Convex linear combination processes for compositions

John Bacon-Shone
Social Sciences Research Centre,
University of Hong Kong,
Pokfulam Road,
Hong Kong
johnbs@hku.hk

Abstract

Aitchison and Bacon-Shone (1999) considered convex linear combinations of compositions. In other words, they investigated compositions of compositions, where the mixing composition follows a logistic Normal distribution (or a perturbation process) and the compositions being mixed follow a logistic Normal distribution. In this paper, I investigate the extension to situations where the mixing composition varies with a number of dimensions. Examples would be where the mixing proportions vary with time or distance or a combination of the two. Practical situations include a river where the mixing proportions vary along the river, or across a lake and possibly with a time trend. This is illustrated with a dataset similar to that used in the Aitchison and Bacon-Shone paper, which looked at how pollution in a loch depended on the pollution in the three rivers that feed the loch. Here, I explicitly model the variation in the linear combination across the loch, assuming that the mean of the logistic Normal distribution depends on the river flows and relative distance from the source origins.

Introduction

Aitchison and Bacon-Shone (1999) considered a number of different models for how a composition may depend on a number of independent sources. However, the modeling was done separately for each outcome composition. In practice, we would expect the mixing process to have common features across different outcomes and to be able to model how the mixing process varies across locations or across time. This paper considers some possibilities for how to model the mixing process.

Notation

This follows Aitchison and Bacon-Shone, in using the vector π for the mixing proportions of dimension C, x_i is the vector composition of the ith source of dimension D and y_j is the vector composition at the jth location of dimension D.

Motivating example

The Aitchison and Bacon-Shone paper considered the situation of a Scottish loch supplied by three rivers, with 10 water samples taken at the mouth of each river and analysed into four-part compositions. Also available were 10 samples, taken at each of three fishing locations. In this case, we instead have 10 samples taken from each of five new fishing locations and we know the longitude and latitude of all eight locations (five new fishing locations and three sources). We also know the average flow rate for each river.

Mixing models

Aitchison and Bacon-Shone considered three possibilities for mixing:

1) the mixing proportions are fixed for a location, which they called the fixed-mixture model and there is no perturbation, so all variability can be attributed to variability in the sources;
2) the mixing proportions are fixed, but there is perturbation that increases the variability beyond that due to the sources, this is called the perturbation model;
3) the mixing proportions are not fixed, but vary according to some logistic normal distribution, which is called the convolution model.

Possible model of dependence in the mixing process

The models we are building are very simple compared to a full water flow model of the sort built by engineers. How the waters mix clearly depends on many factors including the depth of the water across the lake, the meteorological conditions etc. However, it seems reasonable that a good, but simple model of the mixing process should be feasible which can apply across the entire lake, rather than modeling each location separately. In particular, if the composition at the fishing sites does depend on the composition of the sources, we would expect the importance of a source to increase with greater flowrates and the closer that the sites are to the source. Thus we should be able to model some dependence of the mixing process on flowrates and distance from source, I refer to this as the source dependence model.

For the source dependence model, we would expect a proportional effect of flowrate and some unknown power of distance from the source. This suggests a model:

$$y_j = (\pi_0 x_0 + \Sigma \pi_{ij} x_i)/(\pi_0 + \Sigma \pi_{ij})$$

where $\pi_{ij} = f_i d_{ij}^{-ß,}$ f_i is the flowrate of the ith source, d_{ij} is the distance of the jth location from the ith source and ß is the unknown power relationship with distance, while x_0 is the unknown background composition. Slow mixing corresponds to ß close to zero, while fast mixing corresponds to ß large. Special cases include $\pi_0 = \infty$ which means that there is no source effect and $\pi_0 = 0$ which means that there is no background effect at all. Note that if one of the d_{ij} is close to 0 for a location (i.e. the location is at a source), then the composition matches the source, while if all d_{ij} are large for a location, then the composition matches the background. If the depth of the lake is quite constant, we might expect ß to be around 2 (effective area), while if the lake depth increases with distance from the source, ß might approach 3 (effective volume).

For the convolution model, we assume that this source dependence model applies for the centre of the mixing proportion distribution.

It is useful to note that the distance can use any metric of choice and also that the location of the sources can be treated as unknowns to be estimated in order to find the effective location of the sources, which may be different from the measured location.

In the example here, we assume that the distance metric is Euclidean.

Results

For simplicity, we just examine mixing model 1) here (i.e. all variability is due to the source variability). All three sources have an approximately equal flow rate. Rather than considering all possible values for ß, we consider the four possibilities for ß of 0.5, 1, 2 and 3 which covers the key situations that are easy to conceptualize.

Likelihood ratio tests show that values of beta other than 2 are easily rejected relative to 2. We cannot reject the possibility that π_0 is zero, suggesting that there is effectively no background composition playing a role.

Extensions

The model here is developed in the context of point sources in a two dimensional mixing space. It is easy to think of similar models to examine in other mixing situations. For example, if we have a river instead of a lake, but with point sources corresponding to tributaries, we could use a very similar model using a distance metric that measures distance from the source along the river.

However, these models arguably do not account well for rainfall or indeed general runoff. If rainfall simply dilutes all pollutants, then it is irrelevant when looking at the composition of pollutants, excluding water from the composition. Runoff is much trickier given that it will include unknown pollutants. Arguably the most interesting application would be to identify potential sources of additional runoff pollution, assuming that they would be largely point sources at unknown locations along the river. This should show up as large unexplained perturbations. Likelihood maximization should be able to indicate likely locations of these point sources and their composition. Runoff pollution that is not from point sources (e.g. fertilizer used over a wide area) would present a more difficult problem and might have to be modeled assuming regions of the river are sources, rather than specific locations.

Reference

J. Aitchison and J. Bacon-Shone, Convex linear combination of compositions, Biometrika, 86,2, 351-364, 1999

Appendix: Dataset

Each location has 10 data points of a 4 part composition
Source 1 is at location (0,5)
0.6541 0.1553 0.1129 0.0777
0.6012 0.2254 0.0825 0.0908
0.4490 0.4170 0.0472 0.0868
0.5354 0.2253 0.1478 0.0915
0.4097 0.3846 0.0938 0.1119
0.6601 0.1962 0.0818 0.0618
0.5033 0.2815 0.1171 0.0980
0.6862 0.1366 0.1158 0.0614

0.5527 0.1895 0.1596 0.0982
0.5420 0.3497 0.0349 0.0734
Source 2 is at location (10,0)
0.2450 0.2924 0.2450 0.2176
0.2194 0.1439 0.4830 0.1537
0.2196 0.1056 0.5004 0.1744
0.1008 0.3021 0.3834 0.2137
0.1706 0.2066 0.4369 0.1860
0.2181 0.2014 0.3389 0.2416
0.2588 0.1933 0.3138 0.2341
0.1598 0.2423 0.4100 0.1879
0.1406 0.2700 0.4488 0.1407
0.2503 0.0420 0.5571 0.1506
Source 3 is at location (0,-5)
0.3334 0.1704 0.2026 0.2936
0.4483 0.0784 0.2192 0.2541
0.3433 0.1295 0.2488 0.2785
0.2750 0.0949 0.3705 0.2595
0.3361 0.2414 0.1324 0.2901
0.3431 0.1688 0.1757 0.3123
0.3577 0.0462 0.3095 0.2866
0.4069 0.0791 0.2585 0.2555
0.3595 0.0699 0.1931 0.3775
0.4332 0.1409 0.1352 0.2907
Fishing location 1 is at (0,2.5)
0.428021 0.394058 0.065237 0.112684
0.515382 0.322413 0.059711 0.102494
0.632126 0.142509 0.131444 0.093912
0.602803 0.155444 0.140328 0.101424
0.399866 0.343484 0.129011 0.12764
0.488818 0.216777 0.180796 0.113599
0.475098 0.271345 0.129096 0.124366
0.517422 0.173598 0.190097 0.118883
0.61043 0.196273 0.102036 0.091177
0.606075 0.146849 0.141673 0.105403
Fishing location 2 is at (0,0)
0.505578 0.133867 0.197467 0.1631
0.448444 0.196678 0.149622 0.205167
0.451356 0.147233 0.176156 0.225211
0.519844 0.154533 0.161 0.164578
0.387333 0.217822 0.212178 0.182667
0.417822 0.252633 0.140422 0.189122
0.347089 0.308222 0.1504 0.1943
0.369678 0.206667 0.189411 0.234244
0.422222 0.234456 0.160956 0.182411
0.464822 0.185178 0.162989 0.186933
Fishing location 3 is at (0,-2.5)
0.419989 0.173613 0.14261 0.263793
0.357212 0.172198 0.206735 0.26386
0.293897 0.11316 0.352965 0.239883

0.35126 0.111463 0.193237 0.34404
0.35577 0.184728 0.177779 0.281637
0.350738 0.23446 0.145366 0.269436
0.373882 0.066206 0.29754 0.262363
0.286551 0.119754 0.353603 0.240007
0.445063 0.095645 0.225683 0.233609
0.358074 0.17698 0.177655 0.287205
Fishing location 4 is at (5,0)
0.3743 0.1593 0.2623 0.2041
0.323975 0.19075 0.2788 0.206425
0.349975 0.126625 0.3191 0.204275
0.357125 0.17735 0.280475 0.1851
0.334025 0.1958 0.3081 0.162075
0.271975 0.2508 0.313575 0.1637
0.343375 0.220675 0.233325 0.202625
0.34825 0.226675 0.216275 0.2088
0.275325 0.250025 0.2778 0.19685
0.35055 0.21105 0.210675 0.227725
Fishing location 5 is at (-5,0)
0.41702 0.214565 0.17509 0.193335
0.4628 0.172845 0.17475 0.18957
0.475455 0.157165 0.18734 0.179995
0.4529 0.24777 0.121425 0.177915
0.411975 0.152385 0.244125 0.191515
0.42823 0.14775 0.23949 0.184575
0.505265 0.15752 0.152645 0.18457
0.360315 0.19959 0.25776 0.182245
0.41843 0.263285 0.131375 0.18691
0.427205 0.16208 0.177905 0.23281