Matt Dickenson `mcd31`
STA561/CS571 — Fall 2013 **Homework 9** Due: 25 November, 2013

Homework Notes: I did not work with anyone else on this homework or refer to resources other than the course notes, textbook, and course Piazza page.

Problem 1

A The generative model for a continuous η, a base distribution G_0, concentration parameter α, and the $\{B_1, ...B_K\}$ partitions ($K = \text{inf}$), is:

$$\begin{aligned}
(G(\eta \in B_1), ..., G(\eta \in B)K) &\sim \text{Dirich}(\alpha G_0(B_1), ...\alpha G_0(B_K)) \\
p(\eta_i \in B_j) &= \int p(\eta_i \in B_j | G) p(G|G_0) dG \\
&= \frac{\alpha G_0(B_j)}{\sum_K \alpha G_0(B_k)} \\
&\propto \alpha G_0(B_j)
\end{aligned}$$

The posterior is

$$G|\eta_{1:n}, \alpha, G_0 \sim DP(\alpha, G_0 + \sum_{i=1}^{n} \delta_{\eta_i}(\eta))$$

A simple choice for the base distribution is G_0 is the Gamma distribution, due to the conjugacy of the Gamma distribution with the Gaussian distribution.

B For the cluster assignment step in the Gibbs sampler, we can exploit exchangeability.

```
cluster = function(x, centroids, alpha){
  # do the restaurant process
  table_counts = restaurant(x, alpha)
  table_props = table_counts/sum(table_counts)
  num_tables = length(table_counts)

  # then exploit permutation
  permuted_x = sample(x)
  n = length(x)
  table_assignments = rep(NA, n)

  # pretend each x is last to arrive
  for(i in 1:length(permuted_x)){
    table_i = sample(c(1:num_tables), 1, prob=table_props)
    table_assignments[i] = table_i
  }
}
```

```

restaurant = function(x, alpha){
  table_counts = c(1) # number of 'customers' at each 'table'
                      # first customer at first table
  for(m in 2:n){
    tmp = c(table_counts, alpha)
    table_props = tmp/sum(tmp)

    # assign each 'customer' to a 'table' according to crp
    table_m = sample(c(1:length(tmp)), 1, prob=table_props)
    if(table_m==length(tmp)){ table_counts[table_m] = 1}
    else{ table_counts[table_m] = table_counts[table_m] + 1}
  }
  return(table_counts) # sufficient statistic
}
```

C This algorithm does not discard empty clusters. Because the number of clusters is potentially infinite (in theory), it is possible for an observation to be assigned to a previously empty cluster at any iteration of the Gibbs sampler.

D Rather than discarding clusters with fewer than γ points assigned to them, we allow the number of potential clusters to be infinite. Thus, this model better addresses the issue of not knowing the number of clusters *a priori*. If we examine the number of clusters at each iteration of the Gibbs sampler (or across multiple runs), we can even get a posterior distribution over the number of clusters.

Problem 2

As one of the most widely used dependent variables in international conflict studies, much effort has been devoted to estimating models of MID onset and duration. However, this work suffers from several common weaknesses that this project attempts to ameliorate: virtually all projects, especially before the present decade, used a fixed functional form (typically from the family of generalized linear models); out-of-sample testing and cross-validation is used only rarely, making claims of 'prediction' somewhat dubious in many cases; and often the independent variables are measured at the annual level with high levels of serial correlation, meaning that there is little temporal variation in the predictors, while the dependent variable tends to exhibit more sudden onsets [1]. A recent shift toward event data has helped to address the latter two of these issues: with frequent updates (often measured at the daily level), there is substantial variation in the independent variables, validation requires only a brief waiting period for new sets of test data [2, 3, 4, 5, 6].

With this transition toward event data as predictors, the political forecasting community has become attune to new challenges and has responded with several established practices. Coding the sentiment of interactions can now be done in near real-time (NRT) using the Tabari system, which aggregates and deduplicates news reports [7, 8]. Sentiment coding can be done according to two widely used systems. The Goldstein scale assigns a score of -10 (highly conflictual) to +10 (highly cooperative) to events, but it is difficult to employ this scale for aggregations or permutations of the data [9]. CAMEO classifies events into a pre-defined schema of material/verbal and cooperative/conflictual actions, that makes aggregation simpler because we can count events within each category [3]. These event classifications provide a principled, automated method for exhaustively categorizing the types of events that may consitute an interstate dispute [10].

The community has also dealt with challenges when aggregating event data up to various temporal levels. Although there is no single best practice, monthly aggregation has become a common strategy [11, 12] and is used in this project. Modfiying the features by transforming the raw counts into month-to-month changes (i.e. first-differencing) and measuring the balance between conflictual and cooperative interactions as a percentage of the total also helped to simplify the feature set [13].

Interpretability is an important concern in this project due to the policy-relevant nature of the problem and the (potential) need to compare the resulting model to the process used by human coders involved in creating the MID dataset [10]. For this reason, "black box" methods such as Support Vector Machines were judged to be inappropriate. Classification trees (and their continuous counterpart, regression trees, collectively known as CART) offer a nice alternative that is more flexible than GLMs and more interpretable than Random Forests (these two methods should provide lower and upper bounds, respectively, on CART) [14]. CART has been used for event data within conflict studies, and in public health where researchers encounter similar issues of unbalanced and missing data [15, 16, 17].

In later stages of this project, several additional tools may help to improve the predictive accuracy of the model. International conflict is a relatively rare event, meaning that in k-fold cross validation it is possible that some subsets will have no instances of conflict; to prevent this, synthetic minority over-sampling (SMOTE) could be used [18]. To incorporate interdependencies not captured at the dyadic level, future iterations could also include lags that measure conflict in social or spatial neighbors [19, 20, 21, 22, 23, 24]. A Bayesian ensemble model of several classification trees could also improve performance while still maintaining more interpretability than is available in random forests [11, 25, 26, 27]. If these methods are successful, the general processing of automating political indicators through the use of event data could also be applied to other widely used indices such as the Polity and Freedom House regime scales (measuring democracy and autocracy).

References

[1] M.D. Ward, B.D. Greenhill, and K.M. Bakke. The perils of policy by p-value: Predicting civil conflicts. *Journal of Peace Research*, 47(4):363–375, 2010.

[2] Deborah J Gerner, Philip A Schrodt, Ronald A Francisco, and Judith L Weddle. The analysis of political events using machine coded data. *International Studies Quarterly*, 38(1):91–119, 1994.

[3] Deborah J. Gerner, Philip A. Schrodt, Yilmaz Ömür, and Rajaa Abu-Jabr. Conflict and Mediation Event Observations (CAMEO): A New Event Data Framework for a Post Cold War World. Boston, MA, August, 29-September 1 2002. Annual Meetings of the American Political Science Association.

[4] Gary King and Will Lowe. An automated information extraction tool for international conflict data with performance as good as human coders: A rare events evaluation design. *International Organization*, 57(03):617–642, 2003.

[5] Andrea Ruggeri, Theodora-Ismene Gizelis, and Han Dorussen. Events data as bismarck's sausages? intercoder reliability, coders' selection, and data quality. *International Interactions*, 37(3):340–361, 2011.

[6] P. Schrodt and K. Leetaru. Gdelt: Global data on events, location and tone, 1979-2012. *International Studies Association*, 2013.

[7] Sean P Obrien. Crisis early warning and decision support: Contemporary approaches and thoughts on future research. *International Studies Review*, 12(1):87–104, 2010.

[8] Philip A Schrodt. Tabari: Textual analysis by augmented replacement instructions, version 0.7. 2009.

[9] Joshua S Goldstein. A conflict-cooperation scale for weis events data. *Journal of Conflict Resolution*, 36(2):369–385, 1992.

[10] Faten Ghosn, Glenn Palmer, and Stuart A Bremer. The mid3 data set, 1993–2001: Procedures, coding rules, and description. *Conflict Management and Peace Science*, 21(2):133–154, 2004.

[11] Bryan Arva, John Beieler, Bejamin Fisher, Gustavo Lara, Philip A Schrodt, Wonjun Song, Marsha Sowell, and Sam Stehle. Improving forecasts of international events of interest. In *EPSA 2013 Annual General Conference Paper*, volume 78, 2013.

[12] James E. Yonamine. Working with event data: A guide to aggregation choices. *Ph.D. Thesis*, 2013.

[13] George E. P. Box and Gwilym M. Jenkins. *Time Series Analysis: Forecasting and Control*. Holden-Day, San Francisco, 1976.

[14] Beata Beigman Klebanov, Daniel Diermeier, and Eyal Beigman. Lexical cohesion analysis of political speech. *Political Analysis*, 16(4):447–463, 2008.

[15] Philip A Schrodt. Predicting interstate conflict outcomes using a bootstrapped id3 algorithm. *Political Analysis*, 2(1):31–56, 1990.

[16] N Speybroeck. Classification and regression trees. *International journal of public health*, 57(1):243–246, 2012.

[17] Robert Trappl, Johannes Fürnkranz, and Johann Petrak. Digging for peace: Using machine learning methods for assessing international conflict databases'. In *ECAI*, pages 453–457. PITMAN, 1996.

[18] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16(1):321–357, 2002.

[19] Kristian S Gleditsch and Michael D Ward. War and peace in space and time: The role of democratization. *International Studies Quarterly*, 44(1):1–29, 2000.

[20] Kristian S Gleditsch and Michael D Ward. Measuring space: A minimum-distance database and applications to international studies. *Journal of Peace Research*, 38(6):739–758, 2001.

[21] Peter D Hoff and Michael D Ward. Modeling dependencies in international relations networks. *Political Analysis*, 12(2):160–175, 2004.

[22] Michael D Ward and Kristian S Gleditsch. Democratizing for peace. *American Political Science Review*, pages 51–61, 1998.

[23] M.D. Ward, R.M. Siverson, and X. Cao. Disputes, democracies, and dependencies: A reexamination of the kantian peace. *American Journal of Political Science*, 51(3):583–601, 2007.

[24] Michael D Ward, Katherine Stovel, and Audrey Sacks. Network analysis and political science. *Annual Review of Political Science*, 14:245–264, 2011.

[25] Jacob M Montgomery, Florian M Hollenbach, and Michael D Ward. Improving predictions using ensemble bayesian model averaging. *Political Analysis*, 20(3):271–291, 2012.

[26] Adrian E. Raftery. Bayesian model selection in social research (with discussion). In Peter V. Marsden, editor, *Sociological Methodology 1995*. Blackwell, Cambridge, MA, 1995.

[27] Adrian E Raftery, Tilmann Gneiting, Fadoua Balabdaoui, and Michael Polakowski. Using bayesian model averaging to calibrate forecast ensembles. *Monthly Weather Review*, 133(5):1155–1174, 2005.