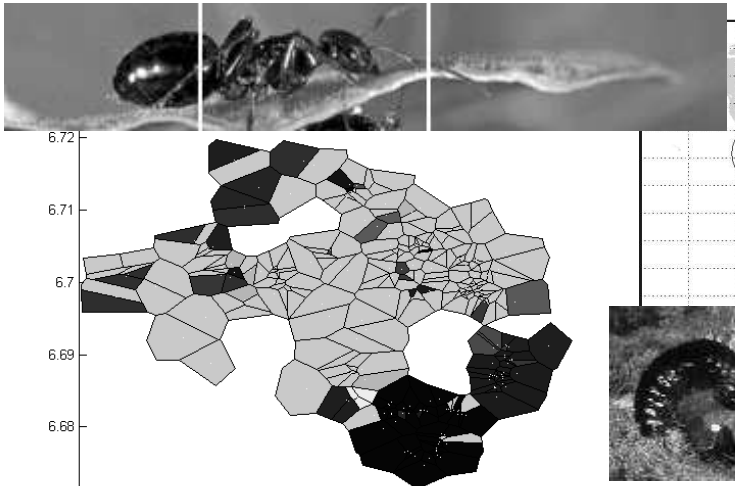
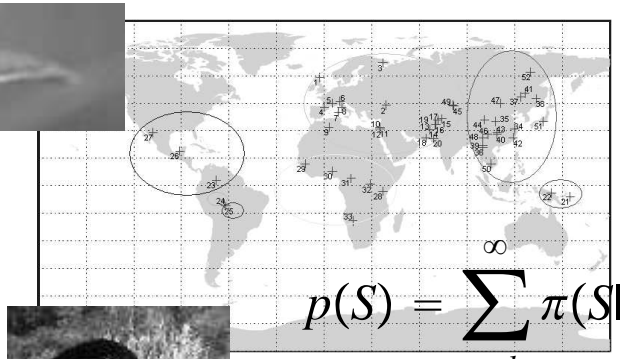
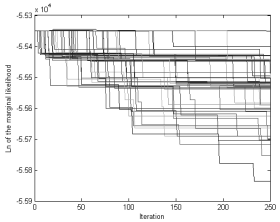
Bayesian analysis of genetic population structure using BAPS: Exercises







$$p(S) = \sum_{u=k}^{\infty} \pi(S|u)p(u), S \in \mathcal{S},$$

Jukka Corander

Department of Mathematics, Åbo Akademi University, Finland

Exercise 1: ’Clustering of groups of individuals’ followed by admixture analysis

- Launch BAPS and set the output file (File-Output File-Set). This txt-file will contain a summary of the results for performed analyses. If the output file is not set, BAPS will automatically create an output file for each analysis based on the name of the analyzed data set.
- Click on ’Clustering of groups of individuals’ and choose ’Preprocessed data’.
- Locate the appropriate data set HumanPopwisePreprocessed10Loci.mat.

- When either BAPS or Genepop data format is used, the program will ask if you wish to save the pre-processed data.
- BAPS converts any data set into another format which supports more efficient computations.
- Conversion takes some time, but enables in turn huge savings in the computational complexity of the estimation algorithms.
- Saving a pre-processed data set is a useful option in particular with large data sets, because multiple replicates of the SAME type of analysis may then be performed without pre-processing them over and over again.

- The current data set contains 52 sample populations with the total of 1056 individuals.
- It is a subset of the famous human data set in Rosenberg et al. (Science 2002).
- In this data set there is a random subset of 10 loci out of the 377 original microsatellites.
- The rather subtle genetic population structure reported in Rosenberg et al. suggests it is not sensible to attempt estimation WITHOUT the information provided by the sampling design.
- For instance, Rosenberg et al. concluded that a large number of loci were required to reliably separate African and European individuals.

- Therefore, we now perform a clustering of the 52 sample populations to investigate what the 10 microsatellites can tell us.
- After this genetic mixture analysis, it is possible to perform an admixture analysis on the level of INDIVIDUALS.
- Such an analysis can reveal migrants and admixed individuals in the inferred genetic mixture.
- BAPS now expects values on the user specified upper limit *K* for the #clusters (=*k*) to be provided.
- For each provided value of *K*, BAPS runs separately the estimation algorithm in the range $1 \leq k \leq K$.
- When multiple values are inserted, results from each run are stored internally in the program and after the last value the estimation results are compared w.r.t. their goodness-of-fit and merged into the final estimate.
- We try here the estimation with the sequence of *K*-values: 2 2 3 3 4 4 5 5 6 6 7 7 8 8 9 9 10 10 15 15 20 20

- For these data we will find evidence for 7 underlying groups.
- When the estimation procedure is completed, a graphical presentation of the inferred genetic mixture appears in a separate window (partition of the sample populations).
- When prompted, save the genetic mixture estimate.
- This enables the partition graphics to be reproduced later, use of descriptive measures (distances and trees of clusters), and also the use of inferred genetic mixture in a subsequent admixture inference.

- We now take a look at the results in text format by opening the output file, e.g. in WordPad.
- The value ‘Log(marginal likelihood) of optimal partition’ is the Bayesian goodness-of-fit measure, by which the different genetic mixture estimates are compared in BAPS.
- The table titled ‘Changes in log(marginal likelihood) if group i is moved to cluster j’ tells about the local peakedness of the posterior distribution around the optimal partition.
- The values in the table are log Bayes factors against the hypothesis of moving a single clustered unit (here a sample population) into an alternative cluster.
- A value close to 0 tells that the data cannot strongly say against the alternative allocation of the clustered unit.
- Values depend on the amount of loci, amount of missing data, level of informativeness of the loci, amount of information in the clustered unit, genetic distances between the clusters...

- The last lines in the result files provide a *crude* estimate of the posterior probability distribution over the #clusters.
- We see that k =7 is indicated as optimal, but k = 6 also has substantial probability mass.
- Therefore, it can be of interest to investigate how the estimate of the genetic mixture would look like conditional on k = 6.
- This can be done by using the ’Fixed K’ option available from Tools-Enable Fixed K clustering.
- Choose this and rerun the ’Clustering of groups’ with k = 6.
- The resulting partition image tells what has changed from the earlier optimum with k = 7.

- By loading a genetic mixture result file, it is possible to obtain various distances between the clusters and view trees of clusters showing the leves of their genetic relatedness.
- Load the first result file (with $k = 7$) through File-Load result-Mixture result.
- Use Graph-Phylogeny-NJ with Nei's distance to see how the 7 clusters are related to each other.
- The appearance of the tree can be changed from Attributes-Visual type.

- To investigate how stable the genetic mixture results are w.r.t to the randomly chosen set of loci, we now repeat the ’Clustering of groups’ analysis by using another preprocessed data set with 20 loci.
- Disable first the ’Fixed K’ option from Tools.
- Use the file HumanPopwisePreprocessed20Loci.mat and the same input as earlier.
- Save and compare the results.

- We now perform an admixture analysis conditional on the genetic mixture estimate for the data set with 20 microsatellite loci.
- Accept otherwise the default inputs, except that use 100 reference individuals.
- The estimation procedure and the calculation of the p-values for this data set with 1056 individuals takes ~5 minutes.
- After the estimation procedure, a window with the admixture graphic opens.
- Notice that this image shows ALL cases where the estimated admixture coefficients deviate from 0, even if they would not be significant according to the simulated p-values.

- Depending on the characteristics of the investigated data, such default admixture image may look noisy.
- An alternative admixture image can be produced by telling the program to show only the significant cases according to a user-specified threshold for the p-values.
- Use File-Load result-Admixture result and then Graph-View admixture results and choose a threshold (default is .05).
- We see that several cases of admixed ancestry have disappeared, as compared to the default image.
- The numerical admixture results can be more closely inspected in the ordinary output file, where the rows correspond to individuals and columns to the clusters in the genetic mixture result file, apart from the last column, which shows the simulation-based *p*-value for each individual.

Exercise 2: ’Clustering of individuals’ followed by admixture analysis

- We now examine a data set by Gasbarra et al. (2007, Theor Pop Biol) which was mentioned in the lectures.
- This data contains genotypes for 15 simulated microsatellites for 90 individuals.
- The individuals represent 30 trios of siblings, such that 3 sets of 10 trios come from 3 different populations separated by genetic isolation for 20 generations.
- As the sampled individuals from each underlying population form closely related subsets, it is expected that a genetic mixture analysis clustering individuals rather discovers the sibling trios than the 3 moderately separated populations.
- Choose ’Clustering of individuals’ and ’Preprocessed data’ and locate the data set HumanDataGasbarraSiblingTrios.mat.

- Run the genetic mixture analysis by providing the following sequence of input values: 5 5 10 10 15 15 20 20 25 25 30 30 35 35 40 40.
- Save the result and examine the inferred genetic population structure.
- It is seen that the estimated partition very accurately reflects the boundaries of the sibling trios.
- Run now an admixture analysis conditional on the genetic mixture estimate.
- Change the default input value for the minimum size of the populations to be included to 1 (default is 5) and set the #reference individuals to 200.
- We see that no admixed cases arise here, i.e. no false positives are obtained (there is no admixture in the data).

Exercise 3: ’Spatial clustering of groups’

- From BAPS v4.14 (went online 2006) version onwards there has been a more refined possibility to utilize geographical information in the estimation of the genetic population structure.
- The method was described in Corander et al (Comp Stat 2008) and it bears similarities with some of the other spatial methods considered in the genetics literature (e.g. GENELAND).
- The spatial model differs from the stochastic partition models considered in the earlier exercises by the assumption that the prior distribution over the space of partitions is not uniform.
- Instead, underlying genetic population structure is a priori assumed to have spatial smoothness, such that spatially organized clustering solutions have higher prior probabilities than spatially random clusterings.
- With relatively weak molecular data, this approach provides more statistical power to correctly infer population structure when it has at least a moderate degree of spatial smoothness.
- With increasing informativeness of the molecular data, the role of the spatial prior diminishes and the spatial model will yield similar results as the uniform prior.

- In ’Spatial clustering of groups’ it is assumed that biologically relevant coordinates are available for a number of sampling units.
- A sampling unit is a flexible construct and may refer to different things in different settings, e.g. it can be an ant nest or a distinct geographical breeding site (an example with butterflies is found in Orsini et al. Mol Ecol 2008).
- Varying numbers of individuals may have been collected from each sampling unit.
- The spatial clustering of groups is particularly helpful when there are very few loci available, or if the data contains a considerable degree of missing genotypes, which reduces the informativeness.
- The spatial model represents the genetic mixture in terms of a colored Voronoi tessellation of the set sampling unit coordinates.

- Choose ’Spatial clustering of groups’ and ’Preprocessed data’ and locate the data set AntdataPreprocessed.mat
- The data is synthetic, but it is created using real data sets as basis (from long-term collaboration with ant biologists).
- The data contains 358 sampling units (ant nests) each from which 5 individuals have been collected.
- Microsatellite data is available at 6 loci for the 1790 individuals.
- 20% of the alleles have been randomly set as missing to create a challenging situation.

- The underlying population structure contains 15 genetically separated subpopulations.
- The #sampling units present in the data per subpopulation varies between 5-50.
- The spatial configuration of a typical subpopulation is fairly smooth, but the shapes and sizes vary considerably.
- The data is rather extensive, so run the clustering only with the input values: 13 14 15 16 17.
- We do this in order to save time, but in a real analysis situation one would of course need to use a more extensive set of values.

- Save the result and investigate the Voronoi tessellation.
- The estimated population structure corresponds in this case exactly to the underlying structure.
- This shows the usefulness of combining both the sampling design and the geographical coordinate information when the molecular data are quite weak.
- Load the spatial clustering result to the program from File-Load result-Spatial mixture.

- From Graph-Voronoi tessellation one can redraw the posterior mode tessellation either with or without sampling unit names.
- From Graph-Local uncertainty one can obtain a graphical represention of the local probability mass concentration in the space of genetic mixtures.
- As the clustering assignment is typically expected to be fairly certain for a sampling unit, this graph shows low bars when the probabilities are near 1 and high bars when the assignment is uncertain.
- Otherwise, high bars would likely obscure the details in the graph.

- The exact definition of the local uncertainty graph is given in Corander et al (2008 Comp Stat).
- For the ant data we see that the posterior distribution is highly peaked in the vicinity of the true population structure, as nearly all tessellation cells are associated with an uncertainty value close to 0.
- An admixture analysis conditional on the genetic mixture (takes ~1 hour with 200 reference individuals simulated) yields a very small number of false positive cases (significant at 5% level).
- Bearing in mind the size of the data set (1790 individuals), this is a very satisfactory result, i.e. maintaining well the false positive rate below the nominal value.

Exercise 4: ’Spatial clustering of individuals’

- We now investigate another simulated spatially organized data set, where 480 individuals with known coordinates are sampled from a structured population (SpatialIndividualDataPreprocessed.mat).
- Microsatellite data over 10 highly polymorphic loci are available for these individuals.
- The underlying population structure contains 10 subpopulations varying in spatial shape, size and the number of sampled individuals available (10-100).
- As the data are again quite extensive and the spatial model is more complex computationally than the models with the uniform prior, we run the clustering with a reduced input to save time. Use single $K = 15$.
- Obviously, in a real situation a more extensive approach would be needed.

- When the estimation procedure is finished, save the spatial genetic mixture result.
- We can now examine the Voronoi tessellation.
- The estimate with 9 clusters reflects very accurately the underlying population structure (two subpopulations were merged and a small #individuals were wrongly assigned.
- Notice that the spatial model also tolerates cases where single individuals are present as migrants in a region corresponding to another subpopulation (see tessellation).
- The local uncertainty image reveals that there is now a bit more uncertainty about the assignments compared to the previous analysis, which is expected as the sampling units consist of single individuals.

Exercise 5: ’Trained clustering’

- We now investigate a scenario with 5 *a priori* known baseline populations from which there is a number of sampled and genotyped individuals available.
- Baseline sample sizes vary between 8-25 and 10 microsatellite loci are used for genotyping.
- The sample data with unidentified individuals contains 3 cases per baseline population.
- Choose ’Trained clustering’ and provide the baseline (prior) and sample datasets to the estimation (TrainedBaselineData.txt and TrainedSampleData.txt, respectively).
- An analysis where the unidentified individuals must be assigned to any of the baseline populations is performed by setting the upper bound *K* equal to the #baseline populations (here 5).

- Run the analysis and check the output.
- The partition image produced in trained clustering shows first the baseline populations in separate colors (from left to the right), whereafter the assignments of the unidentified sample data are shown in the order of the data file.
- Recall that the table titled ‘Changes in log(marginal likelihood) if group i is moved to cluster j’ tells about the local peakedness of the posterior distribution around the optimal partition.
- These values may also be converted to the posterior probabilities of assigning a particular individual to a particular baseline population.
- For each population $c = 1,...,k$, this is given by the formula $\exp(z_c)/\Sigma_c(\exp(z_c))$, where z_c is the value in the cth column of the ’Changes...’ table and $\Sigma_c(\exp(z_c))$ is the sum over the columns.

- The level of genetic separation among the baseline populations is here fairly high, and thus, all unidentified samples are associated with the correct origin in the analysis.
- However, assignment to the correct origin becomes more difficult when only small amounts of baseline data are available.
- This is illustrated by a subset of the earlier data set (stored in the file TrainedBaselineDataSmall.txt), which contains only 5 individuals per baseline population.
- Repeat the trained clustering using the files TrainedBaselineDataSmall.txt and TrainedSampleData.txt as inputs (use $K = 5$).

- It is seen that now some individuals are assigned to a wrong baseline population.
- In a situation like this, with only a small number of baseline observations available, auxiliary biological information may be very useful for the classification model.
- Corander et al. (2006 Fish Bull) suggested that, if some biological information is available, such that it pre-groups certain individuals in the unidentified sample together, then it can be used in the Bayesian model to strengthten the inferences.
- This idea is easily represented by the earlier discussed concept of a 'sampling unit', which may in this case, for instance, be a particular geographical location at a given point in time.
- The sampling unit is based on biological knowledge telling us that every individual in the sampling unit comes from the same (albeit unknown) baseline population.

- When utilized appropriately in Bayesian predictive classification framework, the sampling unit may decrease the false assignment rate considerably.
- Notice that the earlier trained clustering approach is a special case, where each sampling unit consists of a single individual only.
- To illustrate the usefulness of the sampling unit approach, we consider the previous sparse baseline data set, but now analyze it together with the unidentified sample data, where the 3 individuals sharing the same origin always make a sampling unit.
- Thus, the sample data consists of 5 sampling units, which are to be assigned to the baseline populations.
- Notice that this approach allows the sizes to the sampling units vary over any particular data set, while taking it coherently into account.
- Repeat the trained clustering using the files TrainedBaselineDataSmall.txt and TrainedSampleDataPregrouped.txt as inputs (use $K = 5$) and compare the results with the earlier ones.