The Journal of Open Source Software

ChiRP: Chinese Restaurant Process Mixtures for Regression and Clustering

Arman Oganisian[1]

1 Department of Biostatistics, Epidemiology, and Informatics, University of Pennsylvania

DOI: 10.21105/joss.01287

Software

- Review
- Repository
- Archive

Submitted: 27 February 2019
Published: 26 March 2019

Summary

`ChiRP` is a Monte Carlo Markov Chain (MCMC) implementation of **Chi**nese **R**estaurant **P**rocess (CRP) mixtures in `R`. CRP mixtures (Blackwell & MacQueen, 1973; Ferguson, 1973) are a class of Bayesian nonparametric models that can be used for robust regression modeling and clustering problems.

These are common tasks in biomedical research. However, regression often involves parametric assumptions (e.g. normality, linearity, constant variance). Similarly, clustering often involves pre-specifying the number of clusters - typically unknown to the researcher. Flexible machine learning methods exist for such problems, but they focus on predictive accuracy, making them inadequate for biomedical research applications where inference and interval estimation are of interest.

CRP mixtures work by partitioning a dataset into similar clusters - each associated with a locally parametric regression model. Unlike traditional clustering procedures, CRP mixtures allow for infinitely many clusters - thus side-stepping the need to pre-specify the number of clusters. Predictions are formed by ensembling over the local cluster-specific regression models. This fully Bayesian procedure produces an entire posterior distribution for both the cluster assignments and predictions - allowing for both point and interval estimation.

Outcome Types and Model Output

Suppose we are given training data with n subjects $D_T = (Y_i, X_i)_{i=1:n}$. Here, Y_i is the scalar outcome/label and X_i is a $p \times 1$ vector containing either binary or continuous features. `ChiRP` trains a CRP model and yields the following:

1. In-sample posterior mean predictions $(\hat{Y}_i)_{i=1:n}$ from a nonparametric CRP regression of Y on X.
2. Out-of-sample posterior mean predictions on an un-labeled test set $(\tilde{X}_i)_{1:m}$, $(\hat{\tilde{Y}}_i)_{i=1:m}$.
3. Latent posterior mode cluster membership for both training and testing subjects.

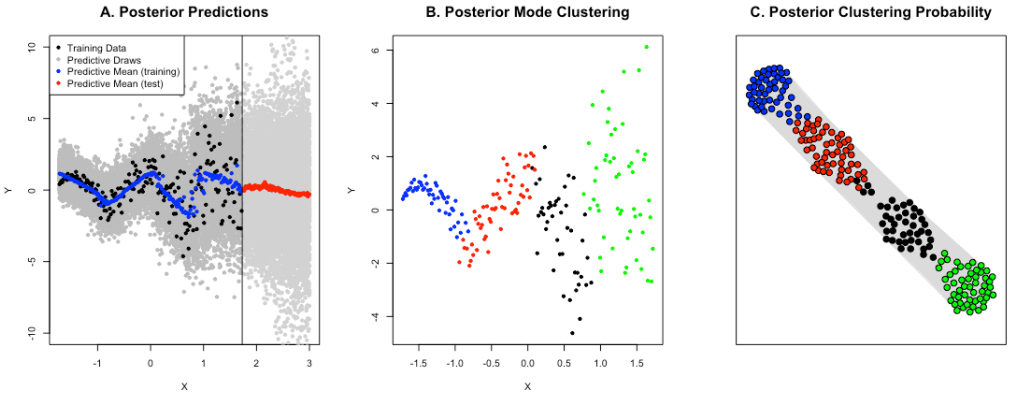
`ChiRP` is not limited to producing the posterior means and modes above. It returns a posterior distribution over each subject's predicted outcome and cluster membership - allowing the user to compute posterior point and interval estimates for any desired estimand of interest.

`ChiRP` implements different local cluster-specific regressions depending on the outcome type:

1. Zero-inflated, semi-continuous outcomes are modeled using cluster-specific zero-inflated regressions (See Oganisian, Mitra, & Roy, 2018)
2. Continuous outcomes are modeled using cluster-specific linear regressions (See Hannah, Blei, & Powell, 2011).
3. Binary outcomes are modeled using cluster-specific logistic regressions.

Simulated Example

The figure illustrates a CRP mixture of linear regressions using outcome data generated from a sine wave. The first panel shows a flexible posterior mean prediction in blue for in-sample data. Posterior mean prediction for the test set is shown in red. We plot 100 posterior draws from the predictive distribution for each X_i to display the uncertainty around the posterior mean prediction. Percentiles of these draws can be used to form credible intervals around the predicted mean.



Panel B gives some intuition about why a locally linear regression works so well with such complex data. The CRP model induces a clustering of points that are similar to each other in terms the linear model parameters. The CRP discovers four distinct clusters - each with its own linear regression. These clusters are indicated by color. Predictions are generated by averaging predictions from these cluster-specific models.

Panel C represents each subject in the training set as a node in a graph. The line connecting any two nodes is inversely proportional to the posterior probability of two subjects being clustered together. Colors indicate the posterior mode assignment. For example, the blue and green points are almost never clustered together. From Panel B it is obvious why this is: the blue and green points are far apart in (X, Y) space. Note that some black points in Panel C are very close to the red points. This indicates that cluster assignment for these subjects are highly uncertain. They could belong to the red group or black group with significant probability. These points are the same points in Panel B around $X = 0$ - the boundary of red and black.

Acknowledgements

This work was supported in part by Grant R01GM112327 from National Institute Of General Medical Sciences. Thanks to Dr. Jason Roy for helpful discussions of underlying MCMC computations and to Nicholas Illenberger and Carolyn Lou for help with package branding.

References

Blackwell, D., & MacQueen, J. B. (1973). Ferguson distributions via polya urn schemes. *Ann. Statist.*, *1*(2), 353–355. doi:10.1214/aos/1176342372

Ferguson, T. S. (1973). A bayesian analysis of some nonparametric problems. *Ann. Statist.*, *1*(2), 209–230. doi:10.1214/aos/1176342360

Gershman, S. J., & Blei, D. M. (2012). A tutorial on bayesian nonparametric models. *Journal of Mathematical Psychology*, *56*(1), 1–12. doi:http://dx.doi.org/10.1016/j.jmp.2011.08.004

Hannah, L. A., Blei, D. M., & Powell, W. B. (2011). Dirichlet process mixtures of generalized linear models. *Journal of Machine Learning Research*, *12*(Jun), 1923–1953. Retrieved from http://www.jmlr.org/papers/volume12/hannah11a/hannah11a.pdf

Hastings, W. K. (1970). Monte carlo sampling methods using markov chains and their applications. *Biometrika*, *57*(1), 97–109. doi:10.2307/2334940

Müller, P., Quintana, F., Jara, A., & Hanson, T. (2015). *Bayesian nonparametric data analysis*. Springer series in statistics. Springer International Publishing. doi:10.1007/978-3-319-18968-0

Neal, R. M. (2000). Markov chain sampling methods for dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, *9*(2), 249–265. doi:10.1080/10618600.2000.10474879

Oganisian, A., Mitra, N., & Roy, J. (2018). A bayesian nonparametric model for zero-inflated outcomes: Prediction, clustering, and causal estimation. *arXiv e-prints*.

Rodríguez, C. E., & Walker, S. G. (2014). Label switching in bayesian mixture models: Deterministic relabeling strategies. *Journal of Computational and Graphical Statistics*, *23*(1), 25–45. doi:10.1080/10618600.2012.735624

Stephens, M. (2000). Dealing with label switching in mixture models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *62*(4), 795–809. doi:10.1111/1467-9868.00265