Churn Modelling: A Perspective on Mixture-Based Clustering of Customer Behaviours

Victor (Sheng) Wang

University of Oxford

Supervisors: Andrew Mellor, Junaid Mubeen

A technical report for

InFoMM CDT Mini-Project 1

in partnership with

Whizz Education

Trinity 2018

1 Introduction to the Problem

We have implemented a finite element method using FEniCS to solve a time-dependent Navier-Stokes equations

$$\frac{\partial}{\partial t}\mathbf{u} + \mathbf{u} \cdot \nabla \mathbf{u} - \nu \Delta \mathbf{u} + \nabla p = \mathbf{f}, \text{ in } \Omega \times (0, T], \tag{1}$$

$$\nabla \cdot \mathbf{u} = 0, \text{ in } \Omega \times (0, T], \tag{2}$$

$$\mathbf{u} = \mathbf{u}_D, \text{ on } \Gamma_D \times (0, T], \tag{3}$$

$$p = p_N, \text{ on } \Gamma_N \times (0, T], \tag{4}$$

which are a system of equations for the velocity $\mathbf{u}$ and pressure p in an incompressible fluid. $\mathbf{f}$ is a given force per unit volume and ν measures the viscosity of the fluid. The boundary of the spatial domain Ω has been split into Γ_D and Γ_N where different boundary conditions are imposed. Next we walk through the procedures of applying the finite element methods, implement them in FEniCS to a specific problem, and finally discuss the resulting numerics.

2 Mixture Model

Generative process

$$\mathbb{P}(\mathbf{y}, \mathbf{c}, \boldsymbol{\theta}) = \prod_{k=1}^{K} G_0(\theta_k) \prod_{n=1}^{N} F(y_n|\theta_{c_n}) P(c_n) \tag{5}$$

- $\mathbf{y} = \{y_1, y_2, \cdots, y_N\}$, observations
- $\mathbf{c} = \{c_1, c_2, \cdots, c_N\}$, cluster assignments
- $\boldsymbol{\theta} = \{\theta_1, \theta_2, \cdots, \theta_K\}$, cluster parameters
- $G_0(\theta)$, a prior over the cluster parameters
- $P(c)$, a prior over the mixing distribution
- $F(y_n|\theta_{c_n})$, a hypothetical distribution over the observations
- Assumptions: each observation is conditionally independent given its latent cluster assignments and the cluster parameters

Posterior probability of assignments

$$\mathbb{P}(\mathbf{c}|\mathbf{y}) = \frac{\mathbb{P}(\mathbf{y}|\mathbf{c}) P(\mathbf{c})}{\sum_{\mathbf{c}} \mathbb{P}(\mathbf{y}|\mathbf{c}) P(\mathbf{c})}, \tag{6}$$

where

$$\mathbb{P}(\mathbf{y}|\mathbf{c}) = \int_{\theta} \left[\prod_{n=1}^{N} F(y_n|\theta_{c_n}) \prod_{k=1}^{K} G_0(\theta_k) \right] \mathrm{d}\theta \tag{7}$$

BNP clustering assume that there is an infinite number of latent clusters (namely $K \to +\infty$), but that a finite number of them is used to generate the observed data. There are an infinite number of clusters, though a finite data set only exhibits a finite number of active clusters.

3 Methodology

4 Data Description and Pre-processing

We denote the feature data by a matrix $\mathbf{X} = (x_{ij}) \in \mathbb{R}^{m \times n}$, which describes n observed values for each of the m features. In addition, we denote the i-th row of $\mathbf{X}$ by $\mathbf{x}_{i,\mathrm{R}} = [x_{i1}\ x_{i2}\ \cdots\ x_{in}]$, and the j-th column by $\mathbf{x}_j = [x_{1j}\ x_{2j}\ \cdots\ x_{mj}]^\top$.

4.1 Data Transformation

In general, learning algorithms benefit from standardisation of the data set. We have employed 3 transformations in sequence to make our data more suitable for mixture model learning task. The transformations are performed for each feature separately, resulting in different sets of transformation parameters for different features. To be specific, for i-th feature we describe the transformations as following.

- Linear transformation
  The linear transformation is applied either to ensure all data to be positive for eligibility of applying the following power transformation, or to adapt to distributional modelling choice. It has the format:
  $$\mathbf{x}'_{i,\mathrm{R}} = a_i \mathbf{1} + b_i \mathbf{x}_{i,\mathrm{R}}, \tag{8}$$
  where a_i and b_i are constants and $\mathbf{1} \in \mathbb{R}^{1 \times n}$ is the row vector of all ones.

- Box-Cox power transformation
  The Box-Cox power transformation is used to modify the distributional shape of a set of data to be more normally distributed so that tests and confidence limits that require normality can be appropriately used. It has the format:
  $$x'_{ij} = \begin{cases} \frac{x_{ij}^{\lambda_i} - 1}{\lambda_i} & \text{if } \lambda_i \neq 0, \\ \ln x_{ij} & \text{if } \lambda_i = 0, \end{cases} \tag{9}$$
  for $j = 1,\ 2,\ \ldots, m$. In Box-Cox transformation, λ_i is estimated by maximizing the likelihood function [reference].

- Standardisation
  We standardise data for different features by scaling them into the same range to ensure the robustness to very small standard deviations of features. We choose to scale all features into range $[1, 100]$. If we denote the maximum and

minimum values of observed feature i as $x_i^{\max}$ and $x_i^{\min}$ respectively, then the standardisation is a linear transformation such that,

$$\mathbf{x}'_{i,\mathrm{R}} = \mathbf{1} + \frac{100-1}{x_i^{\max} - x_i^{\min}} \left(\mathbf{x}_{i,\mathrm{R}} - x_i^{\min} \mathbf{1} \right). \tag{10}$$

Table of transformation parameters

Exhibitions of pre and post transformation

5 Clustering Analysis

6 Markov State Transitional Analysis

7 Conclusion

We derived the general weak form solution for a time-dependent Navier-Stokes equation, and then implemented in FEniCS to seek for the numerical solution for a specific problem. Moreover, we considered a diffusion-reaction process that acts as external force term on the flow, by coupling the governing PDEs into one compound system. The numerical results for flow velocity, pressure and chemicals concentration are displayed in color map. We discussed the evolution of these profiles over time, and also observed how the reaction impacting factor and fluid viscosity influence the velocity profile. Either larger m or smaller ν results in a more irregular dynamics for the velocity in the long term.