Informative Priors in Clustering

Alex Dombowsky

2022-06-04

Motivation

Assume that data $\boldsymbol{y} = \{y_i\}_{i=1}^n$ are conditionally independent given $\boldsymbol{c} = \{c_i\}_{i=1}^n$, a partition of the integers $\{1, \dots, n\}$ into K disjoint groups, as well as component-specific parameters $\boldsymbol{\theta} = \{\theta_h\}_{h=1}^K$. We view $\boldsymbol{c}$ as a random partition, assuming that it follows a prior distribution $p(\boldsymbol{c})$. Examples of possible choices of $p(\boldsymbol{c})$ include a uniform prior on the space of all partitions, or exchangeable partition probability functions (EPPFs) induced by a finite mixture or a nonparametric Bayesian prior such as the Dirichlet and Pitman-Yor processes. The prior distribution $\boldsymbol{p}(\theta)$ is also relevant, as these component specific parameters define the characteristics of each cluster.

Our focus is on creating classes of priors that can take into account different types of prior information on $\boldsymbol{c}$ and/or $\boldsymbol{\theta}$. Prior information in environmental health could take the form of an informed guess by investigators of the partition $\boldsymbol{c}$ of study participants, clusterings found in previous studies, covariates gathered on observations from follow-up visits, and existing groupings of chemical exposures, such as phthalates, based on their structure. Partitioning rare data could be improved by using prior information. Prior information can help lead to more scientifically meaningful clusters that reflect the goals of a study, as opposed to other clustering methods based on mathematical similarities which may conflict with already available information in the field.

Data from the National Birth Defects Prevention Study (NBDPS) helps motivate our goals. The NBDPS enrolled mothers with expected delivery dates in the years 1997-2009, then monitored their children for birth defects across several follow-up visits (Yoon et al. 2001). Though in previous work we have clustered the birth defects themselves using a logistic regression model, another possible clustering objective is to partition mothers by their environmental exposure profiles. The nature of this study includes several possible sources of prior information, including groupings of birth defects specified by investigators, demographic and environmental exposure covariates on the mothers from a telephone interview, follow-up information on the children, and previous birth defect studies.

Preliminary Results: the CP Process

For the setting where we have an informed guess $\boldsymbol{c}_0$ of the underlying partition, we have developed Centered Partition (CP) processes, a class of priors that are shrunk towards $\boldsymbol{c}_0$ (Paganin et al. 2021). The CP process takes the form

$$p(\boldsymbol{c} \mid \boldsymbol{c}_0, \psi) \propto p_0(\boldsymbol{c}) \exp\left\{-\psi d(\boldsymbol{c}, \boldsymbol{c}_0)\right\}. \tag{1}$$

Here, $d(\boldsymbol{c}, \boldsymbol{c}_0)$ denotes a measure of dissimilarity between the partitions $\boldsymbol{c}$ and $\boldsymbol{c}_0$, $\psi > 0$ is a penalization parameter for this dissimilarity, and $p_0(\boldsymbol{c})$ is an EPPF. We implement the CP process using the Variation of Information distance function (Meilă 2007) for $d(\boldsymbol{c}, \boldsymbol{c}_0)$, though the CP process can theoretically be implemented with other partition dissimilarity measures such as Binder's loss (Binder 1978). In Paganin et al. (2021), we also provide guidance on how to choice ψ, how to implement the CP process in an MCMC algorithm, and apply this method to congenital heart defects in the NBDPS data (Figure 1).

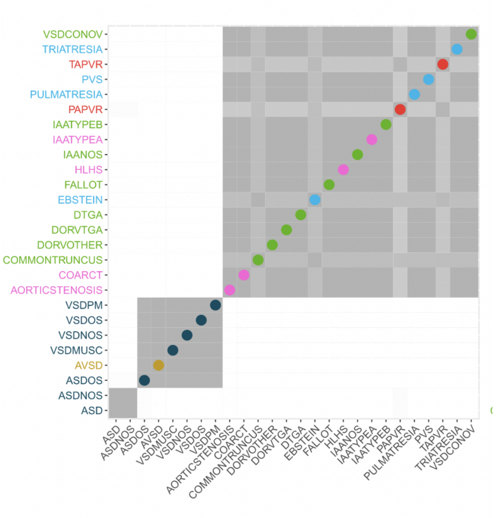

Figure 1: The posterior allocation matrix for the CP process fit to the NBDPS data, where the baseline EPPF is a Dirichlet process with $\alpha = 1$, and $\psi = 80$. The axis labels are congenital heart defects. Colors on the y-axis correspond to an initial clustering $\boldsymbol{c}_0$ given by investigator. From Paganin et al. (2021).

Future Work

Our guidance on choosing ψ becomes significantly more difficult for increasing n, and future work should go into expanding our prior calibration method for larger sample sizes. We also would like to show theoretical results about the CP process and see whether theoretical guarantees from other clustering methods also apply to the CP process. Furthermore, a prior guess $\boldsymbol{c}_0$ may only be feasible for a subset of the observations, or we may be more confident about some cluster labels in our initial guess over others.

We will extend the CP process to allow for the more general setting of having several possible initial guesses $\boldsymbol{c}_{0,1}, \ldots, \boldsymbol{c}_{0,R}$. Having multiple guesses emerges when there are different scientific consensuses on how to group the data. Rigon, Aliverti, Russo, and Scarpa in Paganin et al. (2021) recommend using an additive penalty, such as in Bissiri, Holmes, and Walker (2016), of the form

$$p(\boldsymbol{c} \mid \boldsymbol{c}_{0,1}, \ldots, \boldsymbol{c}_{0,R}) \propto p_0(\boldsymbol{c}) \exp\left\{ -\psi \sum_{r=1}^{R} d(\boldsymbol{c}, \boldsymbol{c}_{0,r}) \right\}. \tag{2}$$

An additional modification would be to provide a weighting structure to the initial guesses, such as the prior proposed by Casa, Fop, and Murphy in Paganin et al. (2021):

$$p(\boldsymbol{c} \mid \boldsymbol{c}_{0,1}, \ldots, \boldsymbol{c}_{0,R}) \propto p_0(\boldsymbol{c}) \exp\left\{ -\psi \sum_{r=1}^{R} \phi_r d(\boldsymbol{c}, \boldsymbol{c}_{0,r}) \right\}. \tag{3}$$

where $\phi_r > 0$, $\sum_{r=1}^{R} \phi_r = 1$.

A further improvement on the CP process would be to use a penalty for clusters $\boldsymbol{c}$ that have undesirable properties, even if they are close to $\boldsymbol{c}_0$ in the distance function. An example, discussed by D'Angelo and Canale in Paganin et al. (2021), is the case where we wish to penalize partitions with more clusters than $\boldsymbol{c}_0$, denoted K_0. This leads to the prior

$$p(\boldsymbol{c} \mid \boldsymbol{c}_0) \propto p_0(\boldsymbol{c}) \exp\left\{ -\left(\psi_1 \mathbf{1}(K_{\boldsymbol{c}} \leq K_0) + \psi_2 \mathbf{1}(K_{\boldsymbol{c}} > K_0)\right) d(\boldsymbol{c}, \boldsymbol{c}_0) \right\} \tag{4}$$

where $K_{\boldsymbol{c}}$ is the number of clusters in $\boldsymbol{c}$. Amongst equidistant partitions, $\psi_2 > \psi_1$ will penalize those with more clusters than K_0 more heavily than partitions with fewer. (4) could be an example of a more general class of priors which penalize equidistant partitions in different ways based off an applied motivation.

Lastly, Antoniano-Villalobos, Villa, and Wade in Paganin et al. (2021) suggest using $\boldsymbol{c}_0$ as a covariate in a dependent normalized random measure. Utilizing $\boldsymbol{c}_0$ as a covariate (and integrating in other covariate information) provides other promising directions for extending the CP process.

Future Directions

Constrained Clustering

As discussed earlier, prior guesses of the partition are only one type of prior information that could be available. A more complicated setting would be that we have pairwise information for each $1 \leq i < j \leq n$. In constrained clustering (Lu and Leen 2004), we assume that we have a symmetric $n \times n$ matrix $\boldsymbol{W}$ of prior information on the observations so that, for each (i, j) pair, W_{ij} expresses our certainty in whether observations i and j should be in the same or different clusters. For clustering birth defects, this may arise by noting etiological similarities and differences between the defects. $\boldsymbol{W}$ is accounted for in the prior of $\boldsymbol{c}$, written as

$$p(\boldsymbol{c} \mid \boldsymbol{W}, \boldsymbol{\pi}) \propto \exp\left\{\sum_{i=1}^{n}\sum_{j\neq i}\mathbf{1}(c_i = c_j)W_{ij}\right\}\prod_{i=1}^{n}\mathrm{Cat}(c_i; 1 : K, \boldsymbol{\pi}) \tag{5}$$

where $\mathrm{Cat}(c_i; 1 : K, \boldsymbol{\pi})$ denotes the categorical distribution with K categories and probabilities $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_K)$, and $h(\boldsymbol{c}, \boldsymbol{W}) = \exp\left\{\sum_{i=1}^{n}\sum_{j\neq i}\mathbf{1}(c_i = c_j)W_{ij}\right\}$ is a weighting function, which increases for partitions that coincide with our prior information. Constrained clustering provides several possible future directions for our aim. One possible direction leverages our already developed CP process by utilizing the weighting function $h(\boldsymbol{c}, \boldsymbol{W})$ in a similar fashion to the clustering distance function $d(\boldsymbol{c}, \boldsymbol{c}_0)$ in (1). This could generalize the CP process to more nuanced prior information. A related direction is to investigate choices of $\boldsymbol{W}$ that could improve clustering, such as functions of additional observation-level covariates obtained at follow-up visits $\{x_i\}_{i=1}^{n}$ or pairwise covariates $\{x_{ij}\}_{1\leq i<j\leq n}$.

Incorporating Classiciation Tree Information

We will also pursue creating a broad class of priors that allow for incorporation of knowledge from a classification hierarchy. Classification hierarchies provide more nuanced information than an initial partition since the different levels, or resolutions, correspond to merges of smaller groups. Each resolution has a different meaning, reflecting various ways to group the data together. Incorporating classification hierarchy information into priors aids our overarching goal of providing scientifically meaningful clusters. Additionally, classification hierarchies are relevant in the NBDPS example, since hierarchical structures have been proposed to group congenital heart defects (Botto et al. 2007).

For example, suppose that, instead of one prior guess $\boldsymbol{c}_0$, we have access to L partitions $\overline{\boldsymbol{c}}_1, \ldots, \overline{\boldsymbol{c}}_L$, which result from merging clusters in the hierarchical tree T. Smith in Paganin et al. (2021) recommends using point mass mixture priors in this case:

$$p(\boldsymbol{c} \mid \boldsymbol{c}_0, \psi) \propto p_0(\boldsymbol{c})\exp\left\{-\psi d(\boldsymbol{c}, \boldsymbol{c}_0)\right\}; \qquad\qquad \boldsymbol{c}_0 \sim \sum_{l=1}^{L} p_l \delta_{\overline{\boldsymbol{c}}_l};$$

where $p_l > 0$, $\sum_{l=1}^{L} p_l = 1$. Another approach is to draw

$$\boldsymbol{c}_0 \sim P_T$$

where P_T is a Markov process on the tree, which selects a partition by choosing a certain resolution and accounts for the correlation between resolutions. Elicitation of $\{p_l\}_{l=1}^{L}$ and P_T that replicate desirable properties of the hierarchy provide possible directions in this area.

Similarly, the prior information matrix $\boldsymbol{W}$ in (5) could be selected from a point mass mixture over different prior information matrices for each possible partition; or a single $\boldsymbol{W}$ could be specified in a way that preserves the hierarchical structure.

Modeling Grouped Data with a Known Hiearchical Structure

We aim to develop approaches for modeling grouped data believed to have been generated from a divisive hierarchical structure. For this example, we focus on a two-level hierarchy:

$$y_1 \sim \sum_{h=1}^{K} \pi_h \mathcal{K}(\lambda_h) \tag{6}$$

$$\lambda_1, \dots, \lambda_K \sim \sum_{l=1}^{L} q_l \mathcal{F}(\theta_l) \tag{7}$$

$$\theta_1, \dots, \theta_L \sim P \tag{8}$$

Testing for Differences in Clustering Structure Across Multiple Studies with Partial Information

Let $\boldsymbol{z} = \{z_l\}_{l=1}^{m}$, where $z_l = (z_{l1}, \dots, z_{lp})$, be a large data set from a previous clustering study, where $\boldsymbol{z}$ consists of the same variables as $\boldsymbol{y}$ but measured on a different population. Assume that the previous study partitioned $\boldsymbol{z}$ into K clusters. Additionally, suppose we do not have access to the raw data $\boldsymbol{z}$, but do have access to within-cluster sample statistics for each variable, such as the sample means $\bar{z}_{\cdot j}^{(h)}$ and variances $s_{\cdot j}^{(h)2}$ for $j = 1, \dots, p$ and $h = 1, \dots, K$.

A natural question to test is whether the cluster structure in $\boldsymbol{y}$ differs from the structure in $\boldsymbol{z}$. That is, we would like to formally test whether the component kernels in $\boldsymbol{z}$ and $\boldsymbol{y}$ are identical. To construct such a test, we fit the model

$$y_i \sim \pi \left(\sum_{h=1}^{K} \omega^{(h)} \mathcal{N}_p(\mu^{(h)}, \Sigma^{(h)}) \right) + (1-\pi) f; \tag{9}$$

$$\Sigma^{(h)} = \operatorname{diag}(\sigma_1^{(h)2}, \dots, \sigma_p^{(h)2}); \tag{10}$$

$$\mu_j^{(h)}, \sigma_j^{(h)2} \sim \mathcal{N} - \mathcal{IG}\left(\xi_j^{(h)}, \lambda_j^{(h)}, \alpha_j^{(h)}, \beta_j^{(h)} \right); \tag{11}$$

$$\omega \sim \operatorname{Dir}(\gamma); \tag{12}$$

$$\pi \sim \operatorname{Beta}(a, a); \tag{13}$$

$$f \sim P \tag{14}$$

where the normal-inverse-gamma priors (11) are chosen to be very highly informative with prior predictive samples that recapture $\bar{z}_{\cdot j}^{(h)}$ and $s_{\cdot j}^{(h)2}$, γ in (12) leads to a weakly informative Dirichlet prior, a in (13) makes the Beta prior cocnentrate on values close to 0, and f in (9) is a non-parametric contamination (Scarpa and Dunson 2009), such as a Dirichlet process or Pitman-Yor process mixture. For the two candidate models in (9), we will calculate the Bayes Factor, creating a formal Bayesian hypothesis test for the mixture model defined by $\boldsymbol{z}$.

A related goal is to develop a localized test, where we evaluate whether there is a mixture component in $\boldsymbol{y}$ that was not captured in the analysis of $\boldsymbol{z}$ due to population differences in the two data sets. These differences may be geographic (e.g. $\boldsymbol{z}$ is measured on European mothers, while $\boldsymbol{y}$ is measured on African mothers) or temporal (e.g. $\boldsymbol{z}$ was measured in a study from 25 years ago). In essence, the localized test determines whether observing $\boldsymbol{y}$ uncovers an additional mixture component. To accomplish this, we set f in (9) to be

$$f = \mathcal{N}(\mu^{(*)}, \Sigma^{(*)}) \qquad\qquad (\mu^{(*)}, \Sigma^{(*)}) \sim \mathcal{N} - \mathcal{IW}_p(\theta^{(*)}, \phi^{(*)}, \Psi^{(*)}, \nu^{(*)}); \tag{15}$$

where$(\theta^{(*)}, \phi^{(*)}, \Psi^{(*)}, \nu^{(*)})$ are chosen so that the prior is weakly informative.

Resources

Binder, David A. 1978. "Bayesian Cluster Analysis." *Biometrika* 65 (1): 31–38.

Bissiri, Pier Giovanni, Chris C Holmes, and Stephen G Walker. 2016. "A General Framework for Updating Belief Distributions." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 78 (5): 1103–30.

Botto, Lorenzo D, Angela E Lin, Tiffany Riehle-Colarusso, Sadia Malik, and Adolfo Correa. 2007. "Seeking Causes: Classifying and Evaluating Congenital Heart Defects in Etiologic Studies." *Birth Defects Research Part A: Clinical and Molecular Teratology* 79 (10): 714–27.

Lu, Zhengdong, and Todd Leen. 2004. "Semi-Supervised Learning with Penalized Probabilistic Clustering." *Advances in Neural Information Processing Systems* 17.

Meilă, Marina. 2007. "Comparing Clusterings—an Information Based Distance." *Journal of Multivariate Analysis* 98 (5): 873–95.

Paganin, Sally, Amy H Herring, Andrew F Olshan, and David B Dunson. 2021. "Centered Partition Processes: Informative Priors for Clustering (with Discussion)." *Bayesian Analysis* 16 (1): 301–70.

Scarpa, Bruno, and David B Dunson. 2009. "Bayesian Hierarchical Functional Data Analysis via Contaminated Informative Priors." *Biometrics* 65 (3): 772–80.

Yoon, Paula W, Sonja A Rasmussen, MC Lynberg, CA Moore, M Anderka, SL Carmichael, P Costa, et al. 2001. "The National Birth Defects Prevention Study." *Public Health Reports* 116 (Suppl 1): 32.