (2.4) in order to describe the sampling procedure of the conditional posterior distributions. Please note that the notation described in each of the four proceeding sections in the Appendix will not necessarily mirror the representative paper's notation. Rather, the purpose of the following sub-sections is to provide an overview for the the reader who is not familiar with the Dirichlet process (3).

3 Simulation Study

To demonstrate the proposed method, the authors perform a simulation study and evaluate the effectiveness of the method using sensitivity and specificity. One hundred Monte Carlo (MC) replicates are generated, each with 10 variables and 400 subjects. The 10 variables are first grouped into M clusters, and within each variable cluster, the subjects are further clustered. The number of variable clusters is determined by optimizing the deviance information criterion (DIC). After running the clustering process for a set of different values of M, the one with the smallest DIC is chosen. In the paper, 73 of the 100 MC replicates were optimized at M = 3. Similar results were found when the process was replicated in R. This is displayed in the figure below.



The final number of joint clusters is decided by identifying an iteration with the least-square-distance. After the MCMC burn-in, MCMC simulation is performed for additional iterations, during which an n ×n ×K matrix composed of indicators of clustering for each iteration is formed. The indicator denotes whether the subject i and j for the kth variable are in one cluster. Then, the final number of joint clusters is selected by comparing the Euclidean distance calculated for each iteration between the matrix formed and the averaged clustering matrix.

4 Major Findings

The quality of clustering formed by the simulation study is assessed using sensitivity and specificity based on pairwise agreements with respect to the true clustering. The authors define sensitivity as SE=TP/(TP+FN) and specificity as SP=TN/(TN+FP), where "TP"

denotes true positives, or correct cluster identification, "FN" denotes false negative, "TN" true negatives, and "FP" false positives. Sensitivity and specificity calculations of the simulation study were replicated in R, and the observed values are listed in Table 1. Similar to the results found in the paper, high sensitivities and specificities are observed for all 6 clusters. The proposed method of joint clustering correctly identifies the true clusters of the variables and the subjects.

Table 1: Means and standard deviations (SD) of each cluster under scenario 1

	Sensitivity		Specificity	
Cluster	Mean	SD	Mean	SD
1	1.0000000	0.00000000	0.9972727	0.02727273
2	1.0000000	0.00000000	0.9986154	0.01384615
3	0.9900000	0.10000000	1.0000000	0.00000000
4	0.9900000	0.10000000	1.0000000	0.00000000
5	0.9900000	0.10000000	1.0000000	0.00000000
6	0.9999545	0.00045455	0.9989888	0.01011236

The results presented in Table 1 demonstrate the robustness of the joint clustering method with respect to different cluster patterns. These results are based on a specific choice of the values of the variance components in the covariance matrices for the three variable clusters. The choice was made to control the impact of large variation in the data on the quality of clustering. In a different scenario, the authors tested such impact by increasing the values of the variance components for the variable clusters from 1, 1, 0.6 to 5, 6, 6, respectively. Keeping the other settings the same, the second scenario was replicated in R, and the results are presented in Table 2. Compared to the first scenario, the means of

the sensitivities and the specificities from the second scenario are not heavily impacted by the larger variations in the data. However, the standard deviations are generally larger in the second scenario than the first. This implies an increased uncertainty that is potentially due to the larger variations in the data.

Table 2: Means and standard deviations (SD) of each cluster under scenario 2

	Sensitivity		Specificity	
Cluster	Mean	SD	Mean	SD
1	0.9700000	0.17144661	0.9972727	0.02727273
2	0.9940000	0.03610062	0.9765385	0.08863434
3	0.9900000	0.10000000	1.0000000	0.00000000
4	0.9900000	0.10000000	1.0000000	0.00000000
5	0.9400000	0.23868326	1.0000000	0.00000000
6	0.9911136	0.04215788	0.9883483	0.03183577

The authors also consider a third scenario in which they assess the sensitivity on the independence assumption between the variable clusters. Due to the lack of information on how the correlation between the variables are setup, this scenario was not replicated. The results from the paper demonstrates the robustness of the joint clustering method by showing high average sensitivities and specificities compared to those under the first scenario. The authors conclude that the increased sensitivity and specificity for some clusters are likely due to the increased homogeneity in the variables and state that the increase is beneficial to the quality of clustering.

The goal of this study was to introduce a novel method for cluster analysis that can be used to examine the patterns of wheal sizes in reaction to different allergens at different

ages and how those patterns are associated with asthma risk. Using the data from a longitudinal study cohort aiming to investigate the history of asthma (4), the authors applied the joint clustering method in real life and were able to identify subsets of allergens as well as patterns of wheal sizes over time that may play a key role in asthma occurrence. According to the authors, the proposed method is ready to be applied to other types of statistical models, such as logistic regressions or log-linear models, with multiple response variables which have different associations with covariate(s) of interest.

5 Implications and Limitations

This paper opens a new a new frontier in methods that can jointly cluster variables and subjects with effect from exogenous covariates. Moreover, the bi-clustering benchmarks' conclusions suggest that this clustering approach may require additional methods for assessing full separation of the clusters.

Finally, because the Dirichlet process plays an important role in the Bayesian nonparametric framework as a prior distribution with wide support and tractable inference. As such, many generalizations and extensions have been explored in the literature. A very popular generalization is the Pitman-Yor process (5; 6). The Pitman-Yor process has an additional parameter $d \in [0, 1)$ and reduces to the Dirichlet process when $d = 0$. If d is close to 1, the clusters generated by the Pitman-Yor process exhibit a power-law behavior that produces few large clusters or many smaller clusters.

Other generalizations of the DP include Pólya trees, stick-breaking priors and Poisson-Kingman models. These can be derived by extending one of the representations of the DP. A different class of extensions uses the Dirichlet process as building blocks to develop more complex models. Two such models are the dependent Dirichlet processes (7) and hierarchi-

cal Dirichlet processes (8). It would be interesting to also see how using the aforementioned procedures has any effect on separability of clusters across all three scenarios.

Beyond extending the DP as it applies to the posterior distribution, recent research has also been focused on exploring more efficient inference methods in Dirichlet process models that do not use the simple Gibbs sampler. Moreover, the complexity of these models and the expense in implementing them poses a significant practical burden within industry and academic research. In conclusion, Bayesian non-parametric methods provide a powerful set of tools for clustering and demonstrate their utility across difference applied settings.

Appendix A

Details of the Dirichlet-Multinomial

The Multinomial (or Categorical distribution) and the Dirichlet Process (DP) are used to represent the probability of finding the kth variable in each of the clusters, M and ξ the hyperprior is chosen, respectively. The DP is detailed in greater detail later in this article. However, for the following generic clustering of variables consider the following set-up:

$$\theta = (\theta_1, \ldots, \theta_m), X_i \in \{1, \ldots, m\}, \sum_i \theta_i = 1.$$

Assume that:

$$X \mid \theta \overset{ind}{\sim} \text{Multinomial}(\theta)$$

or

$$X \mid \theta \overset{ind}{\sim} \text{Categorical}(\theta) P(X_i = j \mid \theta) = \theta_j$$

and

$$\theta \sim \text{Dirichlet}(\alpha)(\theta \mid \alpha) \propto \prod_{j=1}^{m} \theta_j^{\alpha_j - 1}$$

$$\sum_j \theta_j = 1, \theta_i \geq 0 \forall i \in 1...m$$

Likelihood

Define the data as:

$D = (x_1, \ldots, x_n), x_i \in \{1, \ldots m\}$

Using this set-up we have:

$$\begin{aligned} p(D \mid \theta) &= \prod_{i=1}^{n} P(X_i = x_i \mid \theta) \\ &= \prod_{i=1}^{n} \theta_{x_i} \\ &= \prod_{i=1}^{n} \prod_{j=1}^{m} \theta_j^{I(x_i=j)} \\ &= \prod_{j=1}^{m} \theta_j^{\sum_i I(x_i=j)} \\ &= \prod_{j=1}^{m} \theta_j^{c_j} \end{aligned}$$

where $c = (c_1, \ldots c_m) c_j = i : x_i = j\}$.

Likelihood, Prior, and Posterior

Using the likelihood we can then derive the posterior:

$p(D \mid \theta) = \prod_{j=1}^{m} \theta_j^{c_j}$

$P(\theta) \propto \prod_{j=1}^{m} \theta_j^{\alpha_j - 1} I(\sum_j \theta_j = 1, \theta_i \geq 0 \forall i)$

Then we have:

$$P(\theta \mid D) \propto \prod_{j=1}^{m} \theta_j^{c_j} \times \prod_{j=1}^{m} \theta_j^{\alpha_j - 1} I(\sum_j \theta_j = 1, \theta_i \geq 0 \forall i)$$
$$\propto \prod_{j=1}^{m} \theta_j^{c_j + \alpha_j - 1} I(\sum_j \theta_j = 1, \theta_i \geq 0 \forall i)$$

This implies that:

$\theta \mid D \sim \text{Dirichlet}(c + \alpha)$.

1. The Dirichlet is conjugate to the Categorical or Multinomial.
2. Hence, a more simplified posterior can be written as follows:
   $\prod_i \text{Multinomial}(x_i \mid \theta) \times \text{Dir}(\theta \mid \alpha) \propto \text{Dir}(\theta \mid c + \alpha)$.

In addition the covariance Σ_m is produced from the Inverse Wishart distribution with scale parameters S and ν as follows:

Suppose $\Sigma \sim \text{InvWishart}(\nu_o, S_o^{-1})$ where ν_o is a scalar and S_o^{-1} is a matrix.

$$p(\Sigma) \propto \det(\Sigma)^{-(\nu_o + p + 1)/2} \times \exp\{-\text{tr}(S_o \Sigma^{-1})/2\}$$

$$\theta \mid X\Sigma \sim \text{Multivariate Normal}(\theta, \Sigma).$$

Also note that $p(\Sigma \mid X, \theta) = p(\Sigma) \times p(X \mid \theta, \Sigma)$. Taking the previous definition we can then construct the following:

$$
\begin{aligned}
&p(X \mid \theta, \Sigma) \\
&\qquad \propto \det(\Sigma)^{-n/2} \exp\{-\sum_i (\mathbf{X}_i - \theta)^T \Sigma^{-1} (\mathbf{X}_i - \theta)/2\} \\
&\qquad \propto \det(\Sigma)^{-n/2} \exp\{-\mathrm{tr}(S_\theta \Sigma^{-1}/2)\}
\end{aligned}
$$

Now we can calculate $p(\Sigma \mid X, \theta)$:

$$
\begin{aligned}
&p(\Sigma \mid X, \theta) && (1)\\
&\quad = p(\Sigma) \times p(X \mid \theta, \Sigma) && (2)\\
&\quad \propto \det(\Sigma)^{-(\nu_o + p + 1)/2} \times \exp\{-\mathrm{tr}(S_o \Sigma^{-1})/2\} && (3)\\
&\qquad \times \det(\Sigma)^{-n/2} \exp\{-\mathrm{tr}(S_\theta \Sigma^{-1})/2\} && (4)\\
&\quad \propto \det(\Sigma)^{-(\nu_o + n + p + 1)/2} \exp\{-\mathrm{tr}((S_o + S_\theta)\Sigma^{-1})/2\} && (5)
\end{aligned}
$$

This implies that $\Sigma \mid X, \theta \sim \text{InvWishart}(\nu_o + n, [S_o + S_\theta]^{-1} = S_n)$

Suppose that we wish now to take $\theta \mid X, \Sigma \sim$ Multivariate Normal(μ_n, Σ_n) Now let $\Sigma \mid X, \theta \sim \text{InvWishart}(\nu_n, S_n^{-1})$

There is no closed form expression for this posterior. In lieu of producing an analytical solution, we would need to construct an approximation of the posterior via Gibbs sampling (9).

For example, suppose the Gibbs sampler is at iteration m.

1. Sample $\theta^{(m+1)}$ from it's full conditional:

    a) Compute μ_n and Σ_n from X and $\Sigma^{(m+1)}$

    b) Sample $\theta^{(m+1)} \sim$ Multivariate Normal(μ_n, Σ_n)

2. Sample $\Sigma^{(m+1)}$ from its full conditional:

    a) Compute S_n from X and $\Sigma^{(m+1)}$

    b) Sample $\Sigma^{(m+1)} \sim$ InvWishart(ν_n, S_n^{-1})

This aforementioned Gibbs sampling procedure is further detailed in the authors' appendix section.

Appendix B

Additional Details on the DP and GP

For the reader who is not familiar with either the domain of mathematical statistics, nor hierarchical models a natural question may be ask would be: Why even use the DP model? What makes this model particularly useful and so ubiquitously cited in literature? What is its practical significance within data mining techniques? In the following section, an attempt to answer these aforementioned questions is detailed in order to provide the basic setting illustrated in sections 2.1 - 2.3. The Dirichlet model has the form:

$$\begin{aligned} \beta_{i,m}|G &\sim G, \\ G &\sim \mathrm{DP}(\lambda, G_0), \end{aligned}$$

where G_0 is the information provided by cluster variables and subject. Essentially, this measure encapsulates any prior knowledge that might be known about G. Moreover, it can be shown that $\mathbb{E}[G \mid G_0, \lambda] = G_0$. The concentration parameter λ specifies the prior variance and controls the relative contribution that the prior and data make to the posterior.

Part 1: The DP is a conjugate prior that is constructed as follows: if $y_1, \ldots, y_n \sim G$ and $G \sim \mathrm{DP}(G_0, \lambda_0)$, then:

$G \mid y_1, \ldots, y_n \sim \mathrm{DP}\left(\lambda + n, \frac{\lambda G_0 + \sum_{i=1}^{n} \phi y_i}{\lambda + n}\right)$, where ϕy_i denotes a point-mass at y_i. Hence, the posterior distribution of G is a weighted sum of the base measure G_0 and the empirical distribution of the data, with the weighting controlled by λ.

The DP is a prior distribution over the space of probability distributions. As such, samples from a DP are probability distributions. The stick-breaking representation first introduced by (10) shows what such samples can be described as such.

Part 2: Suppose that $G \sim \mathrm{DP}(\lambda, H)$ is a random probability distribution sampled from a DP prior. Then with probability 1, G can be written as:

$$G = \sum_{k=1}^{\infty} w_k \phi_{\psi_k}, \quad \psi_k \sim G_0$$

where

$$w_k = z_k \prod_{i=1}^{k-1} (1 - z_i), \quad z_i \sim \mathrm{Beta}(1, \lambda).$$

Practically, a random probability distribution can be sampled from a DP by first drawing a collection of samples z_i from a Beta distribution, transforming these to produce the weights $\{w_i\}$, and then drawing the associated point masses from G_0. Note that in order for G to represent DP, an infinite number of point masses must be drawn. Practically, the the above summation can be truncated with a finite number (N) draws while still providing a very good approximation to the Dirichlet process.

By combining 1 and 2, we can sample a DP from its posterior distribution $G \mid y_1, \ldots, y_n$ as follows:

Part 3: If $y_1, \ldots, y_n \sim G$ and $G \sim \mathrm{DP}(\lambda, G_0)$ then we can draw a (truncated) sample probability distribution from the posterior $G \mid y_1, \ldots, y_n$ as follows:

$$G = \sum_{k=1}^{N} w_k \phi_{\phi_k}, \quad \phi_k \sim \frac{\lambda G_0 + \sum_{i=1}^{n} \phi_{y_i}}{\lambda + n},$$

where

$$w_k = z_k \prod_{i=1}^{k-1} (1 - z_i), \quad z_i \sim \mathrm{Beta}(1, \lambda + n).$$

and N is a truncation parameter.

Appendix C

Constructing Mixture Models from the Dirichlet Process

The stick breaking representation cited earlier showed that probability distributions sampled from a DP are discrete with probability 1. Thus, the DP is not an appropriate prior for G when G is continuous. Therefore, we specify this heuristic as such:

$$\begin{aligned} y_i &\sim k(y_i \mid \theta_i), \\ \theta_i &\sim G, \\ G &\sim \mathrm{DP}(G_0, \lambda). \end{aligned} \tag{6}$$

In other words, G has a DP prior as before, but rather than the data y_i being drawn from G, it is instead the mixture parameters θ which are draws from G. These θ values then act as the parameters of a parametric kernel function $k(\cdot)$, which is usually continuous. The most commonly used example is the Gaussian mixture model where $\theta_i = (\mu_i, \sigma_i^2)$ so that $k(y_i \mid \theta_i) = N(y_i \mid \mu_i, \sigma_i^2)$.

Since G is discrete, two independent draws θ_i and θ_j from G can have identical values with a non-zero probability. As such, the mixture presentation step described earlier leads to clusters, corresponding to the mixture components. The above model can hence be written equivalently as the following mixture model, which is infinite dimensional can hence be viewed as a generalization of the finite mixture models:

$$\begin{aligned} y_i &\sim G, \\ G &= \int k(y_i \mid \theta) G(\theta) \mathrm{d}\theta, \\ G &\sim \mathrm{DP}(\theta, \lambda). \end{aligned}$$

Practically, it appears that the authors focused on several different parameters of the posterior distribution. Generally, in some cases, the primary object of interest will be all n of the θ_i parameters from Equation 6 which are associated with the n observations. This is particularly the case in clustering applications, where the goal is to assign similar observations to the same cluster (i.e. to identical values of θ). However in other situations it will be the distribution G which is of primary interest.

The authors instead use a MCMC approach via the Gibbs algorithm based on the representation in Equation 6 above, and draws samples of $\theta_1, \ldots, \theta_n$ from their posterior with the distribution G integrated out.

Part 4: Let θ_{-i} denote the set of θ values with θ_i excluded, i.e. $\theta_{-i} = (\theta_1, \ldots, \theta_{i-1}, \theta_{i+1}, \ldots, \theta_n)$. Then the posterior distribution for θ_i conditional on the other model parameters is:

$$p(\theta_i \mid \theta_{-i}, y_{1:n}, \lambda, G_0) = \sum_{j \neq i} q_{i,j} \phi(\theta_j) + r_i H_i,$$

$$q_{i,j} = bk(y_i, \theta_j), \quad r_i = b\lambda \int k(y_i, \theta) \mathrm{d}G_0(\theta)$$

where b is set such that $\sum_{j \neq i} q_{i,j} + r_i = 1$ and H_i is the posterior distribution of θ based off of the prior base measure G_0.

Based on this result, Gibbs sampling is used to repeatedly draw each value of θ_i in turn from its posterior distribution, with all other variables held constant. An important distinction needs to be made between the conjugate case where the G_0 base measure is the conjugate prior for θ with respect to the kernel $k(\cdot)$, and the non-conjugate case there is not a closed form for the posterior distribution. In the conjugate case, the integral in part 4 can be computed analytically and the resulting distribution is simply the predictive distribution. In this case, the θ_i values can be sampled directly from their true posterior distribution.

In the non-conjugate case the integral in part 4 (above) cannot be evaluated. As such, numerical techniques must be used instead. In practice, this is often slower.

$$p(\theta_i \mid y_j) = \prod_{j=i} k(y_j \mid \theta) G_0, \tag{7}$$

for a conjugate base measure, this posterior distribution is tractable and thus can be sampled directly. For a non-conjugate G_0, a posterior sample is achieved using the Metropolis-Hastings algorithm (11).

For density estimation and non-hierarchical predictive tasks, possessing a posterior sample of $\theta_{1:n}$ will be sufficient for inference. However when the DP is used as part of a hierarchical model such as in the setting of a regression problem and/or point process setting it is necessary to have samples from the posterior distribution of G. These can be obtained using the following property:

Given the model from Equation (6) let $\theta_1, \ldots, \theta_n$ be a sample from the posterior $p(\theta_{1:n} \mid y_{1:n}, \lambda, G_0)$ drawn. Then, $p(G \mid \theta_{1:n}, y_{1:n}, \lambda, G_0) = p(G \mid \theta_{1:n}, \lambda, G_0)$ is conditionally independent of $y_{1:n}$. As such, $\theta_{1:n}$ can be considered as an i.i.d sample from G, and so G can be sampled from its posterior distribution using the construction in part 3. As previously discussed:

$$G = \sum_{i=1}^{N} w_i \phi \theta_i, \quad w_i \sim \text{Beta}(\lambda + n, 1), \quad \phi_i \sim G_0 + \sum_{i=1}^{n} \phi y_i$$

where N is a truncation parameter.

Background on Gaussian Mixture Models

The Gaussian distribution is the most commonly used mixture model. In this case, $\theta = (\mu, \sigma^2)$ for the mean and variance. The kernel function can be defined as:

$$k(y_i \mid \theta) = N(y_i \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(y-\mu)^2}{2\sigma^2}\right).$$

The conjugate prior for θ is the Normal-Gamma distribution, with parameters $\gamma = (\mu_0, k_0, \lambda_0, \beta_0)$

$$G_0(\theta \mid \gamma) = N\left(\mu \mid \mu_0, \frac{\sigma^2}{k_0}\right) \text{Inv-Gamma}\left(\sigma^2 \mid \lambda_0, \beta_0\right).$$

the default setting of the parameters is $\mu_0 = 0, \sigma_0^2 = 1, \lambda_0 = 1, \beta_0 = 1$. Re-scaling the data y such that its mean is 0 and standard deviation is 1. leads to the parameterization of G_0 as an uninformative prior.

Since this prior is conjugate, the predictive distribution for a new observation $\hat{y}$ can be found analytically, and is a location/scale Student-t distribution:

$$p(\tilde{y} \mid \gamma) = \int k(\hat{y} \mid \theta) p(\theta \mid G_0) d\theta = \frac{1}{\tilde{\sigma}} \text{Student-t}\left(\frac{\tilde{y} - \tilde{\mu}}{\tilde{\sigma}} \mid \tilde{v}\right),$$

where $\tilde{v} = 2\lambda_0, \quad \tilde{\mu} = \mu_0, \quad \tilde{\sigma} = \sqrt{\frac{\beta_0(k_0+1)}{\lambda_0 k_0}}$.

Finally the posterior distribution is also a Normal Inverse Gamma distribution due to the conjugate nature of the prior

$$\begin{aligned}
p(\theta \mid y, \gamma) &= N(\mu \mid \mu_n, \frac{\sigma^2}{k_0 + n}) \text{Inv-Gamma}(\sigma^2 \mid \lambda_n, \beta_n), \\
\mu_n &= \frac{\kappa_0 \mu_0 + n\overline{y}}{k_0 + n}, \\
\lambda_n &= \lambda_0 + \frac{n}{2}, \\
\beta_n &= \beta_0 + \frac{1}{2}\sum_{i=1}^{n}(y_i - \overline{y})^2 + \frac{\kappa_0 n(\overline{y} - \mu_0)^2}{2(\kappa_0 + n)}.
\end{aligned}$$

References

[1] Paul H. C. Eilers and Brian D. Marx. Flexible smoothing with b-splines and penalties. 1996.

[2] Cosma Shalizi. Splines: Lecture notes in Advanced Data Analysis, 2011. URL: `https://www.stat.cmu.edu/~cshalizi/402/lectures/11-splines/lecture-11.pdf` Last visited on 2020/03/18.

[3] Zou Y. Terry W. Karmaus W. Zhang, H. and H. Arshad. Joint clustering with correlated variables. *The American Statistician*, 73:296–306, 2018.

[4] Matthews S. Tariq S. Hide, D. and S. Arshad. Allergen avoidance in infancy and allergy at 4 years of age. *Allergy*, 51:89–93, 1996.

[5] Jim Pitman and Marc Yor. The two-parameter poisson-dirichlet distribution derived from a stable subordinator. *Ann. Probab.*, 25(2):855–900, 04 1997.

[6] Y. W. Teh. A Bayesian interpretation of interpolated Kneser-Ney. Technical Report TRA2/06, School of Computing, National University of Singapore, 2006.

[7] Steven N MacEachern. Dependent nonparametric processes. In *ASA proceedings of the section on Bayesian statistical science*, pages 50–55. Alexandria, Virginia. Virginia: American Statistical Association; 1999, 1999.

[8] Yee Whye Teh, Michael I Jordan, Matthew J Beal, and David M Blei. Hierarchical dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581, 2006.

[9] George Casella. Empirical Bayes Gibbs sampling. *Biostatistics*, 2(4):485–500, 12 2001.

[10] Sethuraman J. A constructive definition of Dirichlet priors. *Statistica sinica*, pages 693–650, 1994.

[11] W. K. Hastings. Monte Carlo Sampling Methods Using Markov Chains and Their Applications. *Biometrika*, 57(1):97–109, 1970.