Analyzing concept drift

A case study in the financial sector

Masegosa, Andrés; Martinez, Ana ; Ramos-Lopez, Dario; Langseth, Helge; Nielsen, Thomas Dyhre; Samerón, Antonio

Published in:
Intelligent Data Analysis

DOI (link to publication from Publisher):
10.3233/IDA-194515

Publication date:
2020

Document Version
Accepted author manuscript, peer reviewed version

RESEARCH

Analyzing concept drift: a case study in the financial sector

Andrés R Masegosa[1*], Ana M Martínez[2], Darío Ramos-López[1], Helge Langseth[3], Thomas D. Nielsen[4] and Antonio Salmerón[1]

*Correspondence:
andresmasegosa@ual.es
[1]Department of Mathematics, University of Almería, Ctra. Sacramento, s/n, 04120 Almería, Spain
Full list of author information is available at the end of the article

Abstract

In this paper we present a method for exploratory data analysis of streaming data based on probabilistic graphical models (latent variable models). This method is illustrated by concept drift tracking, using financial client data from a European regional bank. For this particular setting, the analyzed data spans the period from April 2007 to March 2014, and therefore starts before the beginning of the financial crisis of 2008. The implied changes in the economic climate during this period manifests itself as *concept drift* in the underlying data generating distribution. We explore and analyze this financial client data using a probabilistic graphical modeling framework that provides an explicit representation of concept drift as an integral part of the model. We show how learning these types of models from data provides additional insight into the hidden mechanisms governing the drift in the domain. We present an iterative approach for identifying disparate factors that jointly account for the drift in the domain. This includes a semantic characterization of one of the main influencing drift factors. Based on the experiences and results obtained from analyzing the financial data, we discuss the applicability of the framework within a more general context.

Keywords: concept drift; latent variable models; financial data

1 Introduction

Performing data analysis in a streaming context raises several important issues that are often less pronounced when conducting batch data analysis. In particular, the instances in a data stream can often not be assumed independent, and when the data stream exhibits concept drift the underlying data generating distribution may change over time [1]. If the concept drift inherent in the domain is not carefully taken into account, the result can be a deterioration of accuracy when doing classification or, more generally, failure to capture and interpret intrinsic properties of the data during data exploration.

In collaboration with a European regional bank (Banco de Crédito Cooperativo, BCC[1]), we have been conducting data analysis over a subset of their clients based on client-specific financial information captured during the period from April 2007 to March 2014. Specifically, focus has been on real-time analysis, detection, and interpretation of financial changes during this period. Particular attention has been given to two groups of clients, defined by whether or not they will default on their financial obligations within the following 12 months.

[1] www.bcc.es/en

The period during which data has been collected starts before the beginning of the financial crisis, hence the general economic climate exhibits changes during the collection period. This is also directly reflected in the client data, where we, for instance, see drifts in the average monthly account balances as illustrated in Figure 1(d). Note that the drift is more pronounced for the defaulting clients than for the non-defaulting clients, and that we also see a slightly inverse trend for the two client groups. This is the reason why we included the defaulting information in the analysis. Another example of drift can be seen for the unpaid amount in mortgages for the two groups as shown in Figure 1(e).

Generally, when comparing these two financial indicators we see that they exhibit different types of concept drift, and that a common/global contextual cause is not immediately apparent from the data. We would therefore like to go beyond this immediate analysis and instead consider 'broader' types of concept drift that are less variable-specific and which influence and govern several key financial indicators simultaneously and across client groups. We thus adopt the general definition of concept drift from [1], where concept drift is defined as the existence of two consecutive time points for which the joint distribution over the domain variables differ. Since this definition does not rely on a designated target variable, we can position the problem as *unsupervised* concept drift detection and analysis [2].

In this paper, we further explore and extend a recently proposed model for capturing concept drift [3]. The model proposed by [3] is based on probabilistic graphical models and provides a principled approach for capturing concept drift by letting the drift be encoded explicitly within the model class. There are several advantages to this approach. First of all, the model does not rely on any of the standard techniques to deal with concept drift, see [1] and the references therein. Such techniques, including external concept drift detectors that, e.g., would use changes in classification accuracy to imply a form of *supervised concept drift*, often require information that only becomes available after a certain time delay [4, 5] (such as the true class labels). Secondly, by letting concept drift be an explicit and integral part of the modeling framework we have added support for semantic interpretations of potential drifts. In particular, concept drift can immediately be linked to selected model components enabling an analysis of how concept drift affects different parts of the model. The proposed method therefore not only provides an unsupervised way of detecting concept drift, it may also enable a more systematic analysis of the (local) domain specific factors that drive concept drift; this type of insight is not immediately provided when, e,g, involving external concept drift detectors.

In comparison with the analysis conducted in [3], the contributions of the present paper includes an extension of the model class of [3] that enables a more fine-grained concept drift analysis targeting individual variables. Furthermore, we propose an iterative approach for identifying disparate factors that jointly account for the drifts in the domain. We demonstrate the use of the proposed methods based om the financial data set supplied by BCC. The results of the analysis includes the identification and semantic characterization of one of the key factors governing concept drift for this particular domain. Lastly, we discuss the proposed modeling framework in a more general setting, linking model validation and concept drift analysis. The proposed methods are released as part of an open-source toolbox for scalable probabilistic machine learning (`http://www.amidsttoolbox.com`) [6, 7, 8].

The remainder of the paper is organized as follows. In Section 2 we provide a detailed description and analysis of the data set that is used in the study. Section 3 discusses and analyzes the modeling framework introduced by [3], on which the concept drift analysis presented in Section 4 is based. In Section 5 we position our analysis within a more general context, providing a critical discussion of the limitations of our procedure as well as possible extensions to the framework and open research questions. Lastly, we give some concluding remarks in Section 6.

2 Description of the Data

The data set, provided by BCC, contains monthly aggregated information for a set of BCC clients for the period from April 2007 to March 2014. Only "active" clients are considered, meaning that we restrict our attention to individuals between 18 and 65 years of age, who have at least one automatic bill payment or direct debit in their accounts. To make the data set as homogeneous as possible, we only retain clients residing in the region of Almería (a mainly agricultural area in the south of Spain), and excluded BCC employees, since they have special banking conditions. We reduced the resulting data set so that it only includes 50 000 clients each month.

Up until December 2010 there are some clients that only become active every six months (due to periodic fees). From December 2010 and until the end of the period this pattern appears every 3 months [3]. The particular clients involved vary, and removing them from the overall data set is therefore not feasible. Instead, and in order to avoid the seasonal peaks produced by these known patterns, we remove the affected 21 months.[2]

Assisted by BCC's experts, we extracted six variables from the resulting data set that encode monthly aggregated information, and which collectively describe the financial status of a client. Figure 1 shows the evolution of these variables for both defaulting and non-defaulting clients throughout the period. We note that some clients may have missing values for some of the variables for a given month (e.g., because a client was not active during that particular month). However, the generative nature of the models we employ (detailed in the following section) ensures that these missing values are naturally handled within the model and do not need to be treated separately. Finally, each client also has an associated class variable, which indicates if that particular client will default during the following 12 months.

If we take a closer look at the attributes, we observe several characteristics that could further challenge the modeling process. Figure 2 shows the histograms for a couple of the variables. The first thing we notice is the high density of zeros, but also the long-tails of the distributions. The latter results in large variances for most of the attributes.

When also considering the evolution of the means in Figure 1 we can see at least two tendencies in the data. One general trend of gradual and monotonic movement and the other of seasonal changes usually peaking at the end of the year. Thus, the data sets appear to exhibit different types of concept drift.

[2] The analysis of the experiments in this paper are practically the same if we consider the peak months, except that the results are more noisy around these months [3].

The drift in the data indicates the need for a model that takes these changes into account. More specifically, we are interested in a simple density estimator that is able to detect the different tendencies over multiple attributes simultaneously while also maintaining the defaulting/non-defaulting distinction of the clients. The following section introduces and discusses the particular model type we have used for the analysis.

3 Modeling concept drift

Concept drift detection and adaptation has typically been considered within a supervised learning context, where changes in model accuracy are seen as an indication of concept drift [1]. This means that concept drift detection is closely linked to a specific prediction task, which may be too restrictive for an exploratory data analysis setting. For example, labeled streaming data is needed in order to evaluate changes in classification accuracy, but these labels often come with a (significant) time delay. For example, for our financial setting described in Section 2, if the future defaulting status of a client is considered as the class variable, then the true class label will only be revealed after a delay of twelve months.

Instead we consider the framework for detecting and analyzing concept drift presented by [3], in which the key idea is to explicitly represent concept drift as an integral part of the model definition without relying on a designated target variable. Thus, the framework considers concept drift as the existence of an instance $\boldsymbol{x}$ and two consecutive time points t_0 and t_1 such that $p_{t_0}(\boldsymbol{x}) \neq p_{t_1}(\boldsymbol{x})$, where $p_t(\cdot)$ denotes the density of $\boldsymbol{x}$ at time t. Note that this type of concept drift modeling can be considered unsupervised. The *concept* in this context refers to a joint distribution over the class and predictive variables, but with the class being treated as a normal random variable. Incidentally, this is also what we will refer to as *global concept drift*, as opposed to *local concept drift*, that captures concept drift happening at the level of a single variable in the model. A more thorough discussion about the type of concept drift that this framework is able to detect is given in Section 5.

The modeling framework proposed by [3] is illustrated using plate notation in Figure 3 and can be seen as a special type of probabilistic graphical model [9]. In the figure, $(Y_{i,t}, \mathbf{X}_{i,t})$ describes the behavior of client i at time t, where $Y_{i,t}$ represents the defaulting status of client i at time t and $\mathbf{X}_{i,t}$ are the financial indicator variables describing the client.[3] The distributions of $Y_{i,t}$ and $\boldsymbol{X}_{i,t}$ are parameterized using the parameters θ_y and θ_x, respectively. Concept drift is captured in the model through the latent variables $H_1, H_2, \ldots, H_t, \ldots$, which are "shared" across clients and indicator variables. Intuitively, when learning from data subject to concept drift, the model responds by "tweaking" H_t over time, thereby using this sequence of latent variables to aggregate the concept drift of each variable to a "model-global" level. To enforce a smooth drift-model, the conditional distribution $H_t|\{H_{t-1} = h_{t-1}\}$ is defined as a random walk with variance θ_h, which has a priori been given

[3] We explicitly represent the class variable in this context as it is a requirement of the BCC experts to have a clear distinction between defaulter and non-defaulter clients. Since the percentage of defaulter clients is small, this ensures that this group of clients is modeled separately. We stress again, however, that our general concept drift model is not relying on classification accuracy to detect concept drift.

preference to smaller movements over time. Note again that the emphasis here is on the concept drift component and not on the specification of an accurate model; hence the simplistic model structure for $(Y_{i,t}, \mathbf{X}_{i,t})$. The specific parametric families employed by the modeling framework for analyzing this data set are presented in Section 4, where the general model class will be instantiated to the financial data set.

The overall framework is positioned in the Bayesian paradigm, where both parameters and unobserved variables (H_t, θ_x, θ_y, θ_h as well as missing data observations) are treated as random variables in the model. For the data up to and including time T, $\mathcal{D}_{1:T}$, inference amounts to calculating the distribution over the variables of interest given $\mathcal{D}_{1:T}$, most notably $p(H_T|\mathcal{D}_{1:T})$.

Bayesian inference is in general NP-hard [10] and for the type of hybrid dynamic models considered in this paper (detailed in the following section) exact inference is intractable. Thus, we resort to approximate inference/learning based on a *variational Bayes* inference engine [11, 12]. Variational Bayes can be seen as a gradient ascent algorithm, and when constraining the (conditional) distributions to be members of the conjugate exponential family, it can be implemented through variational message passing (VMP) [13].

There are several benefits of this approach, including *i*) having concept drift as an integral part of a holistic model; *ii*) concept drift is explicitly represented and therefore open for investigation; *iii*) immediate model validation; *iv*) good fit in terms of marginal log-likelihood to complex data, even for this rather parsimonious model.

A proof of concept of this modeling framework can be found in Appendix B, where we analyze two synthetic data sets widely employed as benchmarks in the concept drift literature. This analysis verifies the applicability of this framework for modeling concept drift beyond the financial domain, which is the focus of this work.

4 Analyzing Concept Drift with Hidden Variables

In this section we detail the instantiation of the general methodology presented in Section 3 with respect to our financial data set in order to analyze the trend in the evolution of the financial profile of the clients. For this, we use a publicly available toolbox[4], called the AMIDST Toolbox. This toolbox is open source and gives access to a modeling language, where models can be described and combined with inference procedures that support Bayesian learning of the model parameters. Moreover, since the data setup is of a streaming nature, scalability is an important feature of the toolbox. A streaming data set is potentially unbounded, thus inference amounts to doing filtering (also known as the forward pass in dynamic model inference). This means that for any t, only the data $\mathcal{D}_{1:t}$ will influence the posterior estimate of H_t; observations at, e.g., $t+1$, will not be taken into account.

We consider two general types of models. First we explore a model containing a hidden variable $H_{j,t}$ for each attribute $X_{j,t}$ with the purpose of analyzing the drift behavior of the different features independently. Secondly, we use a single hidden variable for all features $\boldsymbol{X}_t$ to capture more global types of concept drift.

[4]The code and models used in this paper can be downloaded from the AMIDST Toolbox webpage (through its GitHub repository): www.amidsttoolbox.com

We subsequently analyze the residuals produced by the global model to identify other factors that jointly account for the drift.

4.1 Local Hidden Variables

As a first step we employ a variant of the models presented in Section 3 to track concept drift of individual attributes. In this way, we examine whether or not the six attributes in our data set exhibit different drift behavior. This simpler setting also allows us to better illustrate how our approach captures the general trend in the data over time.

We make a simple instantiation of the general framework, where each attribute is linearly dependent on a local hidden variable, which enables the use of efficient learning algorithms. More complex (non-linear) dependencies could eventually be used by employing alternative parametrizations of the conditional probability distributions, at the cost of having to use more complex learning algorithms.

More precisely, we use a concept drift model with a hidden variable $H_{j,t}$ for each attribute. This model can be expressed as follows:

$$\begin{aligned} x^{+}_{i,j,t} &= \alpha^{+}_{j} + \beta^{+}_{j} \cdot H_{j,t} + \epsilon^{+}_{i,j,t}, \\ x^{-}_{i,j,t} &= \alpha^{-}_{j} + \beta^{-}_{j} \cdot H_{j,t} + \epsilon^{-}_{i,j,t}, \end{aligned} \tag{1}$$

where $x_{i,j,t}$ denotes the value of the j-th attribute of the i-th client at time t. The superscripts $+$ and $-$ refer to the group of defaulter and non-defaulter clients, respectively. The rest of the parameters are defined as random variables following a Bayesian framework (where we have suppressed the $+$ and $-$ to indicate that the same a priori model is assumed for both groups of customers):

$$\begin{aligned} &\alpha_j, \beta_j, H_{j,0} \sim \mathcal{N}(\mu, \sigma^2), \\ &\epsilon_{i,j,t} \sim \mathcal{N}(0, \sigma^2_j), \\ &\sigma^2_j \sim \text{InvGamma}(\alpha, \beta), \\ &H_{j,t} \sim \mathcal{N}(H_{j,t-1}, \sigma^2). \end{aligned}$$

Using standard properties of the Gaussian distribution, we then have that $X_{i,j,t}|\{\alpha_j, \beta_j, \sigma^2_j, h_{j,t}\} \sim \mathcal{N}(\alpha_j + \beta_j \cdot h_{j,t}, \sigma^2_j)$. Note that in this model, we have a single hidden variable $H_{j,t}$ that jointly tracks the drift of the profile of the defaulter and non-defaulter clients for the j-th attribute. Furthermore, the attribute specific $\beta_{(\cdot)}$ coefficients can account for potential scale differences among the features.

In Figure 4 we plot a detailed result of this analysis for two attributes: Account Balance (AB) and Unpaid Amount in Mortgages (UM), respectively. All means in the normal distributions have been arbitrarily initialized to zero: $\alpha_j, \beta_j, H_0 \sim \mathcal{N}(0, \infty)$, where the variance has been initialized with a sufficiently large number to allow for adaption; $\sigma^2_j \sim \text{InvGamma}(0, 1)$ and $H_{j,t} \sim \mathcal{N}(H_{j,t-1}, 0.1)$. Each figure displays the following series:

- $\{x^{+}_{j,t}\}$ and $\{x^{-}_{j,t}\}$ show the empirical mean of the attribute (for defaulter/non-defaulters clients) at every month, i.e. $x^{+}_{j,t} = 1/N^{+} \cdot \sum_i x^{+}_{i,j,t}$, where N^{+} is the number of defaulter clients at month t ($x^{-}_{j,t}$ is defined analogously). With this series we see how the empirical mean changes over time.

- $\{\mathbb{E}[H_{j,t}]\}$ shows the expected value of the hidden variable $H_{j,t}$, which aims at tracking the drift in the empirical means at each month for attribute j.
- Two series defined by $\{a^+_{j,t}\} \doteq \{\mathbb{E}[\alpha^+_j] + \mathbb{E}[\beta^+_j] \cdot \mathbb{E}[H_{j,t}]\}$ and $\{a^-_{j,t}\} \doteq \{\mathbb{E}[\alpha^-_j] + \mathbb{E}[\beta^-_j] \cdot \mathbb{E}[H_{j,t}]\}$ with the linear combination of the expected value of the variables α^+_j, β^+_j, α^-_j, β^-_j and $H_{j,t}$ at every month. This last series should approximate the series describing the empirical mean of the attribute (cf. Equation 1).

Considering Figure 4 we can make the following tentative conclusions:

- The series $\{a_{j,t}\}$ try to approximate the empirical mean series. The fit is not perfect because we are using a model with a small number of parameters, that is, we aim to fit two series with 126 values (the empirical monthly means of defaulters and non-defaulters) with a model which contains only 67 parameters (the 63 expected values of the variable $H_{j,t}$ plus the α_j and β_j parameters of both client groups). Still, $\{a_{j,t}\}$ is able to capture the general trend of the empirical means series.
- The $\{\mathbb{E}[H_{j,t}]\}$ series aim to capture the drift in both empirical mean series $\{x^+_{j,t}\}$ and $\{x^-_{j,t}\}$. We note that the drifts in the $\{x^+_{j,t}\}$ series are different from the drifts in the $\{x^-_{j,t}\}$ series, as we commented in Section 1. The $\{\mathbb{E}[H_{j,t}]\}$ series try to make a *compromise* between the two different drift trends. This is especially visible at the final stages of both time series (defaulters/non-defaulters) in Figure 4.
- The movements of the time series are scaled by the values of their α_j and β_j parameters for the same $\{\mathbb{E}[H_{j,t}]\}$.[5] If we take a closer look at the α_j and β_j values of both defaulters and non-defaulters, we can understand why the same change in $\{\mathbb{E}[H_{j,t}]\}$ affects the estimated means differently. Intuitively speaking, the value of α_j determines the expected mean value of the variables when β_j is zero, whereas β_j determines the change with respect to $\{\mathbb{E}[H_{j,t}]\}$. If we consider the *Unpaid amount in mortgages* for Figure 4 (b), the ratio $\frac{\alpha_j}{\beta_j}$ for non-defaulters is much higher than for defaulters, which means that the former will be less sensitive to changes in $\{\mathbb{E}[H_{j,t}]\}$.

Finally, in Figure 5 we plot the set of $\{\mathbb{E}[H_{j,t}]\}$ series for all the six attributes analyzed in our financial data set. We note again that each $\{\mathbb{E}[H_{j,t}]\}$ series tries to reflect the joint evolution of the profile of the defaulter and non-defaulter clients with respect to the j-th attribute.

It is interesting to see in Figure 5 how we can clearly identify two groups of attributes with different evolution trends. On the one hand, we have that the attributes "Total Credit Amount", "Unpaid Amount in Mortgages" and "Unpaid Amount in Personal Loans" (Att1, Att5 and Att6 respectively) exhibit a kind of monotonically increasing trend over time, with no seasonality. According to our BCC's experts, they mainly show the financial deterioration of the defaulting clients (c.f. Figure 1): higher unpaid amount in mortgages and higher total credit loans, although for "Unpaid Amount in Personal Loans" we can see a slow reduction across the period. The latter is because personal loans are typically small short term loans with high interest rates, which clients prefer to pay back on time. Another effect

[5] Due to confidentiality reasons, we are unfortunately not able to disclose the α and β values.

comes into play here: during the observation period many weak non-defaulter clients changed to the group of defaulting clients, leaving in the former group those clients that were more robust to changes in the economic climate. This translates into an improvement of the financial profile of the group of non-defaulter clients.

The other group of attributes, "Income", "Expenses" and "Account Balance" (Att2, Att3 and Att4 respectively), identified in Figure 5, presents a yearly seasonal pattern down-peaking at the end of the year, which characterizes the particular financial profile of the BCC's clients. The "Account Balance" attribute seems to have a more complex evolution, which will be discussed in more detail in the next section.

During the analysis above we have deliberately neglected that the estimators $\{\mathbb{E}[\alpha_j^+]\}$, $\{\mathbb{E}[\alpha_j^-]\}$, $\{\mathbb{E}[\beta_j^+]\}$ and $\{\mathbb{E}[\beta_j^+]\}$ can also evolve over time. The time-dependency is a consequence of the definition of the estimators; recall that they are calculated as streaming Bayesian posterior mean values, which in turn are based on *the data seen so far*. In consequence, the analysis of the $\{\mathbb{E}[H_{j,t}]\}$ series could in principle be hiding other types of concept drift: a constant $\{\mathbb{E}[H_{j,t}]\}$ series would, for example, be interpreted as if there was no concept drift, even though a drift could actually be absorbed by the $\boldsymbol{\alpha}_j$ and $\boldsymbol{\beta}_j$ series. We examine this potential issue further in Appendix A. We do so by conducting an off-line analysis to evaluate what happens when the parameters are kept fixed (i.e., we prevent the $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ series to evolve over time), thereby ensuring that the $\{\mathbb{E}[H_{j,t}]\}$ series are the only means for the model to absorb the inherent dynamics. We show that the results in this setting are comparable to those of the procedure outlined above, and therefore conclude that this issue does not invalidate the present analysis.

4.2 Global Hidden Variables

In the previous section we looked at the individual trends of each of the attributes. In this section, we are interested in capturing the joint global trend of all of them. For simplicity, let us start by disregarding the defaulter status of the clients, i.e.,

$$x_{i,j,t} = \alpha_j + \beta_j \cdot H_t + \epsilon_{i,j,t}. \tag{2}$$

We are now employing a single scalar variable to model the drift of the full set of variables defining the profile of the client (as before α_j and β_j do not evolve over time). Despite this simple structure, the model is flexible enough to capture different interesting types of concept drift as exemplified below:

- Let us assume we have two series: $\{x_{1,t}\}$ does not change over time (beyond random white noise) while $\{x_{2,t}\}$ linearly increases over time (beyond random white noise). This can be captured by setting β_1 to 0 and choose a proper positive value for β_2 (both α_1 and α_2 need to be properly fixed to fit the data). $\{H_t\}$ will then linearly increase reflecting the change of $\{x_{2,t}\}$.
- Assume now that $\{x_{1,t}\}$ increases linearly, and that $\{x_{2,t}\}$ decreases linearly at a higher pace. This can be captured by a positive β_1 value and a comparatively larger negative β_2 value. $\{H_t\}$ will then increase linearly reflecting the change of $\{x_{1,t}\}$ and $\{x_{2,t}\}$. $\{H_t\}$ could also decrease linearly if we flip the signs of β_1 and β_2.

Table 1: Person's Correlation Coefficient between the Unemployment Rate (3 months shifted) and the $\{\mathbb{E}[H_{j,t}]\}$ and $\{\mathbb{E}[H_t]\}$ series.

$\{\mathbb{E}[H_{1,t}]\}$	$\{\mathbb{E}[H_{2,t}]\}$	$\{\mathbb{E}[H_{3,t}]\}$	$\{\mathbb{E}[H_{4,t}]\}$	$\{\mathbb{E}[H_{5,t}]\}$	$\{\mathbb{E}[H_{6,t}]\}$
0.926	0.672	0.818	-0.131	0.872	0.857

$\{\sum \mathbb{E}[H_{j,t}]\}$	$\{\mathbb{E}[H_t]\}$
0.935	0.961

By extending the model to also include the defaulter status of the clients, we get

$$
\begin{aligned}
x^{+}_{i,j,t} &= \alpha^{+}_{j} + \beta^{+}_{j} \cdot H_t + \epsilon^{+}_{i,j,t}; \\
x^{-}_{i,j,t} &= \alpha^{-}_{j} + \beta^{-}_{j} \cdot H_t + \epsilon^{-}_{i,j,t},
\end{aligned}
\tag{3}
$$

Again, a single hidden variable H_t will be used to jointly track the drift over time in the profiles of the two client groups. This extended version corresponds to the model described in Section 3, where variable $\boldsymbol{X}$ is conditioned on variable Y.

Figure 6 shows the result of this analysis by plotting the $\{\mathbb{E}[H_t]\}$ series. It is interesting to see how this $\{\mathbb{E}[H_t]\}$ series displays a combination of monotonic increasing trend with a seasonal change, so it seems to aggregate the different individual trends of each of the attributes. Even more interesting is to look at this series when compared to the unemployment rate in the region of the financial institution when the latter is shifted three months to the past. As can be seen, both series are highly correlated. For example, during most of the first two years there is hardly any seasonality in either series. But after this period, starting from February 2009, both the unemployment rate and the $\{\mathbb{E}[H_t]\}$ series show a clear overlapping seasonality pattern.

In Table 1 we show the Pearson correlation coefficient between the unemployment rate (three months shifted) and the $\{\mathbb{E}[H_{j,t}]\}$ and $\{\mathbb{E}[H_t]\}$ series. We also compute the Pearson correlation coefficient with respect to the series $\{\sum_j \mathbb{E}[H_{j,t}]\}$, defined by the sum of all the local hidden variables. As can be seen, the correlation achieved by $\{\mathbb{E}[H_t]\}$ is higher than the correlation obtained by the rest of the series. This indicates that by using the global model defined in Equation 2 we are able to better capture the global trend present in our data, which turned out to be largely driven by the unemployment rate.

Correlation does not imply causation, but common sense tells us that when the unemployment rate in a small region moves from 12% to 30% in less than two years, it is difficult to imagine another factor that could have more impact on the financial situation of the inhabitants of this region. Thus, from this analysis it seems reasonable to postulate that the enormous change in the economic profile of the clients was mainly driven by the changes in the unemployment rate during this period.

We are now interested in exploring if the changes in tendency for all the different variables are entirely explained by the $\{\mathbb{E}[H_t]\}$ series in this time period, and, consequently, fully determined by the unemployment rate. For that, we plot in Figures 7 and 8 the expected monthly values of the predictive variables as a linear combination of the parameters. That is, we plot the series $\{a^{+}_{j,t}\}$ and $\{a^{-}_{j,t}\}$ together

with the empirical means $\{x^+_{j,t}\}$ and $\{x^-_{j,t}\}$ and analyze the goodness of the fit (in Section 4.4 we discuss a formal way to look at this issue). If all variables were perfectly learned, that would mean that a single global variable would be able to capture all the changes. There are some variables, like *Income* (Figure 7) whose trend is very well captured by the global variable, despite the noise. However, if we look at other variables like *Account Balance* and *Unpaid amount in mortgages* (Figure 8), we see that, especially towards the end of the time series, the fit starts to degrade.

In the next subsection we show how we can extend our approach to determine if there are unexplained trends which have not been captured by our single $\{\mathbb{E}[H_t]\}$ series and how we could capture them in a meaningful manner.

As for the local model, we also evaluate the robustness of the $\{\mathbb{E}[H_t]\}$ estimates wrt. changes in the series of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ estimators in Appendix A. Once again we find that the conclusions drawn above are not significantly affected by the potential drift in the parameter estimators.

4.3 Residual Analysis

In order to possibly identify other unexplained trends, we look at the *residuals* defined as the difference between the observed value and the estimated value according to the model specified in Equation 3,

$$r_{i,j,t} = x_{i,j,t} - \mathbb{E}[\alpha_j] - \mathbb{E}[\beta_j] \cdot \mathbb{E}[H_t], \tag{4}$$

where $\mathbb{E}[\alpha_j]$, $\mathbb{E}[\beta_j]$ and $\mathbb{E}[H_t]$ denote the expected value of the random variables at month t.

We then employ the same modeling approach we used in the previous sections, but now focusing on the calculated residuals. Firstly, we generate sequences of hidden variables $H^r_{j,t}$ to track the drift over time of the residuals for each attribute,

$$r_{i,j,t} = \alpha^r_j + \beta^r_j \cdot H^r_{j,t} + \epsilon^r_{i,j,t}. \tag{5}$$

Secondly, we generate another sequence of hidden variables H^r_t to track the drift over time of the residuals for all the attributes jointly,

$$r_{i,j,t} = \alpha^r_j + \beta^r_j \cdot H^r_t + \epsilon^r_{i,j,t}. \tag{6}$$

We want to point out that this residual analysis has a straightforward interpretation in terms of multiple hidden variables. That is, an extension of the models given in Equation 3 or Equation 1 to include two hidden variables corresponding to H^r_t and $H^r_{j,t}$ respectively. This can be seen if we take expectations in Equation 5 or Equation 6 and use the equality of Equation 4,

$$\mathbb{E}[x_{j,t}] = \mathbb{E}[\alpha_j] + \mathbb{E}[\alpha^r_j] + \mathbb{E}[\beta_j] \cdot \mathbb{E}[H_{j,t}] + \mathbb{E}[\beta^r_j] \cdot \mathbb{E}[H^r_{j,t}]$$

for the local model and

$$\mathbb{E}[x_{j,t}] = \mathbb{E}[\alpha_j] + \mathbb{E}[\alpha^r_j] + \mathbb{E}[\beta_j] \cdot \mathbb{E}[H_t] + \mathbb{E}[\beta^r_j] \cdot \mathbb{E}[H^r_t],$$

Series	Att1	Att2	Att3	Att4	Att5	Att6
$Var(\{\mathbb{E}[H_{j,t}]\})$	25.50	8.75	15.50	29.34	41.23	96.93
$Var(\{\mathbb{E}[H^r_{j,t}]\})$	8.00	4.10	3.68	43.17	27.17	3.67
$Var(\{\mathbb{E}[H^{r2}_{j,t}]\})$	0.42	3.74	2.94	2.23	7.14	1.34

Table 2: Variances of the $\{\mathbb{E}[H_{j,t}]\}$, $\{\mathbb{E}[H^r_{j,t}]\}$ and $\{\mathbb{E}[H^{r2}_{j,t}]\}$ series; expected mean of the first global hidden variables, the global hidden variable for the first set of residuals and the global hidden variable for the second set of residuals respectively. The index j corresponds to Attribute Att$j$.

for the global model. In both cases, $\mathbb{E}[x_{j,t}]$ denotes the expected value of the j-th attribute at time t. Consequently, the following residual analysis results can also be interpreted as trying to capture an additional hidden variable modeling the drift behavior of the profile of the clients over time.

In Figure 9 we show the $\{\mathbb{E}[H^r_{j,t}]\}$ series according to Equation 5 for the residuals of all the attributes. When we compare these results with the ones displayed in Figure 5, we can see that the $\{\mathbb{E}[H^r_{j,t}]\}$ series displays, for most of the attributes, a much more constant profile than the $\{\mathbb{E}[H_{j,t}]\}$ series. A quantitative evaluation of this fact is given in Table 2 (Rows 1 and 2), where we compare the variance of the $\{\mathbb{E}[H_{j,t}]\}$ and $\{\mathbb{E}[H^r_{j,t}]\}$ series, defined as $Var(\{\mathbb{E}[H_{j,t}]\}) = \frac{1}{T}\sum_1^T \left(\mathbb{E}[H_{j,t}] - \bar{H}_j\right)^2$, where $\bar{H}_j = \frac{1}{T}\sum_1^T \mathbb{E}[H_{j,t}]$; similarly for $Var(\{\mathbb{E}[H^r_{j,t}]\})$. As can be seen, the variance is reduced for all the attributes, except for Att4 (Account Balance). It is interesting to see how the Account Balance attribute clearly diverges showing that the trend of this attribute could not be captured by the previous hidden variable. Something similar happens, but only at the end of the series, for Att1 and Att5.

The behavior in Figure 9 of the Account balance attribute (Att4) can partly be understood by looking at Figure 8 (a). Here we can see that the Account Balance attribute has a negative trend until the end of 2008. After that time point, the Account Balance has a positive trend, which even seems to accelerate from December, 2012 and until the end of the series. According to BCC's expert, the first phase until the end of 2008 could show the progressive financial deterioration of weak non-defaulter clients at the first years of the financial crisis. The posterior increase in mean account balance would show that the clients that still remain non-defaulters are the ones with higher savings. This first phase seems to be mainly driven by the increase in the unemployment rate. This is the reason why the $\{\mathbb{E}[H^r_{4,t}]\}$ series is largely constant during this period (that is, this change was already explained by $\{\mathbb{E}[H_t]\}$). The second and third phase of the evolution of Account Balance cannot be explained by the evolution of the unemployment rate. This is when the $\{\mathbb{E}[H^r_{4,t}]\}$ series for the Account Balance attribute starts to capture this deviation from the main trend.

Similar conclusions can be extracted for the Unpaid Mortgages attribute (Att5) by looking at Figure 8 (c) and the $\{\mathbb{E}[H^r_{5,t}]\}$ series for this attribute. Again, according to the BCC expert, these three phases are due to the effect of a mortgage restructuring process. During the first years of the crisis, we can again see a financial deterioration of the non-defaulter clients until mid 2009, which is mainly explained by the unemployment rate (that is, the $\{\mathbb{E}[H^r_{5,t}]\}$ series is constant for this phase because the trend was already captured by $\{\mathbb{E}[H_t]\}$). Then, during the period from mid. 2009 to the end 2010 it was common at the bank to allow clients to restructure

their mortgages in order to pay them more easily. This policy aimed to slow down the quick increase in the number of defaulter clients due to the financial crisis. This is the reason why the $\{\mathbb{E}[H^r_{5,t}]\}$ series starts to capture something that cannot be explained by the unemployment rate (or the $\{\mathbb{E}[H_t]\}$ series). However, as bad economic conditions persisted, these non-defaulter clients with restructured mortgages finally started to default and were moved to the group of defaulter clients. Observe that the unpaid amount in mortgages starts to decrease for non-defaulter clients after 2011, but it increases more quickly for defaulter clients after the same date.

With the above comments in mind, we can better interpret the behavior of the $\{\mathbb{E}[H^r_t]\}$ series for the model in Equation 6, plotted in Figure 10. As can be seen in the figure, the $\{\mathbb{E}[H^r_t]\}$ series can again identify three different phases which seems to summarize the behavior of the set of $\{\mathbb{E}[H^r_{j,t}]\}$ series shown in Figure 9.

In the $\{\mathbb{E}[H^r_t]\}$ series displayed in Figure 10, we can also see a rapid increase of the series starting at the beginning of 2013. BCC's expert argues that this coincides in time with the fusion of BCC with another smaller regional bank as part of a big restructuring process of the financial institutions that took place in Spain that year. However, a deeper analysis should be performed in order to corroborate these conclusions.

4.4 Quantitative Model Evaluation

The residual approach presented in the previous sections can obviously be repeated in an iterative fashion. This would be equivalent to trying to add more hidden variables for better modeling the drift of the attributes over time. In Figure 10 we also show the result of this approach by including the $\{H^{r2}_t\}$ and $\{H^{r3}_t\}$ series, where $\{H^{r2}_t\}$ refers to the residuals of $\{H^r_t\}$ and $\{H^{r3}_t\}$ refers to the residuals of $\{H^{r2}_t\}$. It can be seen that, at every new iteration the curve becomes more constant, showing that there is less and less trend to be captured as time evolves. This can also be tested in a quantitative way by looking at the prequential marginal log-likelihood of the data according to the different models (with one hidden variable, two hidden variables, three hidden variables, etc) and comparing them to the simple model $x_{i,j,t} = \alpha_j + \epsilon_{i,j,t}$, i.e., a model without hidden variables.

In Figure 11, we plot the evolution of the marginal log-likelihood[6] of the data for models with none, one, two, three, and four hidden variables. The plot shows that including a few hidden variables is enough to increase the marginal log-likelihood of the model, suggesting that increasing the model complexity beyond that point only yields small improvements.

In Table 2 (Row 3), we also show the variance of the $\{H^{r2}_{j,t}\}$ series which corresponds to the local residual analysis associated to $\{H^{r2}_t\}$ series (when having three hidden variables). This again shows that the local effect of the variables is again strongly reduced and, in consequence, mainly explained by the series $\{H_{j,t}\}$ and $\{H^r_{j,t}\}$.

[6] This is an approximated value by variational methods, which is called the *evidence lower bound (ELBO).*

5 Discussion

In this section we first want to highlight that the models used in this paper are simple instantiations of the general model family described in Section 3. For example, we are making the unrealistic assumption that the attributes are independent conditional on the defaulter status of the clients (although this is partly alleviated by the global latent variable), even when knowing that there is a strong correlation between income, expenses, and the account balance of a client. Moreover, we assume that the attributes are normally distributed, which is also an inaccurate assumption considering the histograms displayed in Figure 2. There exist straightforward ways to alleviate the assumptions about independence and normality in distribution by explicitly linking the observed variables or using extra hidden variables and non-Gaussian distributions. Still, since our goal was to understand the underlying dynamics rather than to find a model that is a perfect fit to the data, we have not pursued this line of investigation. Instead we have seen that even when using this simple model class we are able to obtain important insights about the general trends governing the evolution of the financial profile of the clients. In our opinion, this is a strong point in favor of the robustness of our approach.

As commented above, the proposed probabilistic concept drift model considered here is able to detect both global and local concept drift, depending on the hidden structure of the model. Additionally, the iterative process of analyzing the residuals helps reveal different levels of concept drift that may be present. For instance, different attributes may vary at different rates in opposite directions, so that the trend cannot be captured by simply aggregating the local hidden variables or by using a single global hidden variable.

The presented framework can easily be extended in different directions. Computationally, the only requirement is that we choose distributions s.t. the full model is in the conjugate exponential family. Interesting alternatives include the exponential distribution (for positive real numbers with heavy tails) and the Poisson distribution (for count data). The use of more complex dependency structures between the attributes, reflecting expert knowledge, would also allow us to design more faithful models. Referring to the qualitative characterization of types of concept drift described in [14], the instantiation of the framework employed in Section 4 is specifically targeting gradual, non-reoccurring drift. This model-behaviour is to a large extent defined through the assumed prior distribution for $H_t|\{H_{t-1} = h_{t-1}\}$ and $H_{j,t}|\{H_{j,t-1} = h_{j,t-1}\}$ (for the global and local model, respectively). We chose Gaussian distributions with low a priori variance for these dynamic model when analyzing the financial data set, thereby encoding a preferrence for smooth dynamics in latent space. A larger a priori variance would fit well with situations with incremental or probabilistic drift. Reoccurring drift can be modelled using a mixture-model for the latent variables, where the dynamic model is used to encode an a priori preference for staying in a regime.

Another direction would be to position our concept drift analysis within a batch learning context. In such a setting real-time analysis is not required, which means that future data can be used for making inferences about the past. In contrast, in this paper, we have pursued, following the requirements of our problem, a streaming approach, where inference is only based on past and present data. The Bayesian

framework naturally supports the alternative batch approach, which we can exploit to get a more accurate global picture of our analysis. This is equivalent to the "smoothing" phase usually employed in dynamic systems with hidden variables [9].

The inherent flexibility of Bayesian latent variable models reinforces the importance of *model validation*. In Section 4.4 we applied a simple approach in order to study the suitability of our model. More complex evaluation procedures, which look at temporal dependencies between the residuals, would be a new line of work to validate the faithfulness of our model with respect to the analyzed data.

6 Conclusions

In this paper we have used a novel model to capture different sources of concept drift in financial client data from the Spanish bank BCC. The data covers the period from April 2007 to March 2014. Despite the challenging distributions of the analyzed attributes and the simplicity of the applied model, we have been able to detect different trends that on the one hand relate to the general economic climate and on the other to the particular policies implemented by BCC during the period. The analysis is done in a streaming fashion, meaning that inferences drawn at a given point in time t cannot rely on data observed after t. We show that this filtering approach is sufficient to extract interesting concept drift information, and by comparing the generated results to those obtained by utilizing a computationally more expensive non-streaming technique we conclude that on-line analysis is indeed viable for concept drift detection and analysis.

The expected mean of the global concept drift variable in the model correlates almost perfectly with the unemployment rate in the region of the financial institution. It is thus natural to hypothesize that the main driving factor for concept drift is the unemployment rate, a perspective that was corroborated by a BCC expert. The analysis of the residuals has allowed us to pinpoint the attributes that do not follow the trend of the unemployment rate, mainly *Account balance* and *Unpaid amount in mortgages*. Closer analysis of these and consecutive residuals, have shown different phases in which we can see the deterioration of the non-defaulter clients on the first years of the crisis, a shift of weak non-defaulter clients to the defaulter state, and more specific actions taken by BCC like debt restructuring and possibly a fusion with other smaller regional banks.

We have outlined future lines of research both from the point of view of the concept drift detector model and from the point of view of the practitioners.

Appendix A: Robustness Analysis

In this paper, we pursued an approach which is able to model concept drift in a streaming fashion. This approach is based on two different models: the local model described in Section 4.1 and in Equation (1), and the global model described in Section 4.2 and in Equation (3). In both models, we assumed that the expected values over α and β coefficients (i.e. $\mathbb{E}[\alpha_j^+]$, $\mathbb{E}[\alpha_j^-]$, $\mathbb{E}[\beta_j^+]$, $\mathbb{E}[\beta_j^-]$) were constant over time. We note, however, that this is not entirely accurate, because the expected value is computed from the posterior distribution over the parameters following a Bayesian approach,

$$\mathbb{E}[\alpha_j^+] = \int \alpha_j^+ p(\alpha_j^+ | D_1, \ldots, D_t) \mathrm{d}\alpha_j^+,$$

and this posterior therefore depends on $D_{1:t}$, the data seen so far, and therefore also on time. More precisely, we should therefore have indexed this expected value with time to reflect this dependency.

In this appendix we show that disregarding this temporal dependency of the expected values only has a marginal impact on the conclusions we draw from our interpretation of the evolution of the local and global hidden variables. In consequence, we argue that our approach can be safely used to track concept drift in a streaming fashion.

For this purpose we rerun the same experiments whose results were displayed in Figure 5 and Figure 6 for the local and the global model, respectively. During these new experiments, all the α_j^+, α_j^-, β_j^+, β_j^- values in the local and global models were fixed across time (i.e. they were not considered random variables in the Bayesian model). [7] To find meaningful values for these parameters, we decided to choose the last available estimate of the parameters in the local and global models (i.e. after processing the information for all the months). For example, when rerunning the global model we set the α_j^+ value equal to

$$\bar{\alpha}_j^+ = \int \alpha_j^+ p(\alpha_j^+ | D_{1:T}) \mathrm{d}\alpha_j^+, \tag{7}$$

where $T = 84$ is the last month of data made available to us. The other α and β values were computed in the same way.

In consequence, the modeling equations of the global model (cf. Equation (3)) were rewritten as follows,

$$\begin{aligned} x_{i,j,t}^+ &= \bar{\alpha}_j^+ + \bar{\beta}_j^+ \cdot H_t + \epsilon_{i,j,t}^+; \\ x_{i,j,t}^- &= \bar{\alpha}_j^- + \bar{\beta}_j^- \cdot H_t + \epsilon_{i,j,t}^-, \end{aligned}$$

where the following entities are now assumed to be random variables according to the Bayesian formulation:

$$\begin{aligned} &H_0 \sim \mathcal{N}(\mu, \sigma^2), \\ &\epsilon_{i,j,t} \sim \mathcal{N}(0, \sigma_j^2), \\ &\sigma_j^2 \sim \text{InvGamma}(\alpha, \beta), \\ &H_t \sim \mathcal{N}(H_{t-1}, \sigma^2). \end{aligned}$$

We follow the same approach to re-run the local model.

Notice how Equation (7) takes α_j^+ outside the streaming paradigm; we are utilizing all the data $D_{1:T}$ when estimating H_t, even if $t < T$. The model defined in this appendix is therefore not suitable for online analysis, but will serve as a basis for *post-analysis* of the results presented in displayed in Figure 5 and Figure 6.

[7] This setup is only intended to illustrate the marginal effect the time varying parameters have on the previous analysis. It is not being proposed as an alternative analysis method, which would haven taken the proposed method out of the streaming context.

Series	Global	Att1	Att2	Att3	Att4	Att5	Att6
All Months	0.793	-0.182	0.776	0.855	0.702	0.755	0.867
Last 2/3 Months	0.945	0.703	0.954	0.952	0.984	0.942	0.998

Table 3: Pearson Correlation coefficient between the $\{\mathbb{E}[H_{j,t}]\}$ ($\{\mathbb{E}[H_t]\}$) series according to the standard local (global) model and for the local (global) model with α and β values fixed. First row shows the correlation considering all months. Second row consider the correlation considering only the last two thirds of the months.

To this end, Figure 12 and Figure 13 show the result of this analysis for the local and global model, respectively. In these figures we plot together the output of both approaches (with and without fixed α and β values). In order to appreciate better the comparison, we rescale the series[8]. Note that the absolute values of the hidden variables are not relevant for this analysis, only the relative changes. From a statistical point of view, this is not a problem because Gaussian distributions are translation invariant.

It can be appreciated that in the first months the trend captured the hidden variables (i.e. $\{\mathbb{E}[H_t]\}$ and $\{\mathbb{E}[H_{j,t}]\}$ series) hardly match in some cases, but they tend to be much more overlapped in the rest of the months.

In Table 3 we quantitatively evaluate this assessment by computing the Pearson correlation coefficient between both series (i.e. with and without fixed α and β values), considering all months and, also, after discarding the first third of the months. With this analysis we can observe than the correlation in that last two thirds of the months is high (except for Att1), while when considering all the months the correlation drops. The reason we find for this situation is that at the beginning α and β values are randomly initialized. During the first months α and β values are adjusted, in combination with the hidden variables, to fit the data. The prior distribution on the α and β values is $\mathcal{N}(0, \infty)$, see Section 4.1, which means that large changes in their values are allowed, specially when little data has been observed. Estimates of the hidden variables during these first moths are therefore affected by these earlier estimates of the α and β values and, in consequence, not very reliable. This is akin to a *burn-in phase* where the estimates should be discarded. But this problem vanishes as the time goes on and both series (with and without fixed α and β values) become strongly correlated. This analysis shows that the trends captured by the local and global method without fixed α and β values is reliable after discarding the first time steps.

Appendix B: Synthetic data sets

We show how the concept drift modeling framework detailed in Section 3 can be used to analyse two synthetic data sets, widely employed as benchmarks in the concept drift literature. All the experiments have been performed using MOA [15], where the developed concept drift model (in Fig. 3) has been integrated as a new Bayesian

[8] Series cross zero in the middle of the time series by substracting the original value. And all values are divided by the maximum of the series to guarantee a maximum value of one in each series.

streaming classifier, named *bayes.amidstModels*. The Java code to reproduce the experiments can be downloaded from `http://amidst.github.io/toolbox/`.

B.1 SEA Data set

We first analyse the `SEA` data set [16] containing 60 000 samples, with 3 attributes (x_1, x_2, x_3) and 2 classes ($y = 0$ and $y = 1$). The attributes are numerical and uniformly distributed between 0 and 10. Only two of the attributes are relevant for the class label, y, which is defined as $y^t = 1$ if $x_1^t + x_2^t \leq \epsilon^t$ and $y^t = 0$ otherwise. Concept drift has been created by changing the threshold ϵ^t as a function of t. The data set covers four "phases", each with a duration of 15 000 samples, and with different ϵ^t (9, 8, 7, and 9.5 for the four phases, respectively). Figure 14 (left) shows the results of this analysis for batches of size N_t equal to 1000. The plot illustrates the progress of the expected value of the latent variable (denoted H^t) as well as the prequential accuracies computed using a sliding windows of size 1000 for a simple Naïve Bayes model (NB) and the adaptive Hoeffding tree model (AHT). As can be observed, the output of our model (i.e., the expected value of H^t) detects the drift points and clearly identifies the occurrences of the four different phases in the data, whereas those phases are less easily detected based on the accuracy results.

B.2 Rotating Hyperplane Data set

The second data set considered is the rotating hyperplane [17]. This benchmark data set is widely used to simulate "gradual" concept drift problems. We considered three versions of this data set, denoted `Hyp1`, `Hyp2`, and `Hyp3`, each including 10 000 instances. For each data set, 8 out of 9 attributes are drifting but with different magnitudes of change (i.e., 0.1, 0.5, and 1 for the three data sets, respectively), see [17] for details. Figure 14 (right) shows the evolution of the latent variable H^t for each considered data set using a sliding window of size 1000. Here we see that the different drift magnitudes of the three data sets are directly reflected in the development trends of the latent variables. For instance, for the `Hyp1` data, the curve of the H^t variable presents a stable behavior which correctly illustrates the very low change magnitude for this data set, i.e., 0.1.

Declarations

Funding

This work was performed as part of the AMIDST project. AMIDST has received funding from the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement no 619209. This research has been partly funded by the Spanish Ministry of Economy, Industry and Competitiveness, through projects TIN2013-46638-C3-1-P, TIN2015-74368-JIN, TIN2016-77902-C3-3-P and by ERDF funds.

Availability of data and materials

Data is from Banco de Crédito Cooperative (BCC). Due to privacy consideration regarding subjects in our data set, including European Union regulations and Spanish Data Protection rules, we cannot make our data publicly available. The data contains client-specific financial information captured during the period from April 2007 to March 2014. We make public available is all the software code employed for this analysis which is integrated within the AMIDST Toolbox. All the code and models used in this paper can be downloaded from the AMIDST Toolbox webpage (through its GitHub repository): `www.amidsttoolbox.com`. Experiments performed in Appendix B are made with public available data sets and can be fully reproduced.

Authors contributions

Designed the study: AR, AM, DR, HL, TN, AS. Code Development: AR, AM, DR. Analyzed the data: AR, HL, TN, AS. Wrote the paper: AR, AM, DR, HL, TN, AS. All authors read and approved the final manuscript.

Competing interests
The authors declare that they have no competing interests. This paper is an extended version of [3].

Acknowledgements
The authors would like to thank BCC expert Ramón Sáez for providing valuable insights to the paper.

Author details
[1]Department of Mathematics, University of Almería, Ctra. Sacramento, s/n, 04120 Almería, Spain. [2]Vestas Wind Systems A/S, Ørestads Blvd. 108, 2300 København S, Denmark. [3]Department of Computer Science, The Norwegian University of Science and Technology, Høgskoleringen 1, 7491 Trondheim, Norway. [4]Department of Computer Science, Aalborg University, Fredrik Bajers Vej 5, 9100 Aalborg, Denmark.

References

1. Gama, J., Žliobaitė, I.e., Bifet, A., Pechenizkiy, M., Bouchachia, A.: A survey on concept drift adaptation. ACM Computing Surveys **46**(4), 44–14437 (2014)
2. Masegosa, A., Nielsen, T.D., Langseth, H., Ramos-López, D., Salmerón, A., Madsen, A.L.: Bayesian models of data streams with hierarchical power priors. In: International Conference on Machine Learning, pp. 2334–2343 (2017)
3. Borchani, H., Martínez, A.M.., Masegosa, A., Langseth, H., Nielsen, T.D.., Salmerón, A., Fernández, A., Madsen, A.L.., Sáez, R.: Modeling concept drift: A probabilistic graphical model based approach. In: Proc. of The Fourteenth Int. Symposium on IDA, pp. 72–83 (2015)
4. Gama, J., Sebastião, R., Rodrigues, P.P.: On evaluating stream learning algorithms. Machine Learning **90**(3), 317–346 (2013). doi:10.1007/s10994-012-5320-9
5. Žliobaitė, I., Bifet, A., Read, J., Pfahringer, B., Holmes, G.: Evaluation methods and decision theory for classification of streaming data with temporal dependence. Machine Learning **98**(3), 455–482 (2014). doi:10.1007/s10994-014-5441-4
6. Masegosa, A.R., Martínez, A.M., Borchani, H.: Probabilistic graphical models on multi-core CPUs using Java 8. IEEE Computational Intelligence Magazine **11**(2), 41–54 (2016)
7. Masegosa, A.R., Martínez, A.M., Ramos-López, D., Cabañas, R., Salmerón, A., Langseth, H., Nielsen, T.D., Madsen, A.L.: Amidst: a java toolbox for scalable probabilistic machine learning. Knowledge-Based Systems (2018)
8. Cabañas, R., Martínez, A.M., Masegosa, A.R., Ramos-López, D., Samerón, A., Nielsen, T.D., Langseth, H., Madsen, A.L.: Financial data analysis with PGMs using AMIDST. In: Data Mining Workshops (ICDMW), 2016 IEEE 16th International Conference On, pp. 1284–1287 (2016). IEEE
9. Koller, D., Friedman, N.: Probabilistic Graphical Models: Principles and Techniques. MIT Press, ??? (2009)
10. Cooper, G.F.: The Computational Complexity of Probabilistic Inference Using Bayesian Belief Networks. Artificial Intelligence **42**, 393–405 (1990)
11. Jordan, M.I., Ghahramani, Z., Jaakkola, T.S., Saul, L.K.: An introduction to variational methods for graphical models. Machine Learning **37**, 183–233 (1999)
12. Beal, M.J.: Variational algorithms for approximate Bayesian inference. PhD thesis, Gatsby Computational Neuroscience Unit, University College London (2003)
13. Winn, J.M., Bishop, C.M.: Variational message passing. Journal of Machine Learning Research **6**, 661–694 (2005)
14. Webb, G.I., Hyde, R., Cao, H., Nguyen, H.L., Petitjean, F.: Characterizing concept drift. Data Mining and Knowledge Discovery **30**(4), 964–994 (2016)
15. Bifet, A., Holmes, G., Kirkby, R., Pfahringer, B.: MOA: Massive Online Analysis. Journal of Machine Learning Research **11**, 1601–1604 (2010)
16. Street, N., Kim, Y.: A streaming ensemble algorithm (SEA) for large-scale classification. In: 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), pp. 377–382 (2001)
17. Hulten, G., Spencer, L., Domingos, P.: Mining time changing data streams. In: Proceedings of the Seventh International Conference on Knowledge Discovery and Data Mining, pp. 97–106 (2001)

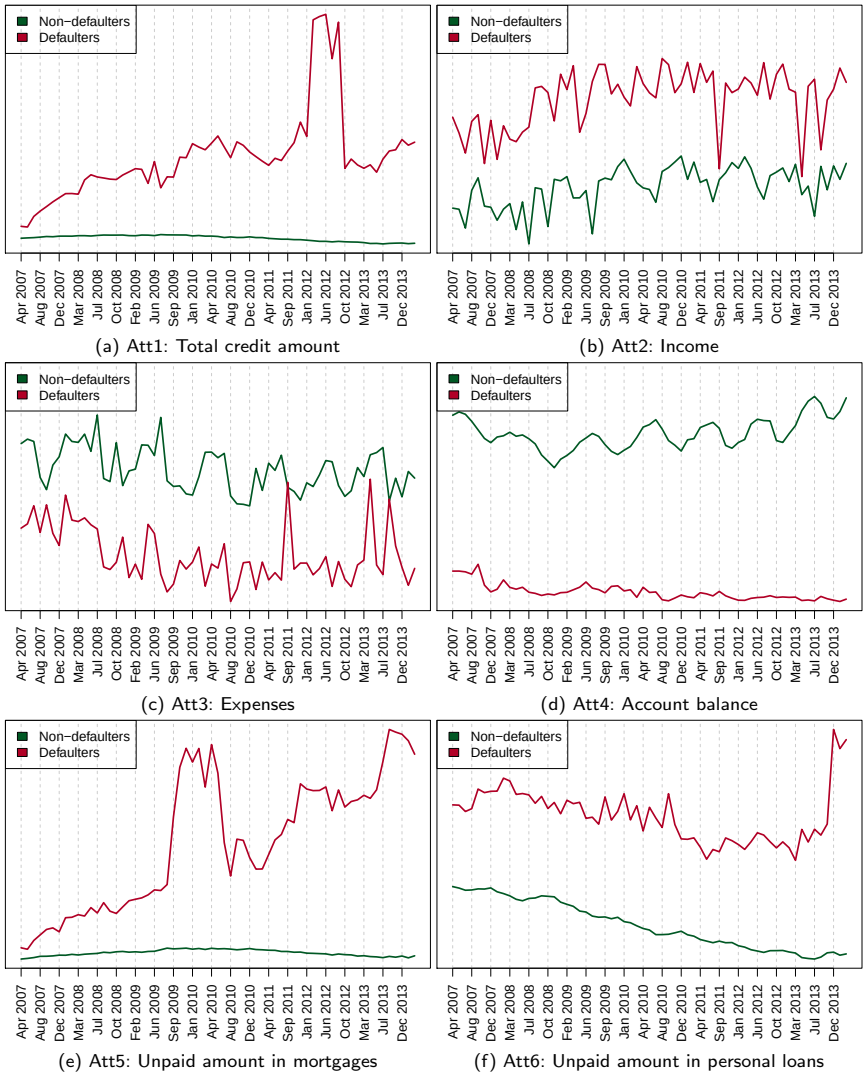


Figure 1: Mean evolution of all predictive variables for defaulting and non-defaulting clients (monthly aggregated). The ranges on the y-axes, both here and in successive figures, have been deliberately removed for confidentiality reasons.

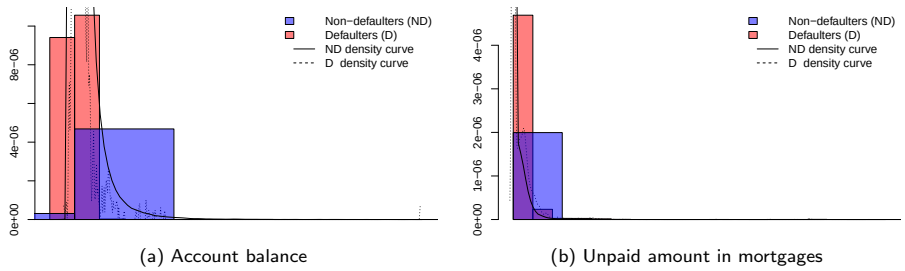


Figure 2: Histograms for two of the predictive attributes. Ranges in the x-axes have been deliberately removed for confidentiality reasons.

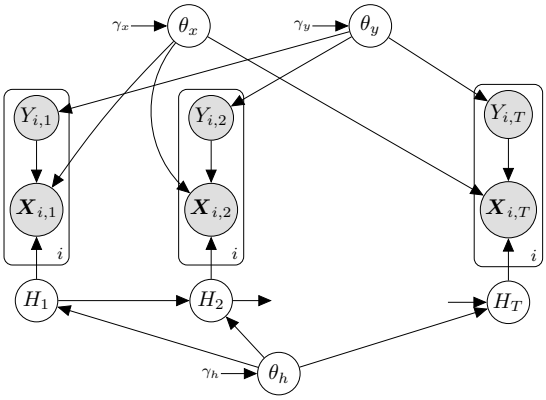
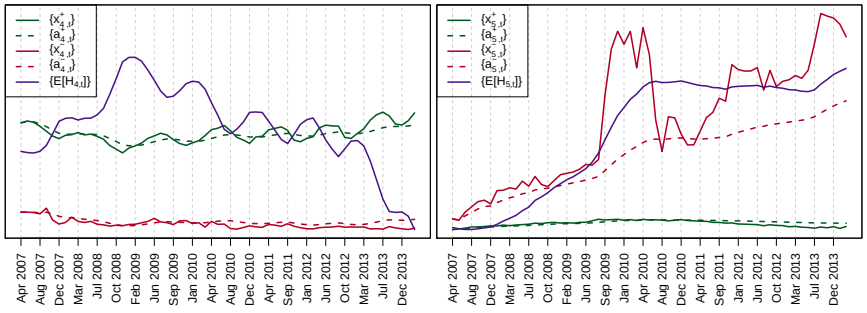


Figure 3: Model of concept drift [3]. In this model structure it is assumed that $X_{i,j,t} \perp X_{i,k,t} \,|\, Y_{i,t};\ \forall t \in \{1,\ldots,T\},\ \forall j,k \in \{1,\ldots,n\}$, where n is the number of attributes.



(a) Att4: Account balance (b) Att5: Unpaid amount in mortgages

Figure 4: Empirical, model means, and the expectation of the local hidden variables for the two feature variables Att4 and Att5. More specifically, $\{x^-_{j,t}\}$ and $\{x^+_{j,t}\}$ are the empirical mean series for defaulter and non-defaulter clients respectively, $\{a^-_{j,t}\}$ and $\{a^+_{j,t}\}$ the learned expected means, and $\{\mathbb{E}[H_{j,t}]\}$ are the expected values of the learned hidden variable $H_{j,t}$ for Attribute Att$j$.

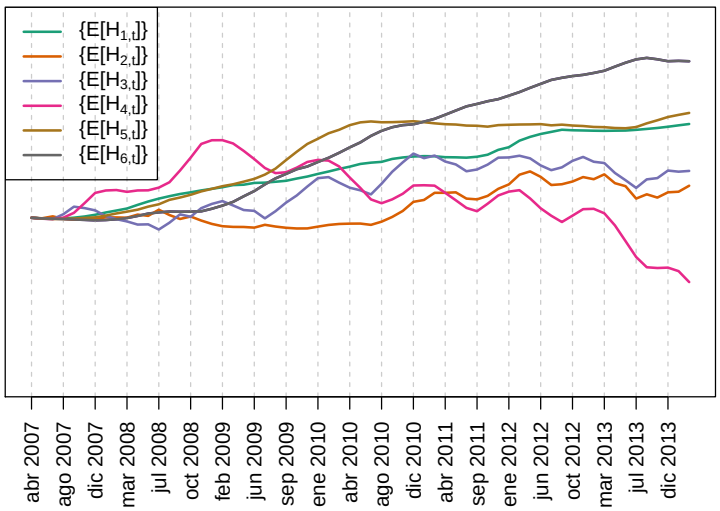


Figure 5: Expected values of the local hidden for all variables.$\{\mathbb{E}[H_{j,t}]\}$ is the expected value of the hidden variable $H_{j,t}$ for Attribute Att$j$.

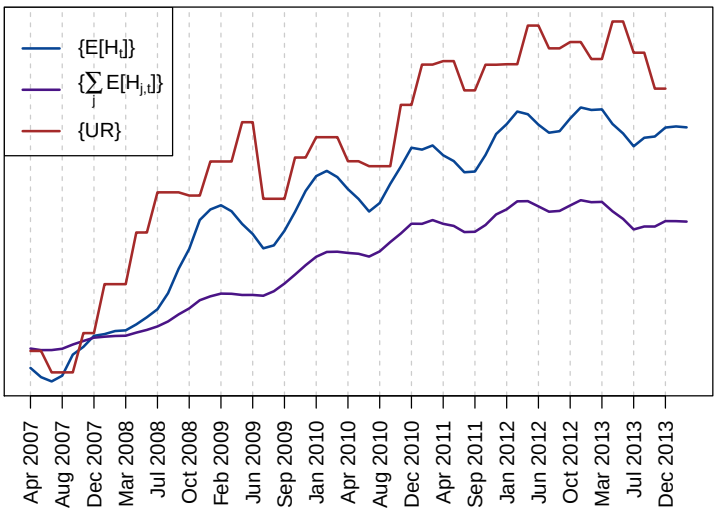
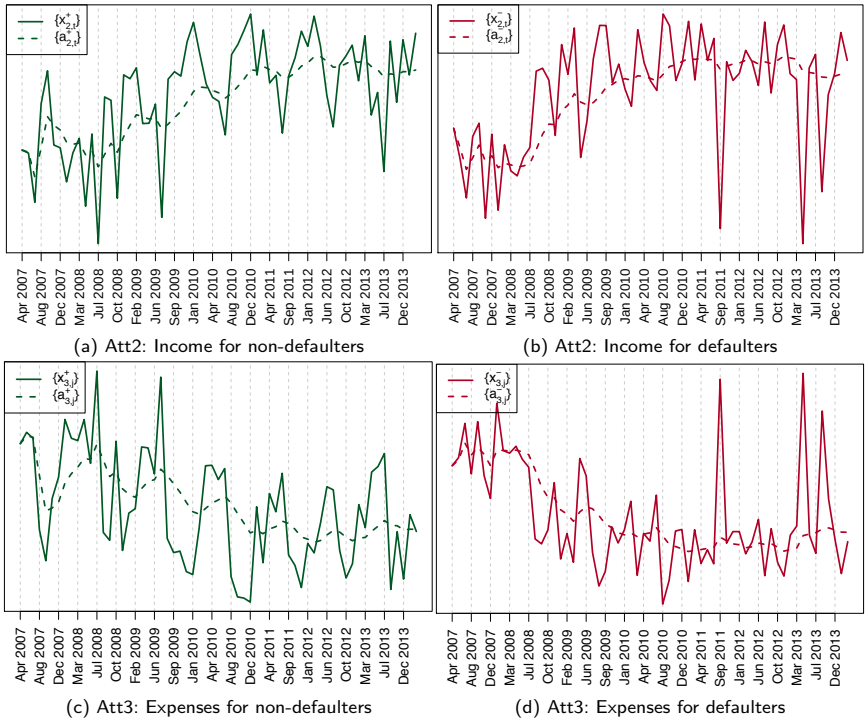


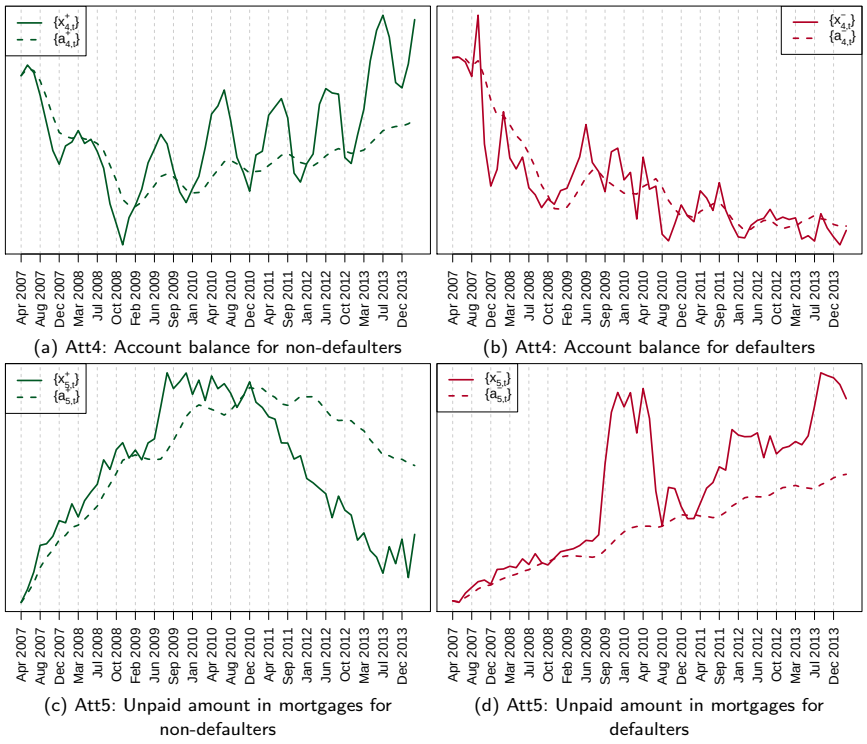
Figure 6: Expected values of the global hidden $\{\mathbb{E}[H_t]\}$, the sum of the expected values for the local hidden variables from Section 4.1 $\{\sum_j \mathbb{E}[H_{j,t}]\}$, and unemployment rate $\{\text{UR}\}$ series.



(a) Att2: Income for non-defaulters (b) Att2: Income for defaulters

(c) Att3: Expenses for non-defaulters (d) Att3: Expenses for defaulters

Figure 7: Empirical $\{x^+_{j,t}\}$ and model mean $\{a^+_{j,t}\}$ (monthly aggregated) for two predictive variables whose trend seems to be captured by the global hidden.



(a) Att4: Account balance for non-defaulters

(b) Att4: Account balance for defaulters

(c) Att5: Unpaid amount in mortgages for non-defaulters

(d) Att5: Unpaid amount in mortgages for defaulters

Figure 8: Empirical $\{x^+_{j,t}\}$ and learned mean $\{a^+_{j,t}\}$ (monthly aggregated) for two predictive variables whose trend is not entirely captured by the global hidden.

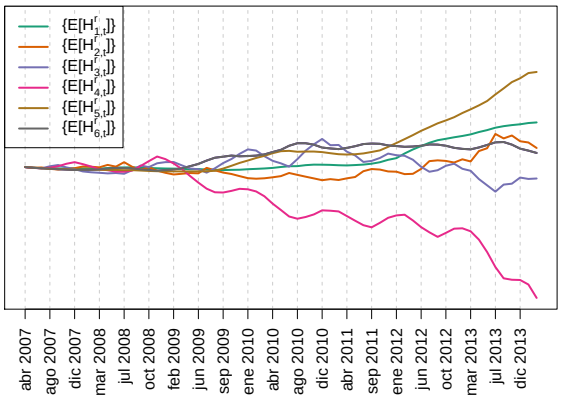


Figure 9: Expected values of the local hidden variables for the residuals $\{\mathbb{E}[H^r_{j,t}]\}$, where the index j corresponds to Attribute Att$j$.

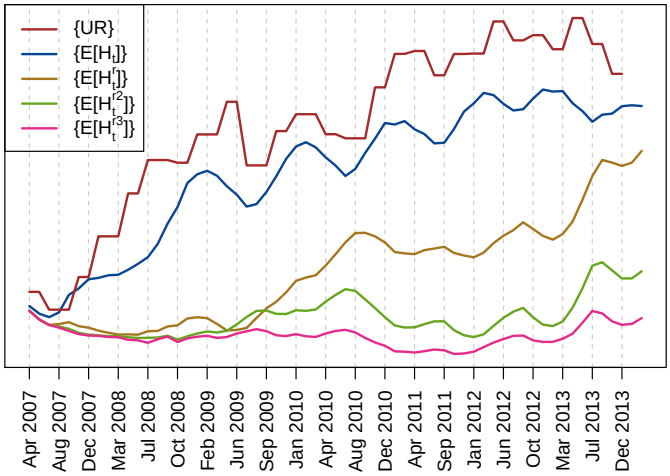


Figure 10: The unemployment rate {UR} as well as the expected value of the global hidden variables for the iterative residuals: $\{\mathbb{E}[H_t]\}$, $\{\mathbb{E}[H_t^r]\}$, $\{\mathbb{E}[H_t^r{}^2]\}$, and $\{\mathbb{E}[H_t^r{}^3]\}$.

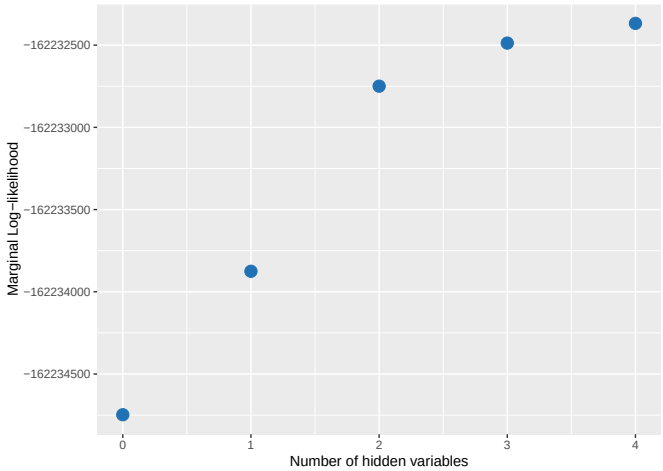
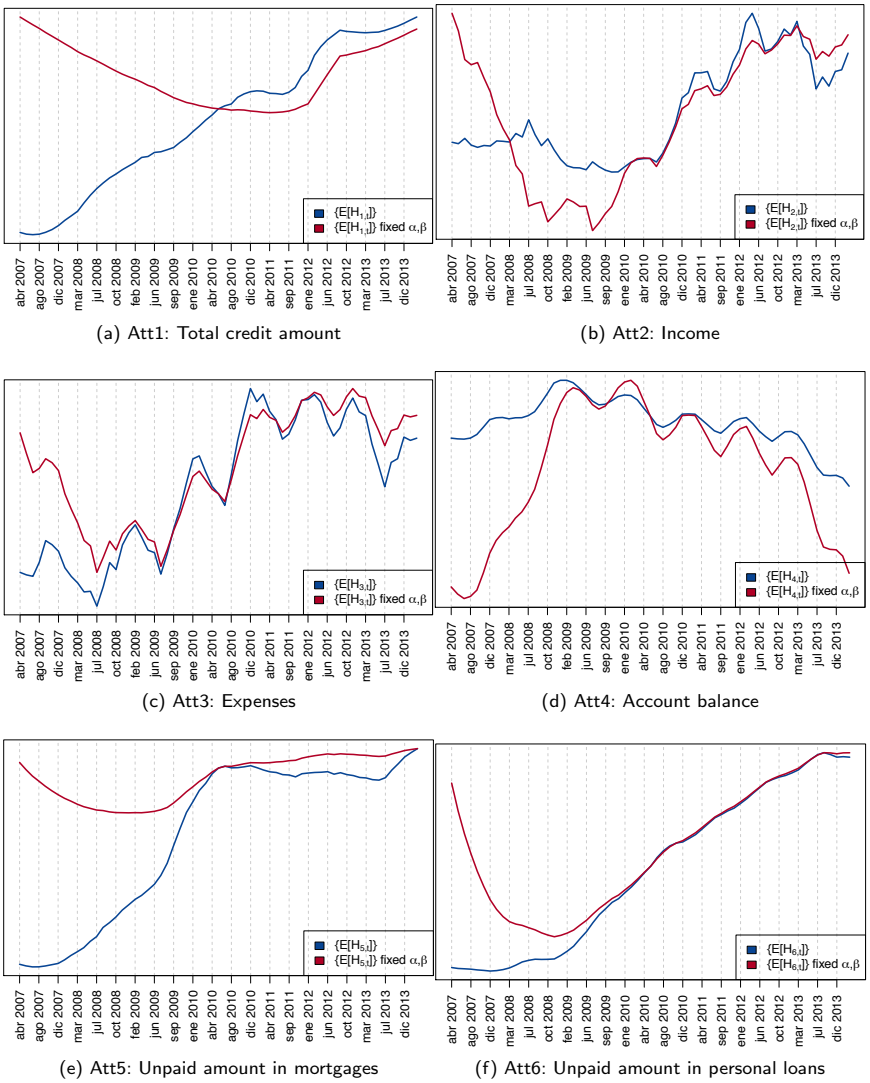


Figure 11: Prequential Marginal LogLikelihood for models with different number of hidden variables.



(a) Att1: Total credit amount

(b) Att2: Income

(c) Att3: Expenses

(d) Att4: Account balance

(e) Att5: Unpaid amount in mortgages

(f) Att6: Unpaid amount in personal loans

Figure 12: Expected values $\{\mathbb{E}[H_{j,t}]\}$ of all the variables according to the standard local model and for the local model with fixed α and β values.

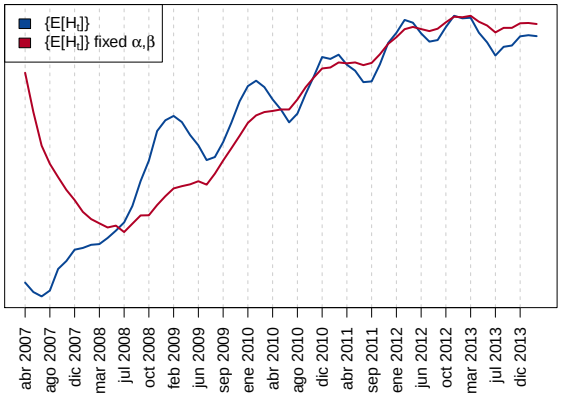


Figure 13: Expected values of the global hidden $\{\mathbb{E}[H_t]\}$ for the standard global model and for the global model with fixed α and β values.

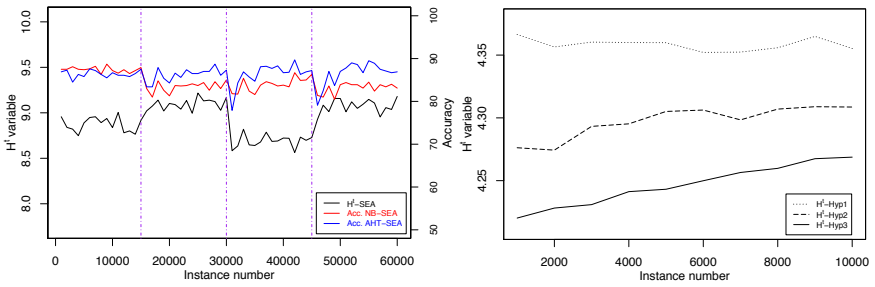


Figure 14: Left: Results for the SEA data set. Right: Results for the hyperplane data sets