Market Definitions in the Real-estate Agents Market
A Data-driven Approach Using Statistical Learning

Adam Lindhe & Johan Orrenius *

October 28, 2024

Abstract

This paper introduces a novel method to define geographic markets using machine learning. Using an unsupervised learning approach we cluster sales based on customers' location such that each cluster represents a market. The novelty of our method is that we leverage each observation's seller identity to capture market structures that are not distance based. We integrate the assumption that sellers focus on a few geographic markets into our Bayesian framework and implement the method empirically using a Gibbs sampler. Estimating the geographic markets for real estate agents in Stockholm, our algorithm does significantly better in correctly classifying sales than the baseline K-means algorithm, achieving a Dice score of 0.78 compared to 0.67. We find that the number of markets each agent works in is distributed more similarlyin our clasicication than in the baseline comparisson. Our method underestimates the market concentration, as measured by the Herfindahl-Hirschman Index (HHI), to a lesser extent than the baseline K-means. Finally, we investigate the correct number of clusters and find that, in our example, it corresponds to the established knowledge of the market's geographic structure.

*We thank Petter Berg for generously providing data access. We thank Richard Friberg and Salil Sharma for their insightful comments. Johan Orrenius thanks Jan Wallanders och Tom Hedelius stiftelse for financial support

1 Introduction

Economic activity takes place in a market, yet there is no consensus on the definition of a market. Numerous economists have made attempts; Cournot, Marshall, and, Walras tried to define the scope of markets, without agreeing on a universally accepted definition. This paper proposes a data-driven unsupervised learning method to define geographic markets. By utilizing only the customers' geographic locations and which seller they bought from, we reveal market structures that are not purely based on a distance metric.

The theoretical definition of a geographic market is a set of locations that are close substitutes. In practice, this is hard to define. In industrial organizations, however, we study how markets function with respect to market power and market shares. For these purposes, an empirically applicable definition of markets is useful, even if it is not an end in itself.

Conceptually, a market should consist of goods or services with high substitution elasticities. These high elasticities need to hold in two dimensions: the product space, where products need to be similar enough to be substitutable, and the geographic dimension, where the area defining the market needs to be such that consumers can viably substitute between sellers.

This paper will focus on defining geographic markets. We will use a technique from machine learning called Gaussian mixture models to group sales into different geographic areas, which allows us to define geographic clusters and treat them as markets. Furthermore, we estimate the number of markets within our chosen setting, the Swedish real estate agents.

Our paper's main contribution to the literature is on the definition of geographic markets. We introduce a novel and straightforward method that captures underlying market structures that have not been addressed before. Geographic markets have been defined historically by raw distance or administrative regions as pointed out in the survey of Elzinga & Howell (2018). The choice of market definition matters as it qualitatively changes the economic results, both in Elzinga & Howell (2018) and in the Genakos & Pagliero (2022). The UK Competition Market Authority case between Poundland and 99p CMA (2015) is a prominent example of when different definitions would give different recommendations. Our new clustering algorithm is a more sophisticated method that could remove some arbitrary elements that policymakers otherwise have to make.

Clustering as a way to define markets is not a novel concept. It has been used by Ellickson *et al.* (2020), Assad *et al.* (2024) Carranza *et al.* (2015) Zwanziger *et al.* (1990) and Lu (2017). Previous attempts to define markets have clustered on the seller's location. We instead cluster on the consumer's location, while incorporating information on the seller's location. Using observational data, our method captures other underlying factors that may limit the geographic scope of the market beyond pure distance or administrative regions. These structures have previously been hard to capture empirically but are emphasized in the new US DOJ merger guidelines.

We also contribute to the literature on competition in the real estate market

Lind & Kopsch (2014).

One problem with all clustering algorithms is to decide on the number of clusters to use. The number of clusters is often set exogenously. Different approaches to endogenize the number of clusters have been made, for example by Carranza *et al.* (2015). In our setting, we use the mathematically coherent Bayesian Information Criterion(BIC) measure to suggest the number of clusters. This allows us toadress the question of at which geographical level competition takes place.

One of the novelties of our paper is that it relies on a minimal number of variables in the data, most notably, we require no prices. This makes it an attractive way to define geographic markets where other methods might be hard to implement. Some examples are markets where consumers do not pay for the service, such as platforms or privately provided public goods that are free of charge to the consumer.

The new merger guidelines for the US DOJ speak to different dimensions of geographic markets as described in the quote:
Factors that may limit the geographic scope of the market includes transportation costs (relative to the price of the good), language, regulation, tariff and non-tariff trade barriers, custom and familiarity, reputation, and local service availability.

Our paper can identify the above-mentioned characteristics in ways that a distance-based algorithm cannot. For example, our method captures natural boundaries, such as rivers and mountains. In other settings, our method could find segregation and discrimination.

Like most observational methods, our method uses sales patterns to define geographic markets. We cannot observe a counterfactual, i.e., how the sales patterns would change if quality or prices change. Instead, we observe the equilibrium outcome which we use for our market definition. These limitations similarly apply to other clustering algorithms cited above.

One standard clustering model is the k-means clustering method MACQUEEN (1967), which is related to Gaussian mixture models. It is a hard clustering method whereas our method is a soft clustering method, using the probabilistic assignment. We benchmark against the k-means method, which we also apply by cluster on the consumer location. The k-means method has been used by Yang (2018) and Ellickson *et al.* (2020), but then on the product space or the producer location. On top of the geographic component, our unsupervised algorithm aligns clusters to follow the sales patterns of different sellers, as we expect sellers to compete in some but not all of the markets.

To demonstrate an application of our method, this paper defines geographic markets for the service of real estate agents for apartments in Sweden. It is a regulated market, where the fee charged by the real estate agent is minimal in comparison to the sales value of the home. The data used will be sales of apartments and the employment of real estate agent services. In 2015 the Swedish Competition Authority stopped a merger of two of the largest real estate agencies, indicating the relevance of competition measures for the real

estate agent market. Various strategies for improving competition in the market have been suggested by Lind & Kopsch (2014). We believe that the setting can be of interest both within and outside of academia.

We find that our algorithm classifies sales into markets significantly better than the baseline of the k-means method. Areas that are separated by natural boundaries and therefore in different markets are classified by k-means as belonging to the same market. Our method finds the natural boundaries without not distorting other correct classifications.

The choice of market definition has an impact on the market concentration, as measured by the Herfindahl–Hirschman index (HHI) in our data. The k-means clustering yields an HHI lower than both the validation set and our method.

We find that our BIC metric indicates a similar number of clusters as there are markets. However, this result is tentative and prone to uncertainty with regard to the specification and data. Following only the BIC metric gives unintuitive results, and should instead be viewed as a guideline on the number of clusters.

Following the introduction, Section 2 will briefly describe our method. A more technical explanation can be found in Appendix A. In Section 3 we will introduce the case study and its settings. Section 4 presents the results, followed by the conclusion in Section 5.

2 Our definition of a market

Clustering is about classifying data points. There are multiple methods available. See ((MACQUEEN, 1967) Hastie et al, other). One split available is that between deterministic and model-based methods. Deterministic methods are hard clustering methods, meaning that each observation is assigned to one cluster. Model-based methods are soft clustering methods, meaning that each observation is assigned a probability of belonging to each cluster.

Previously in market definition, the K-means clustering algorithm has been used (Ellickson *et al.* , 2020; Yang, 2018). It is a deterministic method that minimizes the within-cluster variance under a metric, often Euclidean distance(Athey & Imbens, 2019). Assuming we have clusters $g \in G$, the K-means algorithm classifies each datapoint x_i into a cluster g.

A model-based approach instead treats each cluster g as a probability distribution and assigns the datapoint x_i to the cluster $g \in G$ which it is most likely to belong to (Fraley & Raftery, 2002; Lavine & West, 1992; McLachlan *et al.* , 2019). One common family of model are the mixture models, specifically the Gaussian Mixture Model (McLachlan *et al.* , 2019).

2.1 Gausian Mixture Models

The advantages of a Gaussian mixture model are that it allows for out-of-sample predictions and that it allows for variance in the shape, orientation, and volume

(Bensmail *et al.* , 1997) of the clusters. In our economic setting, each market is a cluster (Ellickson *et al.* , 2020).

A mixture model is teh we define a density function for each g componete that is dependent on the unknown variables Φ, $f_g(x|\Phi)$ which will be a gausian disrubution $f_g(x|\Phi) = \mathcal{N}(\mu_g(\Phi), \Sigma_g(\phi))$ with center μ_g and a covariance matrix Σ_g. . The distribution of the mixture model is the distribution of the components times the wiughts $\pi_g(\Phi)$ giving us

$$f(x|\Phi) = \sum_{j}^{G} \pi_j(\Phi) f_j(x|\Phi)$$

Definition 1 *Each cluster $g \in G$ is characterized by a two-dimensional Gaussian distribution $\mathcal{N}(\boldsymbol{\mu_g}, \Sigma_g)$ with center $\mu_g \in \mathbb{R}^2$ and covariance matrix $\Sigma_g \in PD_2$.*

For our setting we define a point in $x \in \mathbb{R}^2$. It will be classified according to assumtion 1

The distributions are unobserved to the econometrician and are what we want to estimate them. Once we have estimated them we can classify any point in space into a cluster.

Assumption 1 *Given G clusters, an arbitrary x_g will be assigned to cluster g such that $g = \arg\max_{j \in G} p(x_g \in j)$. Each cluster j will be a market.*

Let us now introduce the data that induces the clusters. To estimate the clusters we will use the location of the consumer i and the identity of the sellers of to consumer i which we call y_i.We have N consumers and M sellers. The realization of the data will consist of pairs $\{(x_i, y_i)\}_{i=1}^{N}$ where $x_i \in \mathbb{R}^2$ is the geographical position of the individual and $y_i \in \{1, \ldots, M\}$ is the seller l who sold to i. In this specification, all transactions are treated as different individuals, but it is possible to allow an individual to buy from multiple sellers by allowing i to correspond to a transaction instead. For our data, we will assign them to different clusters. This assignment we call $\{z_i\}_{i=1}^{N}$, $z_i \in \{1, \ldots, G\}$.

Further, each seller l will have a distribution of the probability of being active in each cluster. This probability will allow us to have sellers who sell to more than one market. It is characterized by Θ with $\Theta \in \mathbb{R}^{G \times M}$ with θ_{gl} is seller ls probability to sell in market g. The probabilities are non-negative and sum to one. In our setting, Θ induces the differences in the shape, orientation, and volume of the clusters (Bensmail *et al.* , 1997) and distinguishes from a normally distributed mixture model that converges to the K-means algorithm (McLachlan *et al.* , 2019).

By incorporating the seller's sales pattern in the fitting of the clusters we incorporate the demand-side market measure with supply-side information. This is the novel part of our method compared to the earlier use of clustering for market definition. The theoretical motivation for sellers' concentrating on a few markets in defining markest can be found both in (Elzinga & Hogarty, 1973)

and DoJ merger guidelines (Commision & Trade, 2023). One motivation worth mentioning is search cost for now how on different markets and transportation costs.

2.2 Implementation

As in any Bayesian method, we define a density function for each cluster g and they make up the classification likelihood. As of now, the parameters of the model Φ and the data $\mathbb{X}$ with N observations and G clusters.

$$\mathcal{L}(\Phi, \mathbb{X}) = \prod_{i=1}^{N} \sum_{g=1}^{G} (p(X|\Phi)p(z_i|\Phi)) \tag{1}$$

In appendix **??** we present the full definition including the conjecture prior and the full likelihood.

The aim is to estimate the posterior distribution of the clusters. Since this is intractable we will use a Markov Chain Monter Carlo method to sample from the posterior distribution, as is standard in the literature (Metropolis *et al.* , 1953; Hastings, 1970; Bernado & Smith, 1994). We implement the sampling with a Gibbs sampler (Geman & Geman, 1984; Bernado & Smith, 1994) as standard in (Lavine & West, 1992).

We will use a Gibbs sampler to sample from the posterior distribution. The Gibbs sampler is a way to sample a realization of a Markov chain with $p(z, \mu, \sigma, \Theta|x, y)$ as its invariant distribution. To implement the Gibbs sampler we need to be able to sample from the distributions of $z, \mu, \sigma, \Theta|$, see Appendix **??** for the full derivation.

Once the Gibbs sampler has converged we will have an estimated posterior for the clusters g according to Defenition 1 and any location $x \in \mathbb{R}^2$ is assigned to a cluster according to Defenition 2.

More sections to be added

2.3 The number of clusters

The drawback of any clustering algorithm is that the number of clusters G is set exogenously. Different approaches to endogenize the number of clusters have been made. The most common one is the Bayesian information criteria (BIC) (Bernado & Smith, 1994; Raftery, 2016), based on the Bayes factor. It is derived from the likelihood of the distribution and the number of parameters. The BIC is defined as:

$$\text{BIC} = (5 + M) * G * \ln(N) - 2 * \ln(\mathcal{L}(z|x, y)), \tag{2}$$

where the rule is to select G such that the model estimated minimizes BIC. We iterate the algorithm over G, and benchmark models with a low BIC score as appropriate.

We use the Dice coefficient to evaluate our model. It spans between 0 and 1, where a higher value indicates a better classification. The dice coefficient is

calculated as the overlap between the validation clusters and the outcome of the clustering algorithm. It is a very simple metric but gives us a sense of how well our algorithm is performing. The validation set is created by knowledge of the correct markets in the restricted sample.

3 Case study, the Stockholm real estate agents market

3.1 Setting

We apply our method to the setting of the market for real estate agents in Stockholm. The focus of the data will be apartments in downtown Stockholm. The restriction is due to computational reasons, but an expansion to a larger area is possible.

In our setting, the real estate agent is the seller l and the flat they sell will be considered at the location of the individual i. Real estate agents only work for the seller of the apartment in Sweden, and the fees of selling a flat are low in an international comparison[1].

Stockholm's geography is special as it is a city built on islands and in downtown Stockholm there are four distinct districts on three islands. Two of the districts are their own islands, whereas the other two are on the same island. The four districts are historically administrative regions, although the two districts that are on the same Island were recently merged. The districts are clear identity markers in Stockholm's society.

When listing a real estate, the name of the district is the first thing displayed and buyers usually have clear preferences over districts. For this example, we will use the districts, Kungsholmen, Vasastan, Södermalm and Östermalm.

Real estate agents in Sweden most often work for big franchises. The franchise taker has its own office and coverage area. They have a non-compete agreement with other offices of the same franchise. They are therefore restricted to selling in a designated area.

3.2 Data

The sample used for the analysis is sales of apartments in the inner Stockholm area from 2017.[2] We use 3145 observations, in four districts in downtown Stockholm. We require the agents to sell at least 10 apartments during the period. This is only a subsample of available data, but here we know the market for real estate agents very well, and a good verification set can therefore be created.

Relevant data for the estimation are the geographical position of apartment i, ($x_i \in \mathbb{R}^2$) and the real-estate agent that sold apartment i, (y_i). These are the only data we will use to estimate our model. The final assignment z_i of

[1] 3.5% without tax, compared to 5% and 7.7% in Finland and the US respectively according to Lind & Kopsch (2014)

[2] We are grateful to Petter Berg for supplying the data

appartment i is the output of our method. We will then use the official listing area to validate our method. The validation will be benchmarked by comparing with using K-means clustering to define markets.

4 Results

4.1 Mechanisms of the method

The novelty of our method is the use of a second dimension of information on top of the geographic location to define markets. As markets are defined as Gaussian distribution over $\mathbb{R}^2$ this is equivalent to the movement of the center of each distribution μ_g as well as changes to the covariance matrix Σ_g. To illustrate this we look at our sample data. The validation set is plotted in Figure 1. We now induce 4 clusters using k-means as our baseline. The distributions of each cluster are plotted in Figure 2. On top of that, a sample of observations is plotted in the same figure. Here the different validation clusters are indicated by different shapes. The baseline clustering algorithm corresponds to evaluating the probability of belonging to each of the clusters. The cluster with the highest probability is the one that the observation is assigned to. As we can see there are some mich-match. A validation cluster will have observations that span more than one cluster. One important observation is that the size and shape of the probability distributions are not too different. Also note that the distributions have a large overlap, and visually we see that the blue circles in the middle seem to have a higher probability to be in the lower cluster than in the left cluster, like the rest of the circles. The classifications are shown in Figure 3. Here we see that the the island in the western part of Stockholm, Kungsholmen is split into two markets, such that the eastern part is linked with the southern island of Södermalm. This is an error as seen in the validation set in Figure 1.

We now employ the new algorithm. As said the algorithm takes the information on who buys from which real estate agent and updates the covariance matrix Σ_g of the cluster as well as the centers μ_g. The distributions are plotted in Figure **??**. We have now reshaped them to be more elliptical. As we iterate over the sample we update the distributions such that areas of overlap are minimized, and the cluster with the lower share of common sales, ie where the real estate agents in the overlap do not sell a lot in that cluster will give space to the cluster with a higher common share of sales. This is done by skewing the shape. This leads to a better fit. The classification is shown in Figure 5. Here we see that the island of Kungsholmen is now in one cluster. This is a better fit to the validation set.

Figure 1: Validation

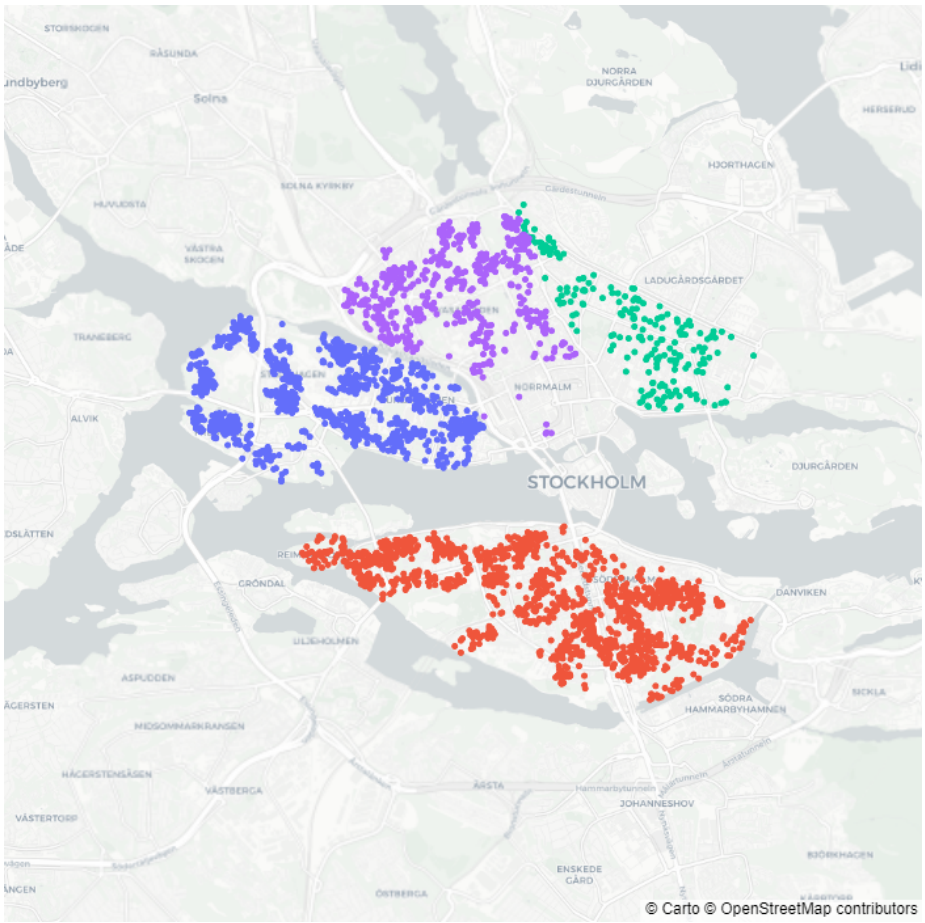


Figure 2: Distributions for baseline

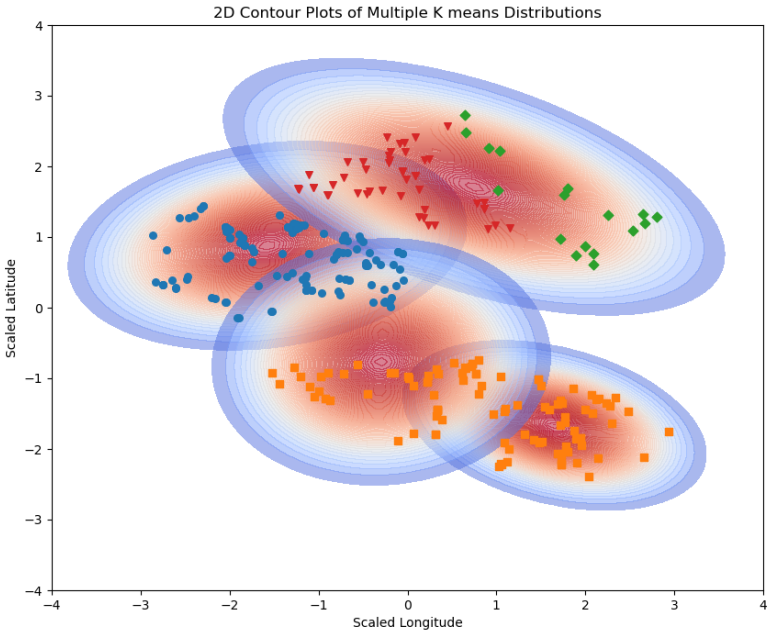


Figure 3: Classification of baseline

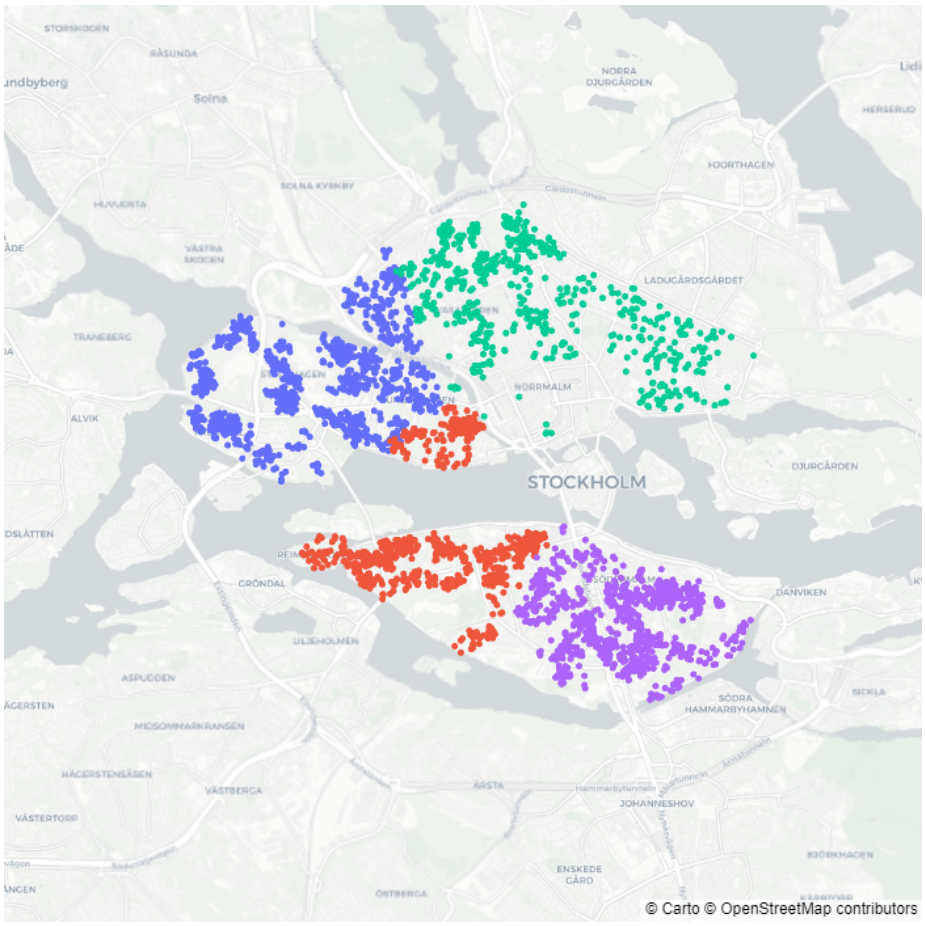


Figure 4: Distributions for Gaussian Mixture Model

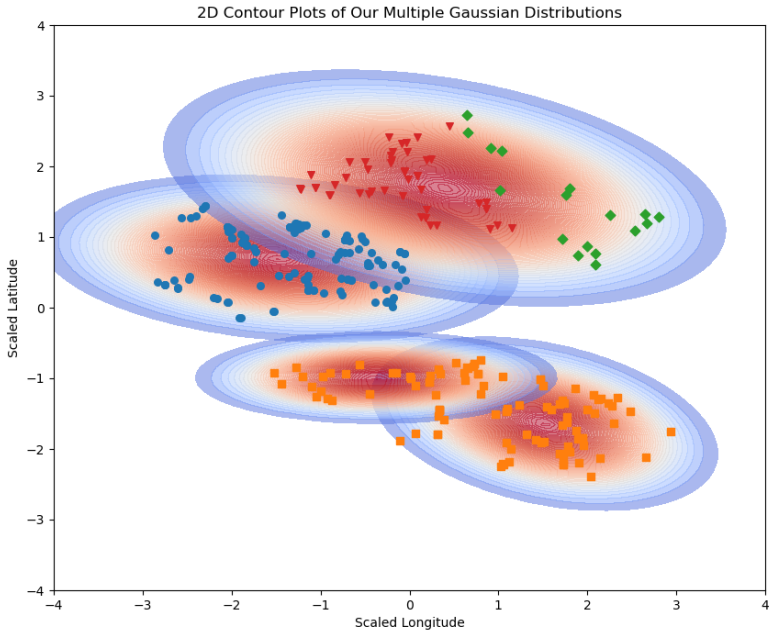
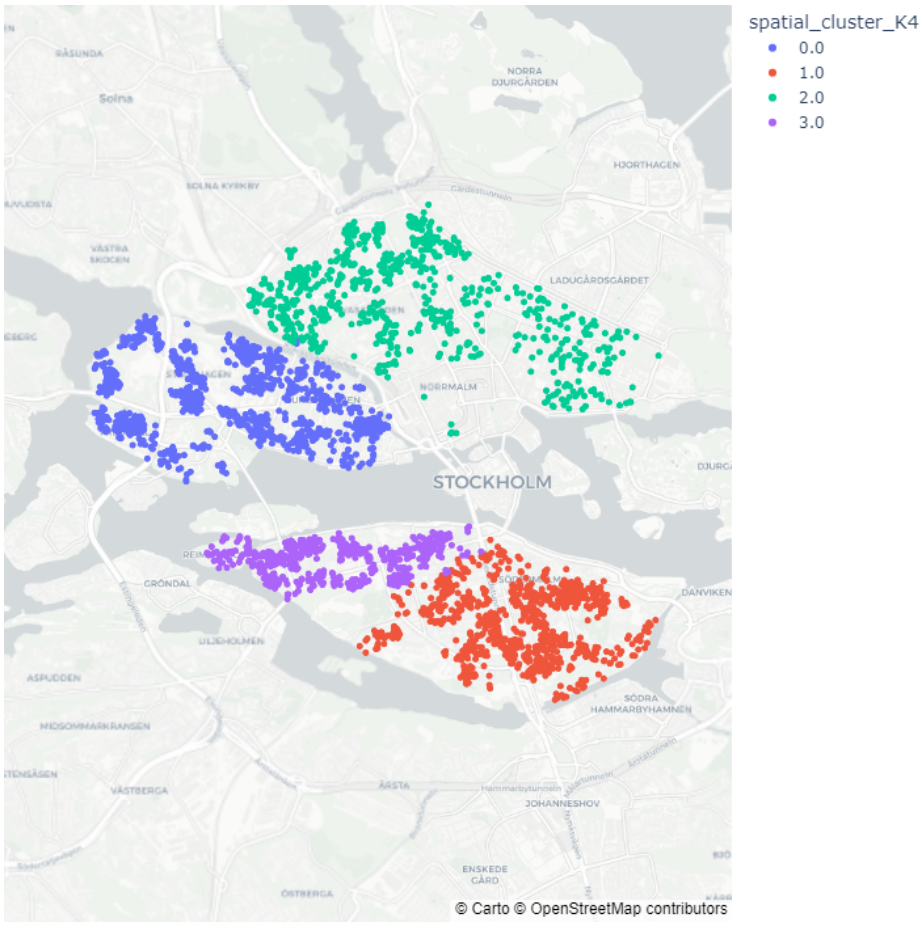
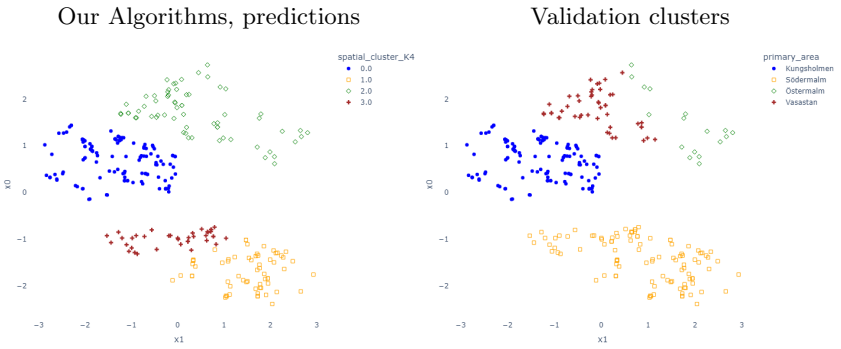


Figure 5: Classification for Gaussian Mixture Model



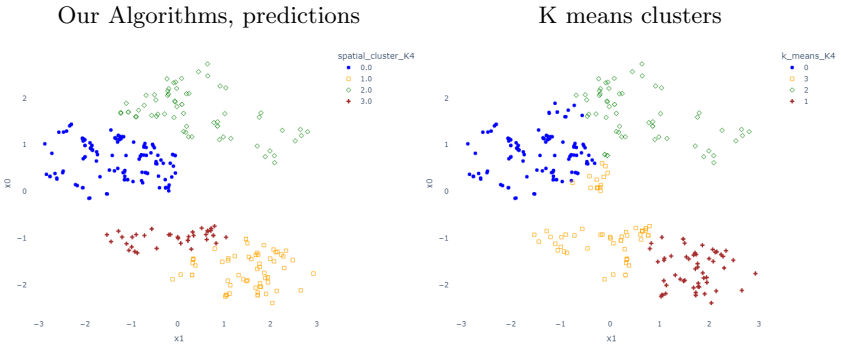
We first plot the clusters as scatter graphs. In Figure 6 we present the results of our method and compare it with the validation clusters. In the validation graph, we see that the southern cluster is on a separate island. Also, the western cluster is on an island, but as the width of the water is small, it is not as obvious. The outcome of our clustering algorithm generates almost the same boundaries as the validation clusters, with the difference that the top clusters are merged and the bottom one is split into two.

Figure 6: The clustering by our algorithm compared to the listing areas



In Figure 7 we plot the difference between our method and k means. In this scatterplot, they look very similar. There are differences but if one of them is better than the other is not indicated. We therefore place a background of a map to see the relationship to geography.

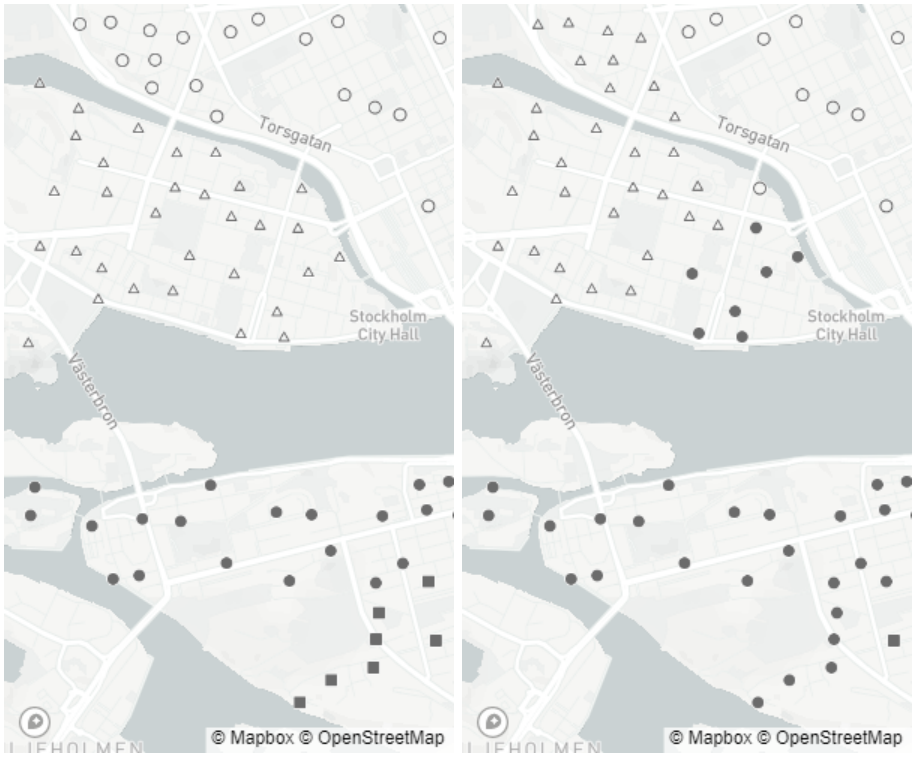
Figure 7: The clustering by our algorithm compared to k means



To understand the geographic component we zoom in on the areas with differences in Figure 8 and see that our algorithm clearly divides the clusters based on the water whereas the pure distance-based k means method thinks that part of the island is in three markets, counter-intuitive to practitioner knowledge.

Figure 8: The clustering by our algorithm compared to k means on a map

Our Algorithms, K means clusters



To check the validity we compute the dice scores of our method and k means with regard to the listing areas which we treat as the validation set. We see that both algorithms are good at finding the underlying market structure but that our score of 0.78 is substantially better than the K means methods' 0.67.

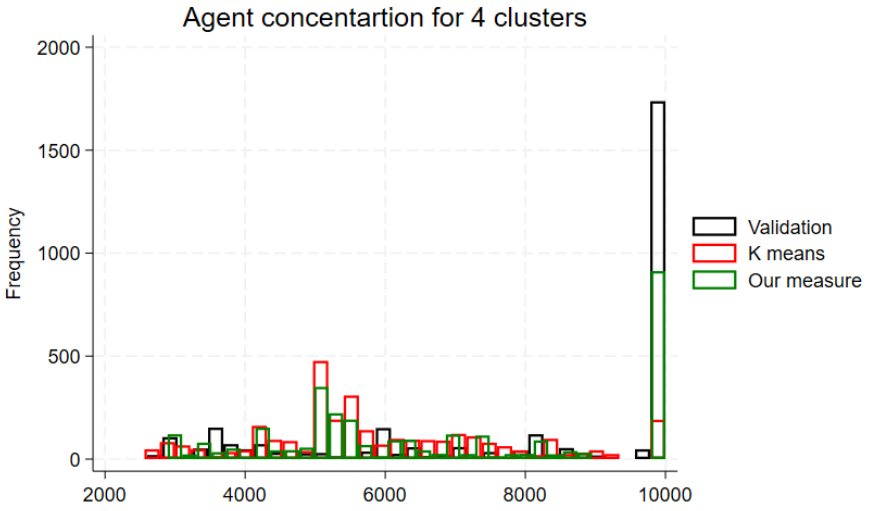
4.2 Economic impact

Our algorithm updates the clusters such that a single agent should be in as few as possible. To see the change we plot the number of markets that each agent is active in if Figure 9 for our three different market definitions. The distribution in clusters corresponds well with the validation set in how many markets each agent is active in, whereas k-means by misclassifying the borders around the island believe the agents work in more than one market.

Figure 9: The number of markets an agent is active in

In Figure 10 we see how concentrated each agent is. Our algorithm finds a higher concentration than k means. The concentration lies much closer to the validation cluster than the k means.

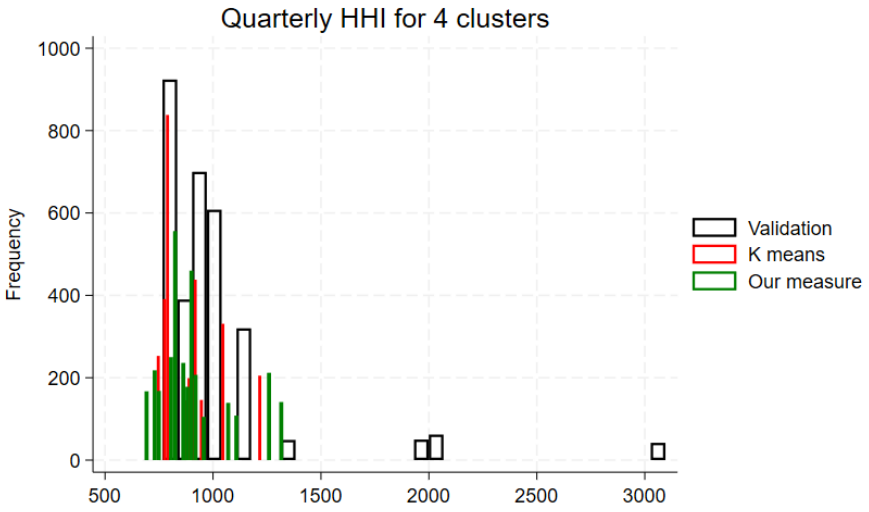
Figure 10: The concentration of each agent, ranging from 0 to 10000



To investigate the economic implications of the definitions we compute an HHI index for each market and quarter and plot the results in Figure 11. Using

the validation set as the market definition we get an average HHI of 999. For our market definition we get 904 and for k means it is 881. We see that our measure comes close to capturing the same market concentration as the validation set whereas k means indicates a lower market concentration. The difference is not large, and should not be extrapolated to other datasets. We further see that the Stockholm real estate agents are not that concentrated.

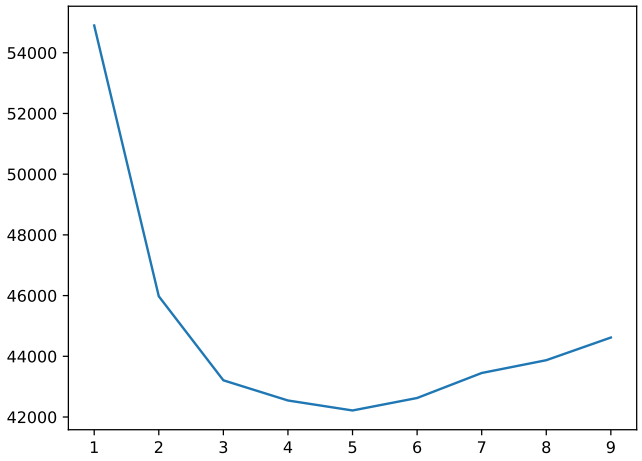
Figure 11: The distribution of HHI for the different markets on a quarterly level



The number of clusters

In Figure 12 we look at the BIC scores for different numbers of clusters. Our assumption based on institutional knowledge of 4 clusters is not the one with the lowest BIC. We therefore look at some of the outcomes for 6 clusters.

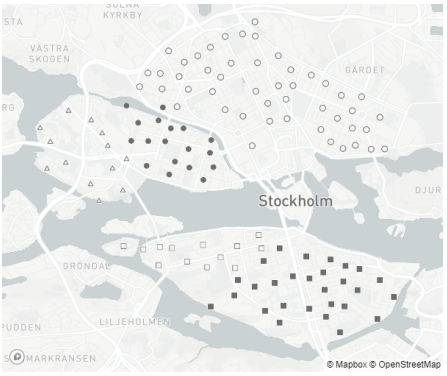
Figure 12: The BIC test indicates that 5 is the correct number of clusters



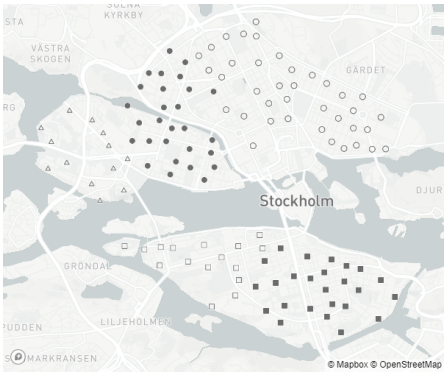
We here uncover a new pattern, where one more of the clusters arises divided into two. Our algorithm once again manages to find natural borders as compared to k-means. We further see that the new cluster is not the one that corresponds to the two validation clusters. Our algorithm here suggests that the relevant market definition would rather be splitting the two islands, as there seem to be many real estate agents who sell in both.

Figure 13: The clustering by our algorithm compared to k means, for 5 clusters

Our Algorithms, predictions



K means



5 Conclusion

In our paper, we propose a method to define geographic markets. It uses an unsupervised clustering algorithm with minimal data requirements to define markets. The sales data allows us to reveal market structures that traditional market definitions would miss. Having a method to suggest geographic market definitions in addition to existing methods would be useful, both for competition authorities and private companies. Our method is very adaptable to different settings and requires very little data, compared to other methods. After tests in different datasets, it could be a complement to the existing methods. In its infancy, the method should be tested in other settings and other industries that are not as special as the real estate agents market. Our framework can also discover discrimination and segregation.

References

Assad, Stephanie, Clark, Robert, Ershov, Daniel, & Xu, Lei. 2024. Algorithmic Pricing and Competition: Empirical Evidence from the German Retail Gasoline Market. *Journal ofPolitical Economy*, **132**(3), 723–771.

Athey, Susan, & Imbens, Guido W. 2019. Machine Learning Methods That Economists Should Know about. *Annual Review of Economics*, **11**(March), 685–725.

Bensmail, Halima, Raftery, Adrian E, & Robert, Christian P. 1997. Bensmail et al., 1997_ Inference in model-based cluster anaylsis. 1–10.

Bernado, J.M., & Smith, A.F.M. 1994. *Bayesian Theory*. Chichester: John Wiely and Sons.

Carranza, Juan Esteban, Clark, Robert, & Houde, Jean François. 2015. Price controls and market structure: Evidence from gasoline retail markets. *Journal of Industrial Economics*, **63**(1), 152–198.

CMA. 2015. *Poundland and 99p, A report on the anticipated acquisition by Poundland Group plc of 99p Stores Limited*. Tech. rept. September. Competition and Markets Authority, London.

Commision, U.S. department of Justice, & Trade, The Federam. 2023. *Merger guidelines*.

Ellickson, Paul B., Grieco, Paul L.E., & Khvastunov, Oleksii. 2020. Measuring competition in spatial retail. *RAND Journal of Economics*, **51**(1), 189–232.

Elzinga, Kenneth G., & Hogarty, Thomas F. 1973. The Problem of Geographic Market Delineation in Antimerger Suits. *The Antitrust Bulletin*, **18**(1), 45–81.

Elzinga, Kenneth G., & Howell, Vandy M. 2018. Geographic Market Definition in the Merger Guidelines: A Retrospective Analysis. *Review of Industrial Organization*, **53**(3), 453–475.

Fraley, Chris, & Raftery, Adrian E. 2002. Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association*, **97**(458), 611–631.

Geman, Stuart, & Geman, Donald. 1984. Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **PAMI-6**(6), 721–741.

Genakos, Christos, & Pagliero, Mario. 2022. Competition and Pass-Through: Evidence from Isolated Markets. *American Economic Journal: Applied Economics*, **14**(4), 35–57.

Hastings, Author W K. 1970. Monte Carlo Sampling Methods Using Markov Chains and Their Applications Published by : Biometrika Trust Stable URL : http://www.jstor.org/stable/2334940. *Biometrika*, **57**(1), 97–109.

Lavine, Michael, & West, Mike. 1992. A Bayesian method for ctassif ication and discrimination*. **20**(4).

Lind, Hans, & Kopsch, Fredrik. 2014. *Konkurrensen på fastighetsmäklarmarknaden*. Tech. rept. Konkurrensverket, Stockholm.

Lu, Anna W. 2017. *Three Essays on Empirical Industrial Organization in Grocery Retailing*. Ph.D. thesis, Heinrich-Heine-Universität.

MACQUEEN, J. 1967. SOME METHODS FOR CLASSIFICATION AND ANALYSIS OF MULTIVARIATE OBSERVATIONS. *Pages 281–297 of: Berkeley Symp. on Math. Statist. and Prob.*

McLachlan, Geoffrey J., Lee, Sharon X., & Rathnayake, Suren I. 2019. Finite mixture models. *Annual Review of Statistics and Its Application*, **6**(1988), 355–378.

Metropolis, Nicholas, Rosenbluth, Arianna W., Rosenbluth, Marshall N., Teller, Augusta H., & Teller, Edward. 1953. Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, **21**(6), 1087–1092.

Raftery, Adriatn E. 2016. Bayesian Model Selection in Social Research Author (s): Adrian E . Raftery Source : Sociological Methodology , Vol . 25 (1995), pp . 111-163 Published by : American Sociological Association Stable URL : http://www.jstor.org/stable/271063 Accessed : 29. **1995**(25), 111–163.

Yang, Yan. 2018. *A New Solution to Market Definition: An Approach Based on Multi-dimensional Substitutability Statistics*. Ph.D. thesis, Washington University.

Zwanziger, Jack, Melnick, Glenn A., & Mann, Joyce M. 1990. Measures of hospital market structure: a review of the alternatives and a proposed approach. *Socio-Economic Planning Sciences*, **24**(2), 81–95.

A Model

B Algorithm

The data consists of pairs $\{(x_i, y_i)\}_{i=1}^N$ where $x_i \in \mathbb{R}^2$ is the geographical position of the house, $y_i \in \{1, \ldots, M\}$ is the real estate agent that sold house i, observe that there are N houses and M different real estate agents. The unobserved variable is $\{z_i\}_{i=1}^N$, $z_i \in \{1, \ldots, K\}$. z_i is the market (or cluster) assignment for house i. Furthermore we have the following random variables

$$\begin{aligned}
&\{\mu_j\}_{j=1}^K,\ \mu_j \in \mathbb{R}^2,\\
&\{\Sigma_j\}_{j=1}^K,\ \Sigma_j \in \mathrm{PD}_2,\\
&\Theta \in \mathbb{R}^{K\times M},\ \text{s.t } \theta_{jl} \geq 0, \sum_{l=1}^M \theta_{jl} = 1.
\end{aligned}$$

μ_j is the center values matrix for cluster j, Σ_j is the covariance matrix for cluster j and θ_{jl} is the probability that a random house in market j is sold by agent l. Furthermore we have the non stochastic hyperpriors

$$\begin{aligned}
\mu_0, \lambda_0, \nu_0, \Psi_0, &\quad \text{hyperpriors to } \mu_j, \Sigma_j,\\
\alpha, &\quad \text{huperpriors to } \theta_j.
\end{aligned}$$

We impose the following

$$\begin{aligned}
&x_i \perp y_i | z_i,\\
&(x_i, y_i) \perp (x_j, y_j) | z_i, z_j, \Theta,\\
&p(z_i | y_i, \Theta) = \theta_{z_i, y_i},\\
&x_i | z_i, \mu, \Sigma \sim \mathrm{N}(\mu_{z_i}, \Sigma_{z_i}),\\
&z_i \sim \mathrm{U}(\{1, \ldots, K\}),\\
&\mu_j, \Sigma_j \sim \mathrm{NIW}(\mu_0, \lambda_0, \nu_0, \Psi_0),\\
&\theta_j \sim \mathrm{Dir}(\alpha).
\end{aligned}$$

The joint distribution can then be written as follows, shorthand notation $x = \{x_1, \ldots, x_N\}, y = \{y_1, \ldots, y_N\}$, $\mu = \{\mu_1, \ldots, \mu_K\}$ etc,

$$p(x, y, z, \mu, \Sigma, z, \Theta) = \prod_{i=1}^N p(x_i | z_i, \mu, \Sigma) p(y_i | z_i, \Theta) p(z_i) \prod_{j=1}^K p(\mu_j | \Sigma_j) p(\Sigma_j) p(\theta_j)$$

$$\prod_{i=1}^N \mathrm{N}(x_i | \mu_{z_i}, \Sigma_{z_i}) \theta_{z_i, y_i} \frac{1}{K} \prod_{j=1}^K \mathrm{N}(\mu_j | \mu_0, \frac{1}{\lambda_0}\Sigma_j) \mathcal{W}^{-1}(\Sigma_j | \nu_0, \Psi_0) \prod_{l=1}^M \frac{1}{B(\alpha)} \theta_{jl}^{\alpha_l - 1}.$$

The generative process to sample a data pair (x, y) is done by, sample z uniform among the clusters, sample θ_z from the Dirichlet distribution with parameters α, sample Σ_z from the Wisehart distribution with parameters ν_0, Ψ_0, sample μ_j from the multivariate normal with mean μ_0 and covariance matrix $\frac{1}{\lambda_0}\sigma_z$. sample y from a categorical distribution with probability vector θ_z and sample x from a normal distribution with mean μ_z and covariance Σ_z.

Now the goal is to find the posterior $p(z|x, y)$. Since this posterior is intractable we find a way to sample from the posterior instead. Here we present a Gibbs sampler, a way to sample a realisation of a Markov chain with $p(z, \mu, \sigma, \Theta|x, y)$ as its invariant distribution. To implemet the Gibbs sampler we need to be able to sample from the following distributions

$$
\begin{aligned}
&p(z|x, y, \mu, \Sigma, \Theta),\\
&p(\Sigma|x, y, z, \mu, \Theta),\\
&p(\mu|x, y, z, \Sigma, \Theta),\\
&p(\Theta|x, y, z, \mu, \Sigma).
\end{aligned}
$$

We start with $p(\Sigma|x, y, z, \mu, \Theta)$, this is a standard conjugate prior so

$$
\begin{aligned}
&\Sigma_j|x, y, z, \mu, \Theta = \Sigma_j|x, z \sim \mathcal{W}^{-1}(\nu'_j, \Psi'_j),\\
&n_j = \sum_{i=1}^{N} 1_{z_i=j},\\
&\bar{x}_j = \frac{1}{n_j}\sum_{i=1}^{N} x_j 1_{z_i=j},\\
&\nu'_j = \nu_0 + n_j,\\
&\Psi'_j = \Psi_0 + \sum_{i=1}^{N}(x_i - \bar{x}_j)(x_i - \bar{x}_j)^T 1_{z_i=j} + \frac{\lambda n_j}{\lambda_0 + n_j}(\bar{x}_j - \mu_0)(\bar{x}_j - \mu_0)^T.
\end{aligned}
$$

In the same way we get the posterior for μ

$$
\begin{aligned}
&\mu_j|x, y, \Sigma_j, \Theta = \mu_j|x, z\Sigma_j \sim \mathrm{N}(\mu'_j, \Sigma_j),\\
&\mu'_j = \frac{\lambda_0\mu_0 + n_j\bar{x}_j}{\lambda_0 + n_j}.
\end{aligned}
$$

The posterior for Θ is also standard conjugate prior calculations

$$
p(\theta_j|x, y, z, \mu, \Sigma) = p(\theta_j|y, z) \propto p(y|\theta_j, z) \propto \prod_{i=1}^{N} \theta_{j,y_i} 1_{z_i=j} \prod_{l=1}^{M} \theta_{jl}^{\alpha_l - 1}
$$

so we conclude that $\theta_j|y, z$ is $\mathrm{Dir}(\alpha'_j)$ distributed, where

$$
\alpha'_{jl} = \alpha_l + \sum_{i=1}^{n} 1_{y_i=l} 1_{z_i=j}.
$$

Finally we have the distribution for z.

$$p(z_i|x_i, y_i, \mu, \Sigma, \Theta) \propto p(x_i|z_i, \mu, \Sigma)p(y_i|z_i, \Theta) = \mathrm{N}(x_i|\mu_{z_j}, \Sigma_{z_j})\theta_{z_i y_i},$$

so the posterior conditional distribution for z_i is categorical dsitributed with probability

$$P(z_i = j) = \frac{\mathrm{N}(x_i|\mu_j, \Sigma_j)\theta_{jy_i}}{\sum_{j'=1}^{K} \mathrm{N}(x_i|\mu'_j, \Sigma'_j)\theta_{j'y_i}}.$$

Now the Gibbs sampler can be implemented as follows.

Input: $x, y, \nu_0, \mu_0, \lambda_0, \Psi_0, \alpha$
Result: Samples z_i^t from the posterior distribution;
Make an initial guess for z_i^0, K-Means for example;
for *t=1... T* **do**
  Sample $\Sigma^t \sim p(\Sigma|x, z^{t-1})$;
  Sample $\mu^t \sim p(\mu|x, z^{t-1}, \Sigma^t)$;
  Sample $\Theta^t \sim p(\Theta|y, z^{t-1})$;
  Sample $z^t \sim p(z|x, y, \mu^t, \Sigma^t, \Theta^t)$;
end
return $\{z^t\}_{t=burn\ in}^{T}$
Gibbs sampler for the posterior distribution of cluster assignment

The burn in period is used because unless the original guess is drawn from the invariant distribution there will be a period until the samples are approximately from the stationary distribution.