

## Chapter 6

# Conclusions

### 6.1 Chapter summary

In chapter one, we defined the context of this work as being the overlap between intelligent systems, product design, and the food industry. The goal was stated as the development of intelligent systems appropriate for food design, and a comparison of such systems with more conventional statistical approaches.

In chapter two, we argued that feature selection and regression should not be separated. The linear and non-linear regression models were found to be very similar in performance, with the best feature selection being a stochastic search combined with cross model validation. We introduced novel semi-supervised approaches for feature selection and to build regression committees. We showed that selecting features from a large initial set can lead to overfitting and very poor generalization.

In chapter three, we described why clustering preference data was necessary, and compared two common algorithms:  $k$ -means clustering and Gaussian mixture models. We described two approaches to estimating the number of clusters present, namely information criteria — and the minimum description length principle, in particular — and cluster consistency. We argued that the former depends heavily on our pre-suppositions about the data, such as what shape Gaussian components we should use. The latter is more robust, in the sense of making fewer assumptions about the data, but may not be useful in cases where very few clusters exist. A non-linear preference map was introduced, and shown to perform at least as well as the more common linear preference map, which is based on (linear) PCA. A combination of clustering and regression was used to compare clustering algorithms, and solutions, with respect to their ultimate usefulness. A novel use of the consistency of clustering solutions was shown to be effective at estimating the number of clusters present.

In chapter four, we described several sources of outliers in preference data, and argued that these records should be removed before clustering is performed. We intro-

duced a sequential outlier rejection algorithm, and compared this with existing techniques. A further study applied sequential outlier rejection to regression, and showed a trade off between modelling a few consumers very accurately, and modelling many consumers less accurately.

In chapter five, we discussed several potential areas of work, including alternative approaches to intelligent food design, and alternative applications of the methods described in this thesis. We also proposed improvements to the data gathering process that would enable better use to be made of the expertise of the trained sensory panel.

## 6.2 Conclusions

Overall, we have argued that the very small data sets that typify food design studies create special problems for the analyst in several areas:

1. Selecting too many features, and so building too complex a regression function;
2. Selecting too many or too few, clusters, and so failing to capture a broad range of preferences; and
3. Identifying too many or too few outliers, by making false parametric assumptions.

With varying degrees of success, these can be overcome with, respectively:

1. Cross model validation, semi-supervised feature selection, or semi-supervised ensemble regression;
2. Information criteria, cluster consistency analysis or predictive accuracy; and
3. A stepwise non-parametric outlier detector (SOR).

Although the regression results are inconclusive, the fact that non-linear models often outperformed linear models reinforces our assumption that consumer preferences are non-linear. The lack of sufficient data may explain why linear models sometimes outperform the non-linear ones. The use of additional unlabelled data has been shown to be a useful enhancement to the small labelled sets, both for regression and for feature selection. At most 3–4 features were found to be useful in each regression model, and often as few as 1–2 features.

Cluster analysis suggests that preference data consists of overlapping clusters. This indicates that there is an underlying model of consumer preferences expressed in terms of latent variables, which is partially revealed through the preference data. Depending on the product in question, typically around 3–8 clusters were found.

Outlier detection within preference data does not yield definitive answers, but the results suggest that we should treat around 10-20% of the preference records as outliers. These panellists should be ignored in subsequent analysis.

Combining these data analysis techniques allows us to develop simple models that capture most consumers' tastes with reasonable accuracy across a range of products. Future data gathering should encourage the sensory panel to consider as many products as possible, which would improve semi-supervised modelling, and to select as few sensory features as possible, which would ease feature selection and increase the accuracy of the regression models.

### 6.3 Contributions

**Semi-supervised feature selection** A method to combine unlabelled and labelled data during feature selection, to prevent too many features being selected (Section 2.8.1).

**Semi-supervised ensemble learning** A method to combine unlabelled and labelled data to build regression ensembles, shown to be superior both to supervised ensemble methods and to semi-supervised non-ensemble methods, when very limited labelled data is available (Section 2.11).

**Consistency in clustering** A method which uses the consistency of solutions to estimate when the correct number of clusters has been identified (Section 3.8).

**Sequential outlier rejection** A method for identifying outliers in clustering solutions, which allows estimation of the number of outliers, as well as identifying the outliers themselves (Section 4.6).