Multivariate Statistics for Analysis of Honey Bee Propolis

Abdulaziz Alghamdi and Alison Gray
Department of Mathematics & Statistics, Strathclyde University
Abdulaziz.Alghamdi@strath.ac.uk & a.j.gry@strath.ac.uk

University of Strathclyde Glasgow

Abstract

This work investigates the use of different statistical methods for analysis of metabolomics data from analysis of propolis samples. Methods studied will include pre-processing methods and multivariate techniques such as principal component analysis (PCA), clustering and partial least squares (PLS) methods.

Background

Honey bees play a significant role ecologically and economically, through pollination of crops. Additionally, honey can be considered as one of the finest products of nature, with a wide range of beneficial uses, including use in cosmetic treatment, eye diseases, bronchial asthma and hiccups. Honey bees also produce beeswax, royal jelly and propolis.

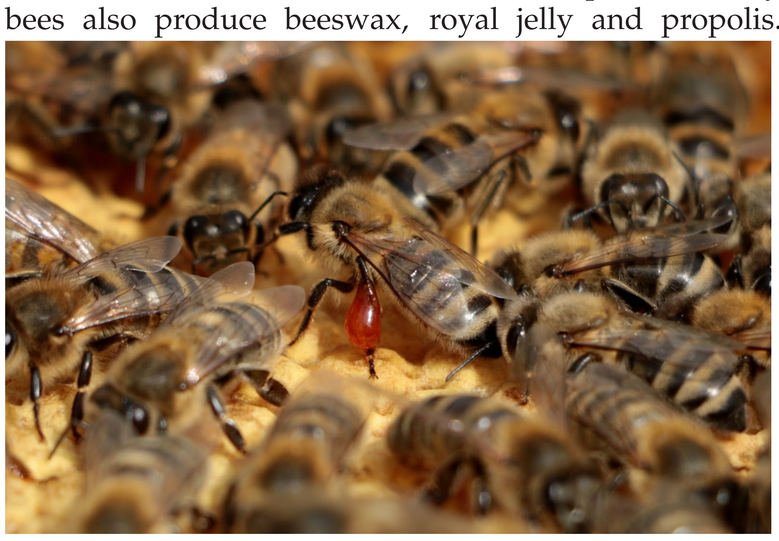

Figure 1: Honey Bees and Propolis

Propolis is a resinous bee product, which consists of a combination of beeswax and resins gathered by honey bees from exudates of various surrounding plants. It is used by the bees to seal and maintain the hives, but is also an anti-infective substance which may protect against disease. Propolis has a highly resinous, sticky gum appearance and its consistency changes depending on the temperature. It becomes elastic and sticky when warm, but hard and brittle when cold. Its colour varies from yellowish-green to dark brown, depending on its age and source.

Metabolomics data analysis

- GOALS
  - Discovery by identifying significant features associated with certain conditions.
  - Diagnosis via classification.
- Challenges
  - Limited sample size.
  - Many metabolites / variables.
- Workflow
  pre-treatment → multivariate analysis

The Data

Propolis was collected from hives on three different sites in Scotland, Aberdeenshire (n=15), Fort William (n=5) and Dunblane (n=3) with triplicate analysis of each sample. The rows are chromatographic peak for identified compounds and the column headings are labels of the hive samples.

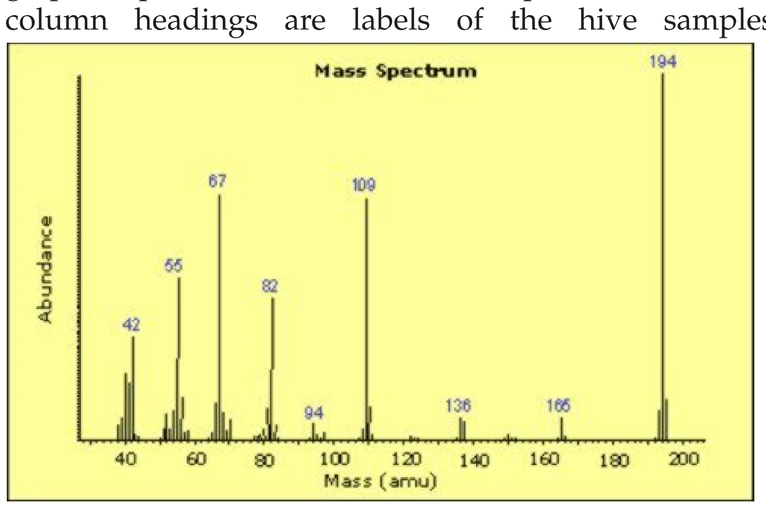

Figure 2: Example of a Mass Spectrum for a compound

RSS 2017 International Conference, Glasgow, 4-7 September 2017

Pre-treatment

Including transformation and scaling.

- GOALS
  - to reduce systematic variation.
  - to separate biological variation from variations introduced in the experimental process.
  - to improve the performance of downstream statistical analysis.
- Approaches
  - Sample normalization: to make samples comparable to each other.
  - Feature/variable normalization: to make features more comparable in magnitude to each other.

Transformation

Both reduce large values relatively more than the small values.

- Log transformation
  - pros: removal of heteroscedasticity.
  - cons: unable to deal with zeroes.
- Power transformation
  - pros: similar to log transformation.
  - cons: not able to make multiplicative effects additive.

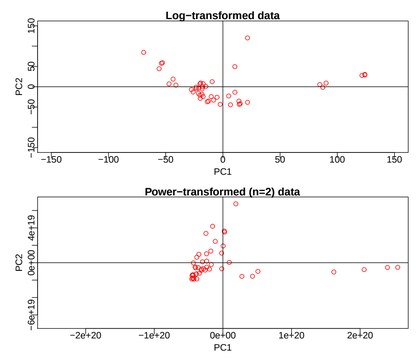

Figure 3: PC1 vs PC2 scores plots for the transformed data sets.

Scaling

- Auto: use the standard deviation as the scaling factor:

$$\tilde{x}_{ij} = \frac{x_{ij} - \bar{x}_i}{s_i}.$$

- Pareto: use the square root of the standard deviation as the scaling factor:

$$\tilde{x}_{ij} = \frac{x_{ij} - \bar{x}_i}{\sqrt{s_i}}.$$

- Range: use (max-min) as scaling factors:

$$\tilde{x}_{ij} = \frac{x_{ij} - \bar{x}_i}{x_{i_{\max}} - x_{i_{\min}}}.$$

- Vast: use standard deviation and the coefficient of variation as scaling factors:

$$\tilde{x}_{ij} = \frac{x_{ij} - \bar{x}_i}{s_i}.\frac{\bar{x}_i}{s_i}.$$

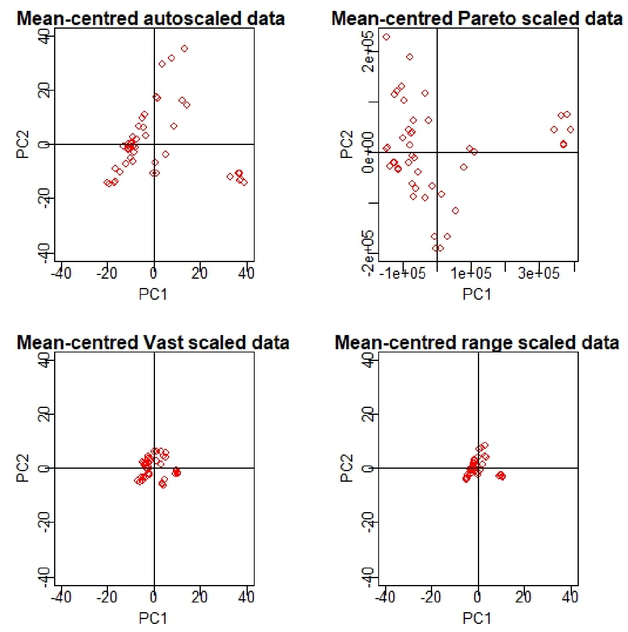

Figure 4: PC1 vs PC2 scores plots for the scaled data sets.

Dimension Reduction: PCA and PLS

- PCA is a statistical procedure to transform a set of correlated variables into a set of linearly uncorrelated variables the principal components.
- The uncorrelated variables are ordered such that the first accounts for as much variability in the data as possible and each succeeding one has the next highest variance.
- By discarding low-variance components, PCA helps us reduce data dimension and visualize the data.
- PLS is a supervised method to find a predictive model that describes the direction of maximum covariance between (X) and (Y) and similar to PCA the original variables are summarized into fewer new variables using weighted averages.

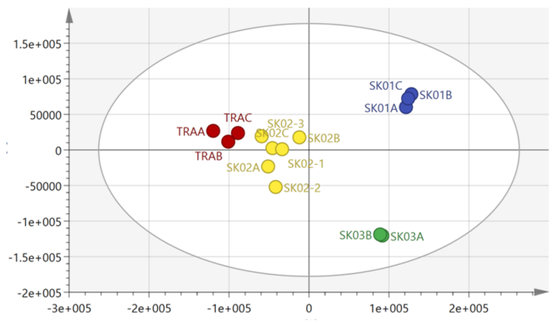

Figure 5: PCA for Fort William data set (Saleh et al., 2015).

Conclusion

- The above shows illustrative results.
- A comprehensive comparison of methods is underway to study best combinations of methods pre-treatment for clustering and classification.

References

[1] Kuropatnicki, A. K., Szliszka, E., and Krol, W. Historical aspects of propolis research in modern times. *Evidence-Based Complementary and Alternative Medicine 2013,2013, Article ID 904149, 11pp*.

[2] Marcucci, M. C. Propolis: chemical composition, biological properties and therapeutic activity. *Apidologie 26*, 2 (1995), 83–99.

[3] Saleh, K., et al. Profiling of scottish propolis from honey bee hives within the same location. *Pre-print* (2015).