Modified abstract of the dissertation titled "Model Based Clustering on High Dimensional Data" submitted for MSc degree in Biostatistics at the National Kapodistrian University of Athens, Medical School.

This thesis concerns the methodology of model-based clustering as an alternative to classical distance-based clustering techniques. In particular, we emphasize to model-based clustering on high-dimensional data using multivariate normal and student distributions. We provide extensive theory together with some examples and applications for the better understanding of the methods.

We first describe some theory around mixtures of multivariate normal distributions and how those are used for model-based clustering. We discuss in detail on the models of the Gaussian Parsimonious Clustering Models (GPCM) family and we commend on some controversial issues such as model selection techniques, proper number of groups, initial values etc. We also make an extensive reference to the use of the expectation-maximization algorithm which is used for parameter estimation.

Then, we present the case of clustering high-dimensional data. The use of models of the GPCM family for high-dimensional data is problematic and the use of factor-analyzers is proposed as an alternative. We present two families of models (i) the Parsimonious Gaussian Mixture Models family (PGMM) and (ii) the Expanded Parsimonious Gaussian Mixture Models family (EPGMM), which are suitable for clustering high-dimensional data based on mixtures of multivariate normal analyzers. We provide examples using models of those families and comment on their advantages and disadvantages. The Alternative Expectation Conditional Maximization (AECM) algorithm is described as well, for parameter estimation in that case.

Then, we explore the benefits of using mixtures of multivariate *t* distributions as an alternative to multivariate normal densities for small data sets and we extend that also to the case of high-dimensional data. We present the Mixtures of Multivariate *t* Factor Analyzers (MMtFA) family of models and we describe in detail the use of the AECM algorithm for parameter estimation.

At the end, we describe an application of the PGMM and MMtFA family models on high-dimensional data from the gene expression study of van 't Veer et al. (2002). The data concern the expression of 24.182 genes from 78 women suffered from breast cancer. We use all models of the PGMM and the UUC of the MMtFA family for clustering the patients using 100 random and 646 genes selected by a technique similar to EMMIX-GENE. R code was written to implement some of the models of the PGMM family and the UUC model of the MMtFA family.