A Model-Based Clustering Approach to Data Reduction for Actuarial Modelling

Dr Adrian O'Hagan and Mr Colm Ferrari, MSc.
School of Mathematical Sciences, University College Dublin

In association with Mr Craig Reynolds (Principal and Consulting Actuary) and Mr Avi Freedman (Principal Actuary) at Milliman, Seattle.

1 Introduction

In the recent past, actuarial modelling has migrated from deterministic approaches towards the use of stochastic scenarios. Such projections are useful to an insurer who wishes to examine the distribution of emerging earnings across a range of future economic and mortality scenarios. The use of nested stochastic processes dramatically increases the required run time for such models. Computational savings are possible using a compressed version of the original data in the stochastic model. This involves the synthesis of "model points": a relatively small number of policies that represent the data at large. Traditionally this has been achieved using variations on the distance-to-nearest-neighbour and k-means nonparametric clustering approaches. The aim of this research is to investigate how model-based clustering can be applied to actuarial data sets to produce high quality model points for stochastic projections.

2 Data

Milliman have provided a data set containing $110,000$ variable annuity policies, each with over 100 variables. As location variables Milliman compiled a set of revenue, expense and benefit present values for each annuity policy, across a range of 5 economic scenarios. The policy size variable is total account value in force.

3 Methods

The weighted distance to nearest-neighbour algorithm used by Milliman is:

1. Define the importance of each policy as its size multiplied by its Euclidean distance to nearest neighbour across its location variables.

2. Identify the least important policy and merge it with its nearest neighbour. The merged policy has size equal to the sum of the merging policy sizes and location variables equal to those of the larger of the merging policies.

3. Recalculate importance values for all policies and repeat the process until the desired number of policies remain.

4. Identify the policies mapped to each cluster and calculate their mean location. The original policy in each cluster nearest to this centre is scaled up for the size of all policies in the cluster as a representative 'model point'.

The nonparametric approach above can be amended to operate within a probabilistic framework. Rather than using weighted distance to nearest neighbour to iteratively merge cells and produce clusters; the clusters are instead identified using mixtures of multivariate Gaussian distributions. This process can be automated to incorporate the policy importance information using the *me.weighted* step within the **R** package **mclust**. The original policy closest to the theoretical mean of each cluster is again scaled up to reflect the size of all policies in the cluster and identified as a representative model point. This model-based clustering approach is initialised using a partial run of the distance to nearest neighbour algorithm to allow for observations with location variables originally valued at 0.

An advantage of the parametric model-based approach is that the resultant clustering has an associated likelihood value. This can be used to control for the presence of strong positive correlation among location variables shared across the 5 economic scenarios present. Rather than analyse the data collectively, the data corresponding to each scenario can be clustered separately and the final model points calculated using Bayesian model averaging across the scenario outcomes.

4 Results and Conclusions

To test the results, the model-based clustering approach is compared with the weighted nearest neighbours Milliman approach at various levels of compression, namely 50, 250, 1000, 2500 and 5000 model points. The model points are employed in a range of stochastic forecasts using Milliman's actuarial pricing model. The model-based clustering approach is demonstrated to provide strong forecast performance, comparable to or better than the Milliman weighted nearest neighbours approach, at all levels of data compression tested. The model-based clustering compressed data forecasts are additionally very close to those generated using the seriatim (full) data. Furthermore, the Bayesian model averaging approach to synthesising model points successfully overcomes the issue of positive correlation among location variables when economic scenarios are analysed collectively.