# A Partial Parameter HMM Based Clustering on Loan Repayment Data: Insights into Financial Behavior and Intent to Repay

Dibu John Philip[1], Nandan Sudarsanam[1,3], Balaraman Ravindran[2,3]
[1]Department of Management Studies
[2]Department of Computer Science and Engineering
[3]Robert Bosch Centre for Data Science and AI (RBC-DSAI)
Indian Institute of Technology Madras

## Abstract

*Financial institutions that provide loans are interested in understanding, as opposed to just predicting, the repayment behavior of its customers. This study applies a modified Hidden Markov Model (HMM) based clustering which clusters repayment sequences across selected subsets of the HMM parameters. We demonstrate that different implementations of this adaptation help us gain an in-depth understanding of various drivers that are hard to observe directly, but nevertheless govern repayment. These include drivers such as the ability to repay, or the intention to repay independent of the ability. Our results are compared to an alternate sequence clustering approach. The study concludes with the observation that the ability to cluster on selective parameters, in conjunction with the structural construct of HMMs, enables the discovery of substantially more meaningful business insights.*

## 1. Introduction

Modeling and understanding repayment behavior using hard-to-observe factors, such as customer financial health, intention to repay and product suitability, offer many real-world advantages. This goes beyond just predicting the repayment likelihood. These driving factors inferred through modeling enable financial institutions to gain business insights, at both individual as well as systemic levels, gauge financial behavior and ultimately improve profitability. For example, a behavior of delinquency can be due to various reasons ranging from poor financial health, willful defaulting, or just oversight. It is critical for the financial institution to understand which of these reasons is responsible for the particular behavior in order to make effective decisions or interventions. Such an exercise in modeling can also help in making inferences about the suitability of a product for a particular customer base, which in-turn can help in fine-tuning features of financial products and aid in cross-selling or designing of new products for a niche market.

One important challenge to such an effort is that the sources of data to model these driving factors of repayment are neither easily available nor authentic. Some of the possible sources of data are surveys and self-reported information. These sources are prone to chronic biases and could also be noisy [1]. The most reliable source of data is repayment (transaction) data which is readily available with the organization. However, much of the past efforts have tried to use this data to predict defaults or other repayment behavior (in a discriminative fashion) [2, 3] rather than link it to financial or behavioral states (which have been established to play an important role in repayment capability [4]). As discussed earlier, this limits the scope for actionable insights that the organization can make. Therefore, it is important to look at a modeling approach that can utilize the temporally sequential repayment data and link it to various latent drivers in a generative framework. These requirements motivate us to look at the use of Hidden Markov Models (HMM) to model repayment behavior, and specifically at an HMM based clustering to segment different demographic and behavioral patterns. HMM is a statistical Markov model where the modeled system is assumed to be a Markov process with hidden states. Here the observations or emissions of the HMMs are repayments, and the hidden states indicate the latent drivers.

However, a direct application of HMM based clustering on the data has some limitations. There are different drivers that affect repayment, which might interact with each other. Also, these drivers might affect different parameters in the HMM model. For instance, in some groups of customers, the link between financial health and repayment might be tenuous. This is because

URI: http://hdl.handle.net/10125/50057
ISBN: 978-0-9981331-1-9

some customers may sacrifice many of their desires to repay their loans and remain credit-worthy [5], or there may exist a group of customers, who, out of principle, would never default under any circumstance [6]. This implies that this group (cluster) of customers could have significantly different behaviour (emissions) under the same conditions (states). In essence, our preliminary work with the data and the practitioner indicate that the intention to repay is better captured through emission related parameters, whereas the predilection to good or bad financial health is captured in the transition parameters. This leads us to look at the proposed partial parameter HMM clustering (PP-HMM) to augment a traditional HMM clustering. We call this partial parameter because we use only a subset of the actual number of parameters available in the original HMM setting. In a traditional HMM based clustering, a separate HMM is built for each cluster of repayment sequences. In other words, all repayments in that sequence will share a common set of emission and transition probabilities. In this study, we propose two implementations of the PP-HMM clustering. In the first, transition probabilities are learned from the entire repayment sequences and the clustering is done only on emission probabilities. This reflects the different behavioral responses in repayment despite identical financial states. In the second, the emission probabilities are learned from the entire set of sequences. The clustering is done only on transition probabilities, which reflects the scenario where repayment response to hardship is modeled globally but the clustering reflects the predilection of different groups to go into hardship and recover from it.

We achieve the partial parametrization in the HMM clustering through two separate calls of the expectation-maximization (EM) algorithm. In the first call, all the repayment sequences (with no clustering structure) are used to learn both emission and transition probabilities. We call this the collective learning phase. In the second call, we fix a subset of parameters (the emission or transition). The remaining parameters are used to form the clusters and simultaneously learn unique values for each cluster in an iterative fashion. Similar to a traditional HMM clustering, the cluster formation and parameter estimation is done in order to maximize the sum of the objective functions of the EM algorithm across all clusters.

This study takes a real-world dataset from a private-sector, publicly listed, retail bank in India that provides personal and business loans. The analysis we conduct compares the clustering formed from the use of PP-HMMs with three baselines: (a) clustering on customer-related and external features only, which are then used to learn a separate HMM for each cluster, (b) results from using only a traditional HMM based clustering and (c) clustering from dynamic time warping, a well established approach in clustering sequential data. In all the comparisons, we look at the potential insights that these approaches provide on the drivers for repayment behavior. These results and the insights are discussed thoroughly in Section 5.

The rest of the paper is organized as follows. Section 2 reviews the work that has been done in this area previously and motivates this research. Section 3 gives a description of the data set. The proposed framework is illustrated in Section 4 and Section 5 includes the results and discussions. Finally, Section 6 concludes the paper.

## 2. Related work

In the recent years, the impact of data analytic techniques on driving business intelligence has been immense [7]. The banking and financial sector too has reaped immense benefits from adopting data driven approaches [8]. Our work is inspired by two broad areas of literature. The first relates to the broad range of techniques used in analyzing repayment data and using it for further insights or applications. The second pertains to clustering time series data.

Repayment data has been largely used in the *prediction* of defaults [2]. Attempts have been made to model loan repayment behavior [9] and to understand factors that affect loan repayment [10]. Studies have also used repayment data to look into the problem of quantifying and predicting prepayments [11]. A drift from the usual default vs. no default is seen in the work of Banasik et al. [12] where the focus is on when default will happen and not on if a borrower will default or not. However, in our work, we are interested more in understanding rather than predicting repayment behavior.

There are many factors, both observable and hidden, that affect the repayment behavior of customers. As explained in the previous section, sacrifices made to repay the loan [5] so as to get better loans next time, repayment by principle [6] and factors leading to over-indebtedness [4] all play an important role in the loan repayment behavior. In addition, the suitability of products on offer also plays a significant role in the repayment capability of the customer [13], and may also result in cross-buying from different vendors, which in-turn could be witnessed as an aggressive prepayment of the loan.

These suggest that there are lot of behavioral dynamics and latent variables behind the decision the customer makes to repay or default, all of which

are hidden from the financial institution. Latent variables were introduced into the problem of analyzing repayments in a study [14] that analyzed repayment performance in group-based credit programs using the Tobit model which describes the relationship between a latent variable and an independent variable. Mofatt in his work [6] used the double-hurdle model for limited dependent variables [15] to study the *extent* of loan default, rather than the probability of default.

Hidden Markov models, as detailed in the introduction, have many advantages in being used to model repayment behavior and product suitability. The use of HMMs in finance [16] has been vastly studied in areas related to default analyses [17], analysis of credit quality [18, 19] and expected credit loss [20]. One of the major uses of HMMs is to detect regime switches [21], which might in this case, be the good financial states and bad financial states.

An HMM built for each customer probably over-fits while a single HMM for the entire data could be a possible under-fit. Thus, it is essential to look into suitable clustering techniques. However, clustering of time series data poses its own unique problems ranging from selecting a suitable distance measure to scalability [22]. Model-based clustering techniques convert a time series into model parameters, and a suitable model distance (generally log-likelihood) and a clustering algorithm is applied to the extracted model parameters. HMM based clustering [23] is a typical example of model based clustering technique. It has been mostly used in biological applications, speaker and motion recognition applications. However, literature related to the use of HMM based clustering in finance, specifically in banking related applications is sparse to the best of our knowledge. While there are some notable exceptions [24, 25], none of these studies uses repayment data to analyze customer behavior and intentions. There are also excellent studies that focus on improving the existing HMM based clustering framework, such as using Dynamic Time Warping to bootstrap the process of fitting HMMs [26] and introducing the Bayesian approach to HMM clustering [27]. A relatively similar work to ours is seen in J. G. Dias et al. [28] where the authors derive insights about groups of stock markets, and how the regime switching dynamics differ across these groups, using an extended multilevel HMM. Our work further looks into how different groups of customers behave under the same global regime switching characteristics, as well as how they switch between regimes under the same global behavior patterns.

The focus of almost all the studies listed here is mainly on discriminating between good vs. bad loans or customers. However, our focus is on understanding the customer behavior and intentions as well, and their transition between different states. We believe this will provide richer and actionable business insights.

## 3. Dataset and preprocessing

The study uses a data set from a small to mid sized retail bank in India (approximately 600 branches and 5 billion USD in total assets) which caters to a wide range of customers and offers a variety of financial products. The data includes complete loan schedule of over one hundred thousand loan accounts that have been opened from the year 2010, along with the actual monthly repayments. The data set consists of only monthly repayment loans, with an average tenure of around 20 months. The data upto and including November 2016 was used for this analysis. This covers six full years of data. No personal information was shared with us for this work.

In addition to the repayment information, the data set also contains various demographic features of the borrower, and other extraneous loan related information, some of which is self-reported. This includes information on internal credit rating of the borrower (for a subset of the loans), purpose of the loan, approved loan amount, branch type (urban/rural), as well as dates of all transactions (which could be meaningful in mining seasonal trends). Around 11% of this information comes from customers from metropolitan areas, while the urban and semi-urban customers together constitute around 75% and 14% comes from rural customers. Around 19% of the approved loans have a principal value between 5000 to 50000 Indian rupees, around 54% have a principal value between 50000 and a million Indian rupees and roughly 27% have a principal value larger than a million Indian rupees.

For every account, a new feature which we call *observation* is generated for each month, where *observation* for the $i^{\text{th}}$ month is defined as

$$\text{observation}_i = \left( \frac{\text{Actual amount paid in month}_i}{\text{Expected amount to be paid in month}_i} - 1 \right) \times 100, \tag{1}$$

where the expected amount to be paid in month$_i$ for a principal amount of $P$ at an interest rate of $r$ % per month (for $\mathcal{N}$ months) is calculated using the standard EMI format $\left( \frac{P \times \frac{r}{100} \times (1+\frac{r}{100})^{\mathcal{N}}}{(1+\frac{r}{100})^{\mathcal{N}}-1} \right)$. That is, the percentage of

money paid in excess or in short of what was actually supposed to be paid for each month, where a positive value indicates prepayment. Further, these observations are binned appropriately to form a new feature called *observation class*.

## 4. Proposed framework

There are three aspects to the proposed methodology, and to HMMs in general that make them suited for modeling repayments. These are (i) the sequential framework (temporally), (ii) the ability to model latent variables, and (iii) the generative construct which enables more meaningful insights between the variables involved. In this section we first present the HMM construct and how it is used to model repayments. We then introduce the HMM based clustering, and provide a detailed description of the proposed methodology.

### 4.1. Prelude: HMMs and repayments

HMMs are used to model systems with unobserved or hidden states that obey Markovian property. The system emits certain symbols (emissions) which are the only outputs (or observations) visible to the outer world. Every HMM with $N$ number of states and $M$ distinct symbols has the following three components: (a) a $N$x$N$ transition probability matrix $\mathbf{A}$ where $A_{ij}$ is the probability of transitioning from state $i$ to state $j$; (b) a $M$x$N$ emission probability matrix $\mathbf{B}$, where $B_{ij}$ is the probability of emitting symbol $i$ in state $j$; (c) A 1x$N$ initial probability vector $\pi$, where $\pi_i$ is the probability of the system starting in state $i$. Thus, an HMM is defined using $\lambda = \{\mathbf{A}, \mathbf{B}, \pi\}$.

Consider this simple example. A banker is interested in assessing the true financial states ({"Good", "Medium","Bad"}) her customers go through during the tenure of their loans. It has to be noted that these states are not easily quantifiable and these are simply tags given to three different regimes or distributions on which the observations or repayments are dependent. The only data that is available shows the delays and advancements in the monthly repayment installments (X = $x_1x_2x_3 \ldots x_T$; where $T$ is the length of the observation sequence). Without losing generality, let us assume that $x_t$ can take only one of 3 possible values (or symbols) for each $t$. Here, $N$ = 3 and $M$ = 3. The structure of HMMs can answer 3 major questions [29]. They are:

1. Given an observation sequence X = $x_1x_2 \ldots x_T$ and an HMM model $\lambda$, how can P(X | $\lambda$) be computed efficiently?

2. Given an observation sequence X = $x_1x_2 \ldots x_T$ and an HMM model $\lambda$, how to select the *optimal* state sequence $Z = z_1z_2 \ldots z_T$?

3. How to adjust the HMM parameters $\lambda$ to maximize P(X | $\lambda$)?

The answers to the above questions are explained in great detail in the seminal paper by Rabiner [29].

### 4.2. Clustering of repayment sequences

Building an HMM for each customer gives the best solutions but possibly over fits. Also, there arises the problem of maintaining and updating a large number of HMMs. On the other hand, learning a single HMM for the entire data might lead to missing out interesting patterns or sequences in the interest of generalization. This motivates the use of clustering the sequences into a number of smaller groups.

Clustering time series data poses numerous challenges, the major one being the lack of a natural distance function between time series. Measures such as Euclidean distance over-emphasizes non-critical variances of the signal, in particular, a delay or a premature cut-off [24]. MLE based clustering with HMMs is very similar to the $k$-means clustering algorithm [30], wherein cluster centres are represented by HMMs and the distance measure becomes the likelihood of the observation belonging to a particular HMM. We direct the reader to P. Smyth et al. and B. Knab et al. [23, 24] for an in-depth algorithmic description of HMM based clustering.

We now introduce the proposed PP-HMM based clustering, which is an adaptation of the traditional HMM based clustering. To initialize the process, an HMM is learned for the entire set of repayment data. We call this the collective learning phase. We then group the customers into $K$ groups based on available features such as the type of branch they belong to and the loan amount applied for. A separate HMM is now learned for the repayment data from each of these groups. The parameters of the HMM learned during the collective learning phase are used to initialize the learning of these $K$ HMMs. We call this the feature based clustering phase. Next we use HMM based clustering to cluster the repayment data, using the parameters from the feature based clustering phase as initial setting. The PP-HMM based clustering then works as follows. A subset of the parameters are set to their values obtained from the collective learning phase. The remaining parameters are initialized with the values obtained from the feature based clustering phase. These are the only parameters that participate in the creation of clusters (of repayments), while iteratively updating

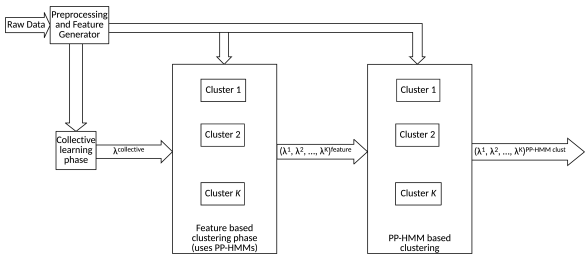



**Figure 1. Proposed framework for PP-HMM based clustering.**

their values based on the EM algorithm as applied to their respective cluster. This entire proposed framework is shown graphically in Figure 1 and discussed through the algorithmic sequence below:

1. Collective learning: Learn an HMM ($\lambda^{\text{Collective}}$) for the entire repayment sequence data of all the customers.

2. Feature based clustering:

   (a) Cluster the customers into $K$ clusters based on available demographic data.

   (b) Based on the choice of constraints, learn $K$ HMMs ($\lambda_{\mathscr{C}_k}$, for $k = 1,2,\ldots,K$) from these $K$ clusters. [Note: $\lambda_{\mathscr{C}_k} \subset \lambda^{\text{Collective}}$ where $\lambda_{\mathscr{C}_k}$ consists of only those HMM parameters whose values are not constrained to the values of parameters in $\lambda^{\text{Collective}}$.]

3. PP-HMM based clustering:

   (a) For each of the $K$ HMMs, compute the log-likelihood of each of the repayment sequences given the model. That is, compute $P(X_l|\lambda_{\mathscr{C}_k})$ for $l = 1,2,\ldots,L$ and $k = 1,2,\ldots,K$, where $L$ is the total number of repayment sequences and $K$ is the number of clusters.

   (b) Next, use the log-likelihood distances to cluster the $L$ sequences into $K$ groups so as to maximize the objective function

   $f(\mathscr{G}) = \prod_{k=1}^{K} \prod_{X_i \in \mathscr{G}_k} L(X_i|\lambda_{\mathscr{C}_k})$, where

   $\mathscr{G} = (\mathscr{G}_1, \mathscr{G}_2, \ldots, \mathscr{G}_K)$ is a partition of the $L$ repayment sequences $\mathbf{X}$.

   (c) Compute new values for the parameters of the $K$ HMMs, $\lambda_{\mathscr{C}_k}^{\tau}$ using the data from the $K$ different clusters and using $\lambda_{\mathscr{C}_k}^{\tau-1}$ as the initial setting, and re-compute the log-likelihood distances, where $\tau$ is the iteration number.

   (d) Repeat steps 3(b) and 3(c) and stop when the improvement in the objective function is below a threshold value or after a pre-determined number of iterations.

We now illustrate two specific implementations of the PP-HMM based clustering which are used in this study.

**4.2.1. Clustering across emission probabilities while keeping the transition probabilities fixed.** In this setting, the transition probabilities are fixed to the values learned during the collective learning phase while the emission probabilities are updated during each iteration. This helps in inferring how different groups of customers respond (in terms of repayment) to global/similar financial environments and seeks to capture the intent to pay independent of state. This could potentially be helpful in understanding the differences in behavior (willful defaults, delinquency due to oversight, repayment even during extreme financial hardship, etc.) under similar conditions.

**4.2.2. Clustering across transition probabilities while keeping the emission probabilities fixed.** Here, the transition probabilities alone are updated during the feature based clustering phase as well as during each iteration of the HMM based clustering. The emission probabilities are fixed to the values learned during the collective learning phase. This enables us to infer the predilection of different groups of customers towards certain financial states, or the nature of their transition between states.

When the results from these clusters are seen along with the traditional HMM based clustering, the potential insights could be more meaningful. We demonstrate this in the next section.

## 5. Results and discussion

This section is organized as follows. We first present the results of the collective learning phase, that is, a single HMM for the entire repayment data. This gives us an idea of the global repayment behavior of the customer base. We then present the results of the three baseline clustering approaches. Next we proceed to illustrate the results of PP-HMM based clustering and show how this approach gives a clearer interpretation of the repayment data and aid in planning suitable interventions.

In all the experiments to follow that include the use of HMMs, the number of states $N$ and the number of unique observations $M$ have been fixed to 3, and the number of clusters in clustering exercises has been fixed

to 2 unless specified otherwise. The $observation_i$ for each month $i$ has been binned into three classes; class '0' if it is between $-10\%$ and $10\%$, class '1' if it is greater than $10\%$ and class '2' if it is less than $-10\%$.

Table 1 shows the parameters learned from the collective learning phase. This illustrates the global behavior or trend of the data in general. Numbers

**Table 1. HMM parameters from the collective learning phase.**

| | A | B | C |
|---|---|---|---|
| **A** | 94.57% | 0.68% | 4.75% |
| **B** | 4.06% | 93.06% | 2.88% |
| **C** | 4.96% | 1.18% | 93.86% |

(a) Transition probabilities

| | A | B | C |
|---|---|---|---|
| **0** | 6% | 2% | 85% |
| **1** | 28% | 91% | 8% |
| **2** | 66% | 7% | 7% |

(b) Emission probabilities

from Table 1b indicate that when in state B, almost all customers repay ahead in time or pay in excess of what is to be paid back each month. However, in state A, customers generally fall back in repayments or pay back less than what they are supposed to, each month. State C seems to be the normal state as it is seen that in general, customers in this state pay back on time, almost the right amount. So in general globally, state C is the *Good* or *Normal* state, state B the *Prepay* state and state A the *Bad* state. The diagonal elements of Table 1a indicate that globally, customers tend to remain in the state that they are, with low probabilities of transition to other states.

## 5.1. Baseline approaches

**5.1.1. Feature based clustering.** Feature based clustering uses the demographic features of the customers to group them into clusters. We then learn an HMM for each of the clusters thus formed. Table 2 shows the parameters of the two HMMs learned in this phase. It is seen that the transition probabilities are not much different from each other and are almost similar to the values learned during the collective learning phase. The emission probabilities for the Prepay state differ between the two clusters. However, the level of actionable insights that can be derived from this observation alone is very limited.

**5.1.2. HMM based clustering.** Results from traditional HMM based clustering are presented in Table 3. It shows the transition and emission probabilities of the 4 clusters (4 so as to account for the $2\times2$ combinations resulting from fixing parameters in PP-HMM clustering) formed from HMM based clustering across both transition and emission probabilities. While the transition probabilities remain almost the same across clusters, there are significant differences across the emission probabilities. Some business insights that could be derived from these observations are mentioned below.

- In cluster 3, customers in state B toggle between prepayments and late payments while customers in cluster 1 actually tend to show delays in payments.
- State A in cluster 1 shows relatively higher chances of late payments, in tune with the global pattern.
- However, clusters 3 and 4 have very high probabilities of late payments in state A while clusters 1 and 2 have relative lower probabilities.

**Table 2. Transition and emission probabilities of the HMMs from the feature based clustering.**

| | A | B | C |
|---|---|---|---|
| **A** | 94.81% | 0.10% | 5.09% |
| **B** | 3.26% | 95.53% | 1.21% |
| **C** | 6.20% | 0.39% | 93.41% |

(a) Cluster 1; transition

| | A | B | C |
|---|---|---|---|
| **0** | 6% | 3% | 84% |
| **1** | 26% | 67% | 9% |
| **2** | 68% | 29% | 6% |

(b) Cluster 1; emission

| | A | B | C |
|---|---|---|---|
| **A** | 94.51% | 0.66% | 4.82% |
| **B** | 4.06% | 93.17% | 2.78% |
| **C** | 5.02% | 1.17% | 93.81% |

(c) Cluster 2; transition

| | A | B | C |
|---|---|---|---|
| **0** | 6% | 2% | 85% |
| **1** | 28% | 91% | 8% |
| **2** | 66% | 7% | 7% |

(d) Cluster 2; emission

**5.1.3. DTW clustering.** DTW is a widely used time-series alignment algorithm that aims in aligning two sequences by warping the time axis iteratively until an optimal match between the two sequences is found and can be used in the clustering of financial time series [31]. Unlike the previous baselines, comparing

**Table 3. Transition (column 1) and emission (column 2) probabilities of the HMMs formed from the HMM based clustering.**

| | A | B | C |
|---|---|---|---|
| **A** | 93.75% | 6.25% | 0.00% |
| **B** | 8.19% | 85.96% | 5.85% |
| **C** | 0.29% | 30.35% | 69.36% |

(a) Cluster 1; transition

| | A | B | C |
|---|---|---|---|
| **0** | 0% | 19% | 85% |
| **1** | 30% | 31% | 5% |
| **2** | 70% | 50% | 10% |

(b) Cluster 1; emission

| | A | B | C |
|---|---|---|---|
| **A** | 89.68% | 8.87% | 1.44% |
| **B** | 3.20% | 94.58% | 2.22% |
| **C** | 0.00% | 6.23% | 93.77% |

(c) Cluster 2; transition

| | A | B | C |
|---|---|---|---|
| **0** | 1% | 3% | 98% |
| **1** | 9% | 85% | 1% |
| **2** | 90% | 12% | 1% |

(d) Cluster 2; emission

| | A | B | C |
|---|---|---|---|
| **A** | 88.33% | 11.67% | 0.00% |
| **B** | 3.56% | 60.36% | 36.08% |
| **C** | 0.00% | 14.02% | 85.98% |

(e) Cluster 3; transition

| | A | B | C |
|---|---|---|---|
| **0** | 0% | 9% | 96% |
| **1** | 0% | 49% | 2% |
| **2** | 100% | 42% | 2% |

(f) Cluster 3; emission

| | A | B | C |
|---|---|---|---|
| **A** | 98.64% | 1.36% | 0.00% |
| **B** | 5.36% | 90.13% | 4.51% |
| **C** | 0.00% | 6.76% | 93.24% |

(g) Cluster 4; transition

| | A | B | C |
|---|---|---|---|
| **0** | 0% | 38% | 99% |
| **1** | 2% | 40% | 1% |
| **2** | 98% | 22% | 0% |

(h) Cluster 4; emission

or reporting emission and transition probabilities corresponding to the different clusters formed through DTW would not be meaningful (since DTW does not use this framework to cluster). Therefore, we perform an in-depth comparison of the results from this approach and the proposed PP-HMM based clustering in section 5.3. Here we report results based on a validation using an extraneous variable which is not originally exposed to both algorithms.

## 5.2. PP-HMM based clustering

Here we present the results of the two proposed settings of the HMM based algorithm and show that it gives richer insights into the data.

**5.2.1. Clustering across transition probabilities.** In this setting, only the transition probabilities are learned during the feature based clustering phase and the HMM based clustering phase while the emission probabilities are fixed to the values obtained from the collective learning phase. From Tables 4a and 4b, it is seen that the clusters are well separated unlike before. The probability of self-transition from state B to B is 0 in cluster 1 while it is 92.62% in cluster 2. Similarly, the transition probability from state B to C is 98.61% in cluster 1 while it is only 2.58% in cluster 2. Also, the transition probability from state C to B is 0% in cluster 1 while that from C to A is 0% in cluster 2. Some key business insights this clustering gives are as listed below.

- When in the Normal state, customers in cluster 2 have a much higher probability of remaining in that state than customers in cluster 1.
- Also, customers in cluster 2, when in the Prepay state have the tendency to remain in that state with high probability while customers in cluster 1 have a higher probability of transitioning to the Normal state.
- It is seen that customers in cluster 2 when paying back normally (state C), never go directly to the fall-back (state A) state, while customers in cluster 1 when paying back normally never go to the Prepay state (state B).
- Hence, any deviation from normalcy among customers in cluster 1 should raise a more serious alarm than for those in cluster 2.

**5.2.2. Clustering across emission probabilities.** Tables 5a and 5b show the emission probabilities of clusters obtained from PP-HMM based clustering keeping the transition probabilities fixed to the values obtained during the collective learning phase. In this

**Table 4. Transition probabilities of the HMMs from the PP-HMM based clustering phase with the emission probabilities fixed to values from the collective learning phase.**

| | A | B | C |
|---|---|---|---|
| **A** | 92.74% | 6.42% | 0.84% |
| **B** | 1.39% | 0.00% | 98.61% |
| **C** | 12.88% | 0.00% | 87.12% |

(a) Cluster 1

| | A | B | C |
|---|---|---|---|
| **A** | 92.99% | 6.03% | 0.98% |
| **B** | 4.81% | 92.62% | 2.58% |
| **C** | 0.00% | 2.01% | 97.99% |

(b) Cluster 2

**Table 5. Emission probabilities of the HMMs from PP-HMM based clustering phase with the transition probabilities fixed to values from the collective learning phase.**

| | A | B | C |
|---|---|---|---|
| **0** | 1% | 3% | 94% |
| **1** | 16% | 94% | 5% |
| **2** | 84% | 2% | 2% |

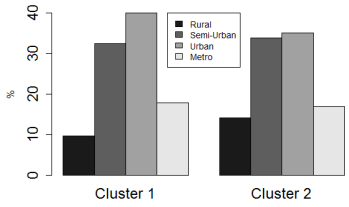
(a) Cluster 1

| | A | B | C |
|---|---|---|---|
| **0** | 8% | 0% | 67% |
| **1** | 37% | 88% | 16% |
| **2** | 55% | 12% | 17% |

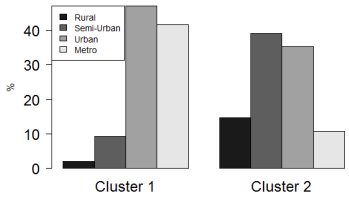
(b) Cluster 2

setting too, we see that the clusters are well separated. Probability of emitting a '1' or a '2' in state A is relatively more equal in cluster 2 than cluster 1. Similarly, while state C in cluster 1 has a very high probability of emitting a '0', it is much less in cluster 2, and a '1' or '2' also has a relatively higher probability of emission in state C in cluster 2. Some key business insights that come from this clustering are listed below.

- Customers in state B (that is, the prepay state) in cluster 2 behave similarly to the global behavior of customers in state B [see Table 1b].
- In state C (the normal state), customers in cluster 1 have a high probability of paying back normally while state C in cluster 2 has a relatively lower probability. In addition, state C in cluster 2 has a relatively higher and equal probability of prepaying as well as falling back in payments.
- Customers in cluster 2 in general show a higher variance in repayment behavior irrespective of the state they are in. This might be a product suitability issue, wherein the customer is striving hard not to default by making extra sacrifices or other loans to pay off the current loan.
- When in state A, customers in cluster 1 have a high probability of falling back (which is the global nature) while customers in cluster 2 have a relatively lower probability, and have a relatively higher probability for prepaying as well.
- Behavior in state B in cluster 2 of the HMM based clustering [see Table 3d] is similar to the behavior in state B in cluster 2 here. However, in the PP-HMM setting, we are able to gain a deeper insight about customer behavior under a global transition probability which might indicate the effect of a global financial phenomenon.




(a) DTW




(b) PP-HMM with fixed emissions

**Figure 2. Customer profile within clusters.**

## 5.3. Validation through an external variable

In this section we perform a quantitative comparison of clustering performance of the proposed PP-HMM using fixed emissions with the DTW approach introduced in section 5.1.3. We do this through the supervised approach for measuring cluster validity, a popular approach which seeks to measure how an external index matches the cluster structure [32]. Here understandably, the external index is in no way presented to the clustering algorithms (or even used in fine-tuning). In working with the practitioner, we identify the external index to be the type of branch, which is categorized as rural, semi-urban, urban, and metro. This could be a meaningful metric associated

with the customer repayment profile since it captures the socio-economic demographic, which includes various parameters affecting repayment such as the access to alternate financial support, job profiles, locational ease of repayment, and perception of delinquent behavior, to name a few. A comparison between the clusters formed from DTW based clustering and those formed from the PP-HMM clustering (fixed emissions) is shown in Figure 2. The figure shows the distribution of branch types across the clusters formed from the two methods. It is seen that the percentage of customers from different segments is almost similar across the two clusters formed using the DTW clustering approach. This is possibly due to the fact that DTW focuses only on the similarities among the sequences ignoring the factors that affect the nature of the sequences. PP-HMM clustering on the other hand forms two well-separated clusters. In terms of segment representation: cluster 1 consists mostly of customers from the metros and urban areas, whereas cluster 2 is mostly constituted by customers from the urban, semi-urban and rural areas. This further reinforces the insights from Table 4. In cluster 1, the chances of transitioning into state B or remaining in it are negligible. This might indicate that in urban areas the phenomenon to prepay is almost non existent, whereas, in rural areas, over-repayment might occur due to financial access from other sources, or because of the hardship to commute large distances to repay. Also, cluster 1 tends to hit the bad states more perhaps because they have no other sources of income once their credit risk goes bad. On the other hand, customers in rural and semi-urban settings might have access to quicker micro loans or alternate informal lenders, and at times also have a moral obligation to repay so as to be able to gain access to larger loans in the future.

## 6. Conclusion and future work

In this work, we introduce the partial parameter HMM (PP-HMM) based clustering to analyze loan repayment data. We demonstrate two specific implementations on a real-world loan repayment data set from a retail bank. Specifically, we show that the cluster structure across transitions provides a greater focus on the predilection of different customer groups towards different financial states. The process of fixing emission probabilities and contrasting it with traditional HMM clustering allows us to account for the different repayment behavior that customers would have in the same financial state. Another key finding of the transition based clustering is that a major differentiator between groups of customers is in their likelihood of hitting states of poor financial health. However, the ability of these groups to return to better states is almost identical. Our analysis on the emission based clustering highlights the different behavioral response (in terms of repayment) that different customer groups have. When this clustering is analyzed in conjunction with the traditional HMM based clustering, we can make statements on the behavioral likelihood of repayment while accounting for the different states of financial well-being that different groups are likely to have. A notable finding of the emission based clustering is that the factor differentiating customers is not a chronic tendency to repay always or be delinquent. It is differentiation along the lines of customers exhibiting consistent behavior (appropriate to their financial state) versus others who show erratic or unpredictable behavior across any given state of financial health. This has a plethora of insights on the erratic nature of cash flows for this group and therefore product suitability. Our overarching conclusion from this study is that PP-HMM provides meaningful and novel insights in understanding repayment behavior, especially, when used in conjunction with a traditional HMM based clustering.

Our future work seeks to broadly advance the techniques for gaining insights in the debt repayment environment. To this end, we intend to look at various frameworks through which extraneous information, both temporal and static, can be integrated in the PP-HMM model (which currently uses repayment data only). These extraneous sources include self-reported data, credit bureau information, and other customer related information mined from social media, phone call records, etc. The integration of this becomes specially challenging because of noisy or potentially biased nature of such sources. Another area of inquiry would be in looking at alternate subsets of parameters to constrain. For instance, clustering only on transition probabilities leaving a poor-financial-health state might specifically speak more of the ability of different customers to recover from distress.

## Acknowledgement

This work was supported by a funding from Robert Bosch Centre for Data Science and Artificial Intelligence (RBC-DSAI) at IIT Madras.

## References

[1] W. E. Deming, "On errors in surveys," *American Sociological Review*, vol. 9, no. 4, pp. 359–369, 1944.

[2] I.-C. Yeh and C.-h. Lien, "The comparisons of data mining techniques for the predictive accuracy of

probability of default of credit card clients," *Expert Systems with Applications*, vol. 36, no. 2, pp. 2473–2480, 2009.

[3] L. Lugovskaya, "Predicting default of russian smes on the basis of financial and non-financial variables," *Journal of Financial Services Marketing*, vol. 14, no. 4, pp. 301–313, 2010.

[4] L. Anderloni and D. Vandone, "Risk of over-indebtedness and behavioural factors," in *Risk Tolerance In Financial Decision Making*, pp. 113–132, Springer, 2011.

[5] J. Schicks, "The sacrifices of micro-borrowers in ghana–a customer-protection perspective on measuring over-indebtedness," *The Journal of Development Studies*, vol. 49, no. 9, pp. 1238–1255, 2013.

[6] P. G. Moffatt, "Hurdle models of loan default," *Journal of the operational research society*, vol. 56, no. 9, pp. 1063–1071, 2005.

[7] H. Chen, R. H. Chiang, and V. C. Storey, "Business intelligence and analytics: From big data to big impact.," *MIS quarterly*, vol. 36, no. 4, pp. 1165–1188, 2012.

[8] S. Moro, P. Cortez, and P. Rita, "Business intelligence in banking: A literature analysis from 2002 to 2013 using text mining and latent dirichlet allocation," *Expert Systems with Applications*, vol. 42, no. 3, pp. 1314–1324, 2015.

[9] J. Paxton, D. Graham, and C. Thraen, "Modeling group loan repayment behavior: New insights from burkina faso," *Economic Development and Cultural Change*, vol. 48, no. 3, pp. 639–655, 2000.

[10] N. Bhatt and S.-Y. Tang, "Determinants of repayment in microcredit: Evidence from programs in the united states," *International Journal of Urban and Regional Research*, vol. 26, no. 2, pp. 360–376, 2002.

[11] N. Sudarsanam and D. J. Philip, "Quantifying and predicting prepayments in the microfinance environment," in *NSE-IFMR Finance Foundation Financial Deepening and Household Finance Research Initiative*, 2016.

[12] J. Banasik, J. N. Crook, and L. C. Thomas, "Not if but when will borrowers default," *Journal of the Operational Research Society*, vol. 50, no. 12, pp. 1185–1190, 1999.

[13] B. C. on Banking Supervision, "Customer suitability in the retail sale of financial products and services." `http://www.bis.org/publ/joint20.pdf`, 2008. [Online; accessed 7-February-2017].

[14] M. Sharma and M. Zeller, "Repayment performance in group-based credit programs in bangladesh: An empirical analysis," *World development*, vol. 25, no. 10, pp. 1731–1742, 1997.

[15] J. G. Cragg, "Some statistical models for limited dependent variables with application to the demand for durable goods," *Econometrica: Journal of the Econometric Society*, pp. 829–844, 1971.

[16] R. S. Mamon and R. J. Elliott, *Hidden markov models in finance*, vol. 104. Springer Science & Business Media, 2007.

[17] G. Giampieri*, M. Davis, and M. Crowder, "Analysis of default data using hidden markov models," *Quantitative Finance*, vol. 5, no. 1, pp. 27–34, 2005.

[18] H. Frydman and T. Schuermann, "Credit rating dynamics and markov mixture models," *Journal of Banking & Finance*, vol. 32, no. 6, pp. 1062–1075, 2008.

[19] D. Lando and T. M. Skødeberg, "Analyzing rating transitions and rating drift with continuous observations," *journal of banking & finance*, vol. 26, no. 2, pp. 423–444, 2002.

[20] E. Gaffney, R. Kelly, F. McCann, *et al.*, "A transitions-based framework for estimating expected credit losses," *Research Technical Paper 16RT14, Central Bank of Ireland*, 2014.

[21] J. D. Hamilton, "Regime switching models," in *Macroeconometrics and Time Series Analysis*, pp. 202–209, Springer, 2010.

[22] S. Aghabozorgi, A. S. Shirkhorshidi, and T. Y. Wah, "Time-series clustering–a decade review," *Information Systems*, vol. 53, pp. 16–38, 2015.

[23] P. Smyth *et al.*, "Clustering sequences with hidden markov models," *Advances in neural information processing systems*, pp. 648–654, 1997.

[24] B. Knab, A. Schliep, B. Steckemetz, and B. Wichern, "Model-based clustering with hidden markov models and its application to financial time-series data," in *Between Data Science and Applied Data Analysis*, pp. 561–569, Springer, 2003.

[25] J. G. Dias and S. B. Ramos, "Dynamic clustering of energy markets: An extended hidden markov approach," *Expert Systems with Applications*, vol. 41, no. 17, pp. 7722–7729, 2014.

[26] T. Oates, L. Firoiu, and P. R. Cohen, "Clustering time series with hidden markov models and dynamic time warping," in *Proceedings of the IJCAI-99 workshop on neural, symbolic and reinforcement learning methods for sequence learning*, pp. 17–21, Sweden Stockholm, 1999.

[27] C. Li and G. Biswas, "A bayesian approach to temporal data clustering using hidden markov models.," in *ICML*, pp. 543–550, 2000.

[28] J. G. Dias, J. K. Vermunt, and S. Ramos, "Clustering financial time series: New insights from an extended hidden markov model," *European Journal of Operational Research*, vol. 243, no. 3, pp. 852–864, 2015.

[29] L. R. Rabiner, "A tutorial on hidden markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, 1989.

[30] J. A. Hartigan and M. A. Wong, "Algorithm as 136: A k-means clustering algorithm," *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, vol. 28, no. 1, pp. 100–108, 1979.

[31] P. Tsinaslanidis, A. Alexandridis, A. Zapranis, and E. Livanis, "Dynamic time warping as a similarity measure: applications in finance," 2014.

[32] A. K. Jain and R. C. Dubes, *Algorithms for clustering data*. Prentice-Hall, Inc., 1988.